第 9 回 続･高橋セミナー

最尤法によるポアソン回帰分析入門

カウント・データに対するポアソン回帰は，一般化線形モデルが扱える統計ソフトの普及により手軽に使えるようになってきた．ところが，実際のデータを用いて解析をしようとすると，通常の回帰分析にはない設定が求められる．分布の種類にポアソン分布を選ぶことは容易なのだが，リンク関数として（恒等・対数・ロジット・プロビット・補 2 重対数・・・）のどれか選択が迫られ，さらに，過分散の有無，オフセット変数の選択も迫られる．一般化線形モデルについての成書もあるが，ポアソン回帰を実施する際の注意点，解析結果の解釈については断片的な記載となっている．そこで，関連する成書などからカウント・データの事例を収集し，ポアソン回帰の基礎と応用について網羅した．稀な現象ではないカウント・データは，過分散となりがちなので，負の 2 項分布を拡張したガンマ・ポアソン回帰について，また，ゼロ・カウントが多発する場合の解析法についても言及した．ポアソン回帰の基礎については，Excel の行列関数による最尤法を主体にし，統計ソフト JMP，SAS，および R による検討結果を加えた．

最尤法によるポアソン回帰分析入門　目 次

続・高橋セミナー　　https://www.yukms.com/biostat/takahasi2/index.htm

高橋セミナーは、生物検定法の再構築をテーマとし 1999 年 12 月に第 1 回目を開催し、2007 年 3 月の第 27 回で終了した。本年 2011 年から統計モデルの活用をテーマにした「続・高橋セミナー」を再開する。テーマは、経時データの解析、繰返しが不揃いの枝分かれ実験データの解析、生存時間データの解析、反復を伴う各種実験データの解析などを対象にする。これらのデータ解析には、最小 2 乗法ではなく最尤法あるいは制限付最尤法（REML）を用いることになる。線形モデルに対する最小 2 乗法について書かれた日本語の教科書は多数あるが、最尤法あるいは制限付最尤法を用いた統計解析の方法を対象にした日本語の教科書は乏しい。英文の専門書は数多くあるものの、実際の解析に際しては SAS などの統計ソフトを使いことを前提にした記述となり、統計解析を行なう当事者たちにとって最尤法あるいは制限付最尤法を用いる統計ソフトを「ブラックボックス」として使わざるを得ない状況となっている。「続・高橋セミナー」では、新たなテーマのみならず、これまでの高橋セミナーの内容をすべての見直し、再編集を行い、追加・訂正を適宜加えて順次公表して行く。目指すのは「用量反応関係の統計モデルの基礎と応用」である。基礎概念の理解のために EXCEL による計算を充実させる。第 1 回は、「JMP による各種分割実験入門 －変量効果モデルの基礎－」としたが、第 7 章は「EXCEL ソルバーによる REML 法入門」である。　2011 年 12 月 27 日　高橋 行雄

第 9 回「最尤法によるポアソン回帰分析入門」公開の経緯

第 9 回に先だって, 2019 年 10 月に第 8 回「最尤法による探索的ポアソン回帰」を公開した. これは第 9 回の第 7 章として準備していた内容を, 先行して取りまとめたものである. 全ての章の原稿が揃い推敲を重ね, 2020 年 4 月に第1章から順次公開し, 7 月の第 13 章で章ごとの公開を完結した. さらに推敲を重ね全章を一括公開する. すでに公開した章ごとにまとめた PDF についても, 細かな不具合を訂正した改訂版を順次追加する.

第 10 回の予定「層別因子を含む探索的な回帰分析入門」

第 9 回の第 13 章「最小 2 乗平均の謎を予測プロファイルで解く」で取り上げた事例を主体にしたセミナーを準備している. コンセプトを以下に示す. 説明変数に質的変数と量的変数の両方を含み, 反応変数を連続量とした場合の代表的な統計モデルとして共分散分析が知られている. 一般的に共分散分析は, 1 元配置型の実験に際して結果に影響をおよぼすことが明らか量的な共変量を考慮した解析法である. 観察研究において層別因子を含んだ回帰分析も“共分散分析”として認識されている. さらに, 一定期間に何らかの処置を継続し, 目的とするある量的な主反応を得た場合に, 副反応が間接的に主反応に影響を与えている場合に, 副反応を含めた解析も“共分散分析”として認識されている. 臨床研究では, 対象集団と介入集団から得られる主反応に対し, 片方の集団に偏って存在する因子が, 主反応に影響を与えるような場合にも“共分散分析”が活用されている. 多くの分野で共通するのが, 経時的に観察されたデータにおける前値を共変量とする場合である. このように多くの分野で共分散分析が活用されているのであるが, 総合的に論じている書物は見当たらない. そこで, 関連する成書のデータを引用しつつ, EXCEL の行列関数による回帰分析および分析ツールを主体にして, “共分散分析”の活用法を論ずる.

第 9 回 続高橋セミナー

最尤法によるポアソン回帰分析入門

高橋 行雄

2021 年 1 月

BioStat 研究所（株）

〒105-0014 東京都 港区 芝 1-12－3 の 1005

takahashi.stat@nifty.com ， FAX : 03-342-8035

最尤法によるポアソン回帰分析入門　目 次

偶数ページ

はじめに

まれな現象から得られるカウント・データがポアソン分布に従うことの例示は多くの統計の成書に取り上げられている．何らかの説明変数に対して，カウント・データを反応とするポアソン回帰は，一般化化線形モデルあるいはカテゴリカルデータを扱う英文の専門書には多くの事例を見いだすことができ，日本語に翻訳されている入門書もあり，ポアソン回帰についていくばくかの知識は得られる．

久保（2012）は,「データ解析のための統計モデリング入門，一般化線形モデル・階層ベイズモデル・MCMC」で,「何でもかんでも正規分布と考えるのはおかしいだろう」というコンセプトで正規分布ではなくポアソン分布を全面的に取り上げて論じている．さらに，第 3 章で植物の種子数を主体した「一般化線形モデル（GLM）－ポアソン回帰－」を展開し，これまでの正規分布を前提とした統計解析とは異なる切り口を提示している．

通常の回帰分析は，反応変数を Y_i，説明変数を X_i としたときに回帰直線 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ を推定する解析法である．回帰パラメータの切片 β_0 および傾き β_1 の推定には，最小 2 乗法が標準的に用いられている. ただし, 回帰分析を適切に行なうためには, 誤差 $\varepsilon_i = Y_i - \hat{Y}_i$ に対し，分散が X_i に関して均一の正規分布に従っていることが前提とされている．だが，Y_i が互いに独立でなくとも，誤差分散が不均一であっても，正規分布を仮定できなくとも，形式的に線形最小 2 乗法が適用できるので，現実的には，これらの前提条件は無視されがちである．

説明変数 X_i が増大するにつれて，しばしば誤差分散が増大することが経験的に知られている．反応変数 Y_i が，0，1，2,... のようなカウント・データの場合は，説明変数 X_i の増加に伴い，推定値 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ も増加し，誤差分散も同程度に増加する．このような場合には，誤差に正規分布を仮定する通常の回帰分析ではなく，ポアソン分布を仮定する回帰分析の適用が必要となる．

多くの成書では，最尤法の原理は述べるものの，実際の計算は統計ソフトに丸投げしてしまうので，読者にとっては実際にどのような計算原理なのかは，ブラックボックスとなってしまう．このことが，統計ソフト依存症の人達を増やす原因となっている．統計ソフトに丸投げの結果として，統計ソフトからの出力結果の範囲内での理解となり，問題の本質に迫るような応用力を奪われてしまいがちになる．

多くの統計の入門書は，その時代の一般的な計算手段を意識しつつ構成されている．手回し計算機の時代，Fortran 言語が使える大型計算機の時代，電卓が普及した時代，パソコン上で Basic 言語が使えるようになった時代，そして R 言語による統計解析がパソコンで手軽にできるようになった現代は，R 言語の使い方が主体の入門書が増えている．

私も 2010 年以前は, 統計ソフト SAS および JMP を主体にした統計解析を行ないつつ, SAS の行列計算のための IML，あるいは，JMP のスクリプトなどの行列計算言語を用いて，統計モデルの計算原理を自ら確認してきた．これらの経験を元に「高橋セミナー」として成果物を 1999 年 12 月の第 1 回目から電子的に公表してきたが，諸般の事情で 2007 年 1 月の第 26 回で終了した．

その後，最尤法にどうしてもなじめない人達を念頭に，レンガを積むがごとく自らの手でじっくりと解析の過程を確認しながら理解を深めるできるように, Excel の行列計算とソルバーを主体にした「続・高橋セミナー」を 2011 年に再開した．そこでは，Excel での計算結果を主体にし，JMP あるいは SAS で検証するというようなスタイルに変えた．一般化線形モデルで取り扱える 2 値データについて，高橋（2017）「一般化線形モデルを Excel で極め活用する ープロビット法・ロジット法・補 2 重対数法ー」で取り上げ，ポアソン回帰については，高橋（2019a）「最尤法による探索的ポアソン回帰」で取り上げた．

カウント・データに対するポアソン回帰分析は，通常の回帰分析と対比しやすいので，一般化線形モデルの入門として適している．第 1 に，得られたデータがどのような分布に従うのかの検討に際して，正規分布があてはまるのか，ポアソン分布があてはまるのか，どちらなのかなどの同定が最初の課題となることも興味深く教育的でもある．第 2 に，単回帰分析に対し，恒等リンクのポアソン回帰分析を対比できることも一般化線形モデルの理解に役に立つ．第 3 に，回帰パラメータを推定し，ポアソン回帰直線の 95%信頼区間をどのように求めるか，などにも共通点がある．第 4 に，1 元配置型のカウント・データに対する平均値の比較に際し，ダミー変数を用いた通常の回帰分析に対し，ポアソン回帰分析でも同様にダミー変数を用いた「平均値」の比較を行なうことも類似点である．

ポアソン回帰について，ドブソン著，田中・森川・山中・富田訳（2008）「一般化線形モデル入門 原著第 2 版」が必読の書である．ただし，「入門」と言っても難解なので，ポアソン回帰に焦点をあてることにより，従来の回帰分析から一般線形モデルへの拡張を段階的に理解が得られるように，Excel の行列関数を活用し，視覚的にも理解できるように配慮した．さ

らに，Excel のソルバーを活用することにより，ポアソン回帰のみならず 2 値データに対する各種の一般化線形モデルのパラメータ推定が容易に求められることも示す.

統計ソフトでできることについて Excel で計算する意義がどこにあるのかと指摘されることもしばしば経験してきた．逆に，統計ソフトで対応できる探索的な解析の結果に対し，何の新規性があるのかとの辛口の指摘も受けたこともあった．自らの責任で行なう統計解析であれば，統計ソフトを誤用しないような最低限の知識は必要であるが，人様にレクチャーする立場になったときに自ら学習した「統計ソフトの使い方」レベルで良いのだろうか. 私も，最小 2 乗法をベースにした一般線形モデルを主体にしていた時代には，統計ソフトを使いこなすための啓蒙活動も大切であると思い実践していたこともあり，複雑な思いである.

その延長線で，複数の誤差構造を前提にする変量効果モデル「線形混合モデル SAS の MIXED プロシジャ」に対しても統計ソフトをいかに使いこなすか啓蒙活動をしてきた．ある時，故芳賀敏郎氏から，線形混合モデルの計算原理である制限付き最尤法（REML）について Excel を使って説明してほしいとの強い要望があった. 数年後に, ようやく Excel による REML 法を実現でき，その結果は，高橋（2011），「JMP による各種分割実験入門 －変量効果モデルの基礎－」に示した.

その経験が，統計ソフトを的確に使いこなすためには，Excel で統計ソフトから出力される各種の統計量を自ら計算できることが必須と考えるようになった．これは，統計ソフトが対応していない応用上の課題に対し，自ら計算して各種の推定値を求める力の根源となる．とはいえ，ただみたいな Excel での計算結果は信用できないとの風評もあり，Excel を使用すること自体の脆弱性も十分に承知しており，私自身も 2010 年以前はできるだけ避けてきた.

最尤法による各種の統計解析法を 2010 年以前に行列計算言語を用いて丁寧に説明しようと試みたこともあった．だが，丁寧に説明しようとすればするほど難解かつ冗長になってしまい断念してしまった．Excel に対する認識が変わったのは，Excel の行列関数とソルバーの存在であった．Web 上で公開されている芳賀（2004）「最小 2 乗法，最尤法，線形モデル，非線形モデル」を参考にしつつ，反復重み付き回帰を用いてプロビット法による 50 パーセント致死量の推定および 95%信頼区間の推定が, Excel シート 1 枚の中で実現できたことに新たな光明を見いだした［高橋（2017)］.

この経験から，最尤法についての入門書を書き始めたのであるが，アラカルト的になり途中で断念した．ドブソン（2008）を改めて読み直し，ポアソン回帰に的を絞ることにより最尤法についての首尾一貫した入門書としてまとめられると確信した．ドブソン（2008）には，

尤法についての首尾一貫した入門書としてまとめられると確信した．ドブソン（2008）には，ポアソン回帰の導入に反復重み付き回帰の行列計算の途中経過が丁寧に示されており，さらに深堀することにより，一般化線形モデルのみならず打ち切りデータのある寿命データの回帰分析などへの展開がスムーズに行えると思われた．

誤差がポアソン分布に従う場合の 2 群間あるいは多群間の平均値の比較，直線あるいは指数曲線（対数リンク）をあてはめるポアソン回帰，複数のポアソン回帰直線の同時あてはめ，実験計画法で取り上げられている要因配置型データの解析など，多様な問題に対して「ポアソン回帰」による応用が可能である．この際に参考になるのは，高橋・大橋・芳賀（1989），「SAS による実験データの解析」であり，一般線形モデルに対する SAS の GLM プロシジャが，回帰分析，重回帰分析，共分散分析，繰り返しが不揃いの多元配置分散分析，各種の直交表解析，など従来は別々解析法として扱われてきたものを統一的に取り扱えることが示されている．一般化線形モデルとしてのポアソン回帰においても SAS の GLM プロシジャと同様に，ありとあらゆる形式のデータ解析が行えるのであるが，その使い方については適当な教科書は見当たらない．

ポアソン回帰を使って，従来の回帰分析と対比して説明しようとしたときに，厄介な問題に直面した．これは，ポアソン回帰直線を最尤法によってあてはめた後，回帰直線の 95%信頼区間，および，個別データ 95%信頼区間（予測区間）を Excel で計算し図示しようとしたときに起きた．通常の回帰直線の信頼区間および予測区間のための計算公式は，ほとんどの教科書で画一的に偏差平方和 S_{xx} などを用いた式が使用されている．しかし，ポアソン回帰では，その考え方が，全く使えないことであり，共通の解析方法としてパラメータの共分散行列の活用に着目した．

ポアソン回帰での 95%信頼区間および予測区間の計算では，切片 $\hat{\beta}_0$ の分散 $Var(\hat{\beta}_0)$，傾き $\hat{\beta}_1$ の分散 $Var(\hat{\beta}_1)$，それの共分散 $Cov(\hat{\beta}_0, \hat{\beta}_1)$ を用いる．通常の回帰直線の場合に，パラメータの共分散行列を用いる方法が一般的になっていれば説明がしやすいのであるが，多くの教科書では，ほとんど見いだすことができなかった．そもそも，通常の回帰分析の解析手順に共分散 $Cov(\hat{\beta}_0, \hat{\beta}_1)$ の計算が含まれていないためである．これは，手計算の時代には，計算をできる限り簡略化することが優先されていたためと理解している．

本書は，3 部構成になっている．第 1 部は，第 1 章から第 6 章までで，ポアソン回帰に関連する基礎的な課題を扱っている．第 1 章は，ポアソン分布に従う各種のカウント・データについての基礎的な解析事例が示されている．第 2 章と第 5 章は，ポアソン回帰のパラメータ推定に用いられる最尤法についての具体的な解析法が示されている．第 3 章は，ポアソン回

帰を使いこなすための尤度比検定とデザイン行列について解説をしている．第 4 章は，ポアソン回帰を扱う上での基礎知識となる通常の行列計算による回帰分析入門である．第 6 章は，カウント・データではあるが，ポアソン分布のあてはめに問題がある過分散に対する方法が示されている．

第 2 部は，第 7 章から第 10 章で，探索的ポアソン回帰の事例となっている．第 7 章の最初の事例は，第 1.7 節の細菌を用いた 2×2 の実験結果を用い，要因配置型のカウント・データに対する探索的ポアソン回帰のアプローチの基本が示されている．第 2 の事例は，第 1.13 節のカブトガニの観察データについての事例で，2 変量ポアソン回帰，繰り返しが不揃いな 2 元配置ポアソン回帰，さらに共分散分析型ポアソン回帰を扱っている．第 3 の事例は，第 1 章で取り上げなかった事例で，2 群比較において，ポアソン分布が仮定できないカウント・データに対し，ゼロ過剰（Zero-Inflated）ポアソン分布などを扱っている．

第 8 章は，2 本の回帰直線をあてはめた後の各種の推定法を扱っている．最初の事例は，共通の切片を持つ回帰直線の傾きの比較で，勾配比検定として知られている解析法であるが，パラメータの共分散行列を用いた新たな解析法が示されている．第 2 の事例は，切片は異なるが共通の傾きをもつ 2 本の回帰直線であり，平行線検定法として知られているのであるが，勾配比検定の場合と同様にパラメータの共分散行列を用いた解析法が，伝統的な解析法よりも見通しが良いことが示されている．第 3 の事例は，第 1.8 節で導入した細菌を用いた用量反応試験データであり，Excel を用いたポアソン回帰による効力比の解析方法を丁寧が解説されている．

第 9 章は，雑草研究誌の論文の事例で，カウント・データに対する R 言語によるポアソン回帰の実施例に対し，さらなる探索的ポアソン回帰を行った結果である．まず，観察されたデータに関連する説明変数をオフセットと見なすか，共変量と見なすか，についての検討が最初に示されている．JMP のポアソン回帰を用いた 2 つの質的変数の交互作用の検討に加え，無償版の SAS の GENMOD プロシジャによる負の二項回帰（ガンマ・ポアソン回帰）の使用法についても，R 言語による解析法と対比して示されている．さらに Excel を使った予測プロファイルの作成法，各種のデビアンスについての詳細な解説がされている．

第 10 章は，Agresti1（2013）に示されている事例を扱っている．これは，貨物船の損傷数に対する観察データであり，オフセットを含む 3 元配置型の探索的ポアソン回帰の実施例である．主効果モデルに対する（0，1）型のデザイン行列と（1，–1）対比型デザイン行列の使い分けに始まり，95%信頼区間を含む効果の推定，それらをグラフ化し評価するの方法として「予測プロファイル」という形式で，探索的な解析結果に対する結果の効率的な提示方法が，

JMP および Excel を使って示されている．さらに，交互作用を考慮した場合の解析法と結果のグラフ表示について新たな提案がなされている．

第 3 部は，第 11 章から第 13 章であり，各種のポアソン回帰に共通するテーマを扱っている．第 11 章は，デビアンス・逸脱度・テコ比・4 種の残差について詳細に示されている．第 1.4 節で用いたドブソン（2008）の人工データについて通常の回帰分析の偏差平方和ベースの結果とポアソン回帰の対数尤度ベースの結果が対比されており，デビアンスについて詳しく説明されている．通常の回帰分析での残差分析に使われているスチューデント化残差に関係するテコ比についても Excel を用いた計算方法が示され，ポアソン回帰で用いられるスチューデント化デビアンス残差についての前振りとなっている．

第 12 章は，パラメータの共分散行列の活用法についてである．観察された多変量データの変数間の共分散行列が Excel の行列関数でスマートに計算でき，さらに相関行列も Excel の行列関数で容易に求められることが最初に示されている．次いで，重回帰分析について伝統的な偏差平方和ベースの解析法とデザイン行列ベースの解析法が比較検討され，各種の推定に必要となるパラメータの共分散行列を得るためには，デザイン行列ベースの解析法が優れていることが強調されている．応用事例として，2 次曲線の 95%信頼区間と予測区間をパラメータの共分散行列を使うことにより手軽に計算できることが示され，2 次式のポアソン回帰についても，通常の 2 次の回帰分析と同様にパラメータの共分散行列を用いることにより 2 次曲線の 95%信頼区間が算出できることが例示されている．

第 13 章は，最小 2 乗平均の謎を予測プロファイルで解く事例研究である．謎ときのためには，質的因子に対し共変量があるような共分散分析の場合が適している．第 13.2 節の事例では，質的因子と共変量間に交互作用がある場合が例示されている．一般的には，共分散分析が適用できるのは，交互作用がないことが前提とされているが，無視できない交互作用があった場合でも，予測プロファイルを使うことにより解析結果に対する適切な考察ができることが示されている．この場合にも，パラメータの共分散行列が活躍し，共変量が 2 変数の場合についても予測プロファイルの有用性が示されている．第 13.2 節では，質的因子が 2 因子ある共分散分析の事例が示されている．R 言語において新たに提供された lsmeans パッケージの使用経験の論文［守屋・広岡（2018)］があり，Excel で最小 2 乗平均（lsmeans）の計算を実現しつつ，JMP での予測プロファイルとの関連が示されている．最後の第 13.3 節では，第 7.2 節のカブトガニの解析に使われている予測プロファイルが，最小 2 乗平均そのものであったとして謎解きが終わっている．

1. ポアソン分布に従う各種のカウント･データ

ポアソン分布に従うカウント・データの特徴について概説する．まず，ポアソン分布の基本的な特徴である平均値と分散が等しいこと，平均値の増加に伴い分散も比例して増加することを例示する．次に，ポアソン分布が稀に発生する事象の確率分布として，2 項分布の極限状態を仮定することにより導出されることを示す．引き続き，ポアソン回帰に関連する各種の文献データから，ポアソン分布の特徴を浮き彫りにするための事例，ポアソン回帰の基本的な事例を取り上げる．それらの事例を通じてポアソン回帰に特徴的な，対数リンク，オフセットの使い分けについても概観し，過分散となり，通常のポアソン分布のあてはめがためらわれる場合の事例なども例示する．

1.1. ポアソン分布の特徴

稀に起きるような現象をある一定期間観察し，それが起きる事象の数が，ポアソン分布に従うことが経験的に知られている．実験研究においては，何らかの刺激 X を加えた場合には，稀に起きる現象 Y が刺激 X の大きさに応じて多発するようになる．シャーレ上の数百万個の細菌の中で，何らかの異常を起こした細菌のコロニーをカウントするような場合，稀に起きる現象なので出現確率は求められないが，異常を起こしカウントされたコロニーの数は，ポアソン分布に従うことことが，経験的に知られている．

ポアソン分布は，分散が平均と同じなので，得られたカウント・データの平均と分散の比を計算して，同程度であれば，ほぼポアソン分布に従うと判断される．分散を平均で割った比が 1 を大きく超えるような場合は，“過分散”（Over Dispersion）が起きていると言い，実験条件が不均一となっていることが原因として疑われる．また，平均が異なるような複数の部分集団が存在するような場合にも過分散が起きやすい．

過分散が起きている場合には，平均に対し分散が大きいことを考慮するために，負の 2 項分布から導出されるガンマ･ポアソン分布を用いることもできる．事象が全く起きないゼロ・カウントの頻度が多い場合にも過分散が起きると判断される．この場合には，ゼロ・カウントの発生割合を新たなパラメータ ω として加味した，ゼロ過剰（Zero-Ifluraited）ポアソン分

布，ゼロ過剰ガンマ・ポアソン分布を仮定することもできる．これらの詳細については，第6章で示す．

ポアソン分布は，位置パラメータとしての平均を μ とし，観測値を y としたときに，ポアソン分布の確率関数 $f(y)$ は，次式で与えられる．

$$f(y)=\frac{\mu^{y}e^{-\mu}}{y!},\quad y=1,2,\ldots \tag{1.1}$$

ポアソン分布の形状の特徴を概観するために Excel でグラフ化を試みる．ポアソン分布の確率関数は，Excel の計算式または Poisson.dist()関数で次のように計算することができる．

$$f(y)=(\mu\,\hat{}\,y)*\exp(-\mu)/\mathrm{Fact}(y)$$
$$=\mathrm{Poisson.dist}(y,\mu,false)$$

3 番目の引数（*false*: 確率 or *true*: 下側確率）

表 1.1 左に Poisson.dist()関数を用いて平均を $\mu=0.5,\ 0.8,\ 1,\ 2$ とし，y を 0，1，2,. . .，11 と変化させた場合のポアソン分布の確率を計算した結果を示す．さらに，表 1.1 右に平均を $\mu=5,\ 10,\ 15,\ 20$ とし，y を 0，3，6,. . .，33 と 3 刻みで確率の計算をした結果を示す．平均

表 1.1　Excel の関数を用いたポアソン分布の確率計算

	平均 μ					平均 μ			
y	0.5	0.8	1	2	y	5	10	15	20
0	**0.6065**	0.4493	**0.3679**	0.1353	0	0.0067	0.0000	0.0000	0.0000
1	0.3033	0.3595	**0.3679**	0.2707	3	0.1404	0.0076	0.0002	0.0000
2	0.0758	0.1438	0.1839	0.2707	6	0.1462	0.0631	0.0048	0.0002
3	0.0126	0.0383	0.0613	0.1804	9	0.0363	0.1251	0.0324	0.0029
4	0.0016	0.0077	0.0153	0.0902	12	0.0034	0.0948	0.0829	0.0176
5	0.0002	0.0012	0.0031	0.0361	15	0.0002	0.0347	0.1024	0.0516
6	0.0000	0.0002	0.0005	0.0120	18	0.0000	0.0071	0.0706	0.0844
7	0.0000	0.0000	**0.0001**	0.0034	21	0.0000	0.0009	0.0299	0.0846
8	0.0000	0.0000	0.0000	0.0009	24	0.0000	0.0001	0.0083	0.0557
9	0.0000	0.0000	0.0000	0.0002	27	0.0000	0.0000	0.0016	0.0254
10	0.0000	0.0000	0.0000	0.0000	30	0.0000	0.0000	0.0002	0.0083
11	0.0000	0.0000	0.0000	0.0000	33	0.0000	0.0000	0.0000	0.0020

$y=0,\mu=0.5$ ： $\mathrm{Poisson.dist}(y,\mu,false)=\mathrm{Poisson.dist}(0,0.5,false)=0.6065$

Excel で表 1.1 のような関数計算を効率良く作成するためには，数式の中のパラメータのセルを設定する際に「相対参照」と「絶対参照」を組み合わせて設定するとよい．列方向の平均 μ，行方向の y を引用する際に =Poisson.dist(**$C6**, **D$5**, *false*) のようなセル参照とする．「$」が付いていると絶対参照となる．「$」の付与は，F4 キーを何回か押すことで設定できる．このようなアドレスにすることにより，数式が縦横に自在にコピーしても表頭・表側のデータを自動的に変化させながら参照させることができる．

を，$\mu=0.5$ とした場合に $y=0$ の確率は，Poisson.dist(0, 0.5, *false*)=0.6065 と計算されている．平均を $\mu=1$ とした場合に $y=0$ と $y=1$ の確率は，0.3679 と同じになり，$y=7$ の場合の確率は，0.0001 となり，それ以後は，0.0000 以下になる．

図 1.1 に表 1.1 で示したポアソン分布の確率を Excel の縦棒グラフで表示する．平均 μ が 1 よりも小さい場合には，指数分布的な片流れ的な形状であり，平均が 2 から 5 ぐらいまでは，右に裾を長く引くような分布である．平均が 10 以上になるとやや右に裾を引くが，左右対称な正規分布に近づく．ただし，分散 σ^2 は，平均と同様に増大して行く．図中に，平均 μ，分散 σ^2，標準偏差 σ を上書きしてある．平均 μ 変化による標準偏差 σ の変化の程度を変動係数 CV（標準偏差/平均）みると，$\mu=1$ の場合 $CV=(1/1)\times100=100.0\%$，$\mu=5$ の場合 $CV=(2.24/5)\times100=44.7\%$，$\mu=20$ の場合 $CV=(4.47/20)\times100=22.4\%$ と減少している．

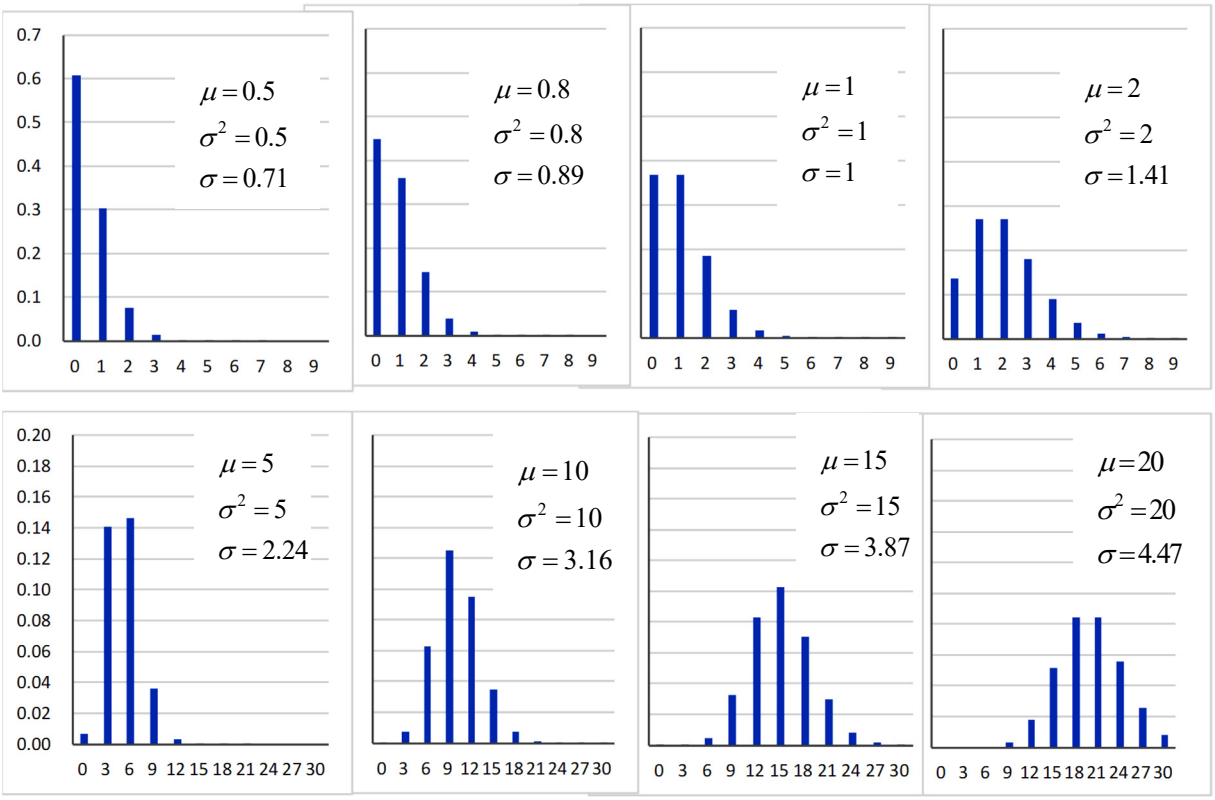


図 1.1　ポアソン分布の形状

1.2. 2 項分布からポアソン分布の導出

ある地域で 1 日に起きる交通事故を考える．1 日を分単位で区切り $n=60\times24=1440$ 分とし，1 分ごとに交通事故が起きれば 1，起きなければ 0 とする．1 日あたりの事故が起きた件数を y とし，1 日あたりの事故件数の平均を μ とする．1 分ごとに事故が起きる確率を $\pi=\mu/n$,起きない確率を $1-\pi$ とする.1 日あたりの事故件数を y とした場合の 2 項分布は，次式で与えられる．

$$\binom{n}{y}\pi^{y}(1-\pi)^{n-y}=\frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\mu}{n}\right)^{y}\left(1-\frac{\mu}{n}\right)^{n-y} \tag{1.2}$$

十分に大きい n に対して y はごく小さいので

$$n(n-1)\cdots(n-y+1)\approx n^{y} \tag{1.3}$$

に置き換えることができる．事故が起きない確率は，

$$\left(1-\frac{\mu}{n}\right)^{n-y} \tag{1.4}$$

であり，十分大きい n に対しては，

$$\left(1-\frac{\mu}{n}\right)^{n-y}\approx\left(1-\frac{\mu}{n}\right)^{n} \tag{1.5}$$

に置き換えることができる．数学の標準的な公式により，

$$\frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\mu}{n}\right)^{y}\left(1-\frac{\mu}{n}\right)^{n}=\frac{n^{y}}{y!}\left(\frac{\mu}{n}\right)^{y}e^{-\mu} \tag{1.6}$$

となる．これは，n が無限大に近付いたときに $(1-\mu/n)^{n}$ は，$e^{-\mu}$ に近付くためである．整理すると n を含まないポアソン分布 P_y となる．

$$\begin{aligned}P_y&=\frac{n^{y}}{y!}\left(\frac{\mu}{n}\right)^{y}e^{-\mu}\\&=\frac{\mu^{y}e^{-\mu}}{y!}\end{aligned} \tag{1.7}$$

ポアソン分布は，μ のみの関数であり，期待値と分散が共に μ となる．期待値は，

$$\begin{aligned}E(y)&=\sum_{y=0}^{\infty}yP_y\\&=\sum_{y=0}^{\infty}\frac{y\mu^{y}e^{-\mu}}{y!}\\&=\sum_{y=1}^{\infty}\frac{\mu^{y}e^{-\mu}}{(y-1)!}\end{aligned} \tag{1.8}$$

と変形でき，さらに，μ を Σ の外に出し，$i=y-1$ と置きなおすと，Σ の中は，i についての

ポアソン分布となり 0 から∞までの和は，分布関数の性質から 1 となり，

$$\begin{aligned} E(y) &= \mu \sum_{y=1}^{\infty} \frac{\mu^{y-1} e^{-\mu}}{(y-1)!} \\ &= \mu \sum_{i=0}^{\infty} \frac{\mu^{i} e^{-\mu}}{i!}, \quad (i = y-1) \\ &= \mu \end{aligned} \tag{1.9}$$

のように期待値が μ となることが確認される．簡便公式 $Var(y) = E(y^2) - E(y)^2$ によって分散を求めるために，まず， $E[y(y-1)] = E(y^2) - E(y)$ を求める．

$$\begin{aligned} E[y(y-1)] &= \sum_{y=0}^{\infty} y(y-1) \frac{\mu^{y} e^{-\mu}}{y!} \\ &= \mu^2 \sum_{y=2}^{\infty} \frac{\mu^{y-2} e^{-\mu}}{(y-2)!} \\ &= \mu^2 \end{aligned} \tag{1.10}$$

これから，$E(y) = \mu$ なので，$E(y^2) - E(y) = \mu^2$ から，$E(y^2) = \mu^2 + \mu$ が得られる．これらから，分散 $V(y)$ は，

$$\begin{aligned} Var(y) &= E(y^2) - E(y)^2 \\ &= \mu^2 + \mu - \mu^2 \\ &= \mu \end{aligned} \tag{1.11}$$

のように期待値 μ と同じになる．つまり， μ に比例して分散が大きくなる．

どのくらいの大きさの n から，ポアソン分布は 2 項分布に近似できるのか検討してみよう．表 1.2 に示すように，事故件数 y の期待値を $\mu = 1.5$ と固定し，母数団の大きさを $n = 10, 100$，

表 1.2　ポアソン分布の 2 項分布に対する近似精度

		n	μ	π	n	μ	π	n	μ	π
		10	**1.5**	0.1500	100	**1.5**	0.0150	1440	**1.5**	0.0010
i	y	二項	ポアソン	差	二項	ポアソン	差	二項	ポアソン	差
1	0	**0.1969**	**0.2231**	**0.0263**	0.2206	0.2231	**0.0025**	0.2230	0.2231	**0.0002**
2	1	0.3474	0.3347	-0.0127	0.3360	0.3347	-0.0013	0.3348	0.3347	-0.0001
3	2	0.2759	0.2510	-0.0249	0.2532	0.2510	-0.0022	0.2512	0.2510	-0.0002
4	3	0.1298	0.1255	-0.0043	0.1260	0.1255	-0.0005	0.1255	0.1255	0.0000
5	4	0.0401	0.0471	0.0070	0.0465	0.0471	0.0005	0.0470	0.0471	0.0000
6	5	0.0085	0.0141	0.0056	0.0136	0.0141	0.0005	0.0141	0.0141	0.0000
7	6	0.0012	0.0035	0.0023	0.0033	0.0035	0.0003	0.0035	0.0035	0.0000
8	7	0.0001	0.0008	0.0006	0.0007	0.0008	0.0001	0.0008	0.0008	0.0000
9	8	0.0000	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0000
10	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

1440 と変化させ，ポアソン分布と 2 項分布の発生件数 y の確率を計算し，その差を求める．ただし，2 項分布の出現確率は $\pi = \mu / n$ で与えられる．　ポアソン分布および 2 項分布の確率は，Excel の関数を用いて

ポアソン分布：　$P_i = \mathrm{Poisson.dist}(y_i, \mu, \mathit{false})$

2 項分布：　$B_i = \mathrm{Binom.dist}(y_i, n, (\mu / n), \mathit{false})$

のように計算する．ポアソン分布の計算には，n が含まれていないので，n を変えても同じ確率である．2 項分布の場合には，n が 10 の場合には $B_1 = 0.1969$ でポアソン分布の場合の $P_1 = 0.2231$ に比べて 0.0263 小さいが，n が 100 の場合は 0.0025 と差は減少し，n が 1440 の場合は 0.0002 と差は更に減少し，2 項分布はポアソン分布に漸近する．

表 1.3 に 1 日に起きる事故件数の期待値を $\mu = 1.5$ としたときに，1 年間 365 日の事故件数の分布を示す．度数 n_i は，ポアソン分布の確率 P_i に 365 日を掛けて四捨五入して整数化し，平均を計算するために積和 $n_i y_i$ を計算する．その合計を 365 日で割って平均 $\mu = 1.5041$ 件が得られている．分散の計算のためには，偏差平方和 $\Sigma_{i=1}^{10} n_i (y_i - \mu)^2 = 547.2438$ を求め，365 日で割った分散が 1.4993 となり，分散と平均がほぼ等しいことが確認される．

表 1.3　ポアソン分布の平均と分散

	事故件数	ポアソン		度数	積和	偏差平方
i	y	P	併合 P	n	$n\,y$	$n\,(y\text{-}\mu)^2$
1	0	0.2231		81	0	183.2500
2	1	0.3347		122	122	31.0034
3	2	0.2510		92	184	22.6235
4	3	0.1255		46	138	102.9337
5	4	0.0471		17	68	105.9010
6	5	0.0141		5	25	61.1062
7	6	0.0035	0.0045	2	12	40.4261
8	7	0.0008		0	0	0.0000
9	8	0.0001		0	0	0.0000
10	9	0.0000		0	0	0.0000
		計		365	549	547.2438
				平均	**1.5041**	**1.4993**
					μ	**分散**

分散は，$\mu = 1.5$ を既知としているので，偏差平方和を 365 で除している．

1.3. 有害雑草の種子の数の分布（1 群）

生物統計の名著として知られている スネデカー・コクラン著, 畑村・奥野・津村訳 (1972) の「統計的方法，第 6 版」，の第 8.14 節に *Phleum praetense*（イチゴツナギ）の 98 副標本に含まれる有害雑草の種子の数が示されている．各副標本の重量は 1/4 オンスで，もちろん沢山の種子を含んでおり, その中のほんの少数が有害雑草のものであった. 表 1.4 に観測度数，期待度数，その差，適合度のカイ 2 乗値などを示す.

表 1.4 有害雑草の種子の数に対するポアソン分布のあてはめ

i	有害種子の数 y	観測度数 n	積和 ny	偏差平方 $n(y-\mu\hat{})^2$	ポアソン確率 P	期待度数 $n\hat{}=NP$	観測−期待 $n-n\hat{}$	適合度 $\frac{(n-n\hat{})^2}{n\hat{}}$	(-2)*対数尤度 $-n\,2\ln(P)$
1	0	3	0	27.3686	**0.0488**	4.7806	-1.7806	0.6632	18.1224
2	1	17	17	69.3948	0.1473	14.4393	2.5607	0.4541	65.1106
3	2	26	52	27.0721	0.2225	21.8062	4.1938	0.8065	78.1441
4	3	16	48	0.0067	0.2240	21.9546	-5.9546	1.6150	47.8717
5	4	18	72	17.2728	0.1692	16.5779	1.4221	0.1220	63.9682
6	5	9	45	35.2691	0.1022	10.0144	-1.0144	0.1028	41.0569
7	6	3	18	26.6339	0.0514	5.0413	-2.0413	0.8265	17.8038
8	7	5	35	79.1858	0.0222	2.1752	2.8248	3.6682	38.0783
9	8	0	0	0.0000	0.0084				0.0000
10	9	1	9	35.7555	0.0028	**1.2104**	-0.2104	0.0366	11.7474
11	10	0	0	0.0000	0.0008	(8以上)			0.0000
12	11	0	0	0.0000	0.0002				0.0000
	計	98	296	317.9592	=平方和	**分散/平均**	**カイ2乗=**	**8.2949**	**381.9035**
	平均 $\mu\hat{}$=	**3.0204**	**3.2779**	**=分散**		**1.0853**	**p=**	**0.3073**	

結果が度数分布で与えられているので，N を観測度数 n_i の和とし，$n_i y_i$ の和 296 を $N=98$ で除して平均 $\hat{\mu}=3.0204$ が得られる.

$$N=\sum_i n_i = 98\,,\quad i=1,\ 2,\ldots,12$$

$$\hat{\mu}=\frac{\sum_i n_i y_i}{N}$$

$$=\frac{296}{98}=3.0204$$

分散 $\hat{\sigma}^2=3.2779$ は，偏差平方和を自由度で割り

$$\hat{\sigma}^2=\frac{\sum_i n_i (y_i-\hat{\mu})^2}{N-1}$$

$$=\frac{317.9592}{98-1}=3.2779$$

が得られる．分散/平均の比は，$3.2779/3.0204=1.0853$ であり，ほぼ 1 に近いと判断される.

ポアソン分布のあてはめが適切かを検討するために，y_i に対するポアソン分布の確率 P_i を求め，観測度数の和 N を掛けて期待度数 $\hat{n}_i = NP_i$ を計算し，観測度数 n_i と比較する．ポアソン分布の確率は，Excel の関数を用いて，

$$P_i = \text{Poisson.dist}\,(\,y_i,\ \hat{\mu}\,, \mathit{false})$$

で計算する．有害種種子の数 $y_1 = 0$ の場合は，

$$P_1 = \text{Poisson.dist}\,(\,0, 3.0204, \mathit{false}) = 0.0488$$

となる．期待度数 $\hat{n}_i$ は，$\hat{n}_i = NP_i$ で求められるが，$\hat{n}_{10} = 1.2104$ の場合 Excel の関数で $y_{10} = 9$ 以上は，7 以下の下側確率から差し引いて

$$\hat{n}_{10} = 98 \times (1 - \text{Poisson.dist}(\,7, 3.0204, \mathit{true})) = 1.2104$$

のように上側確率を計算し観測度数 n_i の合計 $N = 98$ を掛けて期待度数を計算している．これは，期待度数の和が，観測度数の和 $N = 98$ と一致させるためである．観測度数と期待度数との差 $n_i - \hat{n}_i$ は，若干の凸凹があるものの大きな乖離はない．

ポアソン分布に従っているかの適合度の検定の χ^2 値は，

$$\chi^2{}_{9-1-1} = \sum_i \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = 8.2949$$

であり，上側確率 p は，Excel の関数を用いて

$$p = 1 - \text{Chisq.dist}(8.2949,7, \text{true}) = 0.3073$$

となり，ポアソン分布のあてはめは棄却されない．

表の右端の「(–2) *対数尤度 $n[-2\ln(P)]$」は，ポアソン分布の確率 P_i について対数を取り，$-2n_i$ を掛けた結果である．有害種子の数 $y_1 = 0$ のポアソン分布の確率 $P_1 = 0.0488$ なので，

$$\begin{aligned} -2n_1 \ln P_1 &= (-2) \times 3 \times \ln(0.0488) \\ &= 18.1224 \end{aligned}$$

と計算されている．この欄の合計である (–2) 倍の対数尤度 $\ln L$ は，

$$\begin{aligned} \ln L &= \sum_i (-2n_i \ln L_i) \\ &= 18.1224 + 65.1106 + \cdots + 11.7474 \\ &= 381.9035 \end{aligned}$$

となっている．この「(–2) 倍の対数尤度」の和，または，単に「対数尤度」の和は，最尤法による各種の統計的方法の中心的な統計量で，最小 2 乗法での「偏差平方」の和と同等な役割を果す．表 1.5 に統計ソフト JMP の「一変量の分布」でポアソン分布をあてはめた結果の右の欄の「指標」の中の「(–2) *対数尤度」に 381.9035 を見出すことができる．

表 1.5 には，JMP の「一変量の分布」での結果で，観測度数 n_i の縦棒グラフにポアソン分布の観測度数 n_i が上書きされている．「要約統計量の欄」に「平均＝3.0204」と「分散＝3.2779」が表示され，「Poissson 分布のあてはめ」欄に平均と同じ「尺度 λ=3.0204」が推定されている．なお，「尺度 λ」は，「位置 λ」と同義語的に使われている．

「適合度検定」の欄に，χ^2 :「X2＝105.2703」が計算され，自由度（98−1）のカイ 2 乗分布の上側確率が「Prob＝0.2659」と計算されている．表 1.4 での適合度の計算は，出現度数に対する期待度数の関係から求めているのに対し，JMP の「一変量の分布」では，全 98 個の有害種子の数 Y に対して，Y の期待値である $\hat{\mu}$ との差の平方和をその分散 $\hat{\mu}$ で除した和として，次のように計算されている．

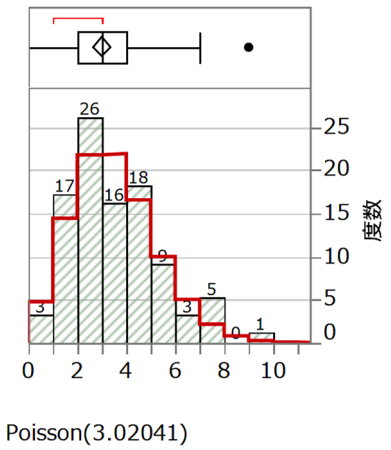
$$\text{Pearsonのカイ2乗} = \sum_i n_i \frac{(y_i - \hat{\mu})^2}{\hat{\mu}}$$

$$= 3 \times \frac{(0-3.0204)^2}{3.0204} + 17 \times \frac{(1-3.0204)^2}{3.0204} + \cdots + 1 \times \frac{(9-3.0204)^2}{3.0204}$$

$$= 105.2703$$

$$p = 1 - \text{Chisq.dist}\,(\,105.2703,\, 98-1,\, \textit{true}) = 0.2659$$

表 1.5　JMP による有害雑草の種子の数に対するポアソン分布のあてはめ



要約統計量	
平均	3.0204
標準偏差	1.8105
平均の標準誤差	0.1829
平均の上側95%	3.3834
平均の下側95%	2.6574
N	98.0000
分散	3.2779

Poisson分布のあてはめ

パラメータ推定値

種類	パラメータ	推定値
尺度	λ	3.0204

指標	
(-2)*対数尤度	381.9035
AICc	383.9451
BIC	386.4884

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
105.2703	0.2659

ここに示したように，統計ソフトの様々な出力結果を Excel で再現することは，様々な統計解析手法の計算理論について理解を深めることになる．その結果として，当該の統計的方法を適切に活用し，様々な応用ができるようになることが期待される．

1.4. 人工データ（恒等リンク，3 水準，回帰）

ドブソン著，田中・森川・山中・富田訳（2008），「一般化線形モデル入門，原著 第 2 版」の第 4.4 節に「ポアソン反応変数に対する回帰分析の例」がある．この節では，一般化線形モデルの計算理論の理解を深めるために表 1.6 に示す人工データが例示され，行列計算による反復重み付き回帰のための手順が示されている．さらに，反復過程での行列計算の結果の一部，回帰パラメータの収束過程が示されている．

表 1.6　ポアソン回帰の例（人工データ）

x	-1	-1	0	0	0	0	1	1	1
y	2	3	6	7	8	9	10	12	15

このデータを用いて，ドブソン（2008）で示されている反復重み付き回帰によるポアソン回帰を Excel の行列関数を用いて再現し，95%信頼区間の計算と結果の表示を追加する．

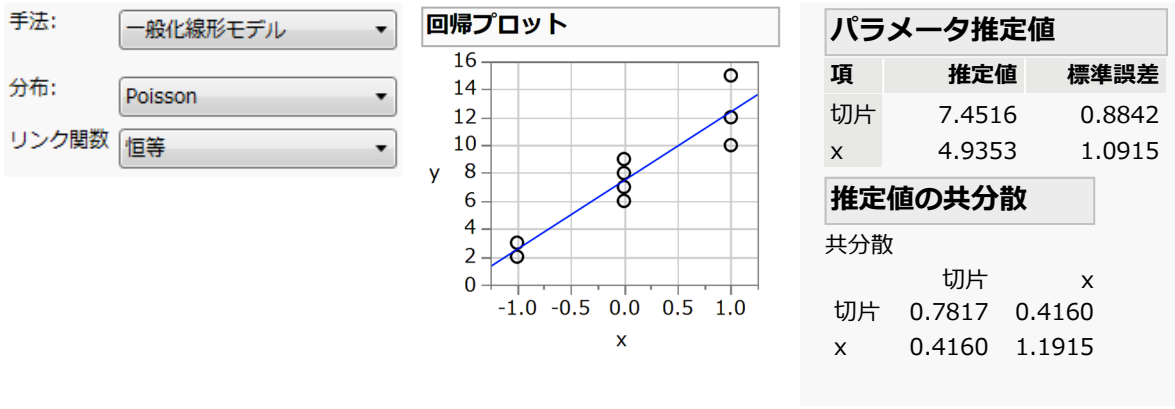
ポアソン回帰の適用

統計ソフト JMP の「モデルのあてはめ」から「一般化線形モデル」を選択し，「分布」で「Poisson」，「リンク関数」で「恒等」を選択し，実行した結果を表 1.7 に示す．回帰式は，

$$\hat{y}_i = 7.4516 + 4.9353 x_i$$

と推定されている．なお，「推定値の共分散」は，回帰直線の 95%信頼区間の計算のために必要となる．

表 1.7　JMP によるポアソン回帰の適用



注）　JMP では，ニュートン・ラフソン法を使用
（推定値の共分散が反復重み付き回帰とは若干異なる）

ポアソン分布を仮定するカウント・データは，0 以上であるが，推定された直線を外挿すると推定値がマイナスになる場合が生ずる．これを防ぐために

$$y_i = \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i \qquad \varepsilon_i \sim Poisson\ 分布 \tag{1.12}$$

のように，指数関数に線形式を含めてポアソン回帰を行い，推定値が 0 以下にならないようにすることが一般的であり，第 1.5 節で導入する．なお，一般化線形モデルでは，両辺に対数を取り推定値としての

$$\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (1.13)$$

$\ln \hat{y}_i$ についての線形式に対して反復重み付き回帰による最尤法が定式化されている．なお，JMP で「一般化線形モデル」を選び，「Poisson」を選ぶと「リンク関数」は自動的に「対数」となる．

ポアソン回帰の実際

統計ソフトでは，どのような計算が実際に行われているのだろうか．統計ソフトを使うユーザの立場であるならば，実際の計算がどのようなものか自ら計算を追試することは必要ないかもしれない．ポアソン回帰を説明する立場になった場合に，原理・原則を説明し，「実際の計算は統計ソフトを使えばいいのだ」と，言い切っていいのだろうか．通常の回帰分析の場合には，散布図上に回帰直線を描き，その 95%信頼区間を加え，さらに個別の 95%信頼区間（予測区間）を描くための数式は，ほとんどの教科書に示されている．統計ソフトでも標準的に出力されている．

ここに示した人工データで，ポアソン回帰直線の 95%信頼区間を描きたいのだけれども，どのようにしたらよいのだろうか．統計ソフトの外部出力の機能で 95%信頼区間の出力し，別途グラフ化することも可能ではある．Excel で奇麗な図を描きたいので，その計算式を知りたいが，どこに書いてあるのだろうか．ほとんどの統計の教科書の回帰分析の説明では，第 4.5 節で示すように偏差平方和を用いた計算式が示されており，この方法ではポアソン回帰直線の 95%信頼区間の計算への応用はできない．ただし，推定された回帰パラメータの共分散行列を用いれば，通常の回帰分析でもポアソン回帰でも，本節の末尾の表 1.10 に示すように同じ考え方で求めることができる．

ポアソン回帰は，どのような計算方法で行なうのだろうか．ドブソン(2008）には，丁寧な記述がある．これに沿って，Excel のシート上で，行列関数を用いて再現してみよう．一般的な統計の教科書での回帰分析は，シグマを用いて記述されており，行列での記述を見いだすことはまれである．Excel の行列関数を用いることにより，煩雑なシグマを使った場合に比べ，すっきりとした回帰分析が可能となる．行列計算に不慣れであれば，この節は飛ばして，第 4 章の「デザイン行列を用いた回帰分析」，次いで第 5 章の「反復重み付き最尤法によるポアソン回帰」を先に読んでもらいたい．

反復重み付き回帰

ポアソン回帰のみならず一般化線形モデル全般の理解と応用には，対数尤度関数の知識が不可欠であり，さらにパラメータに関する偏微分を正確に求める計算力も必要である．最尤法は，推定したいパラメータについて対数尤度関数の 1 階の偏微分を行いベクトル化（スコアベクトル）し，さらに 2 階の偏微分を行い行列化（ヘッセ行列）し，逐次的な計算によってパラメータを推定する（第 2 章で詳細を示す）．反復重み付き回帰は，ヘッセ行列に変えて，（m）回目の反復重み付き回帰式

$$\hat{\boldsymbol{\beta}}^{(m)} = \left[(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{(m-1)} \right]^{-1} (\boldsymbol{X}^T \hat{\boldsymbol{W}} \hat{\boldsymbol{Z}})^{(m-1)} \tag{1.14}$$

の第 1 項を $\mathcal{F}$

$$\mathcal{F} = (\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{(m-1)} \tag{1.15}$$

として，2 階の偏微分行列（ヘッセ行列）の代わりに使う．この方法は，ポアソン回帰を含む一般化線形モデルの各種のモデルに対する計算法として使われている．反復重み付き回帰を用いた最尤法についての詳細は，第 5 章で示すので，ここでは，Excel を使った反復計算の結果を表 1.8 に示す．

Excel による反復重み付き回帰の計算は，Excel の行列関数を活用している．通常の統計ソフトの回帰分析では，一般的に切片に対応する変数は省略している．ただし，行列計算では切片となる変数を明示する必要があり，それを 9×2 のデザイン行列 $\boldsymbol{X}$ とする．反応変数を 9×1 のベクトル $\boldsymbol{Y}$ とし，（m–1）回目の回帰パラメータの初期値を 2×1 のベクトル $\hat{\boldsymbol{\beta}}^{(m-1)}$ として，推定値 $\hat{\boldsymbol{Y}}$ を Excel の行列の関数 Mmult() を使って

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m-1)} = \text{Mmult}(\boldsymbol{X} \text{ の範囲}, \hat{\boldsymbol{\beta}}^{(m-1)} \text{ の範囲}) \tag{1.16}$$

のように計算する．なお，Excel の関数 Mmult() 内の引数（$\boldsymbol{X}$ の範囲）は，Excel シート上で $\boldsymbol{X}$ が存在するセル範囲を選択して設定する．

デザイン行列の i 行目の行ベクトルを $\boldsymbol{x}_i = [x_{0,i}, x_{1,i}]$ とし，分布関数は，表 1.7 に示したように恒等リンク（何も変換しない）を選択する．一般化線形モデルの公式から

$$g(\mu_i) = \mu_i = \boldsymbol{X}_i \hat{\boldsymbol{\beta}}^{(m-1)} = \eta_i \tag{1.17}$$

となる．重みは，恒等リンクなので，$\partial\mu_i / \partial\eta_i = 1$ となり，

$$\begin{aligned} \hat{w}_i &= \frac{1}{Var(\hat{Y}_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \frac{1}{\hat{\beta}_0 X_{0,i} + \hat{\beta}_1 X_{1,i}} = \frac{1}{\hat{Y}_i} \end{aligned} \tag{1.18}$$

重みは分散の逆数で，ポアソン分布の分散は期待値に等しいので$1/\hat{Y}_i$となる．

リンク関数$\eta_i^{(m)}$は，一般化線形モデルの公式から，

$$\eta_i^{(m)} = \mu_i^{(m-1)} + (Y_i - \mu_i^{(m-1)})\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \qquad (1.19)$$
$$= Y_i$$

$\eta_i^{(m)} = Y_i$となる．ここで，他のリンク関数への拡張も考慮して，行列計算の表記上

$$Z_i^{(m)} = \eta_i^{(m)} = Y_i$$

とする．

Excel による反復重み付き回帰

表 1.8 に示した反復重み付き回帰は，次式により段階的に計算する．なお，重みを分散の逆数とするのは，それぞれのX_iでのY_iの分散が明らかに異なる場合に対し分散を基準化するための基本的な方法である．なお，ここに示す行列計算は，第 5 章で詳しく説明する．

$(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(m-1)}$ ： = Mmult (Transpose ($\boldsymbol{X}$ の範囲 * $\hat{\boldsymbol{w}}$ の範囲), $\boldsymbol{X}$ の範囲)

$[(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(m-1)}]^{-1}$ ： = Minverse ($(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(m-1)}$ の範囲)

表 1.8　Excel シート上での反復重み付き回帰による最尤法

	デザイン行列			回帰			重み付回帰	推定値
	$\boldsymbol{X}$		観測値	推定値$^{(m-1)}$	リンク関数	重み	推定値$^{(m)}$	差
i	x_0	x_1	y	$y^\wedge$	$z=\eta=y$	$w^\wedge=1/y^\wedge$	$z^\wedge$	$y^\wedge - z^\wedge$
1	1	-1	2	2.0000	2.0	0.5000	2.5139	-0.5139
2	1	-1	3	2.0000	3.0	0.5000	2.5139	-0.5139
3	1	0	6	7.0000	6.0	0.1429	7.4514	-0.4514
4	1	0	7	7.0000	7.0	0.1429	7.4514	-0.4514
5	1	0	8	7.0000	8.0	0.1429	7.4514	-0.4514
6	1	0	9	7.0000	9.0	0.1429	7.4514	-0.4514
7	1	1	10	12.0000	10.0	0.0833	12.3889	-0.3889
8	1	1	12	12.0000	12.0	0.0833	12.3889	-0.3889
9	1	1	15	12.0000	15.0	0.0833	12.3889	-0.3889
			$\beta_0^\wedge=$	7.0000	$\beta_0^\wedge=$	**7.4514**		**1.4944E+06**
			$\beta_1^\wedge=$	5.0000	$\beta_1^\wedge=$	**4.9375**		平方和x10^6
			初期値 or $\boldsymbol{\beta}^{\wedge(m-1)}$貼り付け			重み付き回帰係数 $\boldsymbol{\beta}^{\wedge(m)}$		
	$(m-1)$		$(m-1)$		$(m-1)$			
	1.8214	-0.7500	0.7292	0.4375	9.8690			
	-0.7500	1.2500	0.4375	1.0625	0.5833			
	$\boldsymbol{X}^T\boldsymbol{W}^\wedge\boldsymbol{X}$		$(\boldsymbol{X}^T\boldsymbol{W}^\wedge\boldsymbol{X})^{-1}$		$\boldsymbol{X}^T\boldsymbol{W}^\wedge\boldsymbol{Z}$			

$(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z})^{(m-1)}$ ： = Mmult (Transpose ($\boldsymbol{X}$ の範囲 $*\hat{\boldsymbol{w}}$ の範囲), $\boldsymbol{Z}$ の範囲)

$[(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z})^{(m-1)}]^{-1}$ ： = Minverse ($(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z})^{(m-1)}$ の範囲)

$\hat{\boldsymbol{\beta}}^{(m)} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z})$: = Mmult ($(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ の範囲, $\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z}$ の範囲)

$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(m)}$ ： = Mmult ($\boldsymbol{X}$ の範囲, $\hat{\boldsymbol{\beta}}^{(m)}$ の範囲)

ただし，ベクトル $\hat{\boldsymbol{w}}$ は，$\hat{\boldsymbol{W}}$ 行列の対角要素，* 演算子は，セル同士の積.

初期値（$m=0$）としては，一般的に重みなしの回帰式

$$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 7.4545 \\ 4.9090 \end{bmatrix}$$

から得られる推定値を使うが，ドブソン(2008）に示されている

$$\hat{\boldsymbol{\beta}}^{(0)} = \begin{bmatrix} 7.0 \\ 5.0 \end{bmatrix}$$

を用いる．表 1.8 に示すように，初期値を入力すると，最初に $\hat{\boldsymbol{Y}} = (\boldsymbol{X}\hat{\boldsymbol{\beta}})^{(0)}$ が計算される．最初の $i=1$ の場合は，$\hat{y}_1 = 1\times 7.0 + (-1)\times 5.0 = 2.0$ となる．次に，重み $w_{ii} = 1/\hat{y}_i = 0.50$ が計算される．

第 1 回目の重み付き回帰のパラメータ $\hat{\boldsymbol{\beta}}^{(1)}$ は，次式のように推定されている．

$$\hat{\boldsymbol{\beta}}^{(1)} = \left[(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{(0)}\right]^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z})^{(0)}$$

$$= \left\{\left[(\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{X}\right]^{(0)}\right\}^{-1}\left[(\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{Z}\right]^{(0)}$$

$$= \begin{array}{|rr|} \hline 0.7292 & 0.4375 \\ 0.4375 & 1.0625 \\ \hline \end{array} \begin{array}{|r|} \hline 9.8690 \\ 0.5833 \\ \hline \end{array}$$

$$= \begin{array}{|r|} \hline \mathbf{7.4514} \\ \mathbf{4.9375} \\ \hline \end{array}$$

ただし，$\boldsymbol{X}*\hat{\boldsymbol{w}}$ は，行列 $\boldsymbol{X}$ とベクトル $\hat{\boldsymbol{w}}$ のセル同士の掛け算である

この推定値 $\hat{\boldsymbol{\beta}}^{(1)}$ を用いて，推定値 $\hat{\boldsymbol{Z}}^{(1)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(1)}$ が計算された結果が示されている．第 1 回目の推定値 $\hat{\boldsymbol{Y}}^{(0)}$ との差 $\hat{\boldsymbol{Y}}^{(0)} - \hat{\boldsymbol{Z}}^{(1)}$ を計算し，平方和を次式で計算している．

$$S_e = \sum_{i=1}^{9}(\hat{y}_i - \hat{z}_i)^2 = 1.4944$$

なお，表には，収束の状況を把握しやすいように 10^6 倍で表示している．

1.4944E+06
平方和x10^6

反復計算

推定値 $\hat{y}_i$ と推定値 $\hat{z}_i$ の差の平方和を計算し，十分に小さくなるまで，推定値 $\hat{\boldsymbol{\beta}}^{(m)}$ の結果をコピーし，“値のみ” を $\hat{\boldsymbol{\beta}}^{(m-1)}$ にペーストする．これは，$\hat{\boldsymbol{\beta}}^{(m)}$ には，計算式が埋め込まれているので，通常のペーストではセルの参照位置が異なりエラーとなるためである．表 1.9 に繰り返しコピー&ペーストした結果を示す．4 回目で平方和が 10^{-6} より小さくなったのでストップする．なお，手作業に換えソルバーで平方和を最小にするように $\hat{\boldsymbol{\beta}}^{(m-1)}$ を変化させることにより解を得ることもできる．

表 1.9　反復重み付き回帰による逐次近似

反復 m		$(m\text{-}1)$	(m)	差	平方和x10^6	共分散行列 $(\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{X}})^{-1}$	
1	$\beta_0\hat{}=$	7	7.451389	0.451389	1.494E+06	0.7292	0.4375
	$\beta_1\hat{}=$	5	4.937500	-0.062500		0.4375	1.0625
2	$\beta_0\hat{}=$	7.451389	7.451632	0.000243	1.581E+01		
	$\beta_1\hat{}=$	4.937500	4.935314	-0.002186		略	
3	$\beta_0\hat{}=$	7.451632	7.451633	0.000001	5.821E-04		
	$\beta_1\hat{}=$	4.935314	4.935300	-0.000013			
4	$\beta_0\hat{}=$	7.451633	7.451633	0.000000	2.142E-08	0.7817	0.4166
	$\beta_1\hat{}=$	4.935300	4.935300	0.000000		0.4166	1.1863

第 3 回目と第 4 回目の推定値の差は 10^{-6} 以下となり，差の平方和は，0.02142×10^{-6} と十分小さくなったので反復を中止し，解が求まったとする．さらに反復を行えば，差の平方和は，ゼロに漸近するが，ある一定の推定値の精度となれば，実用上は問題ないので一般的には，反復の中止基準をあらかじめ設定しておく．

Excel の行列関数を用いた重み付き回帰分析を活用し，反復重み付き回帰による最尤法により解を求めたのであるが，基本は第 4 章に示す通常の回帰分析から初め，段階的に学習することが望ましい．第 5 章で，反復重み付き回帰の基礎について詳しく示す．

パラメータの共分散行列を用いた 95%信頼区間の計算

表 1.9 で計算されているパラメータの推定値および共分散行列を用いてポアソン回帰直線の 95%信頼区間を Excel で計算し図示する．ただし，反復重み付き回帰での結果であるため表 1.7 の JMP による共分散行列と若干異なる．

		パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{X}})^{-1}$					
	$\hat{\boldsymbol{\beta}}$		$\beta_0\hat{}$	$\beta_1\hat{}$			
$\beta_0\hat{}=$	7.4516	$\beta_0\hat{}$	0.7817	0.4166	=	$Var(\beta_0\hat{})$	$Cov(\beta_0\hat{},\beta_1\hat{})$
$\beta_1\hat{}=$	4.9353	$\beta_1\hat{}$	0.4166	1.1863		$Cov(\beta_1\hat{},\beta_0\hat{})$	$Var(\beta_1\hat{})$

推定値 $\hat{y}$ の分散 $Var(\hat{y})$ は，パラメータの共分散行列の対角要素であり，回帰パラメータ $\hat{\beta}_0$ の分散 $Var(\hat{\beta}_0)=0.7817$，$\hat{\beta}_1$ の分散 $Var(\hat{\beta}_1)=1.1863$，$\hat{\beta}_0$ と $\hat{\beta}_1$ の共分散 $Cov(\hat{\beta}_0,\hat{\beta}_1)=0.4166$ を用いて，合成分散の一般式，

$$\begin{aligned} Var(\hat{y}) &= Var(\hat{\beta}_0+\hat{\beta}_1 x) \\ &= Var(\hat{\beta}_0)+2Cov(\hat{\beta}_0,\hat{\beta}_1)x+Var(\hat{\beta}_1)x^2 \\ L95\% &= \hat{y}-1.96\sqrt{Var(\hat{y})} \\ U95\% &= \hat{y}+1.96\sqrt{Var(\hat{y})} \end{aligned}$$

によって求められる．表 1.10 の $x_{1,1}=-2$ の場合については，

$$\hat{y}_1 = 7.4516+4.9353\times(-2) = -2.4190$$

$$Var(\hat{y}_1) = 0.7817+2\times0.4166\times(-2)+1.1863\times(-2)^2 = 3.8607$$

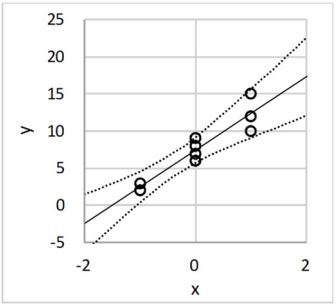
$$L95\% = -2.4189-1.96\sqrt{3.8607} = -6.2701$$

$$U95\% = -2.4189+1.96\sqrt{3.8607} = 1.4322$$

であり，他の $x_{1,i}=(-1,0,1,2)$ についても同様に計算した結果を表 1.10 に示す．散布図上には，この結果に基づきポアソン回帰の回帰の 95%信頼区間を重ね書きしてある．通常の回帰分析とは異なり，x の大きい方が小さい方に比べて広がりが大きくなっている．

表 1.10　ポアソン回帰直線の 95%信頼区間

i	x_0	x_1	y^	$Ver(y$^$)$	SE	$L95\%$	$U95\%$
1	1	-2	-2.4190	3.8607	1.9649	-6.2701	1.4322
2	1	-1	2.5163	1.1349	1.0653	0.4283	4.6043
3	1	0	7.4516	0.7817	0.8841	5.7188	9.1845
4	1	1	12.3869	2.8011	1.6736	9.1066	15.6673
5	1	2	17.3222	7.1931	2.6820	12.0655	22.5790



実際の分散 $Var(\hat{y})$ の計算は，パラメータ共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ が計算されているので，デザイン行列 $\boldsymbol{X}$ の $\boldsymbol{x}_i=(x_{0,i}\ x_{1,i})$ を行ベクトルとした場合に，次の 2 次形式で求められる．

$$\begin{aligned} Var(\hat{y}) &= \boldsymbol{x}_i\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i^T = \boldsymbol{x}_i(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{x}_i^T \\ &= \text{Mmult (Mmult (} \boldsymbol{x}_i \text{ の範囲，} (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1} \text{ の範囲, Transpose(} \boldsymbol{x}_i \text{ の範囲))} \end{aligned}$$

この，2 次形式での分散の計算方法の応用については，第 12 章に詳しく説明しているので参照されたい．ここで示した反復重み付き回帰の方法は，分布が 2 項分布である場合の「ロジット・リンク」，「プロビット・リンク」，補 2 重対数リンクの場合も同様であり，第 5.4 節で取り上げる．

1.5. 冠動脈心疾患の死亡者数（対数リンク，8 水準，オフセット，回帰）

ドブソン(2008) の第 3.5 節にオーストラリアのある地方の年齢 5 歳階級での冠動脈心疾患による死亡者数が示されている．これまでに示した事例は，事象の発現数のみが観測される場合であるが，稀な現象であっても表 1.11 に示すように対象とする部分母集団の大きさ n_i が人口統計学的に得られる場合もある．部分母集団の人数は十分大きく，極端な差がないので，死亡者数 y についてポアソン分布のあてはめが可能ではあるが，分母の大きさも考慮したい．

表 1.11　オーストラリアのある地方の冠動脈心疾患による死亡者数

	年齢層	死亡者数	母集団	死亡率	10万人比
No.	x	y	人数 n	%	人数
1	30	1	17,742	0.0056	5.6
2	35	5	16,554	0.0302	30.2
3	40	5	16,059	0.0311	31.1
4	45	12	13,083	0.0917	91.7
5	50	25	10,784	0.2318	231.8
6	55	38	9,645	0.3940	394.0
7	60	54	10,706	0.5044	504.4
8	65	65	9,933	0.6544	654.4
	全体	205	104,506	0.1962	196.2

年齢層は，30-34, 35-39 のように与えられている．

各年齢層の母集団の人数が分かっており，死亡率を算出すると全体で 0.1962%と小さいので，1 万人あるいは 10 万人あたり (以下，10 万人比とする) の死亡者数にした方が理解しやすい．全体の人数が 104,506 人なので，10 万人比に換算すると 196.2 人となる．各年齢層では，50～54 歳が，231.8 人と平均的である．図 1.2 左に示すように年齢が高くなるにつれて指数関数的に増加する．図 1.2 右の直線のあてはめは，対数変換した 10 万人比に対して通常の

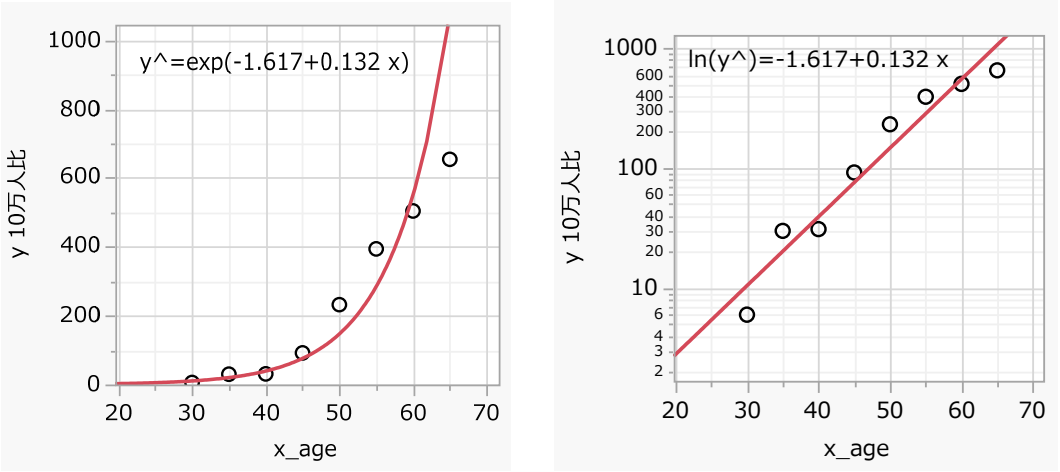


図 1.2　10 万人あたりの死亡者数の対数に対する便宜的な直線のあてはめ

回帰分析をした結果であり，図 1.2 左は，単に元の 10 万人比に戻しただけで，ポアソン回帰の結果ではない．年齢が増加するにつれて頭打ちになる傾向があり，対数変換をして直線をあてはめることは無理そうであり，2 次式のあてはめも考慮する必要がある．

ポアソン分布は，平均と分散が等しいので，説明変数 X の増加に伴い反応変数 Y の平均も直線的に増大するような場合に，通常の回帰分析で仮定する等分散性が成り立たない．さて，反応変数 y に対数を取った場合に，等分散性が成り立つのであろうか．説明変数 x が小さい場合に分散が相対的に大きくなってしましい，残念ながら等分散性が成り立たない．

死亡率が求められているので，ポアソン回帰ではなく 2 値データとしての一般化線形モデルの適用が可能であり，誤差分布を 2 項分布，リンク関数に（プロビット or ロジット or 補 2 重対数）を選択することもできる．リンク関数にロジットを選択した場合には，良く知られたロジスティック回帰と同じ結果となる．ただし，2 値データとしてロジスティック回帰を適用することには，疑問が残る．第 1 は，死亡の原因の一つである冠動脈心疾患の死亡率に対して上限の死亡率を 100%とするシグモイド曲線を仮定すること，第 2 に，1 パーセントに満たない死亡率について推定結果が得られても，そもそも解析する意義があるのか，結果の表示も極めて小さなパーセント表示では，妥当性に欠けると思われる．なお，第 5.6 節では，誤差分布を 2 項分布とした場合について示す．

分母 n_i が分かっている場合のポアソン回帰を次のように指数関数を使って定義し，

$$y_i = n_i \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i \qquad \varepsilon_i \sim \text{ポアソン分布}$$

推定値の形式にし，両辺に対数を取ると，

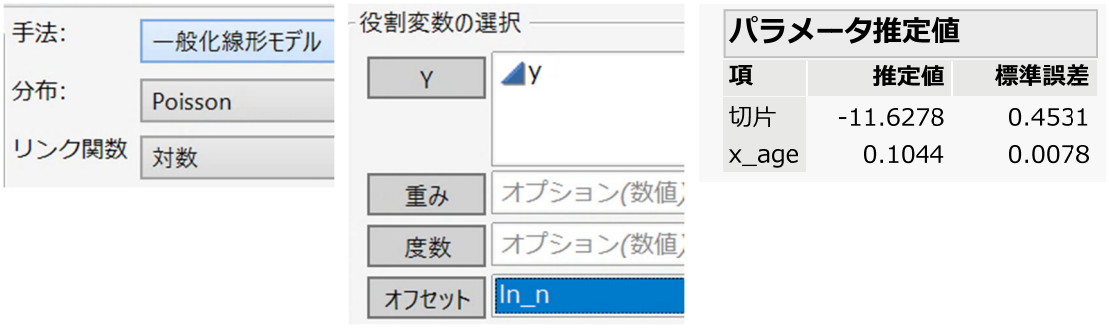
$$\ln(\hat{y}_i) = \ln n_i + \hat{\beta}_0 + \hat{\beta}_1 x_i$$

が得られる．この式で「$\ln n_i$」は，オフセットと言われている．この式で，未知パラメータは，切片 β_0 と傾き β_1 である．推定しようとしている切片 β_0 は，実際には，$(\ln n_i + \beta_0)$ なので β_0 から $\ln n_i$ だけ大きい方にずらした（オフセットした）切片と解釈される．オフセットがゼロとなるのは，$n_i = 1$ の場合である．

JMP でのオフセットの設定について表 1.12 に例示する．オフセットに設定するのは，元の分母 n_i ではなく，自然対数をあらかじめ計算して与える．JMP を用いたオフセットを含んだ予測式は，JMP の計算式の出力で

$$\text{Exp}\left(-11.62782676 + 0.1044379989 \bullet \text{x\_age} + \text{ln\_n}\right)$$

表 1.12　オフセットを用いたポアソン回帰でのパラメータ推定



パラメータ推定値

項	推定値	標準誤差
切片	-11.6278	0.4531
x_age	0.1044	0.0078

となる．一般の式に変換すれば，

$$\hat{y}_i = \exp(-11.6278 \ + \ 0.1044x_i + \ln n_i)$$
$$= n_i \exp(-11.6278 \ + \ 0.1044x_i)$$

となる．推定されたパラメータを用い，オフセットを含んだ予測値を計算した結果を表 1.13 に示す．Excel による解析方法は，第 2.6 節，第 5.5 節で詳細に例示し，第 5.6 節では，Excel の反復重み付き回帰による 2 値データの解析についても例示する．

表 1.13　オフセット含んだ予測式

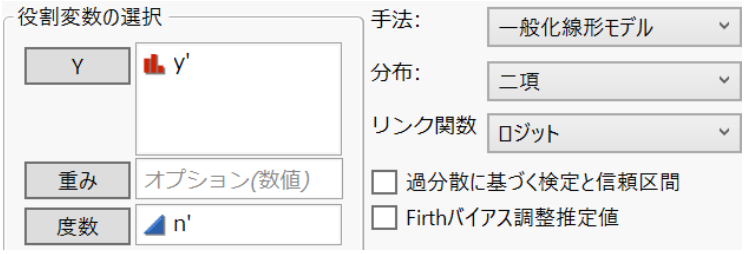
				β^$_0$=	**-11.6278**	
				β^$_1$=	**0.1044**	
	年齢層	死亡者数	母集団	オフセット		推定値
No.	x	y	人数 n	$\ln n$	β^$_0$	y^
1	30	1	17,742	9.7837	-1.8441	3.6
2	35	5	16,554	9.7144	-1.9134	5.7
3	40	5	16,059	9.6840	-1.9438	9.3
4	45	12	13,083	9.4791	-2.1487	12.8
5	50	25	10,784	9.2858	-2.3420	17.8
6	55	38	9,645	9.1742	-2.4536	26.9
7	60	54	10,706	9.2786	-2.3492	50.3
8	65	65	**9,933**	9.2036	-2.4242	**78.6**
オフセット	65	65	**1**	0.0000	-11.6278	**0.00791**

オフセットを 1 とした 65 歳の場合の推定値は，

$$\hat{y}_i = 1 \times \exp(-11.6279 \ + \ 0.1044 \times 65) = 0.00791 \text{ 人}$$

であり，9,933 人当たりでは，78.6 人となる．このように，オフセットを考慮した推定値は，年齢階層別の母集団の人数を考慮した推定値となっている．

JMP で 2 値データとして解析する際には，（n_i / y_i）形式はサポートされていないので，死亡の（あり：0，なし：1）に対し別変数 y_i' とし，人数 n_i' を（死亡あり：$n_i' = y_i$，死亡なし：$n_i' = n_i - y_i$）とするデータ変換が必要である．JMP の一般化線形モデルでは，「分布」を「二項」とし，「リンク関数」をロジットとし，人数 n' を「度数」とする．



2 値データとした場合に表 1.14 に示すような，推定結果を得る．ポアソン回帰での推定値とほぼ同様の結果が得られる．ただし，各種の出力結果は，死亡率のままなので，使い勝手はオフセットありのポアソン回帰に比べて良くない．

表 1.14　ロジスティック回帰を適用した場合の推定値

項	推定値	標準誤差
切片	-11.6395	0.4538
x	0.1047	0.0078

この推定値を用いた死亡率に対するロジスティック曲線は，図 1.3 左に示すように 160 歳でほぼ 100%になるようなシグモイド曲線となる．図 1.3 右に Y 軸の最大の目盛を 1.0%としたシグモイド曲線を示すが，図 1.2 左と同様なあてはめになっており，高年齢層に対する頭打ち現象を捉えていない．なお，第 2.6 節では，シグモイド曲線の上限をパラメータ化する方法を示す．

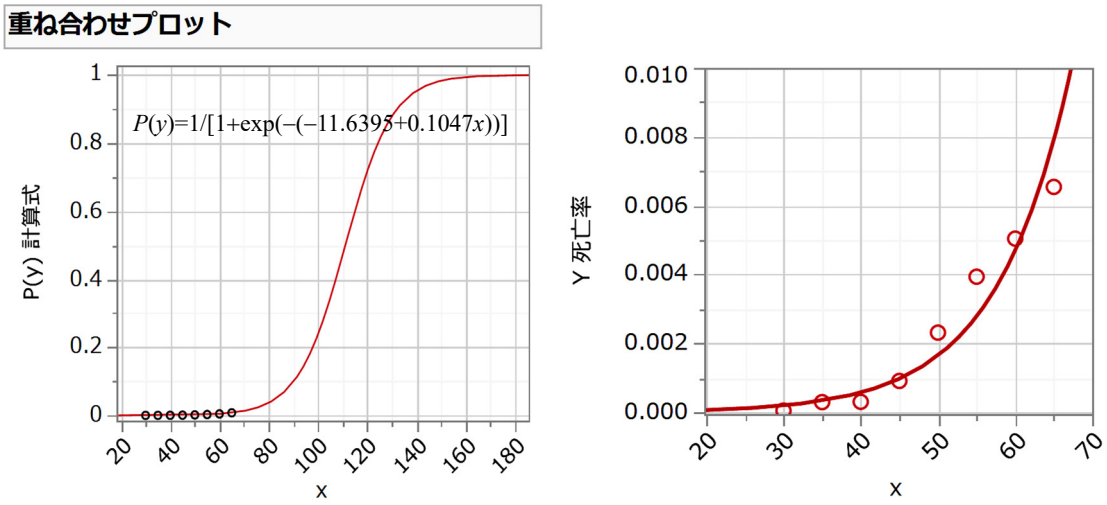


図 1.3 死亡率に対するロジスティック曲線のあてはめ

1.6. 満月と新月の日の犯罪件数に対する尤度比検定（2 群）

アルトマン著，木船・佐久間訳（1999)，「医学研究における実用統計学」の第 4.8 節に「満月と新月の日の犯罪件数」についての比較データが示されている．表 1.15 に示すように，このデータは，インドの 3 地域の 1978～1982 年の間の満月と新月の日の一日当たりの犯罪件数である．満月と新月の日のどちら日に犯罪が多いのかを比較することが目的である．

表 1.15　インドの 3 地域の 1978 年から 1982 年の間の 1 日当たりの犯罪件数

		満月の日 x=0				新月の日 x=1			
	犯罪件数	観測度数	ポアソン	期待値	残差	観測度数	ポアソン	期待値	残差
i	y	n	P	n^	n-n^	n	P	n^	n-n^
1	0	40	0.2469	45.2	-5.2	114	0.6033	112.2	1.8
2	1	64	**0.3453**	**63.2**	**0.8**	56	0.3049	56.7	-0.7
3	2	56	0.2416	44.2	11.8	11	0.0770	14.3	-3.3
4	3	19	0.1126	20.6	-1.6	4	0.0130	2.4	1.6
5	4	1	0.0394	7.2	-6.2	1	0.0016	0.3	0.7
6	5	2	0.0110	2.0	0.0	0	0.0002	0.0	0.0
7	6	0	0.0026	0.5	-0.5	0	0.0000	0.0	0.0
8	7	0	0.0005	0.1	-0.1	0	0.0000	0.0	0.0
9	8	0	0.0001	0.0	0.0	0	0.0000	0.0	0.0
10	9	1	0.0000	0.0	1.0	0	0.0000	0.0	0.0
合計		183		183.0		186		186.0	
平均		1.3989				0.5054			
分散		1.3620				0.5648			
分散/平均		0.9736				1.1177			

なお，平均と分散の比は，それぞれ 0.9736，1.1177 であり，ポアソン分布を仮定することが可能である．ポアソン分布の確率は，満月の日 $y_2^{(満月)}=1$ の場合であれば，Excel の関数で

$$\text{Poisson.dist}(y_2^{(満月)}, \text{平均}, \textit{false}) = \text{Poisson.dist}(1, 1.3989, \textit{false}) = 0.3453$$

で計算され，期待値 $\hat{n}_2^{(満月)}$ は，出現総数は，N=183 なので，

$$\hat{n}_2^{(満月)} = NP_2^{(満月)} = 183 \times 0.3453 = 63.2$$

となる．出現数は $n_2^{(満月)}=64$ なので，その差は，0.8 件である．図 1.4 にはボックス・プロッ

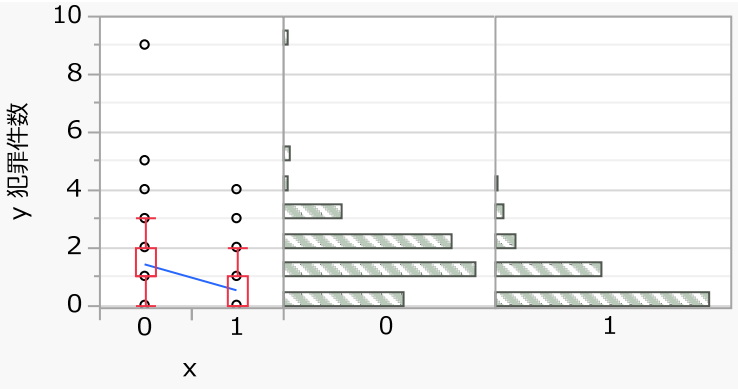


図 1.4　満月 (x=0) と新月 (x=1) の犯罪件数の比較

トと棒グラフの組み合わせた結果を示す．この図は，JMP の「二変量の関係」の「表示オプション」で「ヒストグラム」の選択で作成することができる．

順位和検定

満月の夜の平均犯罪件数は，1.3989 件，新月の夜は，0.5054 件となっていて，満月の夜は，2.8 倍の犯罪件数となっている．稀な現象の場合に 0 件の頻度が多くなるので，犯罪が起きたか否かで 2×2 の分割表にまとめて，カイ 2 乗検定を行えば簡単な検定で済ますこともできる．また，犯罪件数を順序尺度と見なして，表 1.16 に示すように順位和検定を便宜的に適用することも可能である．ただし，有意な差が有るか否かの単純な結果しか得られない．統計量としての「スコア平均」が，出力されているが，順位データの平均による検定であり，これによる考察は無意味である．

表 1.16　JMP の「二変量あてはめ」による順位和検定の便宜的な適用

Wilcoxon/Kruskal-Wallisの検定(順位和)					2標本検定(正規近似)		
水準	度数	スコア和	スコアの期待値	スコア平均	S	Z	p値(Prob>|Z|)
1:満月	183	42252.5	33855.0	230.888	42252.5	8.70452	<.0001*
2:新月	186	26012.5	34410.0	139.852			

JMP のポアソン回帰による 2 群間比較

反応変数がポアソン分布に従うのならば，その特性を生かしたポアソン回帰を統計ソフトで行うことも容易である．しかし，適用事例が身近にある教科書になければ，どのように適用したらよいのか，結果の解釈はどうしたらよいのか迷いが生じ，使用経験がある統計手法で済ませたくなってしまう．なお，2 値反応にまとめ直した場合の Pearson のカイ 2 乗検定，および，尤度比検定については，第 3.1 で詳細に示す．

反応変数がポアソン分布に従うことが確認できれば，表 1.17 に示すように満月を 0，新月を 1 とする説明変数を x として，ポアソン回帰により 2 群間の比較は容易にできる．表 1.18 は JMP の一般化線形モデルのポアソン回帰を用いた結果である．切片は，0 を与えた満月の日の犯罪件数の推定値 1.3989 となり，新月の日に 1 を与えたので “x” の推定値が，満月と新月の平均値差（傾き）が –0.8935 なので，新月の日の犯罪の平均は，1.3989–0.8935＝0.5054 と推定される．この傾き−0.8935 についての検定は，尤度比カイ 2 乗検定を用いて p<0.0001 となり，95%信頼区間は，プロファイル尤度（正確な信頼区間の算出法）により（−1.0968～−0.6969）と計算されている．

表 1.17　JMP のポアソン回帰モデル実行のためのデータセット

	満月_新月	x	y 犯罪件数	n
1	1:満月	0	0	40
2	1:満月	0	1	64
3	1:満月	0	2	56
4	1:満月	0	3	19
5	1:満月	0	4	1
6	1:満月	0	5	2
7	1:満月	0	9	1
8	2:新月	1	0	114
9	2:新月	1	1	56
10	2:新月	1	2	11
11	2:新月	1	3	4
12	2:新月	1	4	1

表 1.18　JMP のポアソン回帰を用いた 2 群間の尤度比による比較

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)	下側信頼限界	上側信頼限界
切片	1.3989	0.0874	256.0000	<.0001*	1.2345	1.5773
x	-0.8935	0.1018	80.5774	<.0001*	-1.0968	-0.6969

説明変数が 2 水準なのでリンク関数は恒等でも対数でも同じ結果となる.

最尤法を用いた 2 群間の比較をポアソン回帰により実施し，表 1.18 に示したように，満月と新月の犯罪件数の検定統計量として尤度比カイ 2 乗値が 80.5774 となっている．もちろん p<0.0001 と有意な差である．この尤度比検定は，満月と新月の犯罪件数の差についての対数用尤度のマイナス 2 倍の対数尤度が，自由度 1 のカイ 2 乗分布に従うとした場合である．

Excel による 2 群間の尤度比検定

表 1.19 に示すように Excel の計算シートで，実際に確認してみる．満月と新月の犯罪件数を加えた場合の平均は，$\hat{\mu}^{(満+新)} = 0.9485$ となり，$y_1 = 0$ の場合のポアソン確率は，

$$P_1^{(満+新)} = \text{Poisson.dist}(0,\ 0.9485,\ false) = 0.3873$$

と計算されている．対数尤度は，

$$\begin{aligned} \ln L_1^{(満+新)} &= n_1^{(満+新)} \times \ln(0.3873) \\ &= 154 \times (-0.9485) \\ &= -146.0705 \end{aligned}$$

であり，全体の対数尤度は，

$$\begin{aligned}\ln L^{(満+新)} &= \sum_{i=1}^{10} n_i^{(満+新)} \,\mathrm{Poisson.dist}(y_i^{(満+新)}, \hat{\mu}^{(満+新)}, \mathit{false}) \\ &= -146.0705 - 120.1648 - \cdots - 14.2261 \\ &= -484.8865\end{aligned}$$

となる．満月と新月それぞれ，

$$\ln L^{(満)} = \sum_{i=1}^{10} n_i^{(満)} \,\mathrm{Poisson.dist}(y_i^{(満)}, \hat{\mu}^{(満)}, \mathit{false}) = -268.4776$$

$$\ln L^{(新)} = \sum_{i=1}^{10} n_i^{(新)} \,\mathrm{Poisson.dist}(y_i^{(新)}, \hat{\mu}^{(新)}, \mathit{false}) = -176.1202$$

なので，

$$\begin{aligned}\chi^2 &= (-2\ln L^{(満+新)}) - [-2(\ln L^{(満)} + \ln L^{(新)})] \\ &= [-2\times(-484.8865)] - [2\times(-268.4776 - 176.1202)] \\ &= 80.5774\end{aligned}$$

と計算され，表 1.18 の JMP の尤度比カイ 2 乗に一致する．

表 1.19　Excel によるポアソン回帰を用いた尤度比検定

		満月の日 x=0			新月の日 x=1			満月＋新月		
i	y	$n_{満}$	$P_{満}$	$\ln L_{満}$	$n_{新}$	$P_{新}$	$\ln L_{新}$	$n_{満+新}$	$P_{満+新}$	$\ln L_{満+新}$
1	0	40	0.2469	-55.9563	114	0.6033	-57.6129	154	**0.3873**	**-146.0705**
2	1	64	0.3453	-68.0458	56	0.3049	-66.5184	120	0.3674	-120.1648
3	2	56	0.2416	-79.5576	11	0.0770	-28.1977	67	0.1742	-117.0747
4	3	19	0.1126	-41.4883	4	0.0130	-17.3780	23	0.0551	-66.6738
5	4	1	0.0394	-3.2342	1	0.0016	-6.4132	2	0.0131	-8.6760
6	5	2	0.0110	-9.0159	0	0.0002	0.0000	2	0.0025	-12.0006
7	6	0	0.0026	0.0000	0	0.0000	0.0000	0	0.0004	0.0000
8	7	0	0.0005	0.0000	0	0.0000	0.0000	0	0.0001	0.0000
9	8	0	0.0001	0.0000	0	0.0000	0.0000	0	0.0000	0.0000
10	9	1	0.0000	-11.1795	0	0.0000	0.0000	1	0.0000	-14.2261
合計		183		**-268.4776**	186		**-176.1202**	369		**-484.8865**
平均		1.3989			0.5054			**0.9485**		
分散		1.3620			0.5648			1.1577		
分散/平均		0.9736			1.1177			1.2205		
						合計	**-444.5978**		(-2)×差	**80.5774**

SAS/GENMOD のポアソン回帰による 2 群間比較

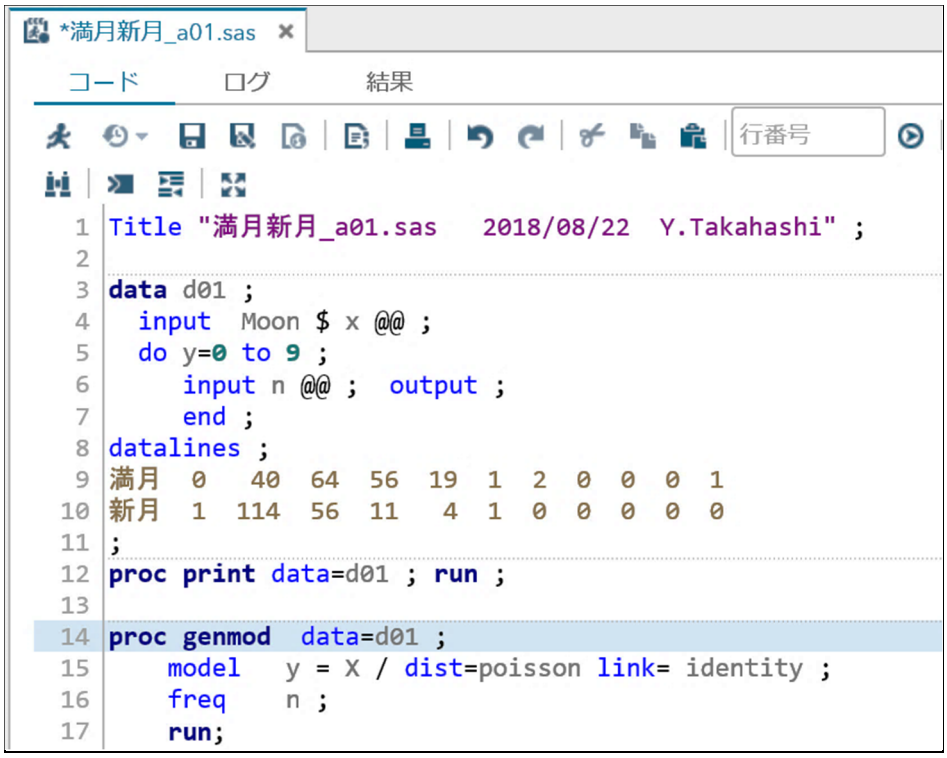
表 1.20 に SAS の GENMOD プロシジャを用いた結果を示す．JMP と同様に Intercept が $x=0$ を与えた満月の日の犯罪件数の推定値 $\hat{\beta}_0 = 1.3989$ となり，$\hat{\beta}_1 = -0.8935$ が $x=1$ を与えた新月の日と新月の日の犯罪件数の差の推定値になっている．傾きについての検定は，JMP

と異なり Wald カイ 2 乗=77.06 で，カイ 2 分布を用いて p<0.0001 となり，95%信頼区間は（−1.0930～−0.6940）と計算されている．

表 1.20　SAS/GENMOD によるポアソン回帰を用いた 2 群間の比較

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr>ChiSq
Intercept	1	1.3989	0.0874	1.2275	1.5703	256.00	<.0001
x	1	−0.8935	0.1018	−1.0930	−0.6940	77.06	<.0001
尺度	0	1.0000	0.0000	1.0000	1.0000		

表 1.21　無償版 SAS の GENMOD プロシジャ実行のための画面



ポアソン回帰を含む一般化線形モデルの計算方法は，2 通りある．JMP の場合は，第 2 章で示すように対数尤度関数をパラメータに関して偏微分した式を用いた方法である．他方，SAS の GENMOD プロシジャの場合は，第 1.4 節で示した反復重み付き回帰によって計算している．どちらの方法でも推定値は一致するが，パラメータの共分散行列には，違いがわずかに生ずる．

また，JMP では，カイ 2 乗統計量もマイナス 2 倍の対数尤度の差から求めた尤度比カイ 2 乗を標準的に計算するが，SAS の GENMOD プロシジャでは，推定値を標準誤差（*SE*）で割って 2 乗した，いわゆる Wald カイ 2 乗を計算している．

1.7. 細菌を用いた試験データ（2×2 要因配置）

吉村・大橋責任編集（1992），「毒性試験データの統計解析」の第 3.3.2 節に復帰突然変異試験（Ames 試験，試験法を確立した人の名前）に関するデータが示されている．TA1537 ネズミチフス菌株を用いた Ames 試験の陰性対照群を 50 枚のシャーレで繰り返して異常が起きたコロニーをカウントした結果である．溶媒として蒸留水 or DMSO（ジメチル スルホキシド，Dimethyl sulfoxide），代謝活性化の（なし or あり）有無を組み合わせた 4 通について 50 枚分シャーレ上で観察されたコロニー数が示されている．表 1.22 に 4 通りの場合について，コロニー数に関する度数分布，および，各種の統計量を計算した結果を示す．

表 1.22 ネズミチフス菌株に関するコロニー数

溶媒	蒸留水		DMOS	
代謝活性化	(−)なし	(＋)あり	(−)なし	(＋)あり
群，*n*，平均	1，50，14.54	2，50，7.54	3，50，12.48	4，50，8.28
平方和，分散	844.42，17.23	310.42，6.34	570.48，11.64	298.08，6.08
分散/平均	**1.19**	**0.84**	**0.93**	**0.73**
カイ2乗値, *p* 値	58.08，**0.176**	41.17，**0.779**	45.71，**0.607**	36.00，**0.917**
コロニー数 3	0	1	0	0
4	1	5	0	2
5	0	4	1	3
6	0	10	1	5
7	2	7	2	12
8	1	6	2	10
9	2	5	3	4
10	3	6	4	6
11	5	3	7	2
12	3	1	7	2
13	2	1	5	1
14	3	1	5	3
15	3	0	1	0
16	6	0	6	0
17	6	0	2	0
18	4	0	2	0
19	5	0	1	0
20	1	0	1	0
21	2	0	0	0
22	1	0	0	0

ポアソン分布のあてはめ

それぞれの分散を平均で割った比は，それぞれ，(1.19，0.84，0.93，0.73)となり，平均と分散が等しいポアソン分布と見なすことができそうである．「カイ 2 乗値, *p* 値」は，文献で示されている結果であるが，表 1.5「JMP による有害雑草の種子の数のポアソン分布のあては

め」で示した JMP の適合度検定の Pearson のカイ 2 乗検定の結果に一致する．この p 値は，(0.176，0.779，0.607，0.917) であり，ポアソン分布のあてはめは棄却されない．なお，吉村ら (1992) では，個別データが示されているが，ここでは，度数表の形でまとめ直している．

図 1.5 に JMP の「一変量の分布」を用いてヒストグラム上にポアソン分布を重ね書きした結果を示す．

群 1（蒸留水，代謝活性化－）の平均は 14.54，分散は 17.23，その比は 1.19 と 1 より大きく，分布に二つの山が見受けられるが，適合度の p 値は 0.176 であり，ポアソン分布のあてはめは棄却されない．

群 2（蒸留水，代謝活性化＋）の平均は 7.45，分散は 6.34，その比は 0.84 と 1 より小さく，右に裾を引く分布であり，適合度の p 値は 0.779 であり典型的なポアソン分布の形状である．

群 3（DMOS，代謝活性化－）の平均は 12.48，分散は 11.64 と同程度であり，その比は 0.93 と 1 より小さいが，適合度の p 値は 0.607 でありポアソン分布の前提がほぼ満たされている．

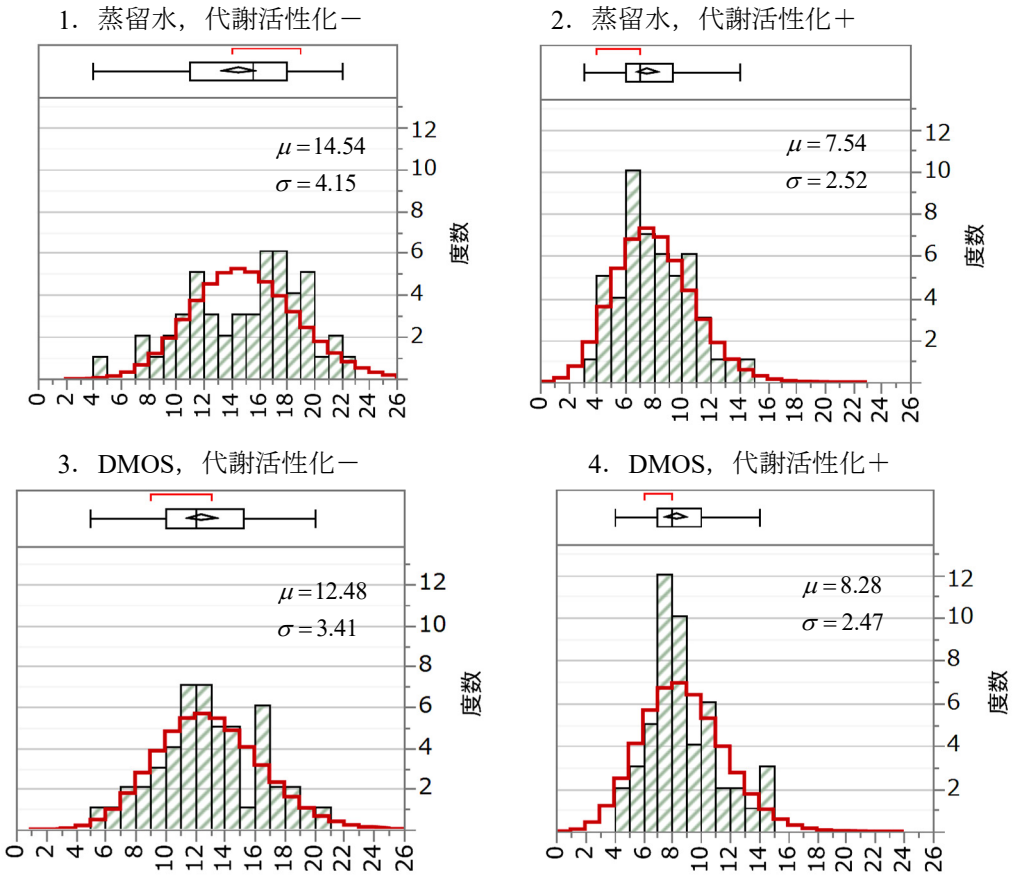


図 1.5　ネズミチフス菌株に関するコロニー数分布に対するポアソン分布のあてはめ

群 4（DMOS，代謝活性化＋）の平均は 8.28，分散は 6.08 と同程度であり，その比は 0.73 と 1 より小さいが，適合度の p 値は 0.917 でありポアソン分布の前提がほぼ満たされている．

正規分布のあてはめ

図 1.5 のみを見た場合には，正規分布を仮定することが可能のようにも思われるかもしれない．それぞれの群について正規性の Shapiro-Wilk の W 検定を行った結果を表 1.23 に示す．ポアソン分布の適合度の検定で第 4 群が $p = 0.917$ と限りなくポアソン分布に適合している場合に正規性が $p = 0.0090$ と疑われる結果となっている．他の群は，正規分布のあてはめもポアソン分布のあてはめも棄却できないとの玉虫色の結果となっている．

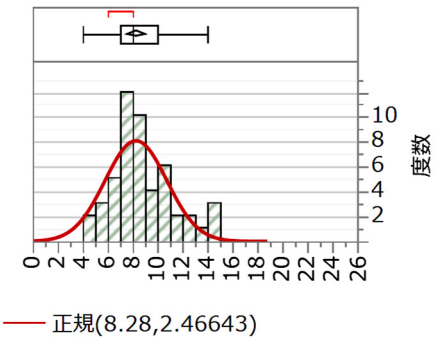
表 1.23　正規性の検定（Shapiro-Wilk の W 検定）

群	1	2	3	4
W 値	0.9687	0.9853	0.9652	0.9355
p 値	0.2054	0.7835	0.1472	0.0090
*				**

表 1.24　群 4 に対する正規性の適合度検定

一変量の分布 A:溶媒=A2, B:活性化=B2

y



正規のあてはめ

適合度検定

Shapiro-WilkのW検定

W	p値(Prob<W)
0.935516	0.0090*

注: Ho = 正規分布からのデータ。 p値が小さい場合はHoを棄却。

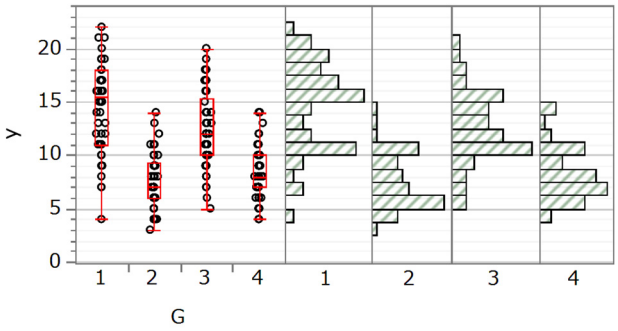
等分散性の検定

これら 4 群を典型的な 2×2 の要因配置実験と見なし，比較の前提となる等分散性について検討する．Bartlett の検定結果は，表 1.25 に示すように $F_3^1 = 6.2279$，$p = 0.0003$ であり，等分散性を仮定することはできない．

観測されたデータがどの様な分布になるかを見極めることは，統計解析の第一歩なのだが，ここに示したように，実際に観測したデータを基にした適合度の検定による推測は，玉虫色

表 1.25　Bartlett の検定

検定	F値	p値(Prob>F)
O'Brien[.5]	6.7607	0.0002*
Brown-Forsythe	6.4493	0.0003*
Levene	7.6019	<.0001*
Bartlett	6.2279	0.0003*



の結果で歯切れがわるい．検定ベースではなく対数尤度をベースにした推測が，この問題に新たな光明を与える．詳しくは，第 3.2 節で示す．

この実験は，被験物質に変異原性があるかの判定のために，各種の溶媒対照群の分布特性について検討する目的で行われたもので，最適条件を見つけるためのものではないが，分布がポアソン分布に従っている場合に，どのような統計解析が良いかを考えるためのデータとしても適している．

表 1.26 に各実験条件について，平均と分散を併記し，実験条件間の平均と差について示した．代謝活性剤が（1:なし）の場合に，（1:蒸留水）に対して（2:DMOS）は，コロニー数が 14.54 から 12.48 と 2.06 減少している．代謝活性剤が（2:あり）の場合には，（1:なし）の場合に比べてコロニー数が 14.54 から 7.54 と半減し，（1:蒸留水）に対して（2:DMOS）は，コロニー数が 7.54 から 8.28 と 0.74 増加している．分散も，平均値の増減を反映していることも確認される．さらなる解析については，第 7.1 節で取り上げる．

表 1.26　二元表による比較

	B:代謝活性化					
	1:なし		2:あり		全体	
A:溶媒	平均	分散	平均	分散	平均の平均	差
1:蒸留水	14.54	17.23	7.54	6.34	11.04	-7.00
2:DMOS	12.48	11.64	8.28	6.80	10.38	-4.20
平均の平均	13.51		7.91			
差	-2.06		0.74			

1.8. 細菌を用いた用量反応試験（恒等リンク，２群，7水準，効力比）

富山・杉本（2004）の「細菌を用いた用量反応試験データ」を用い，ポアソン回帰を適用する事例として取り上げる．表 1.27 に示すデータは，第 1.7 節と同様の Ames 試験の結果であり，陽性対照薬に対する代替物質の減弱の程度を確認することを目的とした実験である．薬物濃度を 7 段階に変化させ，各濃度あたり 3 プレートの変異コロニー数をカウントした結果である．

表 1.27　Ames 試験での変異コロニー数の比較

濃度		陽性対照 S					代替物質 T					
mg/plate	コロニー数			平均	分散	比	コロニー数			平均	分散	比
0	27	33	25	28.3	17.3	0.61	23	26	26	25.0	3.0	0.12
50	68	89	81	79.3	112.3	1.42	68	82	72	74.0	52.0	0.70
75	131	130	117	126.0	61.0	0.48	99	85	115	99.7	225.3	2.26
100	144	157	159	153.3	66.3	0.43	137	131	134	134.0	9.0	0.07
125	199	208	198	201.7	30.3	0.15	189	177	168	178.0	111.0	0.62
150	260	229	228	239.0	331.0	1.38	197	195	220	204.0	193.0	0.95
*300	427	407	456	430.0	607.0	1.41	335	332	348	338.3	72.3	0.21
					平均	0.84					平均	0.70

* この用量は，富山・杉本（2004）には含まれていない

用量ごとにプレートは 3 枚しかないが，分散/平均の比によるポアソン分布しているか検討する．陽性対照 S 薬の場合の比は，0.15～1.42，代替物質 T 薬の場合の比は 0.07～2.26，と大きなバラツキがある．全体で 14 水準の比の平均値は，0.77 でありポアソン分布的ではある．その 95%信頼区間は，(0.40～1.14) と 1 を含むので全体としては，ポアソン分布にほぼ従うものと見なせる．

図 1.6 は，S 薬と T 薬別々に最小 2 乗法による回帰分析を適用した結果で，傾きは異なるが，切片の違いはわずかであり，*dose* = 0 での変異コロニー数は，S 薬も T 薬も含まれていないので，切片を 2 群で共通で，傾きだけが異なる回帰直線のあてはめを行う．

このデータの用量ごとの変異コロニー数は，ポアソン分布に従うことが支持されているので，リンク関数を恒等としたポアソン回帰を行う．切片を共通とするためのデザイン行列に必要なダミー変数を設定する．詳細は，第 3.5 節「2 本の回帰直線に対するデザイン行列」を参照のこと．

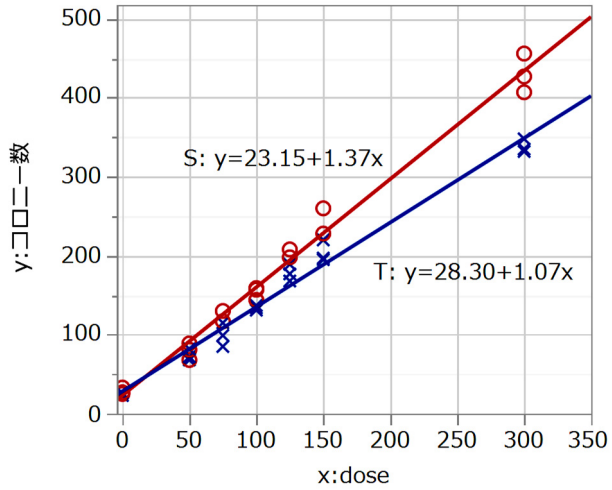


図 1.6　陽性対照薬 S および代替物質 T のコロニー数に対する線形回帰

SAS の GENMOD プロシジャによる解析のために SAS の DATA ステップを用いて SAS データセットを作成する．なお，DATA ステップの詳細は，第 9.3 節を参照のこと．陽性対照 S の場合に $x_S = 1$，それ以外は $x_S = 0$，代替物質 T の場合に $x_T = 1$，それ以外は $x_T = 0$ とする（1，1）型ダミー変数を用い，

$$\begin{matrix} x_S = 1 & x_T = 0 & \text{(S 薬の場合)} \\ = 0 \ , & = 1 & \text{(T 薬の場合)} \end{matrix}$$

解析モデルとして，次のポアソン回帰式を用いる．

$$y_i = \beta_0 + \beta_S x_S dose_i + \beta_T x_T dose_i + \varepsilon_i \qquad \varepsilon_i \sim poisson(y_i, \hat{y}_i)$$

SAS プログラム　変異コロニー_a01.sas

```
Title "変異コロニー_a01.sas  " ;

data d01 ;
  input  dose @@ ;
  xS=1 ;  xT=0 ;
  do k=1 to 3 ;
     input y @@ ; output ;  end ;
  xS=0 ;  xT=1 ;
  do k=1 to 3 ;
     input y @@ ; output ;  end ;
datalines ;
  0    27  33  25    23  26  26
 50    68  89  81    68  82  72
 75   131 130 117    99  85 115
100   144 157 159   137 131 134
125   199 208 198   189 177 168
150   260 229 228   197 195 220
300   427 407 456   335 332 348
;
proc print data=d01 ; run ;

proc genmod  data=d01 ;
    model   y = xS*dose xT*dose /
        dist=poisson link= identity covb ;
    run;
```

表 1.28 および表 1.28 に SAS の GENMOD プロシジャによるリンク関数を恒等リンク（identity）としたポアソン回帰の結果を示す.

表 1.28　切片を共通にする S 薬と T 薬の傾きのポアソン回帰による比較

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept	1	25.0407	1.7813	21.5493	28.5320	197.60	<.0001
xS*dose	1	1.3521	0.0282	1.2968	1.4074	2295.08	<.0001
dose*xT	1	1.0980	0.0263	1.0464	1.1496	1739.59	<.0001
尺度	0	1.0000	0.0000	1.0000	1.0000		

表 1.29　推定されたパラメータの共分散行列

推定値の共分散行列			
	Prm1	Prm2	Prm3
Prm1	3.17318	−0.02262	−0.02248
Prm2	−0.02262	0.0006931	0.0001602
Prm3	−0.02248	0.0001602	0.0007966

陽性対照 S 薬の傾きは $\hat{\beta}_S = 1.3524$，代替物質 T 薬の傾きは $\hat{\beta}_T = 1.0980$ なので，効力比 $\hat{\rho}$ は，

$$\hat{\rho} = \frac{\hat{\beta}_T}{\hat{\beta}_S} = \frac{1.0980}{1.3521} = 0.8121$$

となる．95%信頼区間を求めるために，効力比 ρ を β_T と β_S で偏微分し，共分散行列を用いて，2 次形式によるデルタ法で効力比 $\hat{\rho}$ の分散を計算する．

$$d_1 = \frac{\partial \rho}{\partial \beta_T} = \frac{1}{\beta_S} = \frac{1}{\hat{\beta}_S} = \frac{1}{1.3521} = 0.7396$$

$$d_2 = \frac{\partial \rho}{\partial \beta_S} = \frac{-\beta_T}{\beta_S{}^2} = \frac{-\hat{\beta}_T}{\hat{\beta}_S{}^2} = \frac{-1.0980}{1.3521^2} = -0.6006$$

デルタ法は，推定されたパラメータの比などの分散を計算するための汎用的な統計手法なので，第 4.6 節に詳しく示すが，比についての偏微分および行列計算が必要であり，きちっと説明されている日本語の成書は見当たらない．そこで，偏微分の結果を縦ベクトル $[d_1 \quad d_2]^T$ とし，SAS/GENMOD で得られた共分散行列を用いて，効力比 $\hat{\rho}$ の分散を求め，効力比 $\hat{\rho}$ の 95%信頼区間を Excel シート上での計算する方法を表 1.30 に示す.

効力比 $\hat{\rho}$ の分散は，

$$Var(\hat{\rho}) = \boldsymbol{d}^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{d} = 0.0005$$
$$SE = \sqrt{0.0005} = 0.0229$$

から，

$$\begin{aligned} 95\%CL &= \rho \pm 1.96SE \\ &= 0.8121 \pm (1.96 \times 0.0229) \\ &= (0.7672, 0.8569) \end{aligned}$$

が得られる．結果の解釈は，効力比 $\hat{\rho} = 0.8121$ であり，効力比の 95%信頼区間が 1.0 を含まないので，統計的に有意な差である．なお，効力比は，少ない用量で同等の効果を示す場合は，1 以上であり，多い用量で同等の効果を示す場合は，1 未満である．代替物質 T 薬は，陽性対照 S 薬に比べて同程度の変異原起こすための用量は，$1/\hat{\rho} = 1/0.8121 = 1.23$ 倍が必要であり，変異原性の観点からでは望ましい結果となっている．効力比の推定と応用については，第 8 章で詳しく解説し，本節の SAS の GENMOD プロシジャによるポアソン回帰に引き続き第 8.3 節では，Excel による詳細な解説が示されている．

表 1.30　効力比 ρ の推定およびデルタ法による分散から 95%信頼区間の計算シート

		推定値		Prm1	Prm2	Prm3	二次形式による ρ の分散の計算式				
Intercept		25.0407	Prm1	3.1732	−0.0226	−0.0225		$\boldsymbol{d}^T$	$\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})$		$\boldsymbol{d}$
xS*dose	β_S	1.3521	Prm2	−0.0226	0.0007	0.0002	0.7396	-0.6006	0.0007	0.0002	0.7396
dose*xT	β_T	1.0980	Prm3	−0.0225	0.0002	0.0008			0.0002	0.0008	-0.6006
			パラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})$								
		$\boldsymbol{d}$						$\beta\hat{}_T/\beta\hat{}_S$	95%L	95%U	
$1/\beta\hat{}_S$ =		0.7396	$\rho\hat{}$ の分散 $\boldsymbol{d}^T\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})\boldsymbol{d}$ =			0.0005	$\rho\hat{}$=	0.8121	0.7672	0.8569	
$\beta\hat{}_T/\beta\hat{}_S^2$ =		-0.6006			SE=	0.0229		95%CL=	$\rho\hat{}\pm 1.96\ SE$		

$\rho\hat{}$の分散　$\boldsymbol{d}^T\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})\boldsymbol{d}$ = Mmult (Mmult （ $\boldsymbol{d}^T$, $\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})$, $\boldsymbol{d}$)

1.9. 植物の体サイズに関連した種子数（対数リンク，2 群，回帰）

久保（2012），「データ解析のための統計モデリング入門，一般化線形モデル・階層ベイズモデル・MCMC」の第 3 章「一般化線形モデル －ポアソン回帰－」で取り上げられている人工データ（data3a.csv）を用いる．表 1.31 に施肥処理（なし，あり）の 2 群に対して，それぞれ 50 個体について体サイズ x_i に対する植物の種子数 y_i が示されている．

表 1.31 植物の体サイズと種子数のデータリスト

	C: 施肥処理なし					T: 施肥処理あり				
No	x	y	No	x	y	No	x	y	No	x	y
1	8.31	6	26	10.21	6	51	10.14	14	76	10.24	6
2	9.44	6	27	9.45	7	52	9.05	6	77	11.76	8
3	9.50	6	28	10.44	9	53	9.89	7	78	9.52	9
4	9.07	12	29	9.44	3	54	8.76	9	79	10.40	9
5	10.16	10	30	10.48	10	55	12.04	6	80	9.96	6
6	8.32	4	31	9.43	2	56	9.91	7	81	10.30	7
7	10.61	9	32	10.32	9	57	9.84	9	82	11.54	10
8	10.06	9	33	10.33	10	58	11.87	13	83	9.42	6
9	9.93	9	34	8.50	5	59	10.16	9	84	11.28	11
10	10.43	11	35	9.41	11	60	9.34	13	85	9.73	11
11	10.36	6	36	8.96	10	61	10.17	7	86	10.78	11
12	10.15	10	37	9.67	4	62	10.99	8	87	10.21	5
13	10.92	6	38	10.26	8	63	9.19	10	88	10.51	6
14	8.85	10	39	10.36	9	64	10.67	7	89	10.73	4
15	9.42	11	40	11.80	12	65	10.96	12	90	8.85	5
16	11.11	8	41	10.94	8	66	10.55	6	91	11.20	6
17	8.02	3	42	10.25	9	67	9.69	15	92	9.86	5
18	11.93	8	43	8.74	8	68	10.91	3	93	11.54	8
19	8.55	5	44	10.46	6	69	9.60	4	94	10.03	5
20	7.19	5	45	9.37	6	70	12.37	6	95	11.88	9
21	9.83	4	46	9.74	10	71	10.54	10	96	9.15	8
22	10.79	11	47	8.95	10	72	11.30	8	97	8.52	6
23	8.89	5	48	8.74	9	73	12.40	8	98	10.24	8
24	10.09	10	49	11.32	12	74	10.18	7	99	10.86	7
25	11.63	6	50	9.25	6	75	9.53	5	100	9.97	9

久保（2012）には，R の glm() でリンク関数に対数を設定したポアソン回帰の結果が示されている．

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.2917     0.3637    3.55  0.00038
x              0.0757     0.0356    2.13  0.03358
```

久保（2012），P50 より引用

データの吟味

このデータについて，施肥処理の（なし，あり）別に基本統計量および相関係数を表 1.32 に示す．体サイズ x は，平均が 10 前後に対し分散は 1 前後と小さく，種子数 y は，平均が 8 個弱に対し分散は 7 前後であり，ポアソン分布であるための条件を満たしている．体サイズ x と種子数 y の相関は，施肥処理が「ない」場合には，$r=0.39$ と弱い相関関係であるが，施肥処理が「ある」場合には，$r=0.07$ と無相関に近い．

表 1.32 植物の体サイズ x と種子数 y の基本統計量

		x 体サイズ			y 種子数				x vs. y
施肥処理	N	平均	分散	標準偏差	平均	分散	標準偏差	分散/平均	相関係数
C なし	50	9.8076	0.9996	0.9998	7.7800	6.8690	2.6209	0.88	0.39
T あり	50	10.3706	0.8917	0.9443	7.8800	7.0465	2.6545	0.89	0.07
全体	100	10.0891	1.0162	1.0080	7.8300	6.8900	2.6249	0.88	0.23

群間の t 検定をした結果を図 1.7 に示す．体サイズ x については，施肥処理あり群で有意に大きくなっているが，種子数 y については，全く差がないとの結果である．

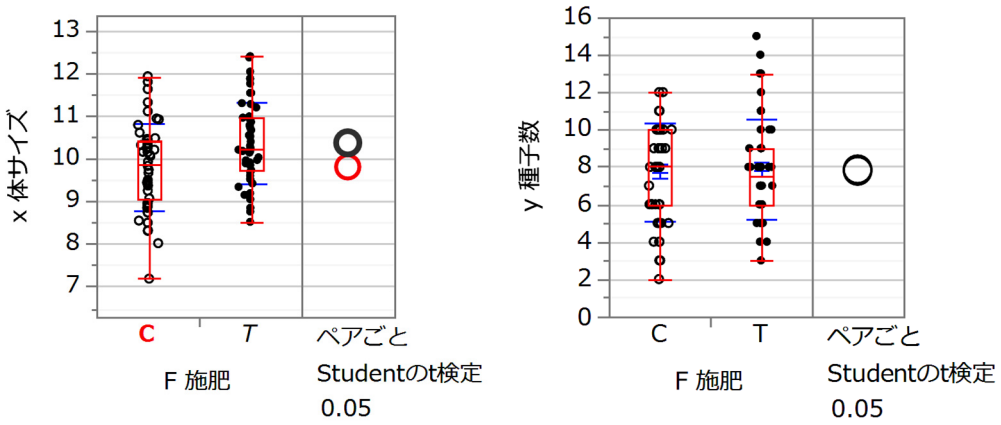


図 1.7　植物の体サイズと種子数について施肥処理の比較

図 1.8 に x と y の散布図上に 95%の確率楕円を上書きした結果を示す．施肥処理が「ある」場合に種子数 y は，体サイズ x との相関関係が見いだせない．施肥処理が「ない」場合には，弱い相関があることから，施肥処理により，体サイズが大きくなったが，それに伴い種子数が増えることはなかったと解される．したがって，x を説明変数とするポアソン回帰は，施肥処理が「ない」場合には意味があるが，施肥処理が「ある」場合にはポアソン回帰を行う必然性はない．

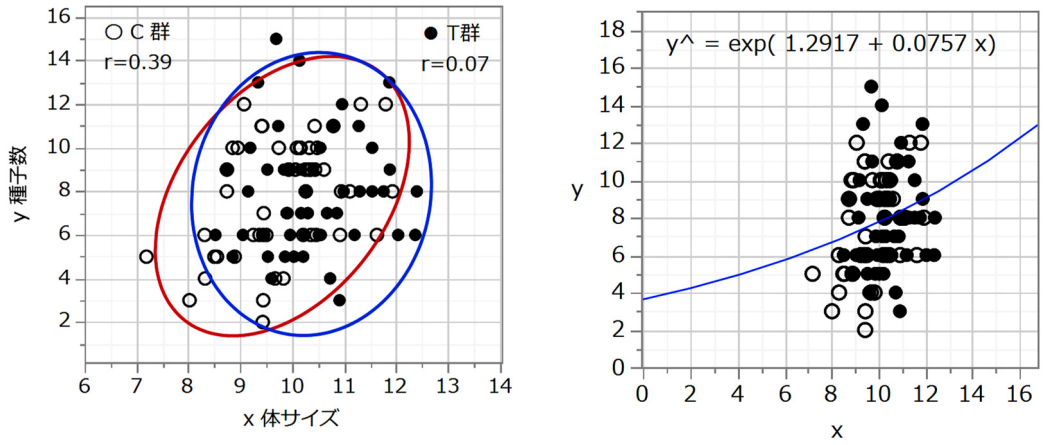


図 1.8　植物の体サイズと種子数の散布図と 95%確率楕円およびポアソン回帰

尤度比検定

施肥処理を込みにした JMP によるポアソン回帰の結果を表 1.33 に示す．対数リンクによるポアソン回帰式は，

$$\log(\hat{y}) = 1.2917 + 0.0757x$$

$$\hat{y} = \exp(1.2917 + 0.0757x)$$

$$\text{Y 切片：}\ \hat{y}_0 = \exp(1.2917) = 3.6390$$

となる．傾き $\hat{\beta}_1 = 0.0757$ に対する尤度比カイ 2 乗=4.5139 であり，p 値=0.0336 と有意な差となっている．この p 値は，「モデル全体の検定」での「差分」の行の「尤度比カイ 2 乗」4.5139 に対する p 値に等しい．

表 1.33 施肥処理を込みにしたポアソン回帰（対数リンク）の結果

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(P>ChiSq)
差分	2.2570	4.5139	1	0.0336*
完全	235.3863			
縮小	237.6432			

適合度統計量	カイ2乗	自由度	p値(P>ChiSq)
Pearson	83.8448	98	0.8452
デビアンス	84.9930	98	0.8226

AICc
474.8962

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(P>ChiSq)
切片	1.2917	0.3637	12.4670	0.0004*
x 体サイズ	0.0757	0.0356	4.5139	0.0336*

「モデル全体の検定」には，各種の尤度比カイ 2 乗値が示されている．「差分」，「完全」「縮小」の意味付けについては，表 1.34 に示すように，3 種類のポアソン回帰の対数尤度 $\ln L$ が必要となる．

モデル	推定式	対数尤度
縮小モデル：	$\hat{y}_i = \exp(\hat{\beta}_0)$	$\ln L^{縮小} = \sum_i \ln(\text{Poisson.dist}(y_i, \exp(\hat{\beta}_0), false))$
完全モデル：	$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$	$\ln L^{完全} = \sum_i \ln(\text{Poisson.dist}(y_i, \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i), false))$
飽和モデル：	$\hat{y}_i = y_i$	$\ln L^{飽和} = \sum_i \ln(\text{Poisson.dist}(y_i, y_i, false))$

縮小モデルは，X 軸に平行な直線のあてはめであり，完全モデルは，図 1.8 右に示した $\exp(1.2917 + 0.0757x)$ のあてはめで，飽和モデルは，100 個の種子数 y_i それ自体を推定値にしたモデルである．それぞれの対数尤度は，パラメータ数が多くなれば大きくなるのであるが，それぞれの対数尤度の差の 2 倍が尤度比カイ 2 乗値となり，それぞれのパラメータ数の差を自由度とするカイ 2 乗分布に従うとして，p 値を算出する．表 1.34 に Excel による各モデルに対する対数尤度の計算結果を示す．

表 1.34 各種のモデルに対する Excel ソルバーを用いたポアソン回帰の結果

				縮小モデル		**完全モデル**		**飽和モデル**	
				y^=exp(β_0^)		y^=exp(β_0^+β_1^x)		y^=y	
				β_0^=	2.0580	β_0^=	1.2917	パラメータ	100 個
						β_1^=	0.0757		
				$\ln L^{縮小}$=	**-237.6432**	$\ln L^{完全}$=	**-235.3863**	$\ln L^{飽和}$=	**-192.8898**
	施肥	体サイズ	種子数	確率	対数尤度	確率	対数尤度	確率	対数尤度
i	F	x	y	P	$\ln(P)$	P	$\ln(P)$	P	$\ln(P)$
1	C	8.31	6	0.1273	-2.0615	0.1525	-1.8806	0.1606	-1.8287
2	C	9.44	6	0.1273	-2.0615	0.1385	-1.9767	0.1606	-1.8287
3	C	9.50	6	0.1273	-2.0615	0.1376	-1.9833	0.1606	-1.8287
4	C	9.07	12	0.0441	-3.1217	0.0308	-3.4796	0.1144	-2.1683
5	C	10.16	10	0.0949	-2.3548	0.0954	-2.3494	0.1251	-2.0786
:									
98	T	10.24	8	0.1393	-1.9709	0.1395	-1.9697	0.1396	-1.9691
99	T	10.86	7	0.1424	-1.9494	0.1343	-2.0077	0.1490	-1.9038
100	T	9.97	9	0.1212	-2.1102	0.1195	-2.1246	0.1318	-2.0268

表 1.33 の「モデル」の欄の

「完全」は $-\ln L^{完全} = -(-235.3863) = 235.3863$,

「縮小」は $-\ln L^{縮小} = -(-237.6432) = 237.6432$,

「差分」は $(\ln L^{完全} - \ln L^{縮小}) = -235.3863 - (-237.6432) = 2.2570$

で計算されている．「差分」の 2 倍が「尤度比カイ 2 乗」となっている．「適合度統計量」の「デビアンス」は，「逸脱度」ともいわれ，カイ 2 乗値は，

$$\begin{aligned}\text{デビアンス}&=2\times(\ln L^{\text{飽和}}-\ln L^{\text{完全}})\\&=2\times[-192.8898-(-235.3863)]\\&=2\times 42.4965=84.9930\end{aligned}$$

で計算されている．$\ln L^{\text{飽和}}$ の自由度は 100，$\ln L^{\text{完全}}$ は 2 なので，デビアンスの自由度は，$100-2=98$ となり，デビアンスのカイ 2 乗値=84.9930 は，自由度 98 のカイ 2 乗分布に従うことから，

$$p=1-\text{Chisq.dist}\left(84.9930,98,true\right)=0.8226$$

と，有意な差ではない．このことは，パラメータ数が 100 個のモデルに対し，パラメータ数が 2 個の回帰モデルは，あてはめとして悪くはないことが示されてる．

表 1.33 の適合度統計量」の「Pearson」のカイ 2 乗値は，通常の Pearson 残差で，

$$\chi^2_{Pearson}=\sum_i\frac{[y_i-\exp(1.2917+0.0757\times x_i)]^2}{\exp(1.2917+0.0757\times x_i)}=83.8448$$

として計算されたものである．

回帰式の妥当性

ポアソン回帰で 1 次式でのあてはめで十分であることは，2 次式をあてはめて，1 次式と 2 次式の対数尤度の差の 2 倍で評価できる．表 1.34 に更に 2 次式を加えた結果から，

回帰モデル・1 次式：　$\hat{y}_i=\exp(\hat{\beta}_0+\hat{\beta}_1x_i)$　　$\ln L^{\text{縮小}}=-235.3863$

回帰モデル・2 次式：　$\hat{y}_i=\exp(\hat{\beta}_0+\hat{\beta}_1x_i+\hat{\beta}_2x_i^{\,2})$　　$\ln L^{\text{完全}}=-234.3971$

$2\times(\ln L^{\text{完全}}-\ln L^{\text{縮小}})=2\times 0.9892=1.9784$

が求められ，2 次式のあてはめの必要性はなく，1 次式で十分と言える．良い統計モデルの選択の基準には，AIC による選択が標準的であるが，ここでは割愛する．

個別の 95%信頼区間

施肥処理がある T 群の場合に体サイズ x との相関がないということは，ポアソン回帰を行う前提を満たしていないので，C 群の場合に限定し，対数リンクによるポアソン回帰を行い，回帰式の 95%信頼区間，個別データの 95%信頼区間を JMP ファイルに出力し，「重ね合わせプロット」を用い，ポアソン回帰直線の 95%信頼区間および個別の 95%信頼区間を散布図上に上書きした結果を表 1.35 に示す．

パラメータの推定値は，対数変換した場合の結果なので，実目盛り上では，指数を取って

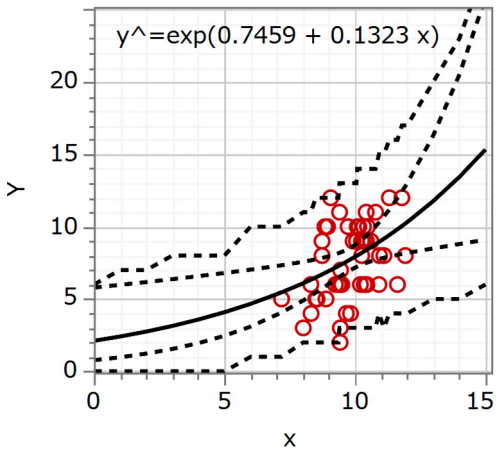
$$\hat{y}_i = \exp(0.7459 + 0.1323 x_i)$$

となる．切片は，$\hat{y}_0 = \exp(0.7459) = 2.1083$ であることが，確認される．また，個別データの 95%信頼区間が，ほぼ全データを包含しているので，適切なポアソン回帰となっていることが裏付けられる．なお，Excel による 95%信頼区間の求め方については，第 4.5 節，第 5.4 節を参照してもらいたい．

表 1.35　C 群に対するポアソン回帰のあてはめと 95%信頼区間および予測区間

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗
切片	0.7459	0.5154	2.0730
x	0.1323	0.0516	6.5990



個別データの 95%信頼区間（予測区間）の中に，ほとんど全ての観測データが含まれており，対数リンクのポアソン回帰のあてはめの妥当性が示されている．また，図 1.9 に示した「スチューデント化デビアンス残差」も，ほとんどが（−2～＋2）の範囲に入っていることからも，あてはめの妥当性が示されている．なお，各種の残差については，第 11 章を参照のこと．

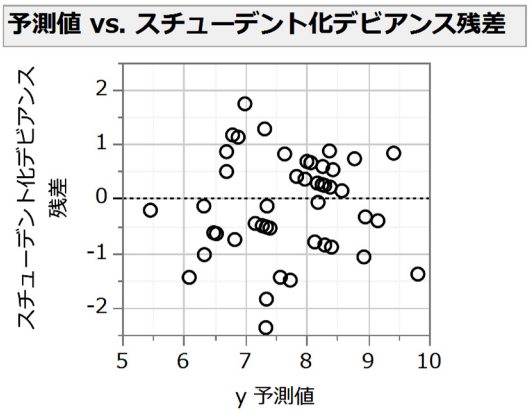


図 1.9　スチューデント化デビアンス残差プロット

1.10. 退役軍人における癌の発生（対数リンク，2 群，11 水準，オフセット）

アーミテージ・ベリー著，椿・椿共訳（2001)，「医学研究のための統計的方法」の第 12.8 節の退役軍人の癌の発生数に関するデータを表 1.36 に示す．このデータは，20 年の期間にわたって追跡された退役軍人に対して，実戦経験が（なし，あり）で，癌の発生数を調べたものである．

各軍人は，20 年の追跡期間中に，いくつか年齢階層に重複して記録されている．研究は実戦経験のない群と実戦経験のある群間の発癌リスクに差が存在するか否かを評価するために行われた．アーミテージら（2001）では，年齢階層を名義尺度として取り扱っているが，ここでは，連続尺度として扱う．

表 1.36　退役軍人の実践経験の有無での癌の発生数の比較

			実戦経験なし:x=0				実戦経験あり:x=1		
age	年齢階層	i	癌の数 y	人年 n	10万人比	i	癌の数 y	人年 n	10万人比
20	-24	1	18	208,487	8.6	12	6	60,840	9.9
25	25-29	2	60	303,832	19.7	13	21	157,175	13.4
30	30-34	3	122	325,421	37.5	14	54	176,134	30.7
35	35-39	4	191	312,242	61.2	15	118	186,514	63.3
40	40-44	5	108	165,597	65.2	16	97	135,475	71.6
45	45-49	6	88	54,396	161.8	17	58	42,620	136.1
50	50-54	7	74	40,716	181.7	18	56	25,001	224.0
55	55-59	8	120	33,801	355.0	19	54	13,710	393.9
60	60-64	9	141	26,618	529.7	20	34	6,163	551.7
65	65-69	10	108	17,404	620.5	21	9	1,575	571.4
70	70-	11	99	14,146	699.8	22	2	273	732.6
	全体		1129	1,502,660	75.1		509	805,480	63.2

各年代の人年数が大きく異なり，癌の発現数 y_i の加齢に伴なう変化が見えにくいので，10 万人比に換算した結果を加えてある. 年齢別の実践経験の有無による癌の発生数には, 図 1.10 に示すように差がないように判断される.

人年 n_i について対数オフセットを $\ln(n_i)$ とし，リンク関数を対数とした場合のポアソン回帰を行い，表 1.37 に示す結果を得た.

$$\begin{aligned}\hat{y}_i &= n_i \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i + \beta_2 age_i) \\ &= n_i \exp(-10.5316 + 0.0362 x_i + 0.0851 age_i)\end{aligned}$$

$$\begin{aligned}\ln(\hat{y}_i) &= \ln(n_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i + \beta_2 age_i) \\ &= \ln(n_i) + (-10.5316 + 0.0362 x_i + 0.0851 age_i)\end{aligned}$$

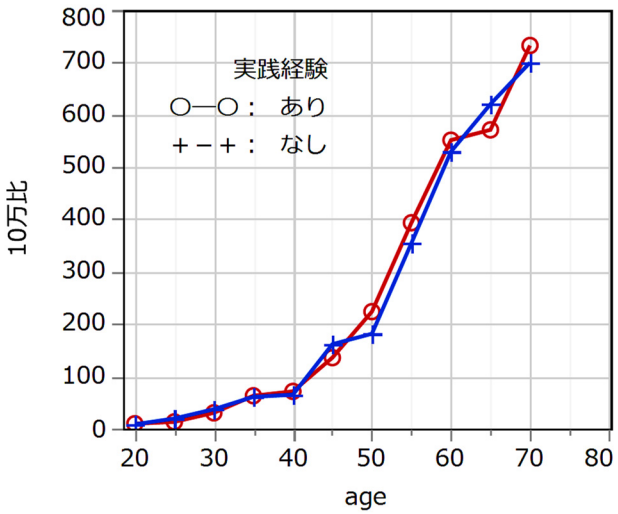


図 1.10　実践経験のなし・あり別 10 万人比での癌の発生数

パラメータの推定結果は，表 1.37 に示すように，実践経験の有無 x の尤度比カイ 2 乗値が 0.4367 であり，有意な差とはならなかった．オフセットの解釈については第 2.6 節を参照してもらいたい．

表 1.37　JMP のポアソン回帰による退役軍人の癌の発生ケースの比較

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(P>ChiSq)	下側信頼限界	上側信頼限界
切片	-10.5316	0.0899	18144.850	<.0001*	-10.7090	-10.3567
実践経験x	0.0363	0.0547	0.4367	0.5087	-0.0717	0.1430
age	0.0851	0.0018	1981.2547	<.0001*	0.0816	0.0886

オフセットを除いた推定値は，部分集団が 1 人とした場合の推定値であり，10 万人比などのように部分集団の大きさを固定して比較検討することが必要である．図 1.11 に，実践経験の（なし，あり）別に，10 万人比に換算した 10 万人比での癌の発生数にポアソン回帰で推定した回帰直線を上書きした結果を示す．

実践経験なしで，年齢が 60 歳での癌の発現数の 10 万人比での推定値は，

$$\text{実践経験なし}\ \ x_9=0\ \ :\ \begin{cases} \hat{y}_9^{(10万人比)} = 100{,}000\times\exp(\hat{\beta}_0+\hat{\beta}_1 x_9+\hat{\beta}_2 age_9) \\ \qquad\qquad = 100{,}000\times\exp(-10.5316+0.0362\times 0+0.0851\times 60) \\ \qquad\qquad = 440.9 \end{cases}$$

実践経験がある場合には，

$$\text{実践経験あり}\quad x_{20}=1\quad:\quad\begin{cases}\hat{y}_{20}^{(10\text{万人比})}=100{,}000\times\exp(\hat{\beta}_0+\hat{\beta}_1x_{20}+\hat{\beta}_2age_{20})\\ \qquad\qquad=100{,}000\times\exp(-10.5316+0.0362\times1+0.0851\times60)\\ \qquad\qquad=457.2\end{cases}$$

であり，その差は，$457.2-440.9=16.3$ 万人比とわずかである．

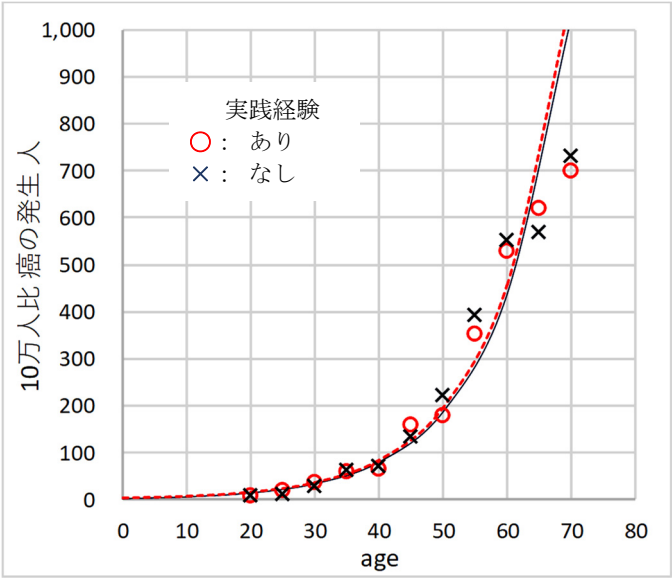


図 1.11　実践経験のありなし別の 10 万人比での癌の発生数と推定値

加齢と共に 10 万人比の癌の発生数は，指数関数的に増大するのであるが，60 歳を超えるあたりから頭打ちになっている．したがって，実践経験の有無による統計的に差は検出されなかったが，対数リンクのポアソン回帰をあてはめることに無理がある．どのような統計モデルが適切なのか，さらなる検討が必要である．第 3.6 節，第 5.4 節，第 12.5 節，および，12.7 節に 2 次式のあてはめの事例が示されている．

1.11.　喫煙による冠動脈心疾患による死亡（対数リンク，2 群，5 水準，オフセット）

ドブソン（2008）の第 9.2 節に，英国の男性医師を対象に喫煙習慣が 10 年間の間に冠動脈心疾患の死亡に対する影響を調べてた結果があり，表 1.38 にデータを示す．このデータは，前節と同様に対数リンクでオフセットがある事例である．

表 1.38　年齢階層別の喫煙習慣と冠動脈心疾患による死亡数の関係

年齢層		非喫煙者（$x=0$）			喫煙者（$x=1$）		
歳	範囲	死亡	人年	10万人比	死亡	人年	10万人比
40	35-44	2	18,790	10.6	32	52,407	61.1
50	45-54	12	10,673	112.4	104	43,248	240.5
60	55-64	28	5,710	490.4	206	28,612	720.0
70	65-74	28	2,585	1083.2	186	12,663	1468.8
80	75-84	31	1,462	2120.4	102	5,317	1918.4

各年齢層の部分母集団の人年が喫煙者と非喫煙者で大きく異なるので，10 万人比に換算した結果を図 1.12 に示す．年齢が 70 歳代までは，喫煙者の死亡数が多いが，80 歳代でやや逆転している．喫煙者で 80 歳を超えて生存しているの人達は，タバコに対して強い耐性を持った人達と推測され，非喫煙者と同様の冠動脈心疾患による死亡数となっている．

ドブソン（2008）では，対数リンクでのポアソン回帰に年齢の 2 乗の項，年齢と喫煙の交互作用を含めたモデルが示されているが，結果の解釈が複雑になるので，ここでは，80 歳代を除いたモデルする．ポアソン回帰の解析は，これまで，第 1.4 節では，Excel による反復重み

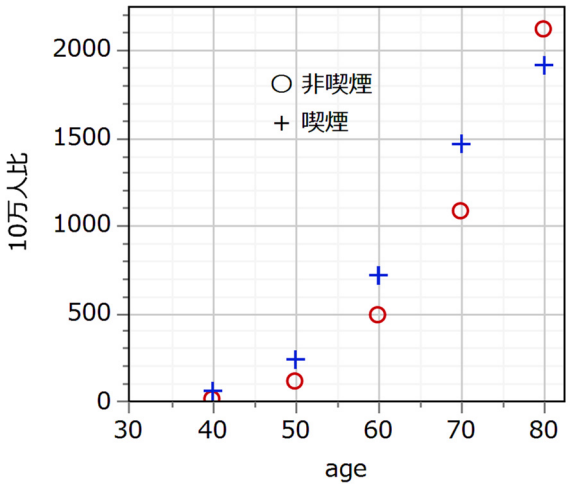


図 1.12　喫煙習慣による冠動脈心疾患による 10 万人あたりの死亡数の比較

付き回帰，第 1.5 節および第 1.6 節では JMP による回帰，第 1.6 節では SAS の GENMOD プロシジャによるポアソン回帰を示してきた．本節では，第 1.9 節と同様に Excel のソルバーを用いて，対数変換することなく，ポアソン回帰のパラメータ推定を行う方法を示す．詳しくは，第 2 章で取り上げるが，統計ソフトと Excel を互いに補完的に使う例として示す．なお，80 歳代を含めた解析については，第 3.6 節で取り上げる．

ポアソン回帰式は，非喫煙者を $x=0$，喫煙者を $x=1$，人年を n_i としたときに，

$$y_i = n_i \exp(\beta_0 + \beta_1 x_i + \beta_2 age_j) + \varepsilon_i \quad \varepsilon_i \sim poisson(y_i; \hat{y}_i)$$

$$\text{ただし，} \quad i = 1, 2, ..., 8$$

である．死亡者数 y_i の対数尤度 $\ln L_i$ を，

$$\ln L_i = \ln[\text{Poisson.dist}(y_i, \hat{y}_i, false)]$$

としたときに対数尤度 $\ln L$ は，

$$\ln L = \sum_i \ln L_i$$

となる．表 1.39 に示すように適当なパラメータ（$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$）を設定し，$\ln L$ を最大化するように Excel のソルバーでパラメータを変化させると，最尤解が得られる．

表 1.39　Excel による年齢と喫煙習慣によるポアソン回帰（初期値）

	切片	喫煙	年齢層	死亡	人年	10万人比	推定値	確率	対数尤度		
i	x_0	x	age	y	n	y'	y^	P	$\ln L_i$		最尤解
1	1	0	40	2	18,790	10.6	9.4	0.0036	-5.6145	β_0^=	**-10.0000**
2	1	0	50	12	10,673	112.4	9.7	0.0895	-2.4141	β_1^=	**2.0000**
3	1	0	60	28	5,710	490.4	9.5	0.0000	-14.3780	β_2^=	**0.0600**
4	1	0	70	28	2,585	1,083.2	7.8	0.0000	-18.1065		
5	1	1	40	32	52,407	61.1	193.8	0.0000	-106.8144		
6	1	1	50	104	43,248	240.5	291.4	0.0000	-83.4919		
7	1	1	60	206	28,612	720.0	351.3	0.0000	-38.9192		
8	1	1	70	186	12,663	1,468.8	283.3	0.0000	-22.5645		
								$\ln L$ =	**-292.3030**		

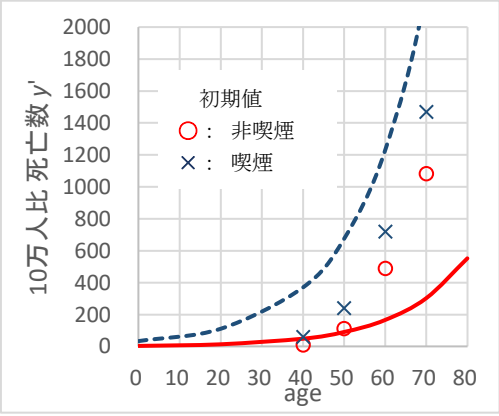


図 1.13　10 万人あたりの死亡数の予測（初期値）

一般化線形モデルは，指数関数で与えられた推定値に関するモデル式の両辺に対数を取り，

$$\ln(\hat{y}_i) = \ln(n_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 age_j)$$

のように線形化して反復重み付き回帰によって，最尤解を求める方法である．ポアソン回帰によらず，他の指数型分布族についても，同じ計算手順で最尤解を求められる．しかし，打ち切りデータがあるような場合への拡張性に難点がある．対数尤度関数 2 階の偏微分行列を使ったニュートン・ラフソン法が万能の計算手段であり，その入門として，Excel のソルバーを用いた最尤法は，ここに示したように，対数尤度関数を示しさえすれば，一気に最尤解を求めてくれる．

また，統計ソフトで出力される結果のグラフは，便利な面もあるが，図 1.13 で示したように，元のデータの散布図に 2 本の推定直線を上書きすることは容易ではない．解析結果を適切な図表に示すことは，結果を解釈するために不可避であり，オフセットがあるような場合は，10 万人比などに換算することにより結果の解釈が容易になる．

表 1.39 の計算シート内に 10 万人比に換算した結果があり，これを図 1.13 の散布図として用いている．ただし，上書きされている 10 万人比でのポアソン回帰の推定値は，表 1.39 の外側で $age = 0, 20, ..., 80$ などを加えて別途推定したものである．

表 1.40 は，Excel のソルバーで，対数尤度 $\ln L$ を最大にするように，パラメータを変化させた結果であり，$\hat{\beta}_0 = -11.7285$，$\hat{\beta}_1 = 0.5182$，$\hat{\beta}_2 = 0.1019$ が推定され，$\ln L = -33.1370$ が得られている．表 1.39 の初期値における $\ln L = -292.3030$ に比べて大きくなっていることが確認できる．表の下側は，推定されたパラメータを用い，10 万人比に換算した推定値 $\hat{y}'$ を計算し，右の図の指数曲線の重ね書きに使われている．なお，初期値に対する図 1.13 は，表 1.40 の図を見ながら，適度な広がりを持つ値として（$\hat{\beta}_0 = -10$，$\hat{\beta}_1 = 2$，$\hat{\beta}_2 = 0.06$）を設定した結果である．

Excel のソルバーを用いた方法は，手軽ではあるが推定したパラメータについての標準誤差が出力されないので，検定統計量の計算ができない．第 2 章で示すように，対数尤度関数の 2 階の偏微分行列を使ったニュートン・ラフソン法を適用すれば，共分散行列が計算過程に含まれるので，この対角要素の平方根が標準誤差として得られる．実用的には，使い慣れた統計ソフトにより，表 1.41 に示すように検定統計量およびパラメータの共分散行列を出力して，結果を Excel に取り込み，新たな追加の計算あるいは必要な図表の作成に使うことが望ましい．

表 1.40　Excel による年齢と喫煙習慣によるポアソン回帰（最尤解）

	切片	喫煙	年齢層	死亡	人年	10万人比	推定値	確率	対数尤度		
i	x_0	x	age	y	n	y'	y^	P	$\ln L_i$		最尤解
1	1	0	40	2	18,790	10.6	8.9	0.0053	-5.2468	β_0^=	**-11.7285**
2	1	0	50	12	10,673	112.4	14.1	0.0975	-2.3275	β_1^=	**0.5182**
3	1	0	60	28	5,710	490.4	20.8	0.0249	-3.6945	β_2^=	**0.1019**
4	1	0	70	28	2,585	1,083.2	26.2	0.0705	-2.6516		
5	1	1	40	32	52,407	61.1	41.8	0.0200	-3.9126		
6	1	1	50	104	43,248	240.5	95.7	0.0275	-3.5949		
7	1	1	60	206	28,612	720.0	175.4	0.0022	-6.1107		
8	1	1	70	186	12,663	1,468.8	215.1	0.0037	-5.5985		
								$\ln L$ =	**-33.1370**		

x_0	x	age	y	n	y'^
1	0	0		100,000	0.8
1	0	20		100,000	6.2
1	0	40		100,000	47.5
1	0	50		100,000	131.8
1	0	60		100,000	365.1
1	0	70		100,000	1011.8
1	0	80		100,000	2804.0
1	1	0		100,000	1.4
1	1	20		100,000	10.4
1	1	40		100,000	79.8
1	1	50		100,000	221.2
1	1	60		100,000	613.0
1	1	70		100,000	1698.8
1	1	80		100,000	4707.8

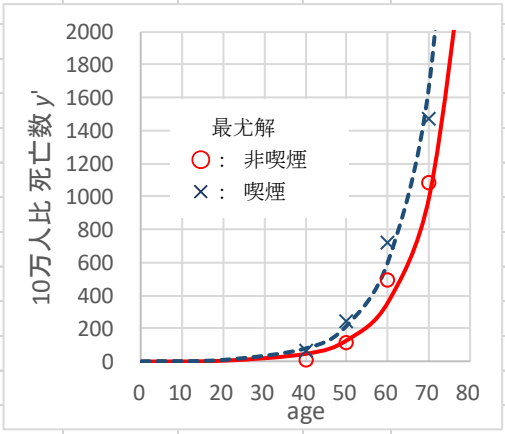


表 1.41　JMP による年齢と喫煙習慣によるポアソン回帰

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	-11.7283	0.5660	679.8171	<.0001*
x_Somok	0.5181	0.2584	4.5802	0.0323*
age	0.1019	0.0086	156.8010	<.0001*

共分散

	切片	x_Somok	age
切片	0.3204	-0.0522	-0.0044
x_Somok	-0.0522	0.0668	-0.0001
age	-0.0044	-0.0001	0.0001

JMP の出力から，非喫煙者に対する喫煙者の推定値は，$\hat{\beta}_1 = 0.5181$ で p 値が $p = 0.0323$ と統計的に有意である．この結果の解釈は，どうしたら良いのであろうか．表 1.42 に示すように推定された回帰式は，

$$\hat{y} = n \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 age)$$
$$= n \cdot \exp(-11.7283 + 0.5181x + 0.10193\,age)$$

なので，10 万人比に換算する場合は， $age=50$ と固定し， $x=0$ or 1 として計算すると，

$$x=0 \quad \hat{y}'=100,000\times\exp(-11.7283+0.5181\times0+0.10193\times50)=131.8$$

$$x=1 \quad \hat{y}''=100,000\times\exp(-11.7283+0.5181\times1+0.10193\times50)=221.2$$

となる．既に，年齢を変えて計算した結果は，表 1.40 の推定曲線の作成のための計算結果に等しく，喫煙者の非喫煙者に対するリスク比は，1.68 倍となる．この様な計算はせずとも，

$$\exp(\hat{\beta}_1)=\exp(0.5181)=1.68$$

として直接計算可能である．

表 1.42　喫煙者の非喫煙者に対するリスク比

	非喫煙	喫煙	リスク
age	x=0	x=1	比
20	6.2	10.4	1.68
40	47.5	79.8	1.68
50	131.8	221.2	1.68
60	365.1	613.0	1.68
70	1011.8	1698.8	1.68
80	2804.0	4707.8	1.68

なお，年齢区分の 80 歳まで含め，2 次式のあてはめについては，第 12.6 節「オフセットを含むポアソン回帰の 95%信頼区間」で取り上げ，さらに，2 値反応とした場合の上限を持つシグモイド曲線のあてはめについても示される．

1.12. 医院への通院回数（過分散）

ポアソン分布は，稀な現象がカウントできる場合の分布として知られている．しかし，稀ではない現象でのカウント・データに対しても適用できるのではないかと期待されるが，ゼロ・カウントが異常に多い，分散が平均値よりも数倍も大きい，などポアソン分布を仮定することに躊躇される事例に直面する．Cameron and Trivedi (1998)，Regression Analysis of Count Data の第 3 章の「Table 3.1 医院への通院回数」を表 1.43 に示す．

表 1.43 医院への通院回数

カウント	0	1	2	3	4	5	6	7	8	9	N	平均	分散	分散/平均
度数	4,141	782	174	30	24	9	12	12	5	1	5,190	0.3017	0.6370	2.1112

平均が 0.3017，分散が 0.6370，分散/平均が 2.11 倍と過分散が起きている．JMP の「一変量の分布」を用いて，ポアソン分布および負の 2 項分布から導出されたガンマ・ポアソン分布のあてはめを行う．ポアソン分布のパラメータは，観察されたデータの平均だけであるが，ガンマ・ポアソン分布のパラメータは，位置パラメータとしての平均，過分散パラメータの 2 つのパラメータからなる．

過分散パラメータの推定には，第 6.3 節で示すように最尤法により推定するのであるが，ここでは，JMP の「一変量の分布」で推定された $\hat{\sigma}'=1.7992$ を使い，JMP のスクリプトで提供されている関数 Gamma Poisson Probability() を用いて確率の計算を行い，表 1.44 に示すように Excel の表とした結果のみを示す．

表 1.44 ガンマ・ポアソン分布（負の 2 項分布）のあてはめ

カウント	度数	ポアソン分布			ガンマ･ポアソン分布		
y	n	確率	期待度数	差	確率	期待度数	差
0	4,141	0.7395	3838.3	302.7	0.8011	4157.8	-16.8
1	782	0.2231	1158.0	-376.0	0.1344	697.3	84.7
2	174	0.0337	174.7	-0.7	0.0411	213.3	-39.3
3	30	0.0034	17.6	12.4	0.0145	75.1	-45.1
4	24	0.0003	1.3	22.7	0.0054	28.2	-4.2
5	9	0.0000	0.1	8.9	0.0021	11.0	-2.0
6	12	0.0000	0.0	12.0	0.0008	4.4	7.6
7	12	0.0000	0.0	12.0	0.0003	1.8	10.2
8	5	0.0000	0.0	5.0	0.0001	0.7	4.3
9	1	0.0000	0.0	1.0	0.0001	0.3	0.7
全体	5,190	μ=0.3017			μ=0.3017,	過分散σ'^=1.7992	

図 1.14 に棒グラフにポアソン分布およびガンマ・ポアソン分布の確率関数をあてはめた結果を示す.

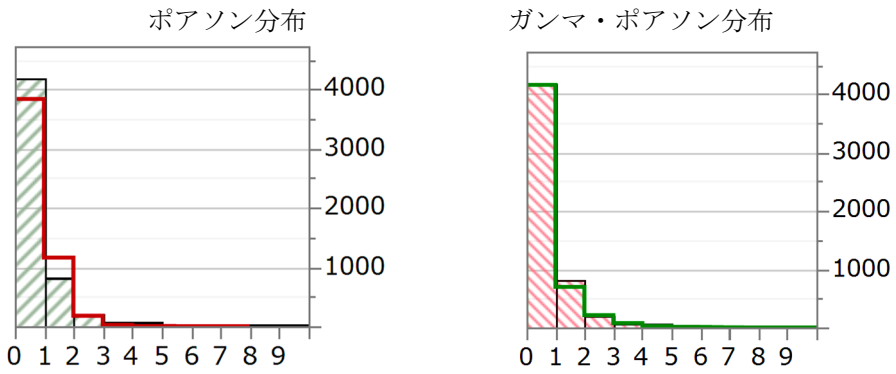


図 1.14　通院回数に対するポアソン分布およびガンマ・ポアソン分布のあてはめ

ポアソン分布の期待度数は，ゼロ・カウントに対して低めとなり，カウント 1 に対しては，やや高めとなっている．ガンマ・ポアソン分布（負の 2 項分布）の場合は，きれいにあてはまっているようである. 棒グラフでは, カウントが 5 以上のあてはまりを検討できないので, 表 1.44 に推定されたパラメータを用いてそれぞれの分布の確率関数を計算し，期待度数を求めて，観測度数との差を求めた.

ポアソン分布の場合には，カウント 4 以上から観測度数に対して期待度数が大幅に落ち込んでいるのに対し，ガンマ・ポアソン分布の場合には，やや低めではあるものの期待度数も追従している．稀な現象ではないカウント・データについては，ガンマ・ポアソン分布のあてはめが適しているように思われる．なお，ガンマ・ポアソン分布については，第 6 章を参照のこと.

1.13.　雌のカブトガニに連結する雄の数(2 因子, 2 変数, 対数リンク, 過分散)

アグレスティ著，渡邉・菅波・吉田・角野・寒水・松永訳（2003）の「カテゴリカルデー解析入門」の第 4.3 節の「計数データに対する一般化線形モデル：ポアソン回帰」に雌のカブトガニに連結する雄のサテライト数（Satellite 数）について，対数リンクによるポアソン回帰が例示されている．表 1.45 に示すようにデータには、173 匹のカブトガニについて説明変数として名義尺度（甲羅の色，後体部の棘の状態）の 2 変数，連続尺度（甲羅の幅，体重）の 2 変数，反応変数としてサテライト数が含まれている．

アグレスティ（2003）では，まず甲羅の幅をＸ軸，サテライト数をＹ軸とした散布図と共に対数リンクによるポアソン回帰の結果が示されている．引き続き，甲羅の幅を 8 区分とし，区分内のカブトガニの数とサテライト数の合計を算出し，カブトガニの数をオフセットとした解析を主体にしている．さらに，第 5 章では，サテライト数が（0，1 以上）の 2 値データとして，ロジスティック回帰を主体にして様々な角度からの解析方法が提示されている．

表 1.46 に示すように，サテライト数の平均は 2.9191，分散は 9.9120 であり，その比は 3.40 と過分散になっている．ポアソン分布を棒グラフ上に上書きした結果を見ても，誤差分布にポアソン分布を仮定することは絶望的とも思われる．もちろん，適合度検定でも $\chi^2 = 584.0436$，$p < 0.0001$ でポアソン分布があてはまるとは言えない．さらなる探索的な解析は，第 7.2 節に示す．

全データが過分散となる場合でも，何らかの条件によりサテライト数の平均が大きく異なる部分集団が複数存在するとも考えられる．甲羅の色によってサテライト数の平均が大きく異なり，それに伴って過分散が解消されるのであろうか．あるいは，甲羅の幅，あるいは，体重によりサテライト数の平均が大きく異なり，過分散が解消されるのだろうか．検討すべき事項である．

アグレスティ（2003）と同様に甲羅の幅を説明変数とし，サテライト数を反応変数としたときの対数リンクのポアソン回帰

$$Satellite_i = \exp(\beta_0 + \beta_0 \cdot width_i) + \varepsilon_i, \quad \varepsilon_i \sim Poisson$$

の結果を表 1.47 に示す．Pearson の適合度のカイ 2 乗値は 544.1570 と自由度の 171 に対して 3.1822 倍と過分散となっていて，表 1.46 から得られた分散を平均で割った比 9.9120/2.9191 =3.3956 であった過分散が，甲羅の幅を説明変数としても解消されていない．ポアソン回帰か

表 1.45　雌のカブトガニに連結する雄のサテライト数

color	spine	width	weight	satell	color	spine	width	weight	satell	color	spine	width	weight	satell	color	spine	width	weight	satell
2	3	28.3	3.050	8	3	1	28.5	3.250	9	4	3	23.5	1.900	0	2	1	28.0	2.900	4
3	3	22.5	1.550	0	3	3	28.9	2.800	4	2	2	24.0	1.700	0	4	3	25.8	2.250	10
1	1	26.0	2.300	9	2	3	28.2	2.600	6	2	1	29.7	3.850	5	2	3	27.9	3.050	7
3	3	24.8	2.100	0	2	3	25.0	2.100	4	2	1	26.8	2.550	0	2	3	24.9	2.200	0
3	3	26.0	2.600	4	2	3	28.5	3.000	3	4	3	26.7	2.450	0	2	1	28.4	3.100	5
2	3	23.8	2.100	0	2	1	30.3	3.600	3	2	1	28.7	3.200	0	3	3	27.2	2.400	5
1	1	26.5	2.350	0	4	3	24.7	2.100	5	3	3	23.1	1.550	0	2	2	25.0	2.250	6
3	2	24.7	1.900	0	2	3	27.7	2.900	5	2	1	29.0	2.800	1	2	3	27.5	2.625	6
2	1	23.7	1.950	0	1	1	27.4	2.700	6	3	3	25.5	2.250	0	2	1	33.5	5.200	7
3	3	25.6	2.150	0	2	3	22.9	1.600	4	3	3	26.5	1.967	1	2	3	30.5	3.325	3
3	3	24.3	2.150	0	2	1	25.7	2.000	5	3	3	24.5	2.200	1	3	3	29.0	2.925	3
2	3	25.8	2.650	0	2	3	28.3	3.000	15	3	3	28.5	3.000	1	2	1	24.3	2.000	0
2	3	28.2	3.050	11	2	3	27.2	2.700	3	2	3	28.2	2.867	1	2	3	25.8	2.400	0
4	2	21.0	1.850	0	3	3	26.2	2.300	3	2	3	24.5	1.600	1	4	3	25.0	2.100	8
2	1	26.0	2.300	14	2	1	27.8	2.750	0	2	3	27.5	2.550	1	2	1	31.7	3.725	4
1	1	27.1	2.950	8	4	3	25.5	2.250	0	2	2	24.7	2.550	4	2	3	29.5	3.025	4
2	3	25.2	2.000	1	3	3	27.1	2.550	0	2	1	25.2	2.000	1	3	3	24.0	1.900	10
2	3	29.0	3.000	1	3	3	24.5	2.050	5	3	3	27.3	2.900	1	2	3	30.0	3.000	9
4	3	24.7	2.200	0	3	1	27.0	2.450	3	2	3	26.3	2.400	1	2	3	27.6	2.850	4
2	3	27.4	2.700	5	2	3	26.0	2.150	5	2	3	29.0	3.100	1	2	3	26.2	2.300	0
2	2	23.2	1.950	4	2	3	28.0	2.800	1	2	3	25.3	1.900	2	2	1	23.1	2.000	0
1	2	25.0	2.300	3	2	3	30.0	3.050	8	2	3	26.5	2.300	4	2	1	22.9	1.600	0
2	1	22.5	1.600	1	2	3	29.0	3.200	10	2	3	27.8	3.250	3	4	3	24.5	1.900	0
3	3	26.7	2.600	2	2	3	26.2	2.400	0	2	3	27.0	2.500	6	2	3	24.7	1.950	4
4	3	25.8	2.000	3	2	1	26.5	1.300	0	3	3	25.7	2.100	0	2	3	28.3	3.200	0
4	3	26.2	1.300	0	2	3	26.2	2.400	3	2	3	25.0	2.100	2	2	3	23.9	1.850	2
2	3	28.7	3.150	3	3	3	25.6	2.800	7	2	3	31.9	3.325	2	3	3	23.8	1.800	0
2	1	26.8	2.700	5	3	3	23.0	1.650	1	4	3	23.7	1.800	0	3	2	29.8	3.500	4
4	3	27.5	2.600	0	3	3	23.0	1.800	0	4	3	29.3	3.225	12	2	3	26.5	2.350	4
2	3	24.9	2.100	0	2	3	25.4	2.250	6	3	3	22.0	1.400	0	2	3	26.0	2.275	3
1	1	29.3	3.200	4	3	3	24.2	1.900	0	2	3	25.0	2.400	5	2	3	28.2	3.050	8
1	3	25.8	2.600	0	2	2	22.9	1.600	0	3	3	27.0	2.500	6	4	3	25.7	2.150	0
2	2	25.7	2.000	0	3	2	26.0	2.200	3	3	3	23.8	1.800	6	2	3	26.5	2.750	7
2	1	25.7	2.000	8	2	3	25.4	2.250	4	1	1	30.2	3.275	2	2	3	25.8	2.200	0
2	1	26.7	2.700	5	3	3	25.7	1.200	0	3	3	26.2	2.225	0	3	3	24.1	1.800	0
4	3	23.7	1.850	0	2	3	25.1	2.100	5	2	3	24.2	1.650	2	3	3	26.2	2.175	2
2	3	26.8	2.650	0	3	2	24.5	2.250	0	2	3	27.4	2.900	3	3	3	26.1	2.750	3
2	3	27.5	3.150	6	4	3	27.5	2.900	0	2	2	25.4	2.300	0	3	3	29.0	3.275	4
4	3	23.4	1.900	0	3	3	23.1	1.650	0	3	3	28.4	3.200	3	1	1	28.0	2.625	0
2	3	27.9	2.800	6	3	1	25.9	2.550	4	4	3	22.5	1.475	4	4	3	27.0	2.625	0
3	3	27.5	3.100	3	2	3	25.8	2.300	0	2	3	26.2	2.025	2	2	2	24.5	2.000	0
1	1	26.1	2.800	5	4	3	27.0	2.250	3	2	1	24.9	2.300	6					
1	1	27.7	2.500	6	2	3	28.5	3.050	0	1	2	24.5	1.950	6					
2	1	30.0	3.300	5	4	1	25.5	2.750	0	2	3	25.1	1.800	0					

注釈:　color=色（1=やや明るい，2=中くらい，3=やや暗い，4=暗い）;

Spine=後体部の棘の状態（1=いずれも正常，2=一方が摩耗または破損している，3=いずれも摩耗または破損している）;

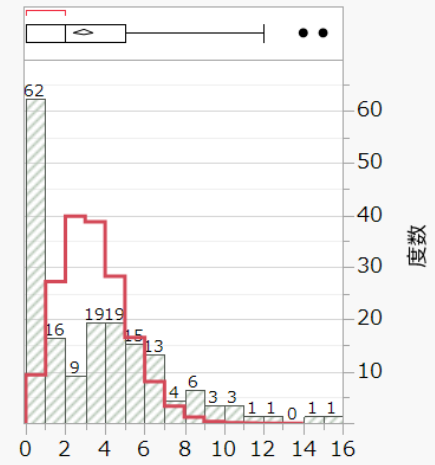
width=甲羅の幅(cm) ;　weight =重さ（kg）;　satell＝サテライト数.

出典: Web at　http://lib.stat.cmu.edu/datasets/agresti　2020 年 4 月 17 日アクセス

表 1.46　サテライト数へのポアソン分布のあてはめ

一変量の分布

サテライト数



要約統計量

平均	2.9191
標準偏差	3.1483
平均の標準誤差	0.2394
平均の上側95%	3.3915
平均の下側95%	2.4466
N	173.0000
分散	9.9120
修正平方和	1704.8671

Poisson分布のあてはめ

パラメータ推定値

種類	パラメータ	推定値
尺度	λ	2.9191

指標	
(-2)*対数尤度	988.0893
AICc	990.1127
BIC	993.2426

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
584.0436	<.0001*

表 1.47　甲羅の幅 width についての対数リンクによるポアソン回帰

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	32.4565	64.9131	1	<.0001*
完全	461.5881			
縮小	494.0447			

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)
Pearson	544.1570	171	<.0001*
デビアンス	567.8786	171	<.0001*

AICc
927.2468

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(P>ChiSq)
切片	-3.3048	0.5422	36.8670	<.0001*
甲羅の幅	0.1640	0.0200	64.9131	<.0001*

推定値の共分散

共分散

	切片	甲羅の幅
切片	0.2940	-0.0108
甲羅の幅	-0.0108	0.0004

ら得られた過分散パラメータを $\varphi = 3.1822$ とし，得られた共分散行列を φ 倍して標準誤差を調整する方法が知られていて，JMP のポアソン回帰でもサポートされている．

過分散を調整したポアソン回帰の結果を表 1.48 に示す．

表 1.47 の甲羅の幅の標準誤差は，$SE = 0.0200$ なので，調整後の SE' は，

$$SE' = \sqrt{\varphi SE^2} = \sqrt{3.1822 \times 0.0200^2}$$

となり，尤度比カイ 2 乗値は，64.9131 から 20.3988 と激減する．

表 1.48　甲羅の幅 width についての対数リンクの過分散調整済みのポアソン回帰

手法: 一般化線形モデル
分布: Poisson
リンク関数: 対数
☑ 過分散に基づく検定と信頼区間
☐ Firthバイアス調整推定値

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	10.1993888	20.3988	1	<.0001*
完全	145.05293			
縮小	155.252319			

適合度統計量	カイ2乗	自由度	p値	過分散
Pearson	544.1570	171	<.0001*	3.1822
デビアンス	567.8786	171	<.0001*	

AICc
296.2479

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	-3.3048	0.9673	11.5854	0.0007*
甲羅の幅	0.1640	0.0356	20.3988	<.0001*

過分散の調整を甲羅の幅の標準誤差で例示したが，調整は共分散行列をφ倍することにより，回帰直線の 95%信頼区間に対しても調整されることになる．

表 1.49　過分散係数を用いた共分散の調整

	パラメータの共分散			過分散		調整後の共分散			標準誤差 *SE*	
	切片	甲羅の幅		φ		切片	甲羅の幅		切片	甲羅の幅
切片	0.2940	-0.0108	*	3.1822	=	0.9357	-0.0343		0.9673	
甲羅の幅	-0.0108	0.0004				-0.0343	0.0013			0.0356

過分散の係数を用いた方法は，過分散となるカウント・データに対する万能の方法とも思われるかもしれないが，表 1.46 に示したヒストグラムに重ね書きしたポアソン分布から，このデータにポアソン分布を仮定することは全くできない．甲羅の幅に対するポアソン回帰によって過分散が解消するのであれば嬉しいのであるが，残念ながら過分散は解消されなかった．

ポアソン回帰を行っても過分散が解消していないことを視覚化するために散布図に個別データの 95%信頼区間を重ね書きしてみると，図 1.15 左に示すよう外側に多数の点がはみ出ていることによりポアソン回帰のあてはめには無理があることを実感できる．図 1.15 右に示すように予測値に対する pearson 残差をプロットすることにより，Pearson 残差が 3 以上の飛び離れデータが多数存在することからも，ポアソン分布を誤差分布とする回帰分析について否定的な結果となっている．

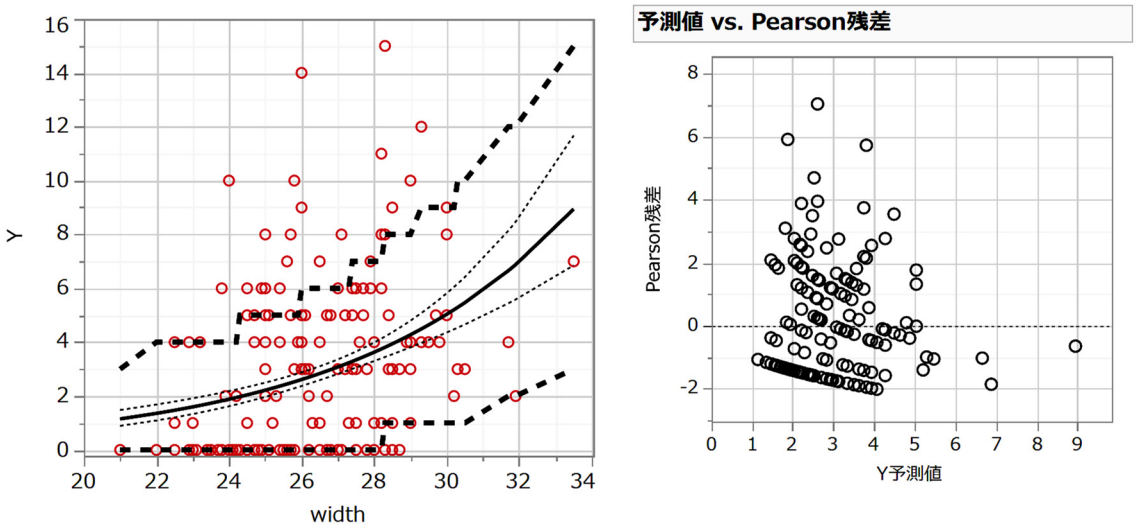


図 1.15　ポアソン回帰に対する 95%信頼区間および予測値に対する pearson 残差

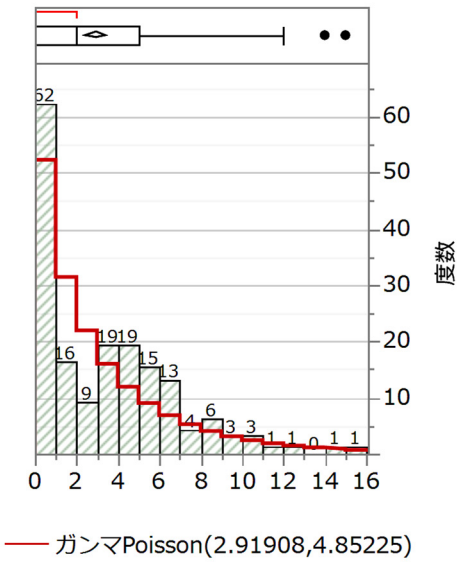
他の変数を加えてポアソン回帰を行っても過分散が解消されないのであれば，ポアソン回帰を行う前提がないことになる．この原因は，173 個体に対してサテライト数がゼロに 62 件と全体の 35.8%を占め，サテライト数が 3 と 4 あたりに分布の山があることから，アグレスティ（2003）にあるサテライト数を（0，1）反応での解析が望ましいのか，あるいは，3 区分程度の順序データとした解析を行うことが望ましいかも知れない．

JMP には，過分散を考慮した負の 2 項分布から導出されたガンマ・ポアソン分布をあてはめることができるので，表 1.50 に結果を示す．結果は，位置 $\lambda = 2.9191$，過分散 $\sigma' = 4.8522$ となる．表 1.46 に示したポアソン分布のあてはめでは，サテライト数が 0 の場合について大きな乖離があったが，過分散を考慮したガンマ・ポアソン分布のあてはめでは，適当なあてはめが行われているように思われる．ただし，サテライト数が 1 および 2 については，元々のデータの出現頻度が小さいため，ガンマ・ポアソン分布のあてはめも無理なのかも知れない．なお，過分散を考慮した解析方法については，第 6 章で取り上げる．

雌のカブトガニに雄が多数連結することは確かなようであるが，単独でいる雌のカブトガニの数が多いことから，雄が雌に連結する場合も合わせてポアソン分布を仮定することは無理のようである．（連結していない，数匹が連結している，かなり連結してる）などの順序尺度としての扱が適してるかもしれない．第 7.2 節で，このデータを基づいて探索的な解析を行った結果を示す．

表 1.50　サテライト数へのガンマ・ポアソン分布のあてはめ（Ver. 14）

サテライト数



ガンマPoissonのあてはめ

パラメータ推定値

種類	パラメータ	推定値	下側95%CL	上側95%CL
位置	λ	2.9191	2.4136	3.5571
過分散	σ	4.8522	3.6559	6.6859

指標	
(-2)*対数尤度	767.4092
AICc	771.4798
BIC	777.7158

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
120.3656	0.9990

注: Ho = ガンマPoisson分布からのデータ。 p値が小さい場合はHoを棄却。

表 1.46 で「尺度 λ」となっているが，ここでは「位置 λ」となっていることに注意．

SAS の GENMOD プロシジャには，ゼロの割合を考慮するポアソン回帰，ガンマ・ポアソン回帰の 2 種類の Zero-Inflated Model が用意されている．これらについても第 6 章で取り上げる．

偶数ページ

2. ニュートン・ラフソン法によるポアソン回帰

「最尤法は，どのような方法なのか」との質問に「尤度を最大にする方法である」との紋切り型な説明がはびこっている. 最小 2 乗法であれば,「誤差平方和を最小にする方法である」との説明と同様であり，質問を単に言い換えた回答でしかない．尤度とは何か，最大化するための方法はどのようなものかについて具体的な事例を用い，さまざまな角度から繰返しをいとわず説明を試みる．第 1 章で取り上げた事例を用い，最尤法の基本となる対数尤度の計算を Excel の関数を補助的に用い，手作業で最大化するための手順を示す．次に，対数尤度関数をパラメータに関して偏微分した偏微分ベクトル, 更に偏微分をした 2 階の偏微分行列（ヘッセ行列）の導出を丁寧に示す．それらを Excel シート上に展開し，ニュートン・ラフソン法による対数尤度の最大化の方法について示す．対数尤度の最大化の方法について，1 群の場合，回帰の場合，対数リンクでの回帰の場合，オフセットがある場合について Excel を用いることにより計算過程を可視化しつつ示す.

2.1. 手作業による逐次的な対数尤度の最大化

表 1.4 の［有害雑草の種子の数」では，算術平均（ポアソン分布の位置パラメータ $\hat{\mu}$）を用いて有害雑草の種子の数に対するポアソン分布の確率を計算した［スネデカーら（1972)］. 本節では，算術平均を用いずに，探索的に対数尤度を最大化するような位置パラメータの推定値 $\hat{\mu}$ を変化させて推定する方法を提示する．理論的には自明であり，ナンセンスかもしれないが，一足飛びにポアソン回帰から始めるよりも，レンガを積むがごとく段階を追って理解を深めることは，応用力の根源となる.

表 2.1 に示すように $\hat{\mu}$=2.0 とし，種子の数 y_i に対するポアソン分布の確率 P_i を Excel の Poisson.dist()関数

$$\begin{aligned} P_i &= \frac{\hat{\mu}^{y_i} e^{-\hat{\mu}}}{y_i!} \\ &= \frac{2.0^{y_i} e^{-2.0}}{y_i!} \\ &= \text{Poisson.dist}(y_i,\ 2.0,\ \textit{false}), \quad i = 1, 2, ..., 10 \end{aligned} \tag{2.1}$$

で計算した結果を示す．なお，引数の *false* は確率の計算のための引数で，*true* とすると下側

確率のための引数となる．それぞれの種子の数 y_i に対する観測度数を n_i とすると，同時に起きる確率は $P_i^{n_i}$ （尤度 L_i）となる．それらを全て掛け合わせた尤度 L は，

$$\begin{aligned} L &= \prod_{i=1}^{10} L_i \\ &= \prod_{i=1}^{10} P_i^{n_i} \\ &= 0.1353^3 \times 0.2707^{17} \times \cdots \times 0.0002^1 = 3.2034 \times 10^{-93} \end{aligned} \tag{2.2}$$

となり，その積は極めて小さく，数値計算に適していない．そこで，尤度の対数を取っても大小関係は保たれるので，対数尤度 $\ln L$ として扱うことが一般的である．

$$\begin{aligned} \ln L &= \sum_{i=1}^{10} \ln L_i \\ &= \sum_{i=1}^{10} \ln(P_i^{n_i}) \\ &= -6.0000 - 22.2165 - \cdots - 8.5635 = -212.9762 \end{aligned} \tag{2.3}$$

表 2.1　初期値 $\hat{\mu} = 2.0$ に対する対数尤度の計算シート

	μ^=	**2.0000**	$\ln L$=	**-212.9761953**	
i	有害種子の数 y	観測度数 n	ポアソン確率 P	n回の出現確率 P^n：尤度L_i	対数尤度 $\ln(P^n)$ $\ln L_i$
1	0	3	0.1353	2.4788E-03	-6.0000
2	1	17	0.2707	2.2465E-10	-22.2165
3	2	26	0.2707	1.7517E-15	-33.9782
4	3	16	0.1804	1.2636E-12	-27.3971
5	4	18	0.0902	1.5695E-19	-43.2984
6	5	9	0.0361	1.0385E-13	-29.8958
7	6	3	0.0120	1.7409E-06	-13.2611
8	7	5	0.0034	4.7968E-13	-28.3657
9	8	0	0.0009	1.0000E+00	0.0000
10	9	1	0.0002	1.9095E-04	-8.5635
	計	98	1.0000	3.2034E-93	**-212.9762**
				積	和

なぜ，確率ではなく“尤度”と言い換えるのか．表 2.2 左に示すように位置パラメータが $\mu = 2.0$ （推定値ではないので $\hat{\mu}$ としない）と固定されている場合に，ポアソン分布の確率関数を用いて観測データ y（正の整数，0 を含む）が起きる確率が計算されていて，$y = 1$ の場合 $P_{(y=1|\mu=2)} = 0.2707$ となっている．逆に，表 2.2 右に示すように観測データがある値 $y = 1$ と固定されている場合に，位置パラメータ μ（正の実数，0 を含まない）を変化させ $\mu = 2.0$ となった場合にも $P_{(\mu=2|y=1)} = 0.2707$ と同じ確率となっている．

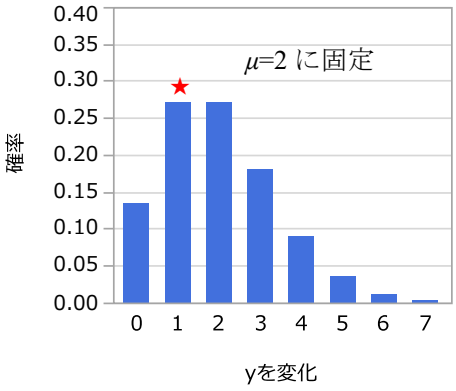
位置パラメータを $\mu = 2.0$ に固定し，$y = 0, 1, 2, \ldots$ と変化させた場合のポアソン確率を全て加えた場合は，分布関数の性質から

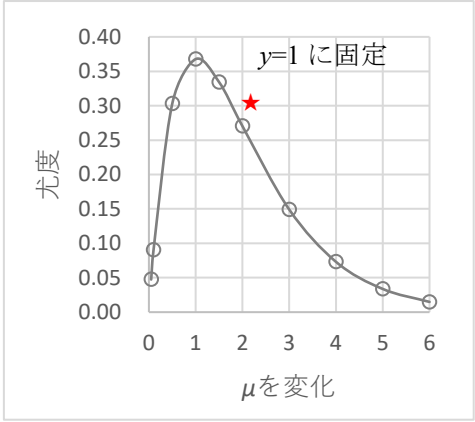
$$\sum_{y=0}^{\infty} P_{(y|\mu=2)} = 0.1353 + 0.2707 + 0.2707 + \cdots = 1.0000$$

表 2.2　確率関数と尤度関数の違い

μを固定	yを変化	確率 P
2	0	0.1353
2	**1**	**0.2707**
2	2	0.2707
2	3	0.1804
2	4	0.0902
2	5	0.0361
2	6	0.0120
2	7	0.0034

μを変化	yを固定	尤度 L_i	対数尤度
0.05	1	0.0476	-3.0457
0.1	1	0.0905	-2.4026
0.5	1	0.3033	-1.1931
1	1	0.3679	-1.0000
1.5	1	0.3347	-1.0945
2	**1**	**0.2707**	**-1.3069**
3	1	0.1494	-1.9014
4	1	0.0733	-2.6137
5	1	0.0337	-3.3906
6	1	0.0149	-4.2082





となることは自明である．反応変数 y を 1 に固定したときに，正の実数である位置パラメータ μ を $\mu = 0.05,\ 0.1,\ 0.5,\ 1,\ 1.5,\ 2,\ldots$ と変化させた場合にポアソン確率が，

$$P_{(\mu|y=1)} = 0.0476,\ 0.0905,\ 0.3033,\ 0.3679,\ 0.3347,\ 0.2707,\ldots$$

と計算できるが，離散型のポアソン分布ではなく連続関数の数値となっている．これらは，ポアソン分布の確率として計算はされるが，まったく別物の連続関数の数値となっている．この連続関数は，一般的に尤度関数と言われている．

表 2.1 に示したように，仮の推定値を $\hat{\mu} = 2.0$ とした場合の種子数が $y_1 = 0$ となるポアソン確率は $P_{(y1=0|\hat{\mu}=2)} = 0.1353$ である．見方を代えた場合には，尤度関数の“尤度”ともなる．観測度数は $n_1 = 3$ なので，$y_1 = 0$ が 3 回起きる確率は，$P_1^3 = 0.1353^3 = 2.4788\times10^{-3}$ であるが，もはやポアソン分布の確率とは言えないので，尤度 $L_1 = 2.4788\times10^{-3}$ と言うことにする．また，尤度の自然対数を取ると $\ln L_1 = -6.0000$ となる．さらに（$y_2 = 1$，$n_2 = 17$）の場合の対数尤度は $\ln L_2 = \ln(0.2707^{17}) = -22.2165$ となり，それらの和としての対数度 $\ln L$ は，

$$\begin{aligned}\ln L &= \ln L_1 + \ln L_2 + \cdots + \ln L_{10} \\ &= -6.0000 - 22.2165 - \cdots - 8.5635 \\ &= -212.9762\end{aligned}$$

となる．

実際のデータの分布は，種子の数 $y=2$ よりも大きい方に裾を引いているので，表 2.1 の $\hat{\mu}=2.0$ に代えて $\hat{\mu}=3.0$ を入力すると対数尤度が，$\ln L=-190.9585233$ と計算される．表 2.3 に値のみをコピー&ペーストして $\hat{\mu}=2.0$ の $\ln L$ との差を求めると 22.0176720 の増加となる．更に $\hat{\mu}=3.1$ に増加すると $\ln L=-191.0527358$ となり −0.0942124 の減少となる．したがって，$\ln L$ を最大にする $\hat{\mu}$ は（3.0，3.1）の範囲に存在することがわかる．このように $\ln L$ の差分を観察しつつ $\hat{\mu}$ を挟み撃ち的に適宜増減することにより，最尤解として $\hat{\mu}=3.0204$ が得られる．

表 2.3　逐次的な挟み撃ち法による最尤解

μ^	μ^増減	対数尤度 $\ln L$	差分	
2.0		-212.9761953		
3.0	+	-190.9585233	22.0176720	
3.1	+	-191.0527358	-0.0942124	
3.05	-	-190.9658499	0.0868858	
3.02	-	-190.9517387	0.0141113	
3.025	+	-190.9520777	-0.0003390	
3.023	-	-190.9518449	0.0002328	
3.021	-	-190.9517416	0.0001032	
3.0205	-	-190.9517361	0.0000055	
3.0204	**-**	**-190.9517360**	**0.0000001**	**最尤解**
3.0203	+	-190.9517362	-0.0000002	

図 2.1 を参照することにより増減の判断の可視化となる．

表 2.1 の対数尤度の計算欄を右方向に伸ばし，表 2.4 に示すように位置パラメータ $\hat{\mu}$ をある範囲に限定し，計算された対数尤度のグラフを眺めながらその範囲を絞るような方法でも最尤解を得ることができる．表頭に示すように $\hat{\mu}$ を，2.950 から 0.025 刻みで 3.100 まで $j=1,2,...,7$ と 7 段階に変化させ，それぞれの y_i と度数 n_i

$$\ln L_{ij}=\ln[\text{Poisson.dist}(y_i,\hat{\mu}_{\cdot j},false)\,\hat{}\,n_i]$$
$$\ln L_{\cdot j}=\sum_{i=1}^{9}\ln L_{ij} \qquad (2.4)$$

表 2.4　対数尤度の最大化のための 0.025 刻みでの計算（$n_i=0$ の場合を除く）

		j	1	2	3	**4**	5	6	7
		母数 μ^	2.950	2.975	3.000	**3.025**	3.050	3.075	3.100
		$\ln L$	-191.0334	-190.9855	-190.9585	**-190.9521**	-190.9658	-190.9995	-191.0527
i	y	n	$\ln L_{\cdot 1}$	$\ln L_{\cdot 2}$	$\ln L_{\cdot 3}$	**$\ln L_{\cdot 4}$**	$\ln L_{\cdot 5}$	$\ln L_{\cdot 6}$	$\ln L_{\cdot 7}$
1	0	3	-8.8500	-8.9250	-9.0000	**-9.0750**	-9.1500	-9.2250	-9.3000
2	1	17	-31.7593	-32.0409	-32.3236	**-32.6075**	-32.8926	-33.1788	-33.4662
3	2	26	-38.4680	-38.6791	-38.8940	**-39.1124**	-39.3345	-39.5600	-39.7889
4	3	16	-23.9415	-23.9364	-23.9348	**-23.9364**	-23.9414	-23.9495	-23.9609
5	4	18	-32.4150	-32.2574	-32.1049	**-31.9574**	-31.8148	-31.6770	-31.5440
6	5	9	-20.9562	-20.8014	-20.6499	**-20.5014**	-20.3561	-20.2137	-20.0743
7	6	3	-9.1153	-9.0384	-8.9627	**-8.8884**	-8.8152	-8.7433	-8.6725
8	7	5	-19.5126	-19.3423	-19.1744	**-19.0089**	-18.8459	-18.6851	-18.5267
9	9	1	-6.0156	-5.9646	-5.9143	**-5.8646**	-5.8156	-5.7671	-5.7192

を用いて対数尤度 $\ln L_{\cdot j}$ を計算している．対数尤度 $\ln L_{\cdot j}$ が最も大きくなるのは，$j=4$，$\hat{\mu}_{\cdot 4}=3.025$，$\ln L_{\cdot 4}=-190.9521$ である．計算された対数尤度を図 2.1 左に示す．最大となるのは，$\hat{\mu}=3.025$ より少し小さい左側にあることが推測される．

表 2.5 に示すように，更に刻み幅を 3.014 から 0.002 刻みで 3.026 まで変化させた結果を図 2.1 右に示す．この結果から，最大となるのは，$\hat{\mu}=3.020$ より少し大きいところにあると推測される．

表 2.5　対数尤度の最大化のための 0.002 刻みでの計算

		j	1	2	3	**4**	5	6	7
		母数 μ^	3.014	3.016	3.018	**3.020**	3.022	3.024	3.026
		$\ln L$	-190.9524	-190.9521	-190.9518	**-190.9517**	-190.9518	-190.9519	-190.9522
i	y	n	$\ln L_{\cdot 1}$	$\ln L_{\cdot 2}$	$\ln L_{\cdot 3}$	**$\ln L_{\cdot 4}$**	$\ln L_{\cdot 5}$	$\ln L_{\cdot 6}$	$\ln L_{\cdot 7}$
1	0	3	-9.0420	-9.0480	-9.0540	**-9.0600**	-9.0660	-9.0720	-9.0780
2	1	17	-32.4824	-32.5052	-32.5279	**-32.5506**	-32.5734	-32.5961	-32.6189
3	2	26	-39.0159	-39.0334	-39.0509	**-39.0685**	-39.0860	-39.1036	-39.1213
4	3	16	-23.9353	-23.9354	-23.9356	**-23.9358**	-23.9360	-23.9363	-23.9366
5	4	18	-32.0217	-32.0099	-31.9982	**-31.9865**	-31.9748	-31.9632	-31.9516
6	5	9	-20.5664	-20.5545	-20.5427	**-20.5309**	-20.5191	-20.5073	-20.4956
7	6	3	-8.9209	-8.9150	-8.9091	**-8.9031**	-8.8972	-8.8913	-8.8854
8	7	5	-19.0814	-19.0682	-19.0550	**-19.0418**	-19.0286	-19.0155	-19.0024
9	9	1	-5.8864	-5.8824	-5.8785	**-5.8745**	-5.8706	-5.8666	-5.8627

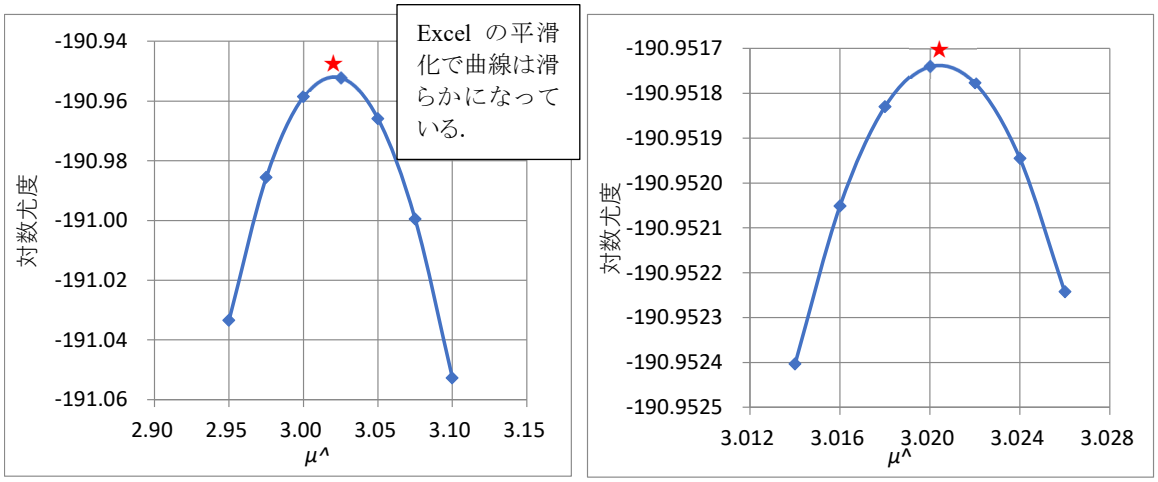


図 2.1　$\hat{\mu}$ を変化させた場合の対数尤度関数

2.2. Excel のソルバーによる対数尤度の最大化

統計ソフトを用いれば，対数尤度を全く意識することなくポアソン回帰の解を求めることができる．このことが，統計モデルの理論を学習し理解する意欲の妨げになっているのではないだろうか．統計ソフトから出力される結果に対して何らかの解釈をすることはできても，その理論的背景を関係者に説明し，納得してもらえることができるのだろうか．前節で示したように，Excel の計算機能を用いた逐次的な方法によって対数尤度を最大化する経験が，学習の第 1 歩であり，説明したい推定値を自ら計算し，解釈することができるようになる．

第 2 歩目は，Excel のソルバーを用いて，対数尤度を最大化する方法の習得である．Excel のソルバーは，手作業による逐次的な手順を代行してくれる．ただし，この Excel のソルバーの最適化の方法は，私にとってもブラック・ボックスである．ただし，対数尤度関数を利用者が，きちっと Excel で計算式として設定できることが必須であり，理論の学習にとって役に立つ．この方法により，統計ソフトがサポートしていない各種の統計モデルについて解析が迅速に行えるようになる．

Excel のソルバーと同様に，R 言語の optim()関数も同様な最適化のための関数である．私は，Excel のソルバーを使う前は，SAS の IML 行列計算言語および JMP のスクリプト言語の maximaize()関数を使っていた．これらに比べ Excel の計算機能は劣っているが，計算過程および結果の可視化の面で総合的に優れている．その結果として Excel のソルバーを多用するようになった．

第 3 歩目が，第 2.3 節に示す対数尤度関数の 2 階の偏微分行列を用いたニュートン・ラフソン法による最大化である．なお，一般化線形モデルに特徴的な，反復重み付き回帰による最尤法については，第 5 章で取り上げる．

表 2.6 左は初期値 $\hat{\mu}=2.0$ に対する有害種子の数 y_i について，表 2.1 と同様に n_i 個のポアソン確率

$$L_i = P_i^{n_i} = \text{Poisson.dist}(y_i, 2, false)^{n_i} \tag{2.5}$$

を計算し，対数を取って

$$\ln L_i = \ln P_i^{n_i} = \ln\left[\text{Poisson.dist}(y_i, 2, false)^{n_i}\right] \tag{2.6}$$

とし，それぞれの i についての対数尤度 $\ln L_i$ を計算している．それらの和

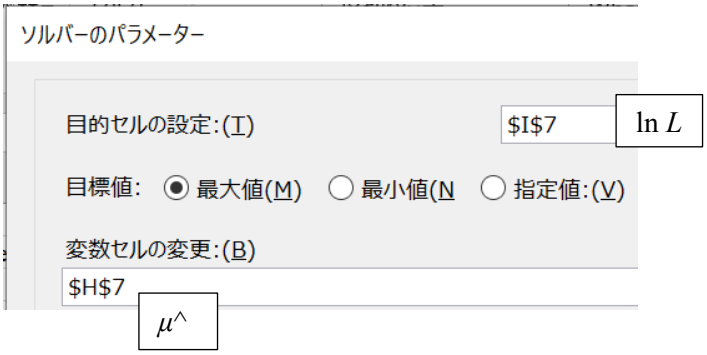
$$\begin{aligned}\ln L &= \sum_{i=1}^{9} \ln L_i \\ &= \sum_{i=1}^{9} \ln P^{n_i} = -6.0000 - 22.2165,...,-8.5635 = -212.9762\end{aligned}$$

表 2.6　ソルバーを用いた対数尤度の最大化

			μ^	$\ln L$		μ^	$\ln L$
			2.0000	**-212.9762**		**3.0204**	**-190.9517**
i	y	n	P^n	$\ln(P^n)$		P^n	$\ln(P^n)$
1	0	3	2.4788E-03	-6.0000		1.1608E-04	-9.0612
2	1	17	2.2465E-10	-22.2165		7.2681E-15	-32.5553
3	2	26	1.7517E-15	-33.9782		1.0745E-17	-39.0721
4	3	16	1.2636E-12	-27.3971		4.0252E-11	-23.9359
5	4	18	1.5695E-19	-43.2984		1.2867E-14	-31.9841
6	5	9	1.0385E-13	-29.8958		1.2151E-09	-20.5285
7	6	3	1.7409E-06	-13.2611		1.3613E-04	-8.9019
8	7	5	4.7968E-13	-28.3657		5.3878E-09	-19.0391
9	9	1	1.9095E-04	-8.5635		2.8124E-03	-5.8737
	計	98	**初期値(表2.1再掲)**			**ソルバーによる最大化**	

が，対数尤度 $\ln L = -212.9762$ と計算されている．

表 2.6 右は，まず表 2.6 左を Excel シート上でコピーし，同じ結果が得られるように計算式のセル位置を調整する．次に，Excel の「データ」タブに含まれている「ソルバー」を起動し，対数尤度の数値セルを「目的セルの設定」ボックスに設定し，［目標値」が最大値に選択されていることを確認し，位置パラメータ $\hat{\mu}$ のセルを「変数セルの変更」ボックスに設定し，「解決」ボタンのクリックで計算されたものである．結果として最適解の $\hat{\mu}$=3.0204 が得られており，対数尤度 $\ln L = -190.9517$ となっている．なお，「ソルバー」は，「分析ツール」と同様に Excel のアドインなので，使用できるような操作を前もって行う必要がある．



Excel のソルバーで，対数尤度を最大化しようとしても，設定した初期値によってはエラーが起きることもある．その場合は，対数尤度が大きくなるように手作業で初期値を変えて実行する必要がある．表 2.6 の例であれば，$\hat{\mu}=8$ までは収束するが，$\hat{\mu}=9$ にするとエラーとなり収束しない．$\hat{\mu}=0$ とすると関数計算が不能となるが，$\hat{\mu}=0.001$ とすれば，解が求まる．このような試行錯誤による計算結果が目の前で瞬時に確認できることは，最尤法を学習するために優れている．

2.3. ニュートン・ラフソン法による対数尤度の最大化

ニュートン・ラフソン法は，ポアソン回帰などの一般化線形モデルのみならず打ち切りデータを含むワイブル回帰など，多くの統計モデルに対する汎用的な解析方法として知られている．「最尤法とはどのような方法なのですか」との質問に，「尤度を最大化する方法です」との質問を単に言い換えた回答がはびこる原因は，ニュートン・ラフソン法を用いて尤度を最大化するための具体的な数値例の例示がないためである．数値例を Web 上でも身近にある統計の教科書を探しているが見出すことができない．

ニュートン・ラフソン法については，Web 上で丁寧な説を見出すことができる．ただし，例示されているのは，2 次式の解を求めるなどの簡単な事例がほとんどである．ポアソン回帰を含む一般化線形モデルの場合には，第 1.4 節で示したように反復重み付き回帰による最尤法で対応可能であるが，より汎用的な解法であるニュートン・ラフソン法をポアソン回帰で習得することは，打ち切りデータを含むワイブル回帰などへの応用力を身に付けることになる．詳しくは，高橋（2015），「寿命試験データの統計解析」，高橋（2018），「正規分布を仮定した打ち切りデータを含む回帰分析入門」を参照してもらいたい．

ニュートン・ラフソン法による最尤法の良さは，対数尤度を最大化する計算過程の中で，パラメータの共分散行列そのものを用いることにある．パラメータの共分散行列が得られれば，ポアソン回帰直線の 95%信頼区間など各種の推定値に対する検定統計量の計算も容易にできる．Excel のソルバーで対数尤度関数を最大化することにより，統計モデルのパラメータの推定値を得ることができるが，各種の推論のためのパラメータの分散，パラメータ間の共分散が残念ながら得られない．そのために，ニュートン・ラフソン法によりパラメータの共分散行列を求める必要がある．

比較的容易なポアソン回帰について対数尤度関数の 2 階の偏微分行列を用いた最尤法で計算することができるようになれば，2 値データに対する一般線化線形モデルに対しても対応できる力を付けることになる．なお，2 値データへの応用は，高橋（2017）「一般化線形モデルを Excel で極め活用するープロビット法・ロジット法・補 2 重対数法ー」を参照してもらいたい．

対数尤度関数の偏微分

ニュートン・ラフソン法を自ら実践するためには，対数尤度関数をパラメータに関して 1 階の偏微分ベクトル $\boldsymbol{U}$（スコアベクトル），さらに偏微分した 2 階の偏微分行列 $\boldsymbol{H}$（ヘッセ行列）を必要とする．一般的な統計の教科書で，偏微分が出てくるのは，回帰分析のパラメータ（β_0，β_1）を推定するための計算式を求める際に，第 4.2 節で示すように誤差平方和 S_e を

$$S_e = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

パラメータ（β_0，β_1）で偏微分して正規方程式を導出する場合に限定されている．

ポアソン回帰の場合は，$\mu_i = \beta_0 + \beta_1 x_i$ とした場合の対数尤度関数

$$\begin{aligned} \ln L &= \sum_i \ln\left(\frac{\mu^{y_i} e^{-\mu_i}}{y_i!}\right) \\ &= \sum_i \ln\left[\frac{(\beta_0 + \beta_1 x_i)^{y_i} e^{-(\beta_0+\beta_1 x_i)}}{y_i!}\right] \end{aligned} \tag{2.7}$$

をパラメータ（β_0，β_1）で偏微分するのであるが，回帰分析の場合に比べて，かなり複雑である．

そこで，より簡単な 1 群のポアソン分布のあてはめについてニュートン・ラフソン法の適用を最初のステップとする．ポアソン分布の確率の対数を対数尤度 $\ln L_i$ とし，それらの合計を $\ln L$ と表記してきた．それぞれの y_i に対するポアソン分布の確率 P_i は，位置パラメータ μ に関して

$$P_i = \frac{\mu^{y_i} e^{-\mu}}{y_i!}, \quad i = 1, 2, \ldots \tag{2.8}$$

と与えられているので，y_i に対して μ を変化させた場合の確率を尤度 $L_i = P_i$ と置き換える．対数尤度 $\ln L$ は，

$$\begin{aligned} \ln L &= \sum_i \ln L_i \\ &= \sum_i \ln\left(\frac{\mu^{y_i} e^{-\mu}}{y_i!}\right) \end{aligned} \tag{2.9}$$

となる．それぞれの y_i に，n_i 個のデータがあるような場合の対数尤度 $\ln L_i$ は，

$$\begin{aligned} \ln L_i &= \ln\left(\frac{\mu^{y_i} e^{-\mu}}{y_i!}\right)^{n_i} \\ &= n_i \left(y_i \ln \mu - \mu - \ln y_i!\right) \end{aligned} \tag{2.10}$$

のように，y_i のポアソン確率に対して n_i 乗となる．対数尤度関数 $\ln L_i$ をパラメータ μ で偏微分すると

$$\begin{aligned} \boldsymbol{U}_i &= \frac{\partial \ln L_i}{\partial \mu} \\ &= n_i \left(\frac{y_i}{\mu} - 1 \right) \\ &= n_i \frac{y_i - \mu}{\mu} \end{aligned} \tag{2.11}$$

となり，さらにパラメータ μ で偏微分すると

$$\begin{aligned} \boldsymbol{H}_i &= \frac{\partial\partial \ln L_i}{\partial \mu \partial \mu} \\ &= n_i \frac{-y_i}{\mu^2} \end{aligned} \tag{2.12}$$

が得られる．これらを i について加え 1 階の偏微分ベクトル $\boldsymbol{U}$ と 2 階の偏微分行列 $\boldsymbol{H}$（ここでは共にスカラーであるが）を求める．

$$\boldsymbol{U} = \sum_i \boldsymbol{U}_i = \sum_i \left(n_i \frac{y_i - \mu}{\mu} \right) \tag{2.13}$$

$$\boldsymbol{H} = \sum_i \boldsymbol{H}_i = \sum_i \left(n_i \frac{-y_i}{\mu^2} \right) \tag{2.14}$$

数理統計の世界では，対数尤度関数の負の 2 階の偏微分行列 $(-\boldsymbol{H})$ を情報行列 $\boldsymbol{I}$（Fisher の情報量）というが，他の分野でも使われている一般的なヘッセ行列を使うことにする．また，対数尤度関数を数理統計の世界では「l」とするのが一般的であるが，尤度関数を L と表記し，対数尤度関数を $\ln L$ と区別して使う．

反復計算

ニュートン・ラフソン法は，位置パラメータ μ の最初の初期値を $\hat{\mu}^{(0)}$ としたときに，

$$\hat{\mu}^{(1)} = \hat{\mu}^{(0)} + (-\boldsymbol{H}^{(0)})^{-1} \boldsymbol{U}^{(0)} \tag{2.15}$$

を繰り返し計算して，対数尤度 $\ln L$ を最大化する方法である．最大化したとの判断は，対数尤度の増分が，10^{-6} 以下になった場合など適宜設定する．

ニュートン・ラフソン法は，対数尤度関数のパラメータで 1 階の偏微分したベクトル $\boldsymbol{U}$，さらに 2 階の偏微分行列 $\boldsymbol{H}$ を用いる方法として定式化されている．ポアソン回帰パラメータを $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \ \ \hat{\beta}_1]^T$ としたときに，

$$\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} + (-\boldsymbol{H}^{(0)})^{-1} \boldsymbol{U}^{(0)} \tag{2.16}$$

と式（2.15）と，全く同じ形式で回帰パラメータを推定することができる．

表 2.7 に示すように，(1) 回目の反復で初期値 $\hat{\mu}^{(0)}=2.0$ をセットすると． $\hat{\mu}^{(1)}=2.6757$ が計算される．Excel のシート上で，最初の（ $y_1=0$ ， $n_1=3$ ）については，

初期値： $\hat{\mu}^{(0)}=2.0$

ポアソン確率 P_i ： $P_1^{(0)}=\text{Poisson.dist}(y_1,\mu^{(0)},false)=0.1353$

尤度 L_i ： $L_1^{(0)}=(P_1^{(0)})^{n_1}=0.1353^3=2.4788\times10^{-3}$

対数尤度 $\ln L_i$ ： $\ln L_1^{(0)}=\ln(2.4788\times10^{-3})=-6.0000$

1 階の偏微分： $\boldsymbol{U}_1^{(0)}=\dfrac{\partial \ln L_1}{\partial\mu}=n_1\left(\dfrac{y_1-\hat{\mu}^{(0)}}{\hat{\mu}^{(0)}}\right)=3\left(\dfrac{0-2.0}{2.0}\right)=-3.0000$

2 階の偏微分： $\boldsymbol{H}_1^{(0)}=\dfrac{\partial\partial \ln L_1}{\partial\mu\partial\mu}=n_1\dfrac{-y_1}{(\hat{\mu}^{(0)})^2}=3\dfrac{-0}{2.0^2}=0.0000$

と計算される．以下， $i=2,3,...,9$ について同様に Excel のフィルハンドルで計算式をコピーする．

表 2.7　ニュートン・ラフソン法（初期値）

		(*m* -1)	変化量	(*m*)	対数尤度	1階の	2階の	負の逆行列
		μ^	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	μ^	ln*L*	偏微分 $\boldsymbol{U}$	偏微分 $\boldsymbol{H}$	$(-\boldsymbol{H})^{-1}$
		2.0000	0.6757	**2.6757**	**-212.9762**	**50.0000**	**-74.0000**	**0.0135**
i	*y*	*n*	ポアソン*P*	L_i	$\ln L_i$	$/\partial\mu$	$/\partial\mu\,\partial\mu$	標準誤差
1	0	3	0.1353	2.4788E-03	-6.0000	-3.0000	0.0000	*SE*
2	1	17	0.2707	2.2465E-10	-22.2165	-8.5000	-4.2500	0.1162
3	2	26	0.2707	1.7517E-15	-33.9782	0.0000	-13.0000	
4	3	16	0.1804	1.2636E-12	-27.3971	8.0000	-12.0000	
5	4	18	0.0902	1.5695E-19	-43.2984	18.0000	-18.0000	
6	5	9	0.0361	1.0385E-13	-29.8958	13.5000	-11.2500	
7	6	3	0.0120	1.7409E-06	-13.2611	6.0000	-4.5000	
8	7	5	0.0034	4.7968E-13	-28.3657	12.5000	-8.7500	
9	9	1	0.0002	1.9095E-04	-8.5635	3.5000	-2.2500	

次に，全ての *i* について以下の計算をする．

対数尤度： $\ln L^{(0)}=\sum_{i=1}^{9}\ln L_i^{(0)}=-6.0000-22.2165\cdots-8.5635=-212.9762$

1 階の偏微分： $\boldsymbol{U}^{(0)}=\sum_{i=1}^{9}\boldsymbol{U}_i^{(0)}=-3.00-8.50+\cdots+3.50=50.0000$

2 階の偏微分： $\boldsymbol{H}^{(0)}=\sum_{i=1}^{9}\boldsymbol{H}_i^{(0)}=0.00-4.25-\cdots-2.25=-74.0000$

$\boldsymbol{H}$ の負の逆行列： $(-\boldsymbol{H}^{(0)})^{-1}=-1/(-74.0000)=0.0135$

$\mu^{(1)}$ の計算：

$$\begin{cases}\hat{\mu}^{(1)}=\hat{\mu}^{(0)}+(-\boldsymbol{H}^{(0)})^{-1}\boldsymbol{U}^{(0)}\\ \quad=2.0+0.0135\times50.0000\\ \quad=2.0+0.6757\\ \quad=2.6757\end{cases}$$

$\hat{\mu}^{(1)} = 2.6757$ をコピーして，値のみを $\hat{\mu}^{(0)}$ セルにペースト

逐次計算： $\hat{\mu}^{(0)} <== \hat{\mu}^{(1)}$

すると，表 2.8 に示すように以上の計算が第 2 回目として $\hat{\mu}^{(2)} = 2.9811$ が計算される．さらに繰返しを行い，5 回目の反復で， $\hat{\mu}^{(4)} = \hat{\mu}^{(5)} = 3.0240$ となったので，解が求まったと判定する．

表 2.8　ニュートン・ラフソン法による反復過程

反復(m)	$\mu^{(m-1)}$	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	$\mu^{(m)}$	
1	2.0000	0.6757	**2.6757**	**初期値**
2	**2.6757**	0.3054	2.9811	
3	2.9811	0.0388	3.0199	
4	3.0199	0.0005	**3.0204**	
5	**3.0204**	0.0000	3.0204	**収束**

表 2.9 に 5 回目の反復で計算された結果を示す. 表の右端が, 負の偏微分 $\boldsymbol{H}$ の逆数 $(-\boldsymbol{H})^{-1} = 0.0308$ である．これが求めたかった推定値 $\hat{\mu}$ の分散の推定値である．平方根を取ると $\hat{\mu}$ に対する標準誤差 0.1756 が求まる．なお，Excel シートに入力した計算式に些細な入力ミスがあっても結果が表示されるのでミスの発見が困難であるが，反復計算をしたときに収束しないので，その時には，丁寧な計算式の見直しを行うことになる．

表 2.9　ニュートン・ラフソン法による対数尤度の最大化

		(m-1)	変化量	(m)	対数尤度	1階の	2階の	負の逆行列
		μ^	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	μ^	ln$L$	偏微分 $\boldsymbol{U}$	偏微分 $\boldsymbol{H}$	$(-\boldsymbol{H})^{-1}$
		3.0204	0.0000	**3.0204**	**-190.9517**	**0.0000**	**-32.4459**	**0.0308**
i	y	n	ポアソンP	L_i	lnL_i	$/\partial\mu$	$/\partial\mu\,\partial\mu$	標準誤差
1	0	3	0.0488	1.1608E-04	-9.0612	-3.0000	0.0000	SE
2	1	17	0.1473	7.2682E-15	-32.5553	-11.3716	-1.8634	0.1756
3	2	26	0.2225	1.0745E-17	-39.0721	-8.7838	-5.7000	
4	3	16	0.2240	4.0252E-11	-23.9359	-0.1081	-5.2615	
5	4	18	0.1692	1.2867E-14	-31.9841	5.8378	-7.8923	
6	5	9	0.1022	1.2151E-09	-20.5285	5.8986	-4.9327	
7	6	3	0.0514	1.3613E-04	-8.9019	2.9595	-1.9731	
8	7	5	0.0222	5.3878E-09	-19.0391	6.5878	-3.8365	
9	9	1	0.0028	2.8124E-03	-5.8737	1.9797	-0.9865	

JMP による切片のみのポアソン回帰の適用

JMP の一般化線形モデルで,「リンク関数」を恒等,「分布」を Poisson,「モデルの構成」は空白のまま（切片のみのモデル）とした結果を表 2.10 に示す.

表 2.10　JMP のポアソン回帰による推定結果

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	3.0204	0.1756	296	<.0001*

推定値の共分散

共分散	切片
切片	0.0308

元のデータ y_i の分散は，推定された $\hat{\mu}$ に等しいので，$Var(y)=\hat{\mu}=3.0204$ となるが，$\hat{\mu}$ の標準誤差 SE は，0.1756 と推定されている．これは，表 2.9 で計算されていた負の 2 階の偏微分行列 $\boldsymbol{H}$ の逆数

$\boldsymbol{H}$ の負の逆行列：　$(-\boldsymbol{H})^{-1}=1/32.4459=0.0308$,

$Var(\hat{\mu})=0.0308$ は，理論的なポアソン分布の分散は期待値 $\hat{\mu}$ に等しいことから，分散をデータ数で割って

$$Var(\hat{\mu})=\frac{\hat{\mu}}{98}=\frac{3.0204}{98}=0.0308$$

求めたものと等しくなり，その平方根は，$SE(\hat{\mu})=\sqrt{0.0308}=0.1756$ として計算され，表 2.10 の標準誤差に一致する．

一般的に，算術平均の SE は，不偏分散を自由度（$N-1$）で割った平方根から得るのであるが，ポアソン分布を仮定した場合の SE とは異なる．なお，標本データの場合の SE は，Excel の SumProduct()関数などを用いて $SE=0.1829$ と計算される．

N=	98	=Sum($\boldsymbol{n}$ の範囲)
平均 μ=	3.0204	=SumProduct($\boldsymbol{y}$の範囲, $\boldsymbol{n}$の範囲)/ N
偏差平方和 S=	317.9592	=SumProduct($\boldsymbol{n}$の範囲, ($\boldsymbol{y}$の範囲 – 平均)^2)
分散 V=	3.2779	=S / (N-1)
μ の SE=	0.1829	=Sqrt(V / N)

JMP による対数尤度関数の偏微分

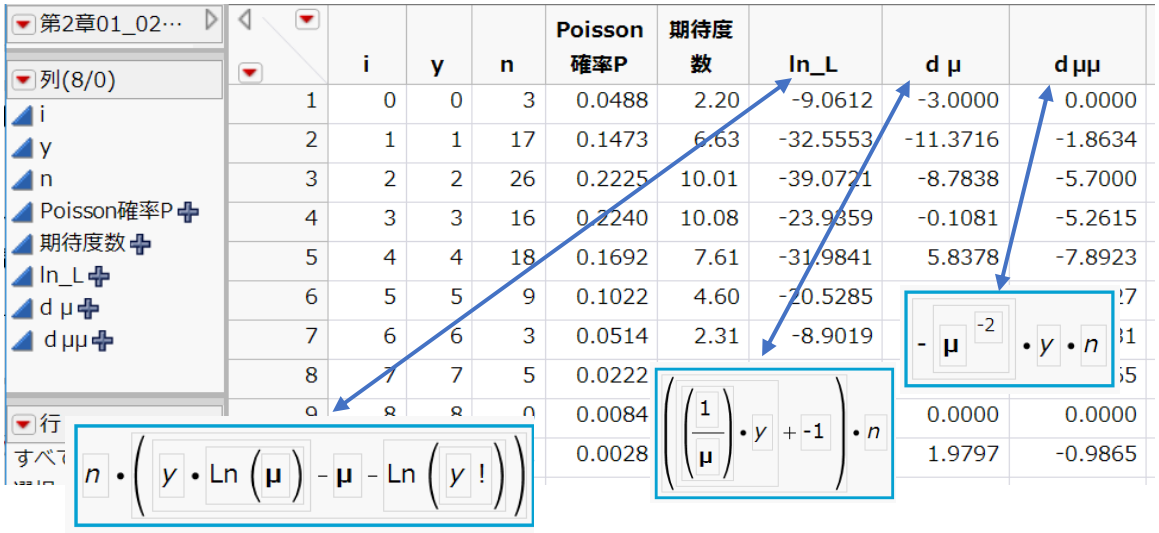
ニュートン・ラフソン法の計算では，対数尤度関数をパラメータに関する偏微分式を正確に解いて，その数式を Excel シートに入力する必要がある．少しでも数式にミスがあると解が収束しない．偏微分式のプラス・マイナスに入れ間違いミスもよく経験する．そのために，JMP の偏微分の機能を使い，Excel シート上での偏微分式の計算結果が一致することを常に確認している．表 2.9 と同様な JMP シートを表 2.11 に示す．「対数尤度 ln$L$」が「ln_L」に，「1 階の偏微分 $\boldsymbol{U}$」が「dμ に」，「2 階の偏微分 $\boldsymbol{H}$」が「dμμ」に対応する．

「Poisson 確率 P」に計算式が埋め込まれ，その引数 μ に対して，パラメータとして 3.0204

計算式：Poisson Probability(μ, y)，　パラメータ：μ = 3.0204

が設定されている．その計算式を用いて $y=0,1,...,9$ に対するポアソン分布の確率が計算され，

表 2.11　対数尤度の JMP 計算式エディタを用いた偏微式



	i	y	n	Poisson 確率P	期待度数	ln_L	d μ	d μμ
1	0	0	3	0.0488	2.20	-9.0612	-3.0000	0.0000
2	1	1	17	0.1473	6.63	-32.5553	-11.3716	-1.8634
3	2	2	26	0.2225	10.01	-39.0721	-8.7838	-5.7000
4	3	3	16	0.2240	10.08	-23.9859	-0.1081	-5.2615
5	4	4	18	0.1692	7.61	-31.9841	5.8378	-7.8923
6	5	5	9	0.1022	4.60	-20.5285		
7	6	6	3	0.0514	2.31	-8.9019		
8	7	7	5	0.0222				
9	8	8	0	0.0084			0.0000	0.0000
				0.0028			1.9797	-0.9865

$n \cdot \left(y \cdot \mathrm{Ln}(\mu) - \mu - \mathrm{Ln}(y!) \right)$

$\left(\left(\frac{1}{\mu} \right) \cdot y + -1 \right) \cdot n$

$-\mu^{-2} \cdot y \cdot n$

表 2.9 の Excel と同じ結果となっている．対数尤度 ln_L には，計算式

$$\text{ln\_L}：\ln L_i = n_i\left(y_i \ln \mu - \mu - \ln y_i!\right) = n \cdot \left(y \cdot \mathrm{Ln}(\mu) - \mu - \mathrm{Ln}(y!) \right)$$

が設定されている．偏微分 U に対する dμ には，

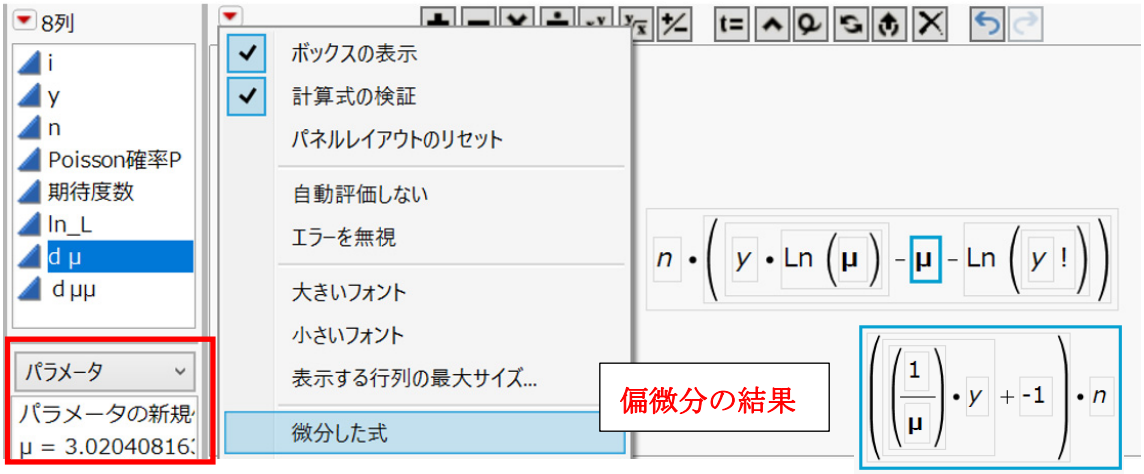
$$\text{d}\,\mu：U_i = \frac{n_i(y_i-\mu)}{\mu} = \left(\left(\frac{1}{\mu} \right) \cdot y + -1 \right) \cdot n$$

が計算式として設定されているが，表 2.12 に示すように，対数尤度 ln_L の式を μ について，JMP の微分の機能を用いて微分した結果で，自ら計算式を入力したものではない．偏微分 dμμ は，対数尤度 ln_L の式を μ について，2 回続けて微分した結果である．

$$\text{d}\,\mu\mu：H_i = n_i\frac{-y_i}{\mu^2} = -\mu^{-2} \cdot y \cdot n$$

表 2.11 の JMP シート上の数値と表 2.9 の Excel シートの数値が一致していれば，計算式の Excel シートへの入力ミスがないことが検証される．

表 2.12　JMP の計算エディタによる偏微分

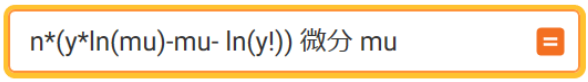


WolframAlpha による対数尤度関数の偏微分

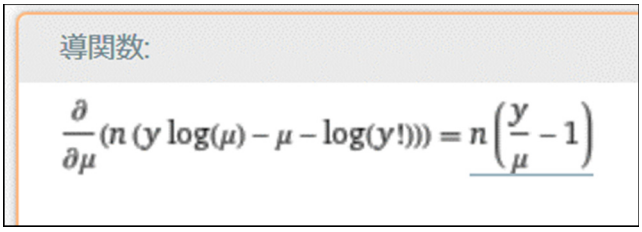
数学を支援するソフト無償版の WolframAlpha を用いて偏微分式を得ることもできる．

https://ja.wolframalpha.com/　（2020 年 4 月 27 日アクセス）

計算したい式に対数尤度関数を次のように入力すると



mu (μ) で微分した式が表示される．



JMP の偏微分の方法に比べ，洗礼された数式を得ることができる．このような数学ソフトなどの助けも借り，対数尤度関数の偏微分式を得ることにより，Excel による最尤法にチャレンジしてもらいたい．その経験が，最尤法は，どのような方法なのか」との質問に「尤度を最大にする方法である」との紋切り型な説明に引き続き，「実際には，...」との説明ができるようになることを願っている．

2.4. ポアソン回帰のバリエーション

ポアソン回帰には，いくつかのバリエーションがある．前章の第 1.4～1.13 節のタイトルを次に示す．

仮定する分布は，全てポアソン分布であるが，（恒等リンク or 対数リンク），（オフセットなし or オフセットあり），（1 群 or 2 群 or 要因配置），（過分散なし，あり）などの組み合わせになっている．基本は，「オフセットなし」の「恒等リンク」でのポアソン回帰である．なお，第 2.5 節では，「オフセットなし」で「対数リンク」の場合，第 2.6 節では，「オフセットあり」で「対数リンク」の場合を取り上げる．過分散がある場合は，第 6 章で取り上げる

恒等リンクにおけるポアソン回帰の対数尤度

ポアソン回帰のモデル式は，ポアソン分布のパラメータ μ を回帰式に置き換えた式

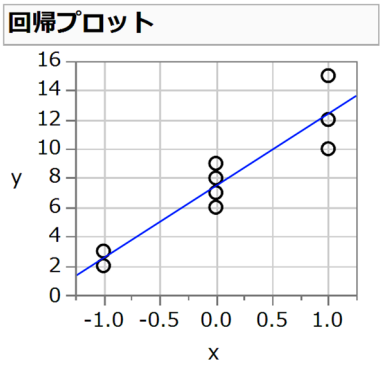
$$\mu_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, ..., n \tag{2.17}$$

$$\left.\begin{aligned} P_i &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \\ &= \frac{(\beta_0 + \beta_1 x_i)^{y_i} e^{-(\beta_0 + \beta_1 x_i)}}{y_i!} \end{aligned}\right\} \tag{2.18}$$

である．第 1.4 節の人工データについてニュートン・ラフソン法によるポアソン回帰を行う．表 2.13 に表 1.8 のデザイン行列およびデータ y を再掲する［ドブソン（2008）］．

表 2.13　人工データに対するポアソン回帰のあてはめ（表 1.8 再掲）

i	x_0	x_1	y
1	1	-1	2
2	1	-1	3
3	1	0	6
4	1	0	7
5	1	0	8
6	1	0	9
7	1	1	10
8	1	1	12
9	1	1	15



それぞれの i について対数尤度 $\ln L_i$ は，

$$\left.\begin{aligned}\ln L_i &= \ln\left[\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}\right] \\ &= \ln\left[\frac{(\beta_0+\beta_1 x_i)^{y_i} e^{-(\beta_0+\beta_1 x_i)}}{y_i!}\right] \\ &= y_i \ln(\beta_0+\beta_1 x_i) - (\beta_0+\beta_1 x_i) - \ln(y_i!)\end{aligned}\right\} \tag{2.19}$$

であり，対数尤度 $\ln L_i$ をパラメータ β_0 および β_1 で偏微分すると

$$U_{1i} = \frac{\partial \ln L_i}{\partial \beta_0} = \frac{y_i}{\beta_0+\beta_1 x_i} - 1 = \frac{y_i-\mu_i}{\mu_i} \tag{2.20}$$

$$U_{2i} = \frac{\partial \ln L_i}{\partial \beta_1} = \frac{\partial \ln L_i}{\partial \beta_0} x_i = \frac{y_i-\mu_i}{\mu_i} x_i \tag{2.21}$$

となり，さらに β_0 および β_1 で偏微分すると

$$\left.\begin{aligned}H_{1,1,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} = \frac{-y_i}{(\beta_0+\beta_1 x_i)^2} = \frac{-y_i}{\mu_i^2} \\ H_{1,2,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_1} = \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} x_i = \frac{-y_i}{\mu_i^2} x_i \\ H_{2,1,i} &= H_{1,2,i} = \frac{-y_i}{\mu_i^2} x_i \\ H_{2,2,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_1\partial\beta_1} = \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} x_i^2 = \frac{-y_i}{\mu_i^2} x_i^2\end{aligned}\right\} \tag{2.22}$$

となる．これらを i について加えスコアベクトル $\boldsymbol{U}$ とヘッセ行列 $\boldsymbol{H}$ にまとめる．

$$\boldsymbol{U} = \begin{bmatrix}\sum_i U_{1i} \\ \sum_i U_{2i}\end{bmatrix} = \begin{bmatrix}\sum_i (y_i-\mu_i)/\mu_i \\ \sum_i x_i (y_i-\mu_i)/\mu_i\end{bmatrix} \tag{2.23}$$

$$\boldsymbol{H} = \begin{bmatrix}\sum_i H_{1,1,i} & \sum_i H_{1,2,i} \\ \sum_i H_{2,1,i} & \sum_i H_{2,2,i}\end{bmatrix} = \begin{bmatrix}\sum_i -y_i/\mu_i^2 & \sum_i -x_i y_i/\mu_i^2 \\ \sum_i -x_i y_i/\mu_i^2 & \sum_i -x_i^2 y_i/\mu_i^2\end{bmatrix} \tag{2.24}$$

ニュートン・ラフソン法は，パラメータ $\boldsymbol{\beta}$ の最初の初期値を $\hat{\boldsymbol{\beta}}^{(0)}=[\hat{\beta}_0^{(0)} \quad \hat{\beta}_1^{(0)}]^T$ としたときに，次式により

$$\hat{\boldsymbol{\beta}}^{(1)}=\hat{\boldsymbol{\beta}}^{(0)}+(-\boldsymbol{H}^{(0)})^{-1}\boldsymbol{U}^{(0)} \tag{2.25}$$

繰り返し計算して，対数尤度 $\ln L$ を最大化する方法である．この式は，第 2.3 節でスカラーの場合の式（2.15）

$$\hat{\mu}^{(1)}=\hat{\mu}^{(0)}+(-\boldsymbol{H}^{(0)})^{-1}\boldsymbol{U}^{(0)} \tag{2.26}$$

と本質的に同じである．最大化したとの判断は，対数尤度の増分が，10^{-6} 以下になった場合など適宜設定する．

反復計算の実際

ニュートン・ラフソン法による計算方法は，第 2.3 節で位置パラメータが μ のみの場合について詳細に示した．ポアソン回帰の場合は，μ に変えて $\mu_i=\beta_0+\beta_1 x_i$ のように 2 つのパラメータなので，2×1 のスコアベクトル $\boldsymbol{U}$，2×2 のヘッセ行列 $\boldsymbol{H}$ となるが，基本の計算原理は全く同じである．

表 2.14 に示すように初期値には，第 1.4 節の反復重み付き回帰の場合と同じ初期値

$$\hat{\boldsymbol{\beta}}^{(0)}=\begin{bmatrix}7.0\\5.0\end{bmatrix}$$

を用いて，偏微分式を用いた反復計算を行う．

表 2.14　初期値に対するニュートン・ラフソン法による計算結果

			元の(m-1)	変化量	新たな(m)	1階の	2階の偏微分		負の逆行列	
			パラメータ	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	パラメータ	偏微分$\boldsymbol{U}$	$\boldsymbol{H}$		$(-\boldsymbol{H})^{-1}$	
		β_0^=	**7.0000**	0.4058	**7.4058**	0.8690	**-2.1192**	**0.9931**	0.6827	0.4499
		β_1^=	**5.0000**	-0.0091	**4.9909**	-0.4167	**0.9931**	**-1.5069**	0.4499	0.9601
i	x	y	μ^	P	$\ln L_i$	$\partial\beta_0$	$\partial\beta_1$	$\partial\beta_0\partial\beta_0$	$\partial\beta_0\partial\beta_1$	$\partial\beta_1\partial\beta_1$
1	-1	2	2.0000	0.2707	-1.3069	0.0000	0.0000	-0.5000	0.5000	-0.5000
2	-1	3	2.0000	0.1804	-1.7123	0.5000	-0.5000	-0.7500	0.7500	-0.7500
3	0	6	7.0000	0.1490	-1.9038	-0.1429	0.0000	-0.1224	0.0000	0.0000
4	0	7	7.0000	0.1490	-1.9038	0.0000	0.0000	-0.1429	0.0000	0.0000
5	0	8	7.0000	0.1304	-2.0373	0.1429	0.0000	-0.1633	0.0000	0.0000
6	0	9	7.0000	0.1014	-2.2886	0.2857	0.0000	-0.1837	0.0000	0.0000
7	1	10	12.0000	0.1048	-2.2553	-0.1667	-0.1667	-0.0694	-0.0694	-0.0694
8	1	12	12.0000	0.1144	-2.1683	0.0000	0.0000	-0.0833	-0.0833	-0.0833
9	1	15	12.0000	0.0724	-2.6257	0.2500	0.2500	-0.1042	-0.1042	-0.1042
				計	**-18.2021**	0.8690	-0.4167	**-2.1192**	**0.9931**	**-1.5069**

最初の $i=1$ の場合は，次のように計算されている．

$$\hat{\mu}_1^{(0)} = \hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_1 = 7.0000 + 5.0000 \times (-1) = 2.0000$$
$$P_1^{(0)} = \text{Poisson.dist}(y_1, \hat{\mu}_1^{(0)}, \mathit{false}) = 0.2702$$
$$\ln L_1^{(0)} = \ln(P_1^{(0)}) = -1.3069$$

$$U_{1,i=1}{}^{(0)} = \frac{\partial \ln L_1}{\partial \hat{\beta}_0^{(0)}} = \frac{y_1 - \hat{\mu}_1^{(0)}}{\hat{\mu}_1^{(0)}} = \frac{2 - 2.0000}{2.0000} = 0.0000$$

$$U_{2,i=1}{}^{(0)} = \frac{\partial \ln L_1}{\partial \hat{\beta}_1^{(0)}} = \frac{y_1 - \hat{\mu}_1^{(0)}}{\hat{\mu}_1^{(0)}} x_1 = \frac{2 - 2.0000}{2.0000}(-1) = 0.0000$$

$$H_{1,1,i=1}{}^{(0)} = \frac{\partial\partial \ln L_1}{\partial \hat{\beta}_0^{(0)} \partial \hat{\beta}_0^{(0)}} = \frac{-y_1}{(\mu_1^{(0)})^2} = \frac{-2}{2.0000^2} = -0.5000$$

$$H_{1,2,i=1}{}^{(0)} = \frac{\partial\partial \ln L_1}{\partial \hat{\beta}_0^{(0)} \partial \hat{\beta}_0^{(0)}} x_1 = (-0.5000) \times (-1) = 0.50000$$
$$H_{2,1,i=1}{}^{(0)} = H_{1,2,i=1}{}^{(0)} = 0.50000$$
$$H_{2,2,i=1}{}^{(0)} = \frac{\partial\partial \ln L_1}{\partial \hat{\beta}_0^{(0)} \partial \hat{\beta}_0^{(0)}} x_1^2 = (-0.5000) \times (-1)^2 = -0.5000$$

すべての $i=1,2,...,9$ について逐次計算され，計の行に i について和が

	$\ln L_i$	$\partial\beta_0$	$\partial\beta_1$	$\partial\beta_0\partial\beta_0$	$\partial\beta_0\partial\beta_1$	$\partial\beta_1\partial\beta_1$
計	**-18.2021**	0.8690	-0.4167	**-2.1192**	**0.9931**	**-1.5069**

計算されている．これらは，1 階の偏微分 $\boldsymbol{U}$ および 2 階の偏微分 $\boldsymbol{H}$ の欄に

1階の偏微分 $\boldsymbol{U}$	2階の偏微分 $\boldsymbol{H}$		負の逆行列 $(-\boldsymbol{H})^{-1}$	
0.8690	**-2.1192**	**0.9931**	0.6827	0.4499
-0.4167	**0.9931**	**-1.5069**	0.4499	0.9601

として代入されている．これらから，第 1 回目の反復は，$\hat{\boldsymbol{\beta}}^{(0)} = [7.0 \quad 5.0]^T$ としたときに，

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{(1)} &= \hat{\boldsymbol{\beta}}^{(0)} + (-\boldsymbol{H}^{(0)})^{-1}\boldsymbol{U}^{(0)} \\
&= \begin{bmatrix} 7.0000 \\ 5.0000 \end{bmatrix} + \begin{bmatrix} 0.6827 & 0.4499 \\ 0.4499 & 0.9601 \end{bmatrix} \begin{bmatrix} 0.8690 \\ -0.4167 \end{bmatrix} \\
&= \begin{bmatrix} 7.0000 \\ 5.0000 \end{bmatrix} + \begin{bmatrix} 0.4058 \\ -0.0091 \end{bmatrix} \\
&= \begin{bmatrix} 7.4058 \\ 4.9909 \end{bmatrix}
\end{aligned}
$$

による計算から，第 1 回目反復の $\hat{\boldsymbol{\beta}}^{(1)}=[7.4058, 4.9909]^T$ が得られる．これをコピーして，$\hat{\boldsymbol{\beta}}^{(0)}$ に“値”のみをペーストすると，第 2 回目の反復計算が行なわれる．この繰返しを収束するまで続ける．少々面倒なので，ソルバーを用いて対数尤度 $\ln L$ の **−18.2021** の欄を最大化するように元の(m-1)パラメータ($\hat{\beta}_0$，$\hat{\beta}_1$)を変化させても良い．表 2.15 に示すように，第 4 回目の反復で対数尤度の増分も 0.0000 となり，パラメータの変化量も 0.0000 となり収束する．

表 2.15　ニュートン・ラフソン法による反復過程

		元の(m-1)	変化量	新たな(m)	負の逆行列		対数尤度	
反復		パラメータ	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	パラメータ	$(-\boldsymbol{H})^{-1}$		$\ln L$	増分
1	β_0^=	7	0.4058	7.4058	**0.6827**	**0.4499**	-18.2021	
	β_1^=	5	-0.0091	4.9909	**0.4499**	**0.9601**		
2	β_0^=	7.4058	0.0445	7.4503			-18.0086	0.1934
	β_1^=	4.9909	-0.0532	4.9378				
3	β_0^=	7.4503	0.0013	7.4516			-18.0039	0.0047
	β_1^=	4.9378	-0.0025	4.9353				
4	β_0^=	7.4516	0.0000	7.4516	**0.7817**	**0.4160**	-18.0039	0.0000
	β_1^=	4.9353	0.0000	4.9353	**0.4160**	**1.1915**		

第 4 反復で収束した結果について行列 $(-\boldsymbol{H})^{-1}$ の対角要素について平方根を取り，推定されたパラメータの標準誤差 SE を推定する．

表 2.16　第 4 反復で収束した結果

	元の(m-1)	変化量	新たな(m)	1階の	2階の偏微分		負の逆行列			
	パラメータ	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	パラメータ	偏微分$\boldsymbol{U}$	$\boldsymbol{H}$		$(-\boldsymbol{H})^{-1}$		SE	
β^$_0$=	**7.4516**	0.0000	**7.4516**	0.0000	**-1.5711**	**0.5485**	0.7817	0.4160	0.8842	
β^$_1$=	**4.9353**	0.0000	**4.9353**	0.0000	**0.5485**	**-1.0308**	0.4160	1.1915		1.0915

JMP によるポアソン回帰

表 2.17 に JMP の一般化線形モデルで，リンク関数を「恒等」，分布を「poisson」とした結果を示す．パラメータの推定値および標準誤差が一致していることが確認できる．もちろんパラメータ推定値の共分散は，負のヘッセ行列の逆行列 $(-\boldsymbol{H})^{-1}$ と一致する．

表 2.17　JMP によるポアソン回帰の結果（表 1.7 再掲）

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	7.4516	0.8842	71.0299	<.0001*
x	4.9353	1.0915	16.5260	<.0001*

推定値の共分散

共分散

	切片	x
切片	0.7817	0.4160
x	0.4160	1.1915

複数の共変量をもつポアソン回帰の偏微分式

ここでは，単回帰の場合について示したが，共変量などを含む変数が複数ある場合への拡張は容易である．変数を増やし，$\boldsymbol{x}=[1\ \ x_{1i}\ \ x_{2i}\ \ \cdots\ \ x_{pi}]$ から成る $(p+1)$ 変数に（切片：$x_{0i}=1$）に拡張した場合，$(p+1)$ のスコアベクトル $\boldsymbol{U}^{(p)}$ は，$\boldsymbol{x}$ に要素 $(y_i-\mu_i)/\mu_i$ を掛けた式

$$\boldsymbol{U}^{(p)}=\begin{bmatrix}\sum_i U_{0i}\\ \sum_i U_{1i}\\ \vdots\\ \sum_i U_{pi}\end{bmatrix}=\begin{bmatrix}\sum_i x_{0i}(y_i-\mu_i)/\mu_i\\ \sum_i x_{1i}(y_i-\mu_i)/\mu_i\\ \vdots\\ \sum_i x_{pi}(y_i-\mu_i)/\mu_i\end{bmatrix} \tag{2.27}$$

となり，$(p+1)\times(p+1)$ のヘッセ行列 $\boldsymbol{H}^{(p)}$ は，2 次形式 $\boldsymbol{x}^T\boldsymbol{x}$ に要素 $[-y_i/\mu_i^2]$ を掛けた式

$$\boldsymbol{x}^T\boldsymbol{x}=[1\ \ x_{1i}\ \ x_{2i}\ \ \cdots\ \ x_{pi}]^T\ \ [1\ \ x_{1i}\ \ x_{2i}\ \ \cdots\ \ x_{pi}]$$

で求めることができる．

$$\left.\begin{aligned}\boldsymbol{H}^{(p)}&=\begin{bmatrix}\sum_i H_{0,0,i} & \sum_i H_{0,1,i} & \cdots & \sum_i H_{1,p,i}\\ \sum_i H_{1,0,i} & \sum_i H_{1,1,i} & & \sum_i H_{1,p,i}\\ \vdots & & \ddots & \vdots\\ \sum_i H_{p,0,i} & \sum_i H_{p,1,i} & \cdots & \sum_i H_{p,p,i}\end{bmatrix}\\ &=\begin{bmatrix}\sum_i -y_i/\mu_i^2 & \sum_i -x_{1,i}y_i/\mu_i^2 & \cdots & \sum_i -x_{p.i}y_i/\mu_i^2\\ \sum_i -x_{1,i}y_i/\mu_i^2 & \sum_i -x_{1i}{}^2y_i/\mu_i^2 & & \sum_i -x_{p,i}x_{1,i}y_i/\mu_i^2\\ \vdots & & \ddots & \vdots\\ \sum_i -x_{p,i}y_i/\mu_i^2 & \sum_i -x_{1,i}x_{p,i}y_i/\mu_i^2 & \cdots & \sum_i -x_{p,i}{}^2y_i/\mu_i^2\end{bmatrix}\end{aligned}\right\} \tag{2.28}$$

2.5. 対数リンクの場合のポアソン回帰

対数リンク

変数 x_i が増大するにつれ，観測データ y_i が指数関数的に増加することもしばしば経験する．位置パラメータ μ_i も指数関数的に増加すると仮定すると，回帰式 $\beta_0+\beta_1 x_i$ は，

$$\mu_i=\exp(\beta_0+\beta_1 x_i)+\varepsilon_i,\quad \varepsilon_i\sim\text{ポアソン分布},\quad i=1,2,...,n \tag{2.29}$$

と定義されるのでポアソン確率 P_i は，

$$\left.\begin{aligned} P_i&=\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}\\ &=\frac{\exp(\beta_0+\beta_1 x_i)^{y_i}e^{-\exp(\beta_0+\beta_1 x_i)}}{y_i!}\end{aligned}\right\} \tag{2.30}$$

となる．推定式の形にして両辺に対数をとって，

$$\ln(\hat{\mu}_i)=\hat{\beta}_0+\hat{\beta}_1 x_i,\quad i=1,2,...,n \tag{2.31}$$

のように線形化することができる．一般化線形モデルでは，この線形化する変換をリンク関数と言う．一般化線形モデルに対する統計ソフトでは，元の指数関数ではなく，線形化する関数名を使っていて，この場合のリンク関数は「対数」である．2 値反応が 2 項分布に従うとした場合のリンク関数としては，（プロビット・ロジット・補対数-対数）などがある．

第 1.9 節「植物の体サイズに関連した種子数（対数リンク，2 群，回帰）」で（対数リンク）としたのは，対数を取った場合に線形化できる事例とした［久保（2012）］．表 1.35 から得られるパラメータを用いた回帰式は，対数変換した

$$\ln(\hat{\mu}_i)=\hat{\beta}_0+\hat{\beta}_1 x_i=0.7459+0.1323x_i$$

であるが，グラフでは元のスケールで，

$$\hat{\mu}_i=\exp(\hat{\beta}_0+\hat{\beta}_1 x_i)=\exp(0.7459+0.1323x_i)$$

で表した．自然対数での目盛りは，元のスケールへの目視による換算は厄介なので，常用対数目盛とした結果を図 2.2 右に示す．切片は，$\exp(\hat{\beta}_0)=\exp(0.7459)=2.1083$ となる．

この図 2.2 のポアソン回帰の 95%信頼区間は，JMP のオプションでファイルに出力した結果を，「重ね合わせプロット」で別途作図したものである．

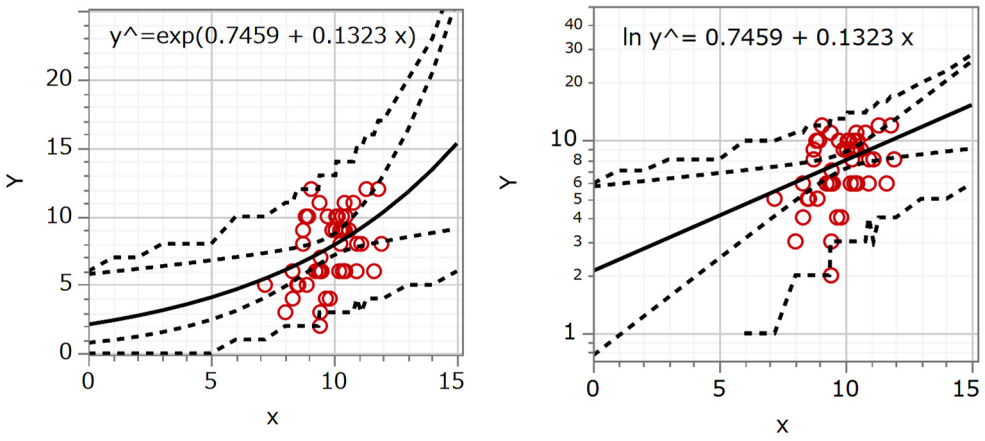


図 2.2　C 群に対するポアソン回帰のあてはめと 95%信頼区間の表示

一般化線形モデルの解法としての反復重み付き回帰では，図 2.2 右の対数変換した線形式を用いて，図 2.2 左の指数関数に対するポアソン回帰を間接的に行っている．詳細は，第 5 章を参照のこと．

対数リンクの場合の偏微分式

ニュートン・ラフソン法による最尤法は，図 2.2 左に示す指数関数に対して直接計算するので，図 2.2 右の対数変換した回帰式を使う必要はない．

$$\mu_i = \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i \qquad i = 1, 2, ..., n \tag{2.32}$$

$$\left.\begin{aligned} P_i &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \\ &= \frac{\exp(\beta_0 + \beta_1 x_i)^{y_i} e^{-\exp(\beta_0 + \beta_1 x_i)}}{y_i!} \end{aligned}\right\} \tag{2.33}$$

である．それぞれの i に対する対数尤度 $\ln L_i$ は，

$$\left.\begin{aligned} \ln L_i &= \ln\left[\frac{\exp(\beta_0 + \beta_1 x_i)^{y_i} e^{-\exp(\beta_0 + \beta_1 x_i)}}{y_i!}\right] \\ &= y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i) - \ln(y_i!) \end{aligned}\right\} \tag{2.34}$$

パラメータ β_0 および β_1 で偏微分すると

$$\left.\begin{aligned} U_{1i} &= \frac{\partial \ln L_i}{\partial \beta_0} = y_i - \exp(\beta_0 + \beta_1 x_i) = y_i - \mu_i \\ U_{2i} &= \frac{\partial \ln L_i}{\partial \beta_1} = y_i x_i - \exp(\beta_0 + \beta_1 x_i) x_i = (y_i - \mu_i) x_i \end{aligned}\right\} \tag{2.35}$$

となり，さらに β_0 および β_1 で偏微分すると

$$\left.\begin{aligned} H_{1,1,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} = -\exp(\beta_0+\beta_1 x_i) = -\mu_i \\ H_{1,2,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_1} = \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} x_i = -\mu_i x_i \\ H_{2,1,i} &= H_{1,2,i} = -\mu_i x_i \\ H_{2,2,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_1\partial\beta_1} = \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} x_i^2 = -\mu_i x_i^2 \end{aligned}\right\} \tag{2.36}$$

となる．これらを i について加えスコアベクトル $\boldsymbol{U}$ とヘッセ行列 $\boldsymbol{H}$ にまとめる．

$$\boldsymbol{U} = \begin{bmatrix} \sum_i U_{1i} \\ \sum_i U_{2i} \end{bmatrix} = \begin{bmatrix} \sum_i (y_i - \mu_i) \\ \sum_i (y_i - \mu_i) x_i \end{bmatrix} \tag{2.37}$$

$$\boldsymbol{H} = \begin{bmatrix} \sum_i H_{1,1,i} & \sum_i H_{1,2,i} \\ \sum_i H_{2,1,i} & \sum_i H_{2,2,i} \end{bmatrix} = \begin{bmatrix} \sum_i -\mu_i & \sum_i -\mu_i x_i \\ \sum_i -\mu_i x_i & \sum_i -\mu_i x_i^2 \end{bmatrix} \tag{2.38}$$

ニュートン・ラフソン法は，これまでも示してきたようにパラメータ $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \ \ \hat{\beta}_1]$ の最初の初期値を $\hat{\boldsymbol{\beta}}^{(0)}$ としたときに，

$$\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} + (-\boldsymbol{H}^{(0)})^{-1}\boldsymbol{U}^{(0)} \tag{2.39}$$

を繰り返し計算して，対数尤度 $\ln L$ を最大化する方法である．最大化したとの判断は，対数尤度の増分が，10^{-6} 以下になった場合など適宜設定する．式（2.39）は，式（2.25）と形式は全く同じである．対数尤度が異なれば，スコアベクトルおよびヘッセ行列も異なるが，対数尤度を最大化する計算式は，全く同じとなることが，ニュートン・ラフソン法の見通しの良さである．

Excel による反復計算

表 2.18 に Excel によるニュートン・ラフソン法で，対数尤度を最大化した結果を示す．計算シートの構成は，表 2.14 と同じであるが，埋め込まれている式が，「対数リンク」を反映している．異なるのは，$\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$，偏微分式 $\partial \ln L_i / \partial\beta_0, \ldots,$ $\partial\partial \ln L_i / \partial\beta_0\partial\beta_1$ の計算である．初期値として，$\hat{\beta}_0 = 0.70$，$\hat{\beta}_1 = 0.10$ を与えて，表 2.18 で示した反復の方法で，第 4 反復で収束した結果である．なお，初期値の見当がつかない場合には，$\ln y_i$ についての単回帰

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \ln \boldsymbol{Y}$$

で求めてもよい．

表 2.18　C 群に対するニュートン・ラフソン法によるポアソン回帰のあてはめ

			元の(m-1)	変化量	新たな(m)	1階の	2階の偏微分		負の逆行列	
			パラメータ	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	パラメータ	偏微分$\boldsymbol{U}$	$\boldsymbol{H}$		$(-\boldsymbol{H})^{-1}$	
		β_0^=	**0.7459**	0.0000	**0.7459**	-0.0019	**-389.0**	**-3865.2**	0.2657	-0.0265
		β_1^=	**0.1323**	0.0000	**0.1323**	-0.0202	**-3865.2**	**-38781**	-0.0265	0.0027
i	x	y	μ^	P	$\ln L_i$	$\partial\beta_0$	$\partial\beta_1$	$\partial\beta_0\partial\beta_0$	$\partial\beta_0\partial\beta_1$	$\partial\beta_1\partial\beta_1$
1	8.31	6	6.3279	0.1592	-1.8373	-0.3279	-2.7245	-6.33	-52.58	-437.0
2	9.44	6	7.3479	0.1408	-1.9607	-1.3479	-12.7243	-7.35	-69.36	-654.8
3	9.50	6	7.4065	0.1392	-1.9716	-1.4065	-13.3613	-7.41	-70.36	-668.4
4	9.07	12	6.9970	0.0263	-3.6384	5.0030	45.3773	-7.00	-63.46	-575.6
5	10.16	10	8.0820	0.1013	-2.2900	1.9180	19.4865	-8.08	-82.11	-834.3
:										
49	11.32	12	9.4222	0.0827	-2.4926	2.5778	29.1811	-9.42	-106.66	-1207.4
50	9.25	6	7.1656	0.1453	-1.9291	-1.1656	-10.7815	-7.17	-66.28	-613.1
			計		-116.2718	-0.0019	-0.0202	-389.00	-3865.20	-38780.7

表 2.19　ニュートン・ラフソン法の反復

反復		元の(m−1)	変化量	新たな(m)
m		パラメータ	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	パラメータ
1	β_0^=	**0.7000**	-0.0179	**0.6821**
	β_1^=	**0.1000**	0.0464	**0.1464**
2	β_0^=	0.6821	0.0565	0.7386
	β_1^=	0.1464	-0.0131	0.1333
3	β_0^=	0.7386	0.0073	0.7459
	β_1^=	0.1333	-0.0010	0.1323
4	β_0^=	**0.7459**	0.0000	**0.7459**
	β_1^=	**0.1323**	0.0000	**0.1323**

JMP による対数リンクでのポアソン回帰

JMP の一般化線形モデルによるポアソン回帰を行い，推定値および共分散行列を表 2.20 に示す．Excel での計算結果と一致していることが確認できる．

表 2.20　JMP による対数リンクでのポアソン回帰

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	0.7459	0.5154	2.0730	0.1499
x	0.1323	0.0516	6.5990	0.0102*

推定値の共分散

共分散

	切片	x
切片	0.2657	-0.026
x	-0.026	0.0027

2.6. 対数リンクでオフセットがある場合のポアソン回帰

第 1.5 節の「冠動脈心疾患の死亡者数」のデータには，年齢階層別の死亡者数をカウントした結果である［ドブソン（2003)］．そのため，人口統計で年齢層ごとの部分母集団の人数を知ることができ，死亡率を計算することが可能である．もちろん 2 項分布を仮定したロジスティック回帰を適用することも可能ではある．しかし，1%以下を対象とした推定となるので結果の解釈が煩わしい．そのために，一般的には，対数リンクを仮定し，分母を考慮するポアソン回帰の適用が好まれる．

表 2.21　オーストラリアのある地方の冠動脈心疾患の死亡者数（表 1.11 再掲）

	年齢層	死亡者数	母集団	死亡率	10万比
i	x	y	人数 n		人数
1	30	1	17,742	0.006%	5.6
2	35	5	16,554	0.030%	30.2
3	40	5	16,059	0.031%	31.1
4	45	12	13,083	0.092%	91.7
5	50	25	10,784	0.232%	231.8
6	55	38	9,645	0.394%	394.0
7	60	54	10,706	0.504%	504.4
8	65	65	9,933	0.654%	654.4
	合計	205	104,506	0.196%	196.2

年齢層は，30-34, 35-39 のように与えられている．

指数関数のあてはめ

モデルは，死亡数が指数的に増えているので，部分母集団の人数 n_i を含んだ指数関数

$$\mu_i = n_i \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i \qquad i = 1, 2, ..., n \tag{2.40}$$

で死亡数の推定をする．推定式の形にして両辺対数を取ると

$$\ln(\hat{\mu}_i) = \ln(n_i) + \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1, 2, ..., n \tag{2.41}$$

オフセット $\ln(n_i)$ を含む線形式となる．これは，両パラメータ β_0 と β_1 を共通とし，切片 β_0 を $\ln(n_i)$ 分かさ上げをした回帰直線を求めることと解される．母集団の人数を 10,000 人と固定すれば，切片 β_0 を $\ln(10,000) = 9.2103$ 分かさ上げすることに対応する．

対数尤度関数の偏微分

対数を取らない式では，

$$\left.\begin{aligned} P_i &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \\ &= \frac{\left[n_i \exp(\beta_0 + \beta_1 x_i)\right]^{y_i} e^{-n_i \exp(\beta_0 + \beta_1 x_i)}}{y_i!} \end{aligned}\right\} \tag{2.42}$$

である．ニュートン・ラフソン法による最尤法では，この式の対数を取った対数尤度関数をパラメータで偏微分する．

それぞれの i に対する対数尤度 $\ln L_i$ は，

$$\left.\begin{aligned}\ln L_i &= \ln\left\{\frac{\left[n_i \exp(\beta_0+\beta_1 x_i)\right]^{y_i}\, e^{-n_i \exp(\beta_0+\beta_1 x_i)}}{y_i!}\right\} \\ &= y_i[\ln(n_i)+(\beta_0+\beta_1 x_i)] - n_i \exp(\beta_0+\beta_1 x_i) - \ln(y_i!)\end{aligned}\right\} \tag{2.43}$$

であり，パラメータ β_0 および β_1 で偏微分すると

$$\left.\begin{aligned}U_{1i} &= \frac{\partial \ln L_i}{\partial \beta_0} = y_i - n_i \exp(\beta_0+\beta_1 x_i) = y_i - \mu_i \\ U_{2i} &= \frac{\partial \ln L_i}{\partial \beta_1} = y_i x_i - n_i \exp(\beta_0+\beta_1 x_i) x_i = (y_i - \mu_i) x_i\end{aligned}\right\} \tag{2.44}$$

ただし， $n_i \exp(\beta_0+\beta_1 x_i) = \mu_i$

となり，さらに β_0 および β_1 で偏微分すると

$$\left.\begin{aligned}H_{1,1,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} = -n_i \exp(\beta_0+\beta_1 x_i) = -\mu_i \\ H_{1,2,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_1} = \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} x_i = -\mu_i x_i \\ H_{2,1,i} &= H_{1,2,i} = -\mu_i x_i \\ H_{2,1,i} &= \frac{\partial\partial \ln L_i}{\partial\beta_1\partial\beta_1} = \frac{\partial\partial \ln L_i}{\partial\beta_0\partial\beta_0} x_i^2 = -\mu_i x_i^2\end{aligned}\right\} \tag{2.45}$$

となる．これらを i について加えスコアベクトル $\boldsymbol{U}$ とヘッセ行列 $\boldsymbol{H}$ にまとめる．

$$\boldsymbol{U} = \begin{bmatrix}\sum_i U_{1i} \\ \sum_i U_{2i}\end{bmatrix} = \begin{bmatrix}\sum_i (y_i - \mu_i) \\ \sum_i (y_i - \mu_i) x_i\end{bmatrix} \tag{2.46}$$

$$\boldsymbol{H} = \begin{bmatrix}\sum_i H_{1,1,i} & \sum_i H_{1,2,i} \\ \sum_i H_{2,1,i} & \sum_i H_{2,2,i}\end{bmatrix} = \begin{bmatrix}\sum_i -\mu_i & \sum_i -\mu_i x_i \\ \sum_i -\mu_i x_i & \sum_i -\mu_i x_i^2\end{bmatrix} \tag{2.47}$$

なお，オフセットがない場合は，$\mu_i = \exp(\beta_0+\beta_1 x_i)$ と n_i を含まないだけで，スコアベクトル $\boldsymbol{U}$ とヘッセ行列 $\boldsymbol{H}$ の形式は同じである．

反復計算

ニュートン・ラフソン法は，パラメータ $\boldsymbol{\beta}=[\beta_0\ \ \beta_1]^T$ の最初の初期値を $\hat{\boldsymbol{\beta}}^{(0)}$ としたときに，

$$\hat{\boldsymbol{\beta}}^{(1)}=\hat{\boldsymbol{\beta}}^{(0)}+(-\boldsymbol{H})^{-1}\boldsymbol{U} \qquad (2.48)$$

を繰り返し計算して，対数尤度 $\ln L$ を最大化する方法である．最大化したとの判断は，対数尤度の増分が，10^{-6} 以下になった場合など適宜設定する．

表 2.22 にニュートン・ラフソン法による繰返し計算の Excel シートを示す．得られた推定値は，$\beta_0=-11.6278$，$\beta_1=0.1044$ である．これは，表 1.12 の JMP でのポアソン回帰で求めた推定値に一致する．

表 2.22　冠動脈心疾患の死亡者数のニュートン・ラフソン法による解

				元の	変化量	新たな	1階の	2階の偏微分		負の逆行列	
				パラメータ	$(-\boldsymbol{H})^{-1}\boldsymbol{U}$	パラメータ	偏微分$\boldsymbol{U}$	$\boldsymbol{H}$		$(-\boldsymbol{H})^{-1}$	
			β^_0=	**-11.6278**	0.0000	**-11.6278**	0.0000	**-205.00**	**-11750**	0.2053	-0.0035
			β^_1=	**0.1044**	0.0000	**0.1044**	0.0000	**-11750**	**-689869**	-0.0035	0.0001
i	x	y	n	μ^	ポアソンP	$\ln L_i$	$\partial\beta_0$	$\partial\beta_1$	$\partial\beta_0\partial\beta_0$	$\partial\beta_0\partial\beta_1$	$\partial\beta_1\partial\beta_1$
1	30	1	17,742	3.629	0.0963	-41519	-2.63	-78.88	-3.63	-109	-3266
2	35	5	16,554	5.708	0.1676	-29568	-0.71	-24.78	-5.71	-200	-6992
3	40	5	16,059	9.335	0.0522	-47429	-4.33	-173.38	-9.33	-373	-14935
4	45	12	13,083	12.819	0.1113	-28718	-0.82	-36.87	-12.82	-577	-25959
5	50	25	10,784	17.812	0.0220	-41180	7.19	359.38	-17.81	-891	-44531
6	55	38	9,645	26.855	0.0083	-46161	11.14	612.97	-26.86	-1477	-81237
7	60	54	10,706	50.250	0.0473	-32670	3.75	224.99	-50.25	-3015	-180900
8	65	65	9,933	78.591	0.0142	-42284	-13.59	-883.43	-78.59	-5108	-332048
					計	**-309529**	0.0000	0.0000	-205.00	-11750	-689869

オフセット

推定された回帰式は，

$$\begin{aligned}\hat{\mu}_i &= n_i\exp(\hat{\beta}_0+\hat{\beta}_1x_i)\\ &= n_i\exp(-11.6278+0.1044x_i)\end{aligned}$$

であり，両辺対数を取ると

$$\begin{aligned}\ln(\hat{\mu}_i) &= \ln(n_i)+\hat{\beta}_0+\hat{\beta}_1x_i\\ &= \ln(n_i)-11.6278+0.1044x_i\end{aligned}$$

のようにオフセット $\ln(n_i)$ を含む線形式となる．推定された回帰式には，オフセット項を含んだ形になっていて，通常の回帰分析の結果と異なる，どのように理解したら良いのだろうか．オフセット $\ln(n_i)$ に実際の数値を入れた場合に，

$$30\text{ 歳：}\begin{cases}\ln(\hat{\mu}_1)=\ln(17,742)-11.6278+0.1044x\\ \quad=9.7837-11.6278+0.1044x\\ \quad=-1.8442+0.1044x\end{cases}$$

：

$$65\text{ 歳：}\begin{cases}\ln(\mu_8)=\ln(9,933)-11.6278+0.1044x\\ \quad=-2.4242+0.1044x\end{cases}$$

のように，各年代の分母の大きさに応じて切片を上下させている．各年代について切片を計算した結果を表 2.23 に示す．

表 2.23　オフセット値を考慮した切片

	年齢層	死亡者数		母集団	オフセット	調整後	傾き
i	x	y	$\ln y$	人数 n	$\ln n$	$\ln n+\beta_0$^	β_1^
1	30	1	0.0000	17,742	9.7837	-1.8441	0.1044
2	35	5	1.6094	16,554	9.7144	-1.9134	0.1044
3	40	5	1.6094	16,059	9.6840	-1.9438	0.1044
4	45	12	2.4849	13,083	9.4791	-2.1487	0.1044
5	50	25	3.2189	10,784	9.2858	-2.3420	0.1044
6	55	38	3.6376	9,645	9.1742	-2.4536	0.1044
7	60	54	3.9890	10,706	9.2786	-2.3492	0.1044
8	65	65	4.1744	9,933	9.2036	-2.4242	0.1044
	合計	205		104,506	β_0^=	**-11.6278**	

図 2.3 に年齢と死亡者数の対数についての散布図上に，各年代別の回帰直線を重ね書きした結果を示す．直線は，切片まで伸ばさずに各年齢の前後に限定して描いた．この例では，母集団の人数に大きな差がなかったので，直線の上下の振れはわずかである．

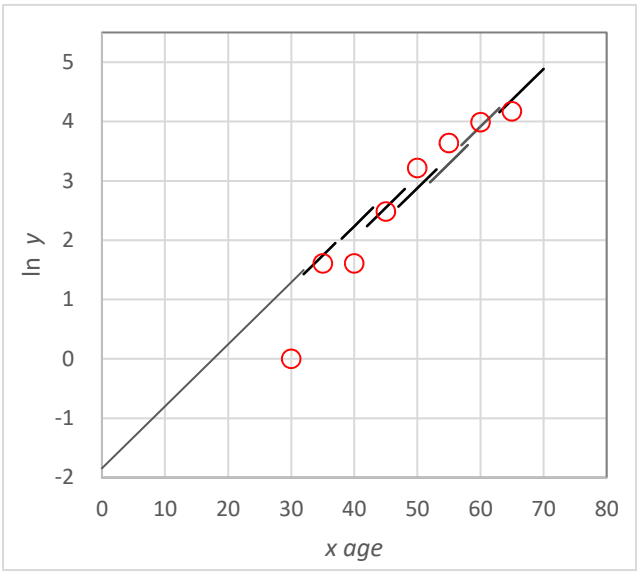


図 2.3　オフセットを含む場合の回帰直線の例示

1 万人比

オフセットがある場合には，母集団の平均的で切れの良い母集団の大きさ，この例であれば 10,000 人と設定し，年齢層 $x_1 = 30$ に対し 1 万人あたりの死亡数を計算し，

$$y_1^{(1万人)} = 10{,}000\frac{x_1}{n_1} = 10{,}000 \times \frac{1}{17{,}742} = 0.56$$

死亡者数の推定値を

$$\begin{aligned}\hat{\mu}_1^{(1万人)} &= n_i \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1) \\ &= 10{,}000 \times \exp(-11.6278 + 0.1044 \times 30) = 2.05\end{aligned}$$

のように計算し，グラフ化することが望ましい．

表 2.24　1 万人比に換算したポアソン回帰の推定値

	年齢層	死亡者数	母集団	死亡者数	推定
i	x	y	n	1 万人比	1 万人比
1	30	1	17,742	0.56	2.05
2	35	5	16,554	3.02	3.45
3	40	5	16,059	3.11	5.81
4	45	12	13,083	9.17	9.80
5	50	25	10,784	23.18	16.52
6	55	38	9,645	39.40	27.84
7	60	54	10,706	50.44	46.94
8	65	65	9,933	65.44	79.12
	$\hat{\beta}_0$ =	-11.6278	$\hat{\beta}_1$ =	0.1044	

死亡者数 1 万人比について図 2.4 に通常の目盛り，および，対数目盛にした場合を示す．

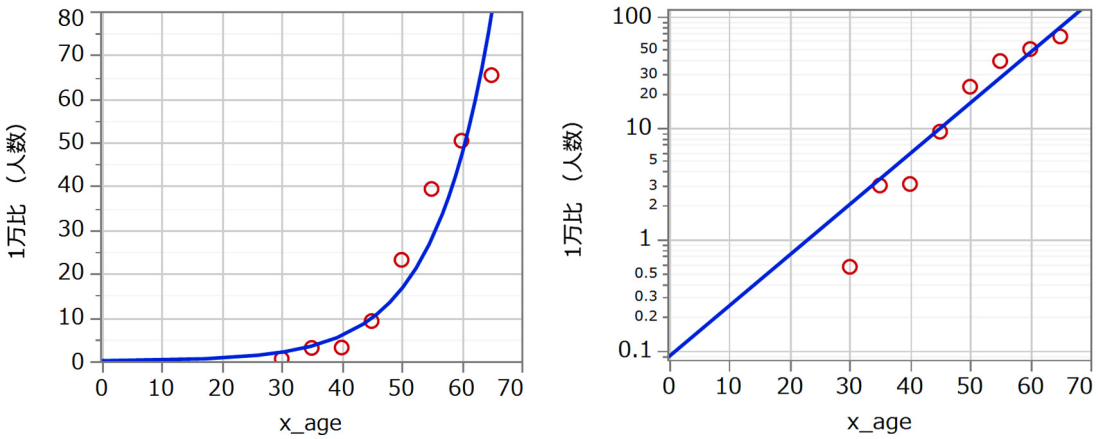


図 2.4　1 万人比で換算したポアソン回帰

オフセットがある場合のポアソン回帰で推定されたパラメータは，オフセットが 1 人の場合の死亡者数（死亡率）を計算していることになる．なお，第 5.5 節では，反復重み付き回帰によるポアソン回帰で 2 次式のあてはめた結果を示す．

2 値反応としたロジスティック回帰

第 1.5 節の表 1.14 に示したように「分布」を「2 項」，「リンク関数」を「ロジット」とした場合の一般化線形モデルの結果から，

$$\hat{\beta}_0 = -11.6395, \quad \hat{\beta}_1 = 0.1047$$

が得られることを示した．これは，いわゆるロジスティック回帰であるが，表 2.25 に示すように，死亡率 π_i に対してロジット変換

$$logit_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2.49}$$

した場合について $logit_i$ の推定値である．推定された $logit$^ から元の死亡率を推定するためには，式（2.49）を π について解いた逆ロジットの次式

$$\left.\begin{aligned}\hat{\pi}_i &= \frac{\exp(logit_i\text{^})}{1+\exp(logit_i\text{^})} \\ &= \frac{\exp(\hat{\beta}_0+\hat{\beta}_0 x_i)}{1+\exp(\hat{\beta}_0+\hat{\beta}_0 x_i)}\end{aligned}\right\} \tag{2.50}$$

で推定することになる．逆ロジットは，死亡率が 0.0 から 1.0 の間でシグモイド曲線をあてはめる．この例では，160 歳強でほぼ 1.0 となることが計算されている．死亡者数が得られた範囲内に限定すれば，オフセット付きのポアソン回帰でもロジスティック回帰でも，ほとんど同じパラメータの推定値が得られるので，どちらが優れた方法とは言い難い．観測された死亡者数 y_i に対して統計モデルをあてはめるという観点からは，オフセット付きのポアソン回帰が望ましいように思われる．

表 2.25　分布を 2 項分布，リンク関数をロジットとした場合の結果

年齢層	死亡者数	母集団	死亡率	ロジット	推定値	逆ロジット
x	*y*	*n*	%	*logit*	*logit*^	%
30	1	17742	0.0056	-9.7836	-8.4985	0.0204
40	5	16059	0.0311	-8.0743	-7.4515	0.0580
50	25	10784	0.2318	-6.0646	-6.4045	0.1651
60	54	10706	0.5044	-5.2845	-5.3575	0.4691
80					-3.2635	3.6845
100					-1.1695	23.6945
120					0.9245	71.5958
140					3.0185	95.3403
160					5.1125	99.4015
		β_0^=	-11.6395	β_1^=	0.1047	

死亡率の上限を新たな変数としたロジスティック回帰

ロジスティック回帰では，死亡率が（0%～100%）の範囲のシグモイド曲線のあてはめを前提にしている．発想を変えて，死亡率の上限をデータから推定することも Excel のソルバーを用いると容易にできる．

式（2.51）は，下限を 0.0，上限を 1.0 とするロジスティック曲線であるが，下限も上限も変数として与えることができる．下限を L_{limit}，上限を U_{limit} とするロジスティック曲線は，

$$\hat{\pi}_i = L_{limit} + \frac{(U_{limit} - L_{limit})\exp(\hat{\beta}_0 + \hat{\beta}_0 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_0 x_i)} \qquad (2.51)$$

で与えられる．一般化線形モデルとしてパラメータの推定はできないが，ソルバーならば推定可能である．表 2.26 に通常のロジスティック曲線のパラメータ推定値と上限付きのロジスティック曲線で，$\hat{U}_{limit} = 0.0074$ として推定されている．図 2.5 にそれぞれの曲線をプロットした結果を示す．対数尤度の比較からも通常のロジスティック曲線雄あてはめは支持されない．

表 2.26　上限を持つロジスティック曲線のあてはめ

					β_0=	**-11.6395**		β_0=	**-10.1994**	
					β_1=	**0.1047**	lnL	β_1=	**0.1862**	lnL
					U_{limit}=	-	**-25.1465**	U_{limit}=	**0.0074**	**-19.0732**
	年齢層	死亡者数	母集団	死亡率	**上限が 1.0 のシグモイド**			**上限が U_{limit} のシグモイド**		
i	x	y	人数 n	p	π^	$P^{binomial}$	lnL_i	π^	$P^{binomial}$	lnL_i
1	30	1	17,742	0.0001	0.0002	0.0971	-2.3317	0.0001	0.3557	-1.0336
2	35	5	16,554	0.0003	0.0003	0.1679	-1.7846	0.0002	0.1001	-2.3014
3	40	5	16,059	0.0003	0.0006	0.0523	-2.9505	0.0004	0.1245	-2.0838
4	45	12	13,083	0.0009	0.0010	0.1114	-2.1949	0.0010	0.1059	-2.2453
5	50	25	10,784	0.0023	0.0017	0.0221	-3.8124	0.0021	0.0738	-2.6062
6	55	38	9,645	0.0039	0.0028	0.0084	-4.7755	0.0038	0.0620	-2.7805
7	60	54	10,706	0.0050	0.0047	0.0475	-3.0464	0.0053	0.0496	-3.0045
8	65	65	9,933	0.0065	0.0079	0.0143	-4.2504	0.0064	0.0489	-3.0178
	70				0.0133			0.0070		
	80				0.0369			0.0073		
	100				0.2374			0.0074		
	200				0.9999			0.0074		

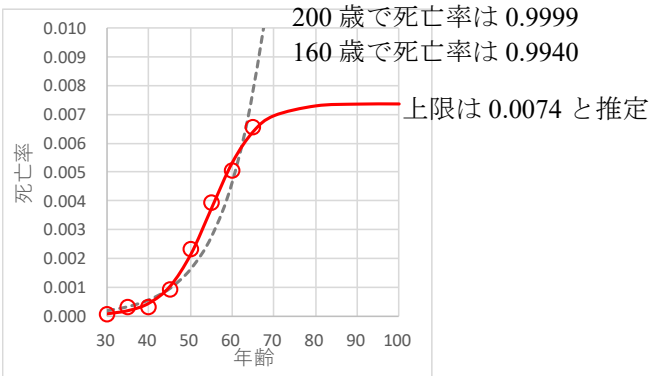


図 2.5　上限をパラメータとしたロジスティック曲線

3. 尤度比検定のためのデザイン行列

ポアソン分布を仮定した 2 群間の尤度比検定については，第 1.6 節でポアソン回帰による方法を示した．尤度比検定は，身近な 2×2 の分割表に対する検定の一つとして良く知られており，ポアソン分布を仮定した 2 群間の尤度比検定の考え方と対比する．ポアソン回帰を使用して様々な解析を行うためには，回帰分析の基礎となるデザイン行列について習熟し，目的とする比較を行うために適合するデザイン行列を自ら生成することが必要である．そのために，2×2 の要因配置型のデータに対してポアソン回帰を行うために必要な各種のデザイン行列について基礎的な考え方に基づいた応用方法を示す．さらに，2 本のポアソン回帰直線をあてはめる場合に，切片が共通の場合，傾きが共通の場合，交互作用を検討する場合，別々の直線をあてはめる場合，などに必要なデザイン行列の型の選択方法について示す．

3.1. 2×2 の分割表に対する尤度比検定の基礎

尤度比検定は，2×2 の分割表に対する Pearson のカイ 2 乗検定と同様に良く知られている．そこで，2×2 の分割表に対するこれらの検定方法を復習し，ポアソン分布を仮定した 2 群間の尤度比検定の基礎となる考え方を示す．表 3.1 に示す 2×2 の分割表は，第 1.6 節の表 1.15 の 1 日当たりの犯罪件数データを犯罪の（0:なし，1:あり）にまとめ直したものである［アルトマン（1999)］.

表 3.1　インド 3 地域の 5 年間の満月と新月の日に起きた犯罪の件数

	犯罪		
	0:なし	1:あり	計
1:満月	40	143	183
2:新月	114	72	186
計	154	215	369

	犯罪		
	0:なし	1:あり	計
1:満月	n_{11}	n_{12}	$n_{1\cdot}$
2:新月	n_{21}	n_{22}	$n_{2\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

分割表に対する 2 種類の検定

JMP の「二変量の関係」を用い，2×2 の分割表としての解析結果を表 3.2 に示す．尤度比のカイ 2 乗値が 60.9396，Pearson のカイ 2 乗値が 58.9845 として出力されている．これらのカイ 2 乗値が，自由度 1 のカイ 2 乗分布に従うことから p 値が求められている．

表 3.2　満月と新月の日に起きた犯罪の有無に対する 2×2 の分割表に対する検定

月	度数 期待値	犯罪 0:なし	犯罪 1：あり	合計
1:満月		40 76.374	143 106.626	183
2:新月		114 77.626	72 108.374	186
合計		154	215	369

検定	カイ2乗	p値(Prob>ChiSq)
尤度比	60.9396	<.0001*
Pearson	58.9845	<.0001*

どちらの検定も，2×2 の分割表に対する期待度数 $\hat{\mu}_{ij}$ を用いて定式化されている．期待度数は，周辺の度数を用いて計算される．ここで，$n_{i\cdot}$ は列方向の和，$n_{\cdot j}$ は行方向の和，$n_{\cdot\cdot}$ はセル全体の和である．期待度数 $\hat{\mu}_{ij}$ は，次のように周辺度数から求められた割合の積から期待割合を算出し，全体の数 $n_{\cdot\cdot}$ を掛けて期待度数を算出している．

$$\hat{\mu}_{ij} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \cdot \frac{n_{\cdot j}}{n_{\cdot\cdot}} \cdot n_{\cdot\cdot} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}, \quad i = 1,2, \quad j = 1,2 \tag{3.1}$$

Pearson のカイ 2 乗検定統計量は，各セルの実現値 n_{ij} と期待度数 $\hat{\mu}_{ij}$ の差の平方を期待度数 $\hat{\mu}_{ij}$ で割り，全てのセルの和として定義されている．

$$\left.\begin{aligned} \chi^2_{Pearson} &= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \\ &= \frac{(40 - 76.374)^2}{76.374} + \cdots + \frac{(72 - 108.374)^2}{108.374} \\ &= 17.3235 + \cdots + 12.2083 \\ &= 58.9845 \end{aligned}\right\} \tag{3.2}$$

簡便な計算公式として，セルの対角要素の積の差を平方し $n_{\cdot\cdot}$ を掛けて，全ての周辺の和で割った計算式

$$\left.\begin{aligned} \chi^2_{Pearson} &= \frac{(n_{11}n_{22} - n_{21}n_{12})^2 n_{\cdot\cdot}}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} \\ &= \frac{(40 \times 72 - 114 \times 143)^2 \times 369}{183 \times 186 \times 154 \times 215} \\ &= 58.9845 \end{aligned}\right\} \tag{3.3}$$

も良く知られている．

尤度比カイ 2 乗値は，各セルの実現値 n_{ij} を期待度数 $\hat{\mu}_{ij}$ で割り，対数を取り，実現値 n_{ij} を掛け，全てのセルの和とする計算手順が知られている．この手順は，多くの統計の教科書でも示されていて，もちろん統計ソフトでの計算手順の解説でもこの方法が示されている．

$$\left.\begin{aligned}\chi^2_{likelihood} &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\ln\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)\\ &= 2\left[40\times\ln\left(\frac{40}{76.3740}\right)+\cdots+72\times\ln\left(\frac{72}{108.3740}\right)\right]\\ &= (-51.7410)+\cdots+(-58.8848)\\ &= 60.9396\end{aligned}\right\} \qquad (3.4)$$

これらの式を用いて Excel によって計算した結果を表 3.3 に示す．この結果は，もちろん表 3.2 に示した JMP による計算結果と一致する．

表 3.3　Excel による期待度数から求めた Pearson および尤度比のカイ 2 乗値

	0:なし	1:あり	計	期待度数		Pearson		尤度比	
1:満月	40	143	183	76.3740	106.6260	17.3235	12.4085	-51.7410	83.9459
2:新月	114	72	186	77.6260	108.3740	17.0441	12.2083	87.6194	-58.8848
計	154	215	369			総和	58.9845	総和	60.9396

期待度数 $\hat{\mu}_{ij}$ を用いた計算手順は，簡潔ではあるが，表 1.19 で示したポアソン分布を仮定した 2 群間の尤度比検定の計算の考え方とは異なる．なお，式（3.4）は，簡便な計算公式としての認識なく広く使われている．

出現確率を用いた尤度比検定

第 1.6 節の表 1.19 で計算したポアソン分布を仮定した尤度比検定と同様に，2×2 の分割表に対しても出現確率を求めて，ポアソン分布と同様な方法を試みる．表 3.4 に示すように満月における 1 日当たりの犯罪が［1:あり］の確率は，$P_{満月,1}=143/183=0.7814$ であり，143 日分の尤度は，$L_{満月,1}=0.7814^{143}$ となる．犯罪が「0:なし」の 40 日分の尤度は，$L_{満月,0}=0.2186^{40}$ である．満月の日の対数尤度の和 $\ln L_{満月}$ は，

$$\left.\begin{aligned}\ln L_{満月} &= \ln(0.2186^{40})+\ln(0.7814^{143})\\ &= -60.8243-35.2697=-96.0940\end{aligned}\right\} \qquad (3.5)$$

表 3.4　出現率に対する対数尤度

	満月			**新月**			**満月＋新月**		
犯罪	$n_{満月}$	$P_{満月}$	$\ln L_{満月}$	$n_{新月}$	$P_{新月}$	$\ln L_{新月}$	$n_{満+新}$	$P_{満+新}$	$\ln L_{満+新}$
0:なし	40	0.2186	-60.8243	114	0.6129	-55.8085	154	0.4173	-134.5720
1:あり	143	0.7814	-35.2697	72	0.3871	-68.3338	215	0.5827	-116.1341
計	183		-96.0940	186		-124.1423	369		-250.7061
	$L_{満月}$		1.8489E-42	$L_{新月}$		1.2181E-54	$L_{満+新}$		1.3174E-109
	完全モデル　−96.0940−124.1423= −220.2363						**縮小モデル**		

である．ここでの対数尤度の計算は，1 日ごとの犯罪が（0:なし，1:あり）の 2 値反応に対するベルヌーイ分布としての確率を用いている．同様に，新月の $\ln L_{新月}$ は，

$$\left.\begin{aligned}\ln L_{新月} &= \ln(0.6129^{114}) + \ln(0.3871^{72}) \\ &= -55.8085 - 68.3338 = -124.1423\end{aligned}\right\} \tag{3.6}$$

であり，満月と新月を合わせた $\ln L_{満月+新月}$ は，

$$\left.\begin{aligned}\ln L_{満月+新月} &= \ln(0.4173^{154}) + \ln(0.5827^{215}) \\ &= -134.5720 - 116.1341 = -250.7061\end{aligned}\right\} \tag{3.7}$$

となる．

尤度比検定は，満月と新月を合わせた全体の犯罪発生率の対数尤度（縮小モデル）と，満月と新月を別々にした犯罪発生率の対数尤度（完全モデル）の差の 2 倍を検定統計量としている．別々の満月と新月の対数尤度を加えた完全モデルは，

$$\left.\begin{aligned}\ln L_{満月,新月} &= \ln L_{満月} + \ln L_{新月} \\ &= -96.0940 - 124.1423 = -220.2363\end{aligned}\right\} \tag{3.8}$$

であり，縮小モデルの対数尤度 $\ln L_{満月+新月} = -250.7061$ との差の 2 倍の $\chi^2_{likelihood}$ は，

$$\left.\begin{aligned}\chi^2_{likelihood} &= 2 \times (\ln L_{満月,新月} - \ln L_{満月+新月}) \\ &= 2 \times [(-220.2363) - (-250.7061)] \\ &= 2 \times 30.4698 = 60.9396\end{aligned}\right\} \tag{3.9}$$

となる．このように，完全モデルと縮小モデルの確率の積の対数を基本にした尤度比検定の結果と，表 3.3 に示した 2×2 の分割表に対する尤度比検定の結果が一致することが確認される．

尤度比検定は「比」と言っているのに，実際は「差」としている．元々は，尤度の比について対数を取り 2 倍した統計量が，カイ 2 乗分布に従うことから，伝統的に「尤度比」と言われている．尤度比の形式では，

$$\left.\begin{aligned}\chi^2_{likelihood} &= 2\ln\left(\frac{\text{完全モデル}}{\text{縮小モデル}}\right) \\ &= 2\ln\left(\frac{L_{満月}L_{新月}}{L_{満月+新月}}\right) \\ &= 2\ln\left(\frac{1.8489\times10^{-42}\times1.2181\times10^{-54}}{1.3174\times10^{-109}}\right) \\ &= 60.9396\end{aligned}\right\} \tag{3.10}$$

となり，2×2 の分割表に対する式（3.4）に示した簡便公式の尤度比検定統計量と一致することが確認される．

分割表に対する簡便公式の尤度比検定統計量の誘導

これまで，出現確率を用いた尤度比検定の考え方を示してきたが，尤度比検定の定義に従った計算式は，

$$\chi^2_{likelihood} = 2\ln\left[\frac{\prod_{i=1}^{2}\prod_{j=1}^{1}\left(\frac{n_{ij}}{n_{i\bullet}}\right)^{n_{ij}}}{\prod_{j=1}^{2}\left(\frac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^{n_{\bullet j}}}\right] \tag{3.11}$$

である．式を変形することにより，分割表の簡便公式による尤度比検定統計量が次のように導出できる．

$$\left.\begin{aligned}\chi^2_{likelihood} &= 2\ln\left[\frac{\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{n_{ij}}{n_{i\bullet}}\right)^{n_{ij}}}{\prod_{j=1}^{2}\left(\frac{n_{\bullet j}}{n_{\bullet\bullet}}\right)^{n_{1j}+n_{2j}}}\right] \\ &= 2\ln\left[\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{n_{ij}}{n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet}}\right)^{n_{ij}}\right] \\ &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\ln\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)\end{aligned}\right\} \tag{3.12}$$

式（3.12）は，式（3.4）に一致することから，ポアソン分布を仮定した 2 群間比較に対する尤度比検定の考え方は，2×2 の分割表における期待度数 $\hat{\mu}_{ij}$ を用いた尤度比検定と同様な方法であることが確認される．

ポアソン分布を仮定した 2 群のカウント・データに対し，第 1.6 節の表 1.8 に示したようにポアソン回帰により尤度比検定の結果が得られ，表 1.19 に示したように Excel による尤度比検定を試み結果が一致した．同様の考え方でベルヌーイ分布を仮定した 2×2 の分割表に対する式（3.11）で示した尤度比検定の結果が，いわゆる 2×2 の分割表に対する尤度比検定の結果に一致することを確認できた．これは，式(3.11）が式(3.4）に等しくなることを式(3.12）によって確認した．

2×2 の分割表に対する尤度比検定の式(3.4）の導出は，尤度比検定の最も基礎的な課題であるとの認識で，参考になる文献を探して見たが，なかなか見出すことができなかった．唯一，Agresti（2013），Categorical Data Analyssis 3rd ed の 3.2.1 Peason and Likelifood-Ratio Chi-Squared Test に分割表の一般式として同様の導出方法が示されていることが見いだされた．

3.2. 一般化線形モデルで 2 項分布を仮定した 2 群間比較

「満月の日：$x_1=0$」,「新月の日：$x_1=1$」,「犯罪のなし：$y=0$」,「犯罪のあり：$y=1$」として一般化線形モデルで，2 項分布を仮定した 2 群間比較を行う．JMP の「一般化線形モデル」を使うために表 3.5 に示すように件数を縦方向に並べ直す．

表 3.5　一般化線形モデルを適用するためのデータリスト

切片	満新	犯罪	件数
x_0	x_1	y	n
1	0	0	40
1	0	1	143
1	1	0	114
1	1	1	72
		計	369

リンク関数は，（ロジット，プロビット，補 2 重対数）の 3 種類あるが，群間比較の場合は，完全フィットするので，尤度比カイ 2 乗値は，どれでも同じ結果となる．表 3.6 にリンク関数を「ロジット」とした場合についての尤度比検定の結果を示す．

表 3.6　犯罪の（なし，あり）に対する一般化線形モデル（ロジット・リンク）

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	30.4698	60.9396	1	<.0001*
完全	220.2363			
縮小	250.7061			

「縮小」の行の「(–1)*対数尤度」列の 250.7061 が，表 3.4 に示した $\ln L_{満月+新月}=-250.7061$ に対応し，「完全」の 220.2363 が，$\ln L_{満月}+\ln L_{新月}=-220.2363$ に対応する．「差分」の行の「尤度比カイ 2 乗」の 60.9369 が，式（3.9）の $\chi^2_{likelihood}=2\times 30.4698=60.9396$ に対応する．

表 3.7 に回帰パラメータについての推定値，標準誤差，尤度比カイ 2 乗の結果を示す．「項」の列の「x」の行の「尤度比カイ 2 乗」の結果も 60.9396 と表 3.6 の結果に一致する．

表 3.7　ロジット変換に対する回帰パラメータの推定

パラメータ推定値			
項	推定値	標準誤差	尤度比カイ2乗
切片	-1.2740	0.1789	61.5039
x1	1.7335	0.2338	60.9396

表 3.7 のロジット変換値を用いた回帰パラメータの推定値は，「満月の日：$x_1=0$」で「犯罪のなし：$y=0$」のロジットが

$$logit\hat{}=\ln\left(\frac{p_{満月,y=0}}{1-p_{満月,y=0}}\right)=\ln\left(\frac{0.2186}{1-0.2186}\right)=-1.2740 \tag{3.13}$$

切片となり，「新月の日：$x_1=1$」とした場合には，満月と新月のロジットの差

$$\left.\begin{aligned} logit\hat{}&=\ln\left(\frac{p_{新月,y=0}}{1-p_{新月,y=0}}\right)-\ln\left(\frac{p_{満月,y=0}}{1-p_{満月,y=0}}\right)\\ &=\ln\left(\frac{0.6129}{1-0.6129}\right)-(-1.2740)\\ &=0.4595+1.2740\\ &=1.7335 \end{aligned}\right\} \tag{3.14}$$

となっている．

Excel ソルバーを用いたロジスティック回帰

一般化線形モデルで分布を「2 項」，リンク関数を「ロジット」と設定し，ロジスティック回帰による 2×2 の分割表に対する尤度比検定を行い，結果の見方について示した．ポアソン回帰のみならず，ロジスティック回帰も Excel のソルバーを用いて簡単にできることを示す．

ロジスティック回帰は，表 3.5 に示すように，満月の日を $x_1=0$，新月の日を $x_1=1$ とし，2 値反応 y に対して式（3.13）および式（3.14）に示したロジット変換（$logit$）し，直線をあてはめる方法である．図 3.1（左）に示すように 2 ポイントしかないのでそれらの点を通る直線となる．もちろん切片は $\beta_0=-1.2740$，傾きはロジットの差 $\beta_1=1.7335$ である．

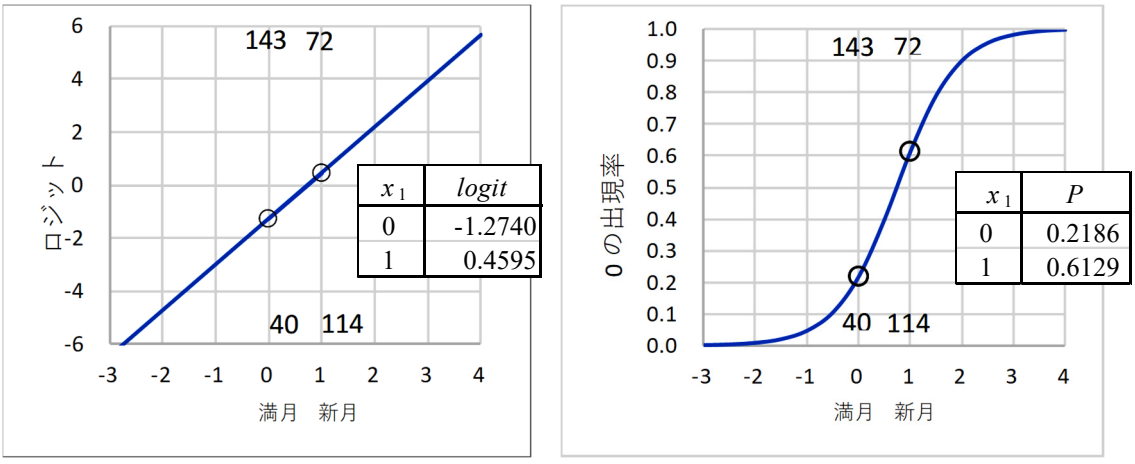


図 3.1　ロジスティック直線およびロジスティック曲線のあてはめ

シグモイド曲線として，ロジスティック分布の累積分布関数 $F(x_1)$

$$\left.\begin{aligned} F(x_1) &= \frac{\exp\left(\dfrac{x_1-\mu}{\sigma}\right)}{1+\exp\left(\dfrac{x_1-\mu}{\sigma}\right)} = \frac{\exp\left(-\dfrac{\mu}{\sigma}+\dfrac{1}{\sigma}x_1\right)}{1+\exp\left(-\dfrac{\mu}{\sigma}+\dfrac{1}{\sigma}x_1\right)} \\ &= \frac{\exp(\beta_0+\beta_1 x_1)}{1+\exp(\beta_0+\beta_1 x_1)} \end{aligned}\right\} \qquad (3.15)$$

ただし，$\beta_0 = -\mu/\sigma$，$\beta_1 = 1/\sigma$

を使う．表 3.7 で得られた回帰パラメータは，$\hat{\beta}_0 = -1.2740$，$\hat{\beta}_1 = 1.7335$ なので，図 3.1（右）に示すように推定値は，

$$\text{満月 } x_1 = 0 \ : \quad \hat{\pi}_0 = F(x_1 = 0) = \frac{\exp(-1.2740 + 1.7335 \times 0)}{1+\exp(-1.2740 + 1.7335 \times 0)} = 0.2186$$

$$\text{新月 } x_1 = 1 \ : \quad \hat{\pi}_1 = F(x_1 = 1) = \frac{\exp(-1.2740 + 1.7335 \times 1)}{1+\exp(-1.2740 + 1.7335 \times 1)} = 0.6129$$

となる．

表 3.8 に Excel ソルバーを用いた最尤法によるロジスティックス回帰の解析法を示す．初期値を $\hat{\beta}_0 = 0.0$，$\hat{\beta}_1 = 0.0$ として

$$logit_i\text{^} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} \qquad (3.16)$$

表 3.8　Excel ソルバーによるロジスティック回帰

		初期値	β_0^=	**0.0000**			
			β_1^=	**0.0000**			
切片	満新	犯罪	件数		二項分布		
x_0	x_1	y	n	$logit$^	π^	P^	$\ln L_i$
1	0	0	40	0.0000	0.5000	0.5000	-27.7259
1	0	1	143	0.0000	0.5000	0.5000	-99.1200
1	1	0	114	0.0000	0.5000	0.5000	-79.0188
1	1	1	72	0.0000	0.5000	0.5000	-49.9066
		計	369			$\ln L$ =	**-255.7713**
		最尤解	β_0^=	**-1.2740**			
			β_1^=	**1.7335**			
切片	満新	犯罪	件数		二項分布		
x_0	x_1	y	n	$logit$^	π^	P^	$\ln L_i$
1	0	0	40	-1.2740	0.2186	0.2186	-60.8243
1	0	1	143	-1.2740	0.2186	0.7814	-35.2697
1	1	0	114	0.4595	0.6129	0.6129	-55.8086
1	1	1	72	0.4595	0.6129	0.3871	-68.3337
		計	369			$\ln L$ =	**-220.2363**

$$\hat{\pi}_i = \frac{\exp(logit_i)}{1+\exp(logit_i)} \tag{3.17}$$

$$\hat{P}_i = \text{Binom.dist}(1-y_i, 1, \hat{\pi}_i, false) \tag{3.18}$$

$$\ln(L_i) = n_i \ln(\hat{P}_i) \tag{3.19}$$

$$\ln L = \sum_i \ln(L_i) = -255.7713 \tag{3.20}$$

が計算されている．Excel ソルバーで，$\ln L$ の最大化するために，$\hat{\beta}_0 = 0.0$，$\hat{\beta}_1 = 0.0$ を変化させると $\hat{\beta}_0 = -1.2740$，$\hat{\beta}_1 = 1.7335$ が得られ，対数尤度 $\ln L = -220.2363$ は，表 3.4 の完全モデルに一致する．

切片に対する尤度比検定に対する補足

さて, 表 3.7 に示したロジット変換に対する回帰パラメータ推定値で, 切片の推定値−1.2740 に対し，尤度比カイ 2 乗値が，61.5039 と高度に有意となっている．どのような計算が行われているのだろうか．尤度比検定は，完全モデルの尤度と縮小モデルの尤度の比であるので，モデル式から切片を除いたモデルであるが，実際にはどのように計算したらよいのであろうか．

切片がない場合のモデルは,「原点を通る直線」をあてはめることになる．ロジット変換した場合にゼロとなる出現率は，0.50 である．確認すると

$$logit = \ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{0.50}{1-0.50}\right) = 0 \tag{3.21}$$

確かにゼロとなる．したがって,「満月の日：$x_1 = 0$」,「新月の日：$x_1 = 1$」，としたので，満月の日の犯罪がない率を 0.50 とし，新月の日の犯罪がない率 0.6129 とした場合が縮小モデルとなる．表 3.9 に「切片 x_0 抜き縮小モデル」の対数尤度が -250.9882 となり，完全モデルとの差の 2 倍が，61.5039 となり，表 3.7 の結果に一致する．

表 3.9　切片抜き縮小モデルの対数尤度

切片	満月新月	犯罪	件数	**完全モデル**		x_1 **抜き縮小モデル**		x_0 **抜き縮小モデル**	
x_0	x_1	y	n	P^	$\ln L_i$	P^	$\ln L_i$	P^	$\ln L_i$
1	0	0	40	0.2186	-60.8243	0.4173	-34.9538	0.5000	-27.7259
1	0	1	143	0.7814	-35.2697	0.5827	-77.2427	0.5000	-99.1200
1	1	0	114	0.6129	-55.8085	0.4173	-99.6182	0.6129	-55.8085
1	1	1	72	0.3871	-68.3338	0.5827	-38.8914	0.3871	-68.3338
			369		-220.2363		-250.7061		-250.9882
		完全モデルを基準とした対数尤度の差					30.4698		30.7519
				2倍の対数尤度			60.9396		61.5039

3.3. ポアソン回帰を用いた 2 群間の比較

第 1.6 節の表 1.19 で Excel によるポアソン分布を仮定した 2 群間の尤度比検定の結果を説明なしで示した．第 3.2 節で 2×2 の分割で尤度比検定の考え方を示したので，ポアソン分布の場合について拡張する．完全モデルの尤度は，満月および新月の日の犯罪件数に対し，それぞれについてポアソン分布を仮定した場合である．縮小モデルは，満月と新月を合わせた場合にポアソン分布を仮定する．尤度比検定は，それらの比として定義され，仮定する分布が 2 項分布でもポアソン分布でも尤度比検定の考え方は同じである．

$$\chi^2_{likelihood} = 2\ln\left(\frac{\text{完全モデルでの尤度}}{\text{縮小モデルでの尤度}}\right) \tag{3.22}$$

満月の日の尤度，新月の日の尤度をそれぞれ $L_{\text{満月}}$ と $L_{\text{新月}}$ とし，満月と新月を合算した場合の尤度を $L_{\text{満月＋新月}}$ とすると尤度比検定は，

$$\left.\begin{aligned}\chi^2_{likelihood} &= 2\ln\left(\frac{L_{\text{満月}}L_{\text{新月}}}{L_{\text{満月＋新月}}}\right)\\ &= 2\left(\ln L_{\text{満月}} + \ln L_{\text{新月}} - \ln L_{\text{満月＋新月}}\right)\\ &= 2\times[(-268.4776)+(-176.1202)-(-484.8865)]\\ &= 80.5774\end{aligned}\right\} \tag{3.23}$$

となることを示した．この式は，2×2 の分割の尤度比検定で用いた式（3.10）と同じであるが，尤度の計算にポアソン分布を仮定した対数尤度を用いている．表 3.10 に JMP のポアソン回帰による縮小モデル，完全モデル，それらの差，および，マイナス 2 倍の対数尤度を示す．

表 3.10　ポアソン回帰による 2 群間比較

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	40.2887	80.5774	1	<.0001*
完全	444.5978			
縮小	484.8865			

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	1.3989	0.0874	256.0000	<.0001*
x1	-0.8935	0.1018	80.5774	<.0001*

表 3.11 にポアソン回帰による 2 群間の比較の Excel の計算シートを示す．「満月の日：$x_1 = 0$」，「新月の日：$x_1 = 1$」とし，犯罪発現件数を y_i，日数を n_i とする．

完全モデルは，

$$\begin{aligned}\hat{y}_i &= (\text{完全}\hat{\beta}_0)x_{0,i} + (\text{完全}\hat{\beta}_1)x_{1,i} \\ &= 1.3989 - 0.8935x_{1,i}\end{aligned}$$

$$\ln L_i = n_i \ln\left[\text{Poisson.dist}(y_i, \hat{y}_i, false)\right] \tag{3.24}$$

$$\begin{aligned}\ln L &= -55.9562 - 68.0458 - \cdots - 6.4132 \\ &= -444.5978\end{aligned}$$

と計算されている．切片のみの縮小モデルは，

$$\begin{aligned}\hat{y}_i &= (\text{縮小}\hat{\beta}_0)x_{0,i} \\ &= 0.9485\end{aligned} \tag{3.25}$$

$$\ln L_i = n_i \ln\left[\text{Poisson.dist}(y_i, \hat{y}_i, false)\right] \tag{3.26}$$

$$\begin{aligned}\ln L &= -37.9404 - 64.0879 - \cdots - 4.3380 \\ &= -484.8865\end{aligned} \tag{3.27}$$

と計算されている．完全モデルの尤度 –444.5978 と縮小モデル –484.8865 との差の 2 倍で，表 3.11 の最後の行に示すように， 80.5774 となり式（3.23）と一致する．

傾きのみの縮小モデルは，計算エラー（#NUM!）となり計算されていない．これは，切片 β_0 がないモデルとなり， $x_1 = 0$ の場合にポアソン分布のパラメータが， $\hat{\mu}_i = x_1 = 0$ となり，1

表 3.11　犯罪件数に対する（0，1）形式でのポアソン回帰

			完全β_0^=	**1.3989**	縮小β_0^=	**0.9485**		**—**		
			完全β_1^=	**-0.8935**		**—**	縮小β_1^=	**0.5054**		
	切片	満新	件数	日数	完全モデル		x_1 縮小モデル		x_0 縮小モデル	
i	x_0	x_1	y	n	y^	lnL_i	y^	lnL_i	y^	lnL_i
1	1	0	0	40	1.3989	-55.9562	0.9485	-37.9404	0.0000	0.0000
2	1	0	1	64	1.3989	-68.0458	0.9485	-64.0879	0.0000	#NUM!
3	1	0	2	56	1.3989	-79.5576	0.9485	-97.8535	0.0000	#NUM!
4	1	0	3	19	1.3989	-41.4883	0.9485	-55.0783	0.0000	#NUM!
5	1	0	4	1	1.3989	-3.2342	0.9485	-4.3380	0.0000	#NUM!
6	1	0	5	2	1.3989	-9.0159	0.9485	-12.0006	0.0000	#NUM!
7	1	0	9	1	1.3989	-11.1795	0.9485	-14.2261	0.0000	#NUM!
8	1	1	0	114	0.5054	-57.6128	0.9485	-108.1301	0.5054	-57.6129
9	1	1	1	56	0.5054	-66.5184	0.9485	-56.0769	0.5054	-66.5184
10	1	1	2	11	0.5054	-28.1977	0.9485	-19.2212	0.5054	-28.1977
11	1	1	3	4	0.5054	-17.3780	0.9485	-11.5954	0.5054	-17.3780
12	1	1	4	1	0.5054	-6.4132	0.9485	-4.3380	0.5054	-6.4132
					ln L=	**-444.5978**	ln L=	**-484.8865**	ln L=	**#NUM!**
							対数尤度の差=	40.2887		#NUM!
							2倍の差=	80.5774		#NUM!

以上のデータが全くないことになり，ポアソン分布が定義できないために

$$y_i = 0 \ : \begin{cases} \ln L_i = n_i \ln\left[\text{Poisson.dist}(y_i = 0, \hat{\mu}_i = 0, \mathit{false})\right] \\ \qquad = n_i \ln[1] = 0 \end{cases} \tag{3.28}$$

$$y_i > 0 \ : \begin{cases} \ln L_i = n_i \ln\left[\text{Poisson.dist}(y_i > 0, \hat{\mu}_i = 0, \mathit{false})\right] \\ \qquad = n_i \ln[0] = \text{\#NUM!} \end{cases} \tag{3.29}$$

計算不能（#NUM!, An invalid NUMber）となるためである．JMP および SAS での計算結果は，表 1.18 および表 1.20 に示したように，尤度比カイ 2 乗が共に，256.00 と出力されている．これは，共分散行列から計算された *SE* を用いたワルド統計量を便宜的に使っているためであろう．

計算不能となったのは，満月の日：$x_1 = 0$」，「新月の日：$x_1 = 1$」としたために起きた現象で，表 3.12 に示すように，「満月の日：$x_1 = 0.5$」，「新月の日：$x_1 = 1.5$」のようなダミー変数とすれば，計算可能となる．

表 3.12　犯罪件数に対する（0.5，1.5）型ダミー変数によるポアソン回帰

			完全β_0^=	**1.8457**	縮小β_0^=	**0.9485**			**—**	
			完全β_1^=	**-0.8935**		**—**	縮小β_1^=		**0.9447**	
	切片	満新	件数	日数	完全モデル		x_1縮小モデル		x_0縮小モデル	
i	x_0	x_1	y	n	y^	lnL_i	y^	lnL_i	y^	lnL_i
1	1	**0.5**	0	40	1.3989	-55.9563	0.9485	-37.9404	0.4723	-18.8934
2	1	**0.5**	1	64	1.3989	-68.0458	0.9485	-64.0879	0.4723	-78.2337
3	1	**0.5**	2	56	1.3989	-79.5576	0.9485	-97.8535	0.4723	-149.2745
4	1	**0.5**	3	19	1.3989	-41.4883	0.9485	-55.0783	0.4723	-85.7716
5	1	**0.5**	4	1	1.3989	-3.2342	0.9485	-4.3380	0.4723	-6.6507
6	1	**0.5**	5	2	1.3989	-9.0159	0.9485	-12.0006	0.4723	-18.0203
7	1	**0.5**	9	1	1.3989	-11.1795	0.9485	-14.2261	0.4723	-20.0248
8	1	**1.5**	0	114	0.5054	-57.6128	0.9485	-108.1301	1.4170	-161.5385
9	1	**1.5**	1	56	0.5054	-66.5184	0.9485	-56.0769	1.4170	-59.8337
10	1	**1.5**	2	11	0.5054	-28.1977	0.9485	-19.2212	1.4170	-15.5437
11	1	**1.5**	3	4	0.5054	-17.3780	0.9485	-11.5954	1.4170	-8.6525
12	1	**1.5**	4	1	0.5054	-6.4132	0.9485	-4.3380	1.4170	-3.2009
					ln L=	**-444.5978**	ln L=	**-484.8865**	ln L=	**-625.6383**
						対数尤度の差=		40.2887		181.0406
							2倍の差=	80.5774		362.0811

この結果が正しいのか，JMP での結果を表 3.13 の最後の行に示すように，切片の尤度比カイ 2 乗は，362.0811 となり，Excel での結果と一致する．この様にポアソン回帰の場合には，名義尺度の与える数値により，結果が異なるので，切片に対する「尤度比カイ 2 乗」は，全く意味をなさないことに注意が必要である．

表 3.13　JPM による犯罪件数に対する（0.5，1.5）型ダミー変数でのポアソン回帰

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	40.2887	80.5774	1	<.0001*
完全	444.5978			
縮小	484.8865			

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	1.8457	0.1337	362.0811	<.0001*
x3	-0.8935	0.1018	80.5774	<.0001*

名義尺度にどのような数値（ダミー変数）を与えるかは，自己責任で自由に設定できる．ただし，切片に対する尤度比カイ 2 乗値は，計算の原理からは計算不能となる場合もあるが，JMP および SAS では，便宜的な対応をしていることが推測される．

3.4. 2×2 の要因配置モデルに対する各種のデザイン行列

デザイン行列に与える変数（ダミー変数）

一般化線形モデルで分布をポアソン分布，リンク関数を恒等と設定してポアソン回帰を実施する方法をこれまで示してきた．基本的なポアソン回帰の場合は，

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \text{ポアソン分布} \tag{3.30}$$

のように，通常の回帰分析と同様な回帰式を用いてきた．この式で β_0 の推定値は，$x=0$ における y の値なので Y 切片となり，β_1 の推定値は，回帰直線の傾きで x のプラス 1 増加した場合における y の増分である．自明ではあるが，そのように言えるのか，数式を使って説明する．回帰パラメータの推定値 $\hat{\beta}_0$ と $\hat{\beta}_1$ が得られ，回帰式で

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.31}$$

と推定値 $\hat{y}_i$ が得られた場合に，$x=0$ とした場合に

$$\begin{aligned} \hat{y}_{x=0} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times 0 = \hat{\beta}_0 \end{aligned} \tag{3.32}$$

推定値は，$\hat{y}_{x=0} = \hat{\beta}_0$ で，いわゆる Y 切片である．次に x_i に適当な値として $x=2$ を代入した場合，

$$\begin{aligned} \hat{y}_{x=2} &= \hat{\beta}_0 + \hat{\beta}_1 \times 2 \\ &= \hat{y}_{x=0} + 2\hat{\beta}_1 \end{aligned} \tag{3.33}$$

となり，$\hat{\beta}_1$ について解くと

$$\hat{\beta}_1 = \frac{\hat{y}_{x=2} - \hat{y}_{x=0}}{2} \tag{3.34}$$

が得られる．これは，x が 2 の場合の $\hat{y}_{x=2}$ から，x が 0 の場合の $\hat{y}_{x=0}$ を差し引いて 2 で割っており，$\hat{\beta}_1$ は，x が 1 増加した場合の $\hat{y}$ の増分であり，回帰分析における「傾き」である．

第 1.6 節の満月と新月の犯罪件数では，2 群間の比較をする際に，名義尺度に対して満月の場合に $x=0$，新月の場合に $x=1$ としてポアソン回帰を行った．もちろん次式で β_0 と β_1 を推定するのであるが，

$$\begin{aligned} &x=0 \text{ の場合：} \quad \hat{\beta}_0 = \hat{y}_{\text{満月}} \\ &x=1 \text{ の場合：} \quad \hat{\beta}_1 = \hat{y}_{x=1} - \hat{y}_{x=0} = \hat{y}_{\text{新月}} - \hat{y}_{\text{満月}} \end{aligned} \tag{3.35}$$

となり，（切片・傾き）とは，言い難くなる．

第 1.4 節で反復重み付き回帰式を，事前の説明なしに，

$$\hat{\boldsymbol{\beta}}^{(m)} = \left[(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{(m-1)} \right]^{-1} (\boldsymbol{X}^T \hat{\boldsymbol{W}} \hat{\boldsymbol{Z}})^{(m-1)} \tag{3.36}$$

として示した．この式の中の行列 $\boldsymbol{X}$ は，計画行列またはデザイン行列と言われており，ポアソン回帰による各種の推定を行う際に中心的な役割を果す．なお，重み $\hat{\boldsymbol{W}}$ の対角要素が全て 1 の場合には，$\boldsymbol{X}^T \hat{\boldsymbol{W}} = \boldsymbol{X}^T$ なので，$\hat{\boldsymbol{Z}}$ を $\boldsymbol{y}$ と置き換えて通常の線形回帰式におけるパラメータの推定式

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{3.37}$$

となる．

第 1.4 節の表 1.8 では．デザイン行列 $\boldsymbol{X}$ を，9×2 の矩形データとし，Excel の行列関数を用いて，回帰パラメータを推定した．

$\boldsymbol{X}$ =

x_0	x_1
1	-1
1	-1
1	0
1	0
1	0
1	0
1	1
1	1
1	1

一般的な回帰分析では，切片を含めない変数のみを設定するが，デザイン行列を用いた回帰式を用いる場合は，

$$y_i = \beta_0 x_{0,i} + \beta_1 x_{1,i} + \varepsilon_i \qquad \varepsilon_i \sim Normal(0, \sigma), \quad i = 1, 2, \ldots, n \tag{3.38}$$

のように切片を説明変数として $x_{0,i} = 1$ を含めた式とする必要がある．

2×2 の要因配置実験

第 1.7 節の表 1.22 に示した Ames 試験におけるコロニー数については，2×2 の要因配置デザインと見なすこともでき，表 3.14 に結果のサマリーを示す．各セルの平均は異なるが，それぞれの分散の比が 1 に近いことからも，全体としてポアソン分布に従うと判断される［吉村ら（1992)］.

この実験を 2 因子実験と見なすならば，主効果を（A : 溶媒，B : 代謝活性化），交互作用として（A×B）を含めて解析することができる．因子 B の効果は，統計的に云々する必要がないくらい明らかに B_1 に比べて B_2 が小さいが，因子 A の効果は，B_1 内では A_1 に対し A_2 は減少傾向であり，B_2 内の場合は，A_1 に対し A_2 は増加傾向で統計的な判断が必要となる．

表 3.14　A:溶媒と B:代謝活性化の組合せ結果（各セル，n=50）

	B:代謝活性化			
	1:なし		2:あり	
A:溶媒	平均	分散	平均	分散
1:蒸留水	14.54	17.23	7.54	6.34
2:DMOS	12.48	11.64	8.28	6.80

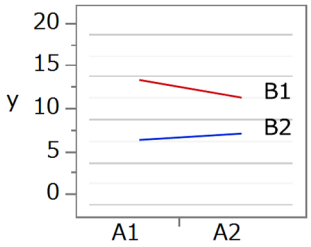
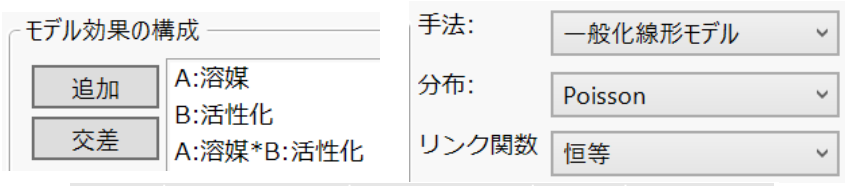


表 3.14 に示した 2×2 の実験データについてポアソン回帰による 2 元配置型の解析を行なう．JMP のモデル効果の構成は，通常の二元配置分散分析に準じて設定する．表 3.15 にパラメータの推定結果を示す．

表 3.15　JMP による対比型のポアソン回帰の結果



モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	78.8560	157.7120	3	<.0001*
完全	507.2925			
縮小	586.1485			

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	10.7100	0.2314	2142.0000	<.0001*
A:溶媒[A1]	0.3300	0.2314	2.0339	0.1538
B:活性化[B1]	2.8000	0.2314	148.1204	<.0001*
A:溶媒[A1]*B:活性化[B1]	0.7000	0.2314	9.1569	0.0025*

ここで，切片= 10.7100 は，何を意味しているのだろうか．A:溶媒[A1] = 0.3300 は何を意味しているのだろうか．表 3.16 には，2×2 表について A:溶媒についての平均，B:代謝活性剤に

表 3.16　セル平均の平均と総平均からの差

	B:代謝活性化			
	1:なし	2:あり	全体	
A:溶媒	平均	平均	平均の平均	効果(差)
1:蒸留水	14.54	7.54	11.04	0.33
2:DMOS	12.48	8.28	10.38	-0.33
平均の平均	13.51	7.91	10.71	
効果(差)	2.80	-2.80		

ついての平均，さらに全体の平均 10.71，全体平均から差を効果（差）の計算結果が示されている．これから，切片= 10.71 は，総平均であると推測でき，

A:溶媒[A1] = 0.33

は，A_1=11.04 から全体の平均 10.71 との差が 0.33 であり，A_1 の平均と A_2 の平均の差の 2 分の 1 と一致する．

$$\text{因子 } A_1 \text{ の効果：} \frac{(14.54+7.54)/2-(12.48+8.28)/2}{2}=\frac{0.66}{2}=0.33$$

同様に，

B:活性化[B1] = 2.80

は，B_1 の平均 13.51 と総平均 10.71 の差であり，B_1 の平均と B_2 の平均の差の 2 分の 1 と一致する．

$$\text{因子 } B_1 \text{ の効果：} \frac{(14.54+12.48)/2-(7.54+8.28)/2}{2}=\frac{5.60}{2}=2.80$$

では，交互作用

A:溶媒[A1]*B:活性化[B1] = 0.70

は，何を意味しているのであろうか．試行錯誤的に検討すると，たすき掛けで加えた平均の差の平均

$$\text{交互作用 1：} \frac{(14.54+8.28)/2-(12.48+7.54)/2}{2}=\frac{1.40}{2}=0.70$$

に一致する．更に，表 3.17 に示すように A_1 内の B_1 と B_2 の差の平均，A_2 内の B_1 と B_2 の差の平均，それらの差の平均の平均で

$$\text{交互作用 2：} \frac{(14.54-7.54)/2-(12.48-8.28)/2}{2}=\frac{1.40}{2}=0.70$$

同じ結果が得られるが，式を変形すれば，互いに同じ式になることが確認できる．

表 3.17　セル平均の差の差

	B:代謝活性化		
	1:なし	2:あり	全体
A:溶媒	平均	平均	差/2
1:蒸留水	14.54	7.54	3.50
2:DMOS	12.48	8.28	2.10
差/2	1.03	-0.37	1.40
		差の差の効果	0.70

交互作用の効果としてプラスの効果 0.70 が得られたが，効果としてはマイナスの効果がないとバランスしない．たすき掛けの（$A_1B_1 \times A_2B_2$）－（$A_1B_2 \times A_2B_1$）差を第 1 項と第 2 項を

入れ替えると

$$\text{交互作用 3：}\frac{(12.48+7.54)/2-(14.54+8.28)/2}{2}=\frac{-1.40}{2}=-0.70$$

マイナスの効果が出てくる．これらから，表 3.18 に示すように交互作用の効果としては，

表 3.18　交互作用の効果

	B:代謝活性化		
A:溶媒	1:なし	2:あり	和
1:蒸留水	0.70	-0.70	0.00
2:DMOS	-0.70	0.70	0.00
和	0.00	0.00	0.00

これらの効果を用いて 2×2 のセル平均を求めてみと，次のようにセル平均を求めることができる．

A_1B_1＝総平均＋A_1 効果＋B_1 効果＋交互作用 $_{11}$ = 10.71＋0.33＋2.80＋0.70＝14.54

A_1B_2＝総平均＋A_1 効果＋B_2 効果＋交互作用 $_{12}$ =10.71＋0.33－2.80－0.70＝ 7.45

A_2B_1＝総平均＋A_2 効果＋B_1 効果＋交互作用 $_{21}$ = 10.71－0.33＋2.80－0.70＝12.48

A_2B_2＝総平均＋A_2 効果＋B_2 効果＋交互作用 $_{22}$ = 10.71－0.33－2.80＋0.70＝ 8.28

このような探索的な見当は，統計ソフトの出力を理解し解釈するために欠かせない．ここでの結果は，JMP 一般線形モデルの出力に対する結果であるが，SAS の GENMOD プロシジャでは，第 13.4 節で示すように全く異なる．統計ソフトの結果の解釈には，統計ソフトの名義尺度に対するダミー変数の与え方を確認することが必須である．

(1, −1)対比型デザイン行列

前項では，JMP のポアソン回帰の解析で出力されるパラメータの推定値に対して探索的な見当により意味づけを行った．この計算プロセスを行列で整理してみよう．セル平均を表 3.19 に示す順番で縦方向に展開し，行と列の入れ替えをする転置記号 $[\cdots]^T$ を用いてセル平均 y_i を列ベクトル

$$\boldsymbol{y}=[y_1 \quad y_2 \quad y_3 \quad y_4]^T=[14.54 \quad 7.54 \quad 12.48 \quad 8.28]^T$$

とする．デザイン行列は，前項でセル平均の計算でプラスとなっている場合を 1 とし，マイナスとなっている場合を −1 と置き換えたものをデザイン行列 $\boldsymbol{X}$ としている．

デザイン行列の列ベクトル $\boldsymbol{X}_0$ は，全て 1 で「切片」ともいわれるが，その役割は他のデザイン行列の設定によって微妙に異なる．列ベクトル $\boldsymbol{X}_1$ は，因子 A の水準 A_1 に対して +1，A_2

表 3.19　対比型（1, −1）のデザイン行列

	B:代謝												
						$\boldsymbol{X}$							
A:溶媒	活性化	平均		$\boldsymbol{y}$		x_0	x_1	x_2	x_3	$\boldsymbol{\beta}$		$\boldsymbol{\varepsilon}$	$\boldsymbol{\beta}$^
1:蒸留水	1:なし	14.54		y_1	=	1	1	1	1	β_0	+	ε_1	10.71
	2:あり	7.54		y_2		1	1	-1	-1	β_1		ε_2	0.33
2:DMOS	1:なし	12.48		y_3		1	-1	1	-1	β_2		ε_3	2.80
	2:あり	8.28		y_4		1	-1	-1	1	β_3		ε_4	0.70

に対して −1 のように足して 0 となる数値のセットを対比型という．列ベクトル $\boldsymbol{X}_2$ は，因子 B の水準 B_1 に対して+1，B_2 に対して −1 と対比型で設定する．列ベクトル $\boldsymbol{X}_3$ は，因子 A と因子 B の交互作用で，$\boldsymbol{X}_3 = \boldsymbol{X}_1 \cdot \boldsymbol{X}_2$ のように積で計算されている．

推定値 $\hat{\boldsymbol{\beta}}$ は，［総平均　A_1 効果　B_1 効果　交互作用効果］で，

$$\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2 \quad \hat{\beta}_3]^T = [10.71 \quad 0.33 \quad 2.80 \quad 0.70]^T$$

のように推定されている．セル平均 $\hat{y}_i$ は，$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ で計算されていて，以下に示すように展開することができる．

$$\hat{y}_1 = +10.71 + 0.33 + 2.80 + 0.70 = 14.54$$
$$\hat{y}_2 = +10.71 + 0.33 - 2.80 - 0.70 = 7.45$$
$$\hat{y}_2 = +10.71 - 0.33 + 2.80 - 0.70 = 12.48$$
$$\hat{y}_4 = +10.71 - 0.33 - 2.80 + 0.70 = 8.28$$

この行列 $\boldsymbol{X}$ を一般的には，デザイン行列または計画行列とも言われている．行列計算になれていない場合は，先に第 4 章で統計計算に必要な行列計算の基礎を学習してもらいたい．

表 1.22 に示した Ames 試験におけるコロニー数の表を並び替えて整理した結果を表 3.20 に示す．平均に対して与えた 4×4 のデザイン行列 $\boldsymbol{X}$ は，それぞれに繰返しあるので，56×4 と拡大されたデザイン行列 $\boldsymbol{X}'$ とする，反応の平均に与えていた y_i に代えて y_i' と区別する．説明変数の並びを行ベクトル　$\boldsymbol{x}_i = [x_{0,i} \quad x_{1,i} \quad x_{2,i} \quad x_{3,i}]$　で表すと $\hat{y}_i'$ は，

$$\begin{aligned}\hat{y}_i' &= \boldsymbol{x}_i\hat{\boldsymbol{\beta}} \\ &= x_{0,i}\hat{\beta}_0 + x_{1,i}\hat{\beta}_1 + x_{2,i}\hat{\beta}_2 + x_{3,i}\hat{\beta}_3 \\ &= 10.71x_{0,i} + 0.33x_{1,i} + 2.80x_{2,i} + 0.70x_{3,i}\end{aligned}$$

として計算されている．推定値 $\hat{\boldsymbol{\beta}}$ は，表 3.15 に示した JMP での推定値を用いたのであるが，Excel のソルバーを用いた方法を示す．

表 3.20　対比型のデザイン行列を用いたポアソン回帰

					— デザイン行列 X' —								
G	A	B	y'	n	x_0	x_1	x_2	x_3	y'^	P_i	$\ln L_i$	最大化	
1	A1	B1	4	1	1	1	1	1	14.54	0.001	-7.01	$\ln L$=	**-507.2925**
1	A1	B1	7	2	1	1	1	1	14.54	0.013	-8.65	β^$_0$=	**10.7100**
1	A1	B1	8	1	1	1	1	1	14.54	0.024	-3.73	β^$_1$=	**0.3300**
1	A1	B1	9	2	1	1	1	1	14.54	0.039	-6.50	β^$_2$=	**2.8000**
1	A1	B1	10	3	1	1	1	1	14.54	0.056	-8.63	β^$_3$=	**0.7000**
1	A1	B1	11	5	1	1	1	1	14.54	0.075	-12.98		
1	A1	B1	12	3	1	1	1	1	14.54	0.090	-7.21		
1	A1	B1	13	2	1	1	1	1	14.54	0.101	-4.58	初期値	
1	A1	B1	14	3	1	1	1	1	14.54	0.105	-6.76	$\ln L$=	**-549.4206**
1	A1	B1	15	3	1	1	1	1	14.54	0.102	-6.86	β^$_0$=	**10.0000**
1	A1	B1	16	6	1	1	1	1	14.54	0.092	-14.29	β^$_1$=	**1.0000**
1	A1	B1	17	6	1	1	1	1	14.54	0.079	-15.23	β^$_2$=	**1.0000**
1	A1	B1	18	4	1	1	1	1	14.54	0.064	-11.00	β^$_3$=	**1.0000**
1	A1	B1	19	5	1	1	1	1	14.54	0.049	-15.09		
1	A1	B1	20	1	1	1	1	1	14.54	0.036	-3.34		
1	A1	B1	21	2	1	1	1	1	14.54	0.025	-7.41		
1	A1	B1	22	1	1	1	1	1	14.54	0.016	-4.12		
2	A1	B2	3	1	1	1	-1	-1	7.54	0.038	-3.27		
2	A1	B2	4	5	1	1	-1	-1	7.54	0.072	-13.19		
2	A1	B2	5	4	1	1	-1	-1	7.54	0.108	-8.91		
2	A1	B2	6	10	1	1	-1	-1	7.54	0.136	-19.98		
2	A1	B2	7	7	1	1	-1	-1	7.54	0.146	-13.47		
2	A1	B2	8	6	1	1	-1	-1	7.54	0.138	-11.90		
2	A1	B2	9	5	1	1	-1	-1	7.54	0.115	-10.80		
2	A1	B2	10	6	1	1	-1	-1	7.54	0.087	-14.65		
2	A1	B2	11	3	1	1	-1	-1	7.54	0.060	-8.46		
2	A1	B2	12	1	1	1	-1	-1	7.54	0.037	-3.28		
2	A1	B2	13	1	1	1	-1	-1	7.54	0.022	-3.83		
2	A1	B2	14	1	1	1	-1	-1	7.54	0.012	-4.45		
3	A2	B1	5	1	1	-1	1	-1	12.48	0.010	-4.65		
3	A2	B1	6	1	1	-1	1	-1	12.48	0.020	-3.91		
3	A2	B1	7	2	1	-1	1	-1	12.48	0.036	-6.67		
3	A2	B1	8	2	1	-1	1	-1	12.48	0.055	-5.78		
3	A2	B1	9	3	1	-1	1	-1	12.48	0.077	-7.69		
3	A2	B1	10	4	1	-1	1	-1	12.48	0.096	-9.37		
3	A2	B1	11	7	1	-1	1	-1	12.48	0.109	-15.52		
3	A2	B1	12	7	1	-1	1	-1	12.48	0.113	-15.24		
3	A2	B1	13	5	1	-1	1	-1	12.48	0.109	-11.09		
3	A2	B1	14	5	1	-1	1	-1	12.48	0.097	-11.67		
3	A2	B1	15	1	1	-1	1	-1	12.48	0.081	-2.52		
3	A2	B1	16	6	1	-1	1	-1	12.48	0.063	-16.59		
3	A2	B1	17	2	1	-1	1	-1	12.48	0.046	-6.15		
3	A2	B1	18	2	1	-1	1	-1	12.48	0.032	-6.88		
3	A2	B1	19	1	1	-1	1	-1	12.48	0.021	-3.86		
3	A2	B1	20	1	1	-1	1	-1	12.48	0.013	-4.33		
4	A2	B2	4	2	1	-1	-1	1	8.28	0.050	-6.01		
4	A2	B2	5	3	1	-1	-1	1	8.28	0.082	-7.49		
4	A2	B2	6	5	1	-1	-1	1	8.28	0.113	-10.88		
4	A2	B2	7	12	1	-1	-1	1	8.28	0.134	-24.10		
4	A2	B2	8	10	1	-1	-1	1	8.28	0.139	-19.74		
4	A2	B2	9	4	1	-1	-1	1	8.28	0.128	-8.23		
4	A2	B2	10	6	1	-1	-1	1	8.28	0.106	-13.48		
4	A2	B2	11	2	1	-1	-1	1	8.28	0.080	-5.06		
4	A2	B2	12	2	1	-1	-1	1	8.28	0.055	-5.80		
4	A2	B2	13	1	1	-1	-1	1	8.28	0.035	-3.35		
4	A2	B2	14	3	1	-1	-1	1	8.28	0.021	-11.63		

それぞれの y_i' に対し Excel の関数で,

$$x_{0,i} = 1$$
$$x_{1,i} = \text{if } ((\text{A}_i = \text{"A1"}), 1, -1)$$
$$x_{2,i} = \text{if } ((\text{B}_i = \text{"Bl"}), 1, -1)$$
$$x_{3,i} = x_{1,i} * x_{2,i} \qquad \text{交互作用は，主効果の積}$$
$$\hat{y}_i' = \text{Mmult}(\boldsymbol{x}_i\text{の範囲}, \hat{\boldsymbol{\beta}}\text{の範囲})$$
$$P_i = \text{Poisson.dist}(y_i', \hat{y}_i', \mathit{false})$$
$$\ln L_i = n_i * \ln P_i$$
$$\ln L = \sum_i \ln L_i$$

のように計算されている．初期値として $\hat{\boldsymbol{\beta}} = [10 \quad 1 \quad 1 \quad 1]^T$ を与え，ソルバーで $\ln L$ を最大化するように $\hat{\boldsymbol{\beta}}$ を変化させた結果を「最大化」の欄で示した．最大化された対数尤度とパラメータの推定値は,

$$\ln L = \sum_i \ln L_i = -507.2925$$
$$\hat{\boldsymbol{\beta}} = [\,10.7100 \quad 0.3300 \quad 2.8000 \quad 0.7000\,]^T$$

と表 3.15 に示した JMP でのポアソン回帰の出力結果に一致する．

(0, 1)型デザイン行列

デザイン行列 $\boldsymbol{X}$ は，回帰パラメータとして何を推定したいのかによって自由に設定することができる．表 3.19 に示したように総平均からの効果を求めたいためには（1，−1)対比型のデザイン行列 $\boldsymbol{X}$ が適している．

さて，(1，−1)対比型のデザイン行列 $\boldsymbol{X}$ ではなく，表 3.21 に示すように(0，1)型のデザイン行列 $\boldsymbol{X}$ とすることは可能なのだろうか. 実はどのような数値のデザイン行列 $\boldsymbol{X}$ を定義するかは，全く制約がない．R の glm() 関数の(0，1)型でも，SAS の(1，0)型としても，(−1，1)逆対比型としても，表 3.12 で用いた(0.5，1.5)型としても何ら問題はない．ただし，その結果として推定された回帰パラメータ $\hat{\boldsymbol{\beta}}$ は異なり，それがどの様な意味を持つのかも異なる．

推定された回帰パラメータ $\hat{\boldsymbol{\beta}}$ をどのように解釈するかは，全て自己責任となるので，何を推定したいのかを明確にし，その目的が達成されるようなデザイン行列を設定する必要がある．さらに，推定された回帰パラメータが，目的通りに推定されているかの検証も欠かしてはならない.

さて，(1，–1)対比型ではなく表 3.21 に示すような（0，1)型のデザイン行列 $\boldsymbol{X}$ とした場合に，回帰パラメータ $\hat{\boldsymbol{\beta}}$ は何を推定することになるのであろうか．結果は,

$$\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2 \quad \hat{\beta}_3]^T = [\,14.54 \quad -0.66 \quad -5.60 \quad -1.40\,]^T$$

となる．表 3.19 の場合は，

$$\hat{\boldsymbol{\beta}} = [\,10.71 \quad 0.33 \quad 2.80 \quad 0.70\,]^T$$

と全く異なっていることに．注意してもらいたい．

表 3.21　(0, 1)型のデザイン行列

	B:代謝					***X***								
A:溶媒	活性化	平均		***y***		x_0	x_1	x_2	x_3	***β***		***ε***		***β^***
1:蒸留水	1:なし	14.54		y_1	=	1	0	0	0	β_0	+	ε_1		14.54
	2:あり	7.54		y_2		1	0	1	1	β_1		ε_2		-0.66
2:DMOS	1:なし	12.48		y_3		1	1	0	1	β_2		ε_3		-5.60
	2:あり	8.28		y_4		1	1	1	0	β_3		ε_4		-1.40

表 3.22 に示すように, 回帰パラメータ $\hat{\beta}_0 = 14.54$ は，A_1B_1 のセル平均であり，$\hat{\beta}_1 = -0.66$ は，A_1 水準のセル平均の平均からの A_2 水準の差であり，$\hat{\beta}_2 = -5.60$ は，B_1 水準のセル平均の平均からの B_2 水準の差である．さて，$\hat{\beta}_3 = -1.40$ は，たすき掛の和の平均の差

$$\hat{\beta}_3 \text{ ：交互作用：} \frac{12.48 + 7.54}{2} - \frac{14.54 + 8.28}{2} = -1.40$$

となっていて，表 3.19 の交互作用 $\hat{\beta}_3 = 0.70$ のマイナス 2 倍となっている．交互作用の意味するところは，主効果 A と主効果 B の効果のずれを計量化した結果である．表 3.17 および表 3.18 を参照しつつ，意味するところを感じ取ってもらいたい．

表 3.22　(0, 1)型でのポアソン回帰の結果

	B:代謝活性化			
	1:なし	2:あり	全体	
A:溶媒	平均	平均	平均の平均	差
1:蒸留水	14.54	7.54	11.04	基準
2:DMOS	12.48	8.28	10.38	-0.66
平均の平均	13.51	7.91	10.71	
差	基準	-5.60		

厳密には，与えられたデザイン行列 ***X*** に対する　$\boldsymbol{y} = \boldsymbol{X\beta}$　を展開し

$$(1) \quad y_1 = \beta_0 \times 1 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times 0$$

$$(2) \quad y_2 = \beta_0 \times 1 + \beta_1 \times 0 + \beta_2 \times 1 + \beta_3 \times 1$$

$$(3) \quad y_3 = \beta_0 \times 1 + \beta_1 \times 1 + \beta_2 \times 0 + \beta_3 \times 1$$

$$(4) \quad y_4 = \beta_0 \times 1 + \beta_1 \times 1 + \beta_2 \times 1 + \beta_3 \times 0$$

これらの式を β_0，β_1，β_2，β_3 について解くことが必要である．(1) 式から β_0 は，

$$y_1 = \beta_0 \times 1$$

$$\hat{\beta}_0 = y_1 = 14.54$$

が得られ，これは A_1B_1 のセル平均となっている．β_1 は，[(3)＋(4)] －[(1)＋(2)] によって

$$(y_3 + y_4) - (y_1 + y_2) = 2\beta_1$$

$$\hat{\beta}_1 = \frac{y_3 + y_4}{2} - \frac{y_1 + y_2}{2} = \frac{12.48 + 8.28}{2} - \frac{14.45 + 7.54}{2} = -0.66$$

が得られ A の水準の差となっている．β_2 は，[(2)＋(4)] －[(1)＋(3)] によって

$$(y_2 + y_4) - (y_1 + y_3) = 2\beta_2$$

$$\hat{\beta}_2 = \frac{y_2 + y_4}{2} - \frac{y_1 + y_3}{2} = \frac{7.54 + 8.28}{2} - \frac{14.54 + 12.54}{2} = -5.60$$

が得られ B_2 水準の差となっている．β_3 は，[(2)＋(3)] －[(1)＋(4)] によって

$$(y_2 + y_3) - (y_1 + y_4) = 2\beta_3$$

$$\hat{\beta}_3 = \frac{y_2 + y_3}{2} - \frac{y_1 + y_4}{2} = \frac{7.54 + 12.48}{2} - \frac{14.54 + 8.28}{2} = -1.40$$

が得られ，たすき掛の和の平均の差であるが，A_1 内の B_1 と B_2 の差の 2 分の 1，A_2 内の B_1 と B_2 の差の 2 分の 1 を求め，それらの差とも解釈される．これらは，交互作用といわれているのであるが，どのように解釈するかについて，ここで示した方法を参考に思考を重ねてもらいたい．

$$\hat{\beta}_3 = \frac{y_3 - y_4}{2} - \frac{y_1 - y_2}{2} = \frac{12.48 - 8.28 +}{2} - \frac{14.54 - 7.54}{2} = 2.10 - 3.50 = -1.40$$

推定された回帰パラメータの意味することは，与えられたデザイン行列 $\boldsymbol{X}$ に対し回帰パラメータ $\boldsymbol{\beta}$ を掛けて，それぞれの β_i について解き，その内容を吟味して後付け的に得られる．

基準との差(0, 1)型の拡張

さて，表 3.23 に示したデザイン行列 $\boldsymbol{X}$ は，2×2 の要因配置ではなく，4 水準の 1 元配置型と見なして，最初の水準を基準として，他の水準との差の検定を行いたい場合などに使われる．

表 3.23　基準との差の推定

	B:代謝					$\boldsymbol{X}$							
A:溶媒	活性化	平均		$\boldsymbol{y}$		x_0	x_1	x_2	x_3	$\boldsymbol{\beta}$		$\boldsymbol{\varepsilon}$	$\boldsymbol{\beta}$^
1:蒸留水	1:なし	14.54		y_1	=	1	0	0	0	β_0^	+	ε_1	14.54
	2:あり	7.54		y_2		1	1	0	0	β_1^		ε_2	-7.00
2:DMOS	1:なし	12.48		y_3		1	0	1	0	β_2^		ε_3	-2.06
	2:あり	8.28		y_4		1	0	0	1	β_3^		ε_4	-6.26

推定される回帰パラメータの意味を考えてみよう．切片 $\hat{\beta}_0$ は，$\hat{\beta}_0 = y_1$ であることは自明であろう．$\hat{\beta}_1$ は，$\hat{\beta}_1 = y_2 - y_1$ となることも容易に求められるであろう．同様に，$\hat{\beta}_2 = y_3 - y_1$，$\hat{\beta}_3 = y_4 - y_1$ が得られる．このように基準となる組合せ水準からの差について尤度比検定がまとめてポアソン回帰により行える．

表 3.20 で示した Excel による最尤法のデータを JMP ファイルとし，（0，1）拡張型のデザイン行列にしてポアソン回帰を行った結果を表 3.35 に示す．尤度比カイ 2 乗検定の結果は，全て有意な差であることが確認される．

表 3.24　JMP による基準との差の推定［(0，1）拡張型］

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値	下側信頼限界	上側信頼限界
切片	14.5400	0.5393	727.0000	<.0001*	13.5085	15.6227
x1	-7.0000	0.6645	112.8979	<.0001*	-8.3116	-5.7046
x2	-2.0600	0.7351	7.8603	0.0051*	-3.5038	-0.6201
x3	-6.2600	0.6756	86.9731	<.0001*	-7.5921	-4.9419

交互作用の吟味

さて，B：代謝活性化が（2:あり）の場合に，溶媒を（1:蒸留水）から（2:DMOS）にした場合にコロニー数がわずかに増加しているが，統計的にはどうであろうか．この検討のために，(0，1）型のデザイン行列を設定することもできるが，簡便的には，因子 B 別に因子 A 内の比較を（1，−1）型で行うこともできる．表 3.25 に示すように，p=0.1882 と有意ではない．推定値は，切片が A_1 と A_2 の平均で，A:溶媒[A1] は，A_1 と A_2 の差の 2 分の 1

$$\text{A:溶媒[A1]} = (7.54 - 8.28)/2 = -0.74/2 = -0.37$$

となっている．

表 3.25　代謝活性がある B_2 場合の A_1:蒸留水と A_2: DMOS との差

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値	下側信頼限界	上側信頼限界
切片	7.9100	0.2812	791.0000	<.0001*	7.3715	8.4741
A:溶媒[A1]	-0.3700	0.2812	1.7314	0.1882	-0.9225	0.1813

3.5. 2 本の回帰直線に対する各種のデザイン行列

前節では 2×2 の要因配置型のデータに対して解析のための幾つかのデザイン行列を例示し，それぞれ異なるパラメータの推定値が得られることを示した．その結果の解釈を行うための考え方も詳細に述べてきた．通常の回帰分析については，第 4 章で詳しく述べるが，ここでは，2 本のポアソン回帰直線をさまざまな観点から同時あてはめを行うためのデザイン行列に焦点をあてる．なお，最小 2 乗法による 2 本の回帰直線の解析でもここに示したデザイン行列の考え方は，全く同じである．

切片を共通とする場合(1, 1)型

第 1.8 節では, 2 本のポアソン回帰直線の傾きは異なるが, 共通の切片を持つ場合を示した．解析に用いたデザイン行列については，SAS の DATA ステップで内部的なダミー変数を用いて作成し，詳しい説明を避けた．どの様なデザイン行列が作成されたのであろうか．実際のデータを用いた例示はサイズが大きくなりすぎるので，最小限に縮約したデータを用いて説明する．

表 3.26 に示したのは，第 1.8 節の表 1.27 で示した細菌を用いた用量反応試験の結果を要約したものである［富山ら（2004)］．薬剤の濃度を（0，50，100）の 3 段階にし，反応を変異コロニー数の平均を整数化し $n=6$ の小さなデータとして，さまざまなデザイン行列をコンパクトに例示できるようにする．

S 薬と T 薬に共通の切片を持ち，それぞれの薬剤に別々の傾きを持つようなデザイン行列を設定する．さらに，ポアソン分布による対数尤度 $\ln L$ を設定し，Excel のソルバーにより，$\ln L$ を最大化するように $\hat{\boldsymbol{\beta}}$ を変化させた結果である．推定された 2 本の回帰式は，

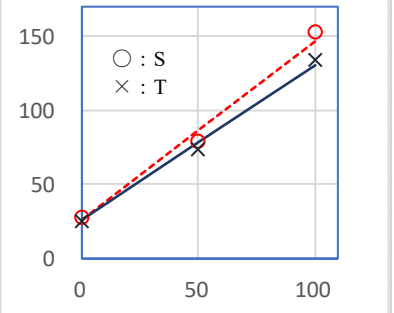
$$\text{S 薬：}\quad \hat{y}^{(S)} = 25.6263 + 1.2121x_1$$

$$\text{T 薬：}\quad y^{(T)} = 25.6263 + 1.0495x_2$$

のように同じ切片を持つが，異なる傾きを持つ回帰式である．

表 3.26　切片を共通にする S 薬と T 薬のポアソン回帰直線

薬剤	濃度	y	y^		デザイン行列 $\boldsymbol{X}$ x_0	x_1	x_2	$\boldsymbol{\beta}$^		対数尤度 $\ln L_i$
S	0	28	25.6263	=	1	**0**	0	**25.6263**	β_0^	-2.6947
	50	79	86.2337		1	**50**	0	**1.2121**	β_1^	-3.4170
	100	153	146.8411		1	**100**	0	**1.0495**	β_2^	-3.5621
T	0	25	25.6263		1	0	**0**			-2.5394
	50	74	78.0996		1	0	**50**			-3.1816
	100	134	130.5729		1	0	**100**			-3.4131
									$\ln L$ =	**-18.8079**



デザイン行列の x_0 は，切片を求めるために全て 1 とし，x_1 には，薬剤 S に対応した行にのみに濃度が設定され，x_2 には，薬剤 T に対応した行にのみに濃度が設定されている．推定値 $\hat{y}_i$ は，デザイン行列 $\boldsymbol{X}$ と回帰パラメータのベクトル $\hat{\boldsymbol{\beta}}$ の積で，次のように計算される．

$$\hat{y}_1 = \hat{\beta}_0 x_{0,1} + \hat{\beta}_1 x_{1,1} + \hat{\beta}_2 x_{2,1} = 25.6263 \times 1 + 1.2121 \times 0 + 1.0495 \times 0 = 25.6263$$
$$\hat{y}_2 = \hat{\beta}_0 x_{0,2} + \hat{\beta}_1 x_{1,2} + \hat{\beta}_2 x_{2,2} = 25.6263 \times 1 + 1.2121 \times 50 + 1.0495 \times 0 = 86.2337$$
$$\vdots$$
$$\hat{y}_6 = \hat{\beta}_0 x_{0,6} + \hat{\beta}_1 x_{1,6} + \hat{\beta}_2 x_{2,6} = 25.6263 \times 1 + 1.2121 \times 0 + 1.0495 \times 100 = 130.5729$$

対数尤度 $\ln L_i$ は，平均を $\hat{y}_i$ とするポアソン分布の確率の対数で，次のように計算される

$$\ln L_1 = \ln[\text{Poisson.dist}(28,\ 25.6263,\ \textit{false})] = \ln[0.0676] = -2.6947$$
$$\ln L_2 = \ln[\text{Poisson.dist}(79,\ 86.2337,\ \textit{false})] = \ln[0.0328] = -3.4170$$
$$\vdots$$
$$\ln L_6 = \ln[\text{Poisson.dist}(134,\ 130.5729,\ \textit{false})] = \ln[0.0329] = -3.4131$$

対数尤度 $\ln L = -18.8079$ は，それぞれの $\ln L_i$ の合計であり，ソルバーにより $\hat{\boldsymbol{\beta}}$ を変化させて最大化した結果である．S 薬について $\hat{y}_i$ の推定値は，$x_{2,i}$ が全て 0 なので

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_{1,1} = 25.6236 + 1.2121 \times 0 = 25.6263$$
$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_{1,2} = 25.6263 + 1.2121 \times 50 = 86.2337$$
$$\hat{y}_3 = \hat{\beta}_0 + \hat{\beta}_1 x_{1,3} = 25.6263 + 1.2121 \times 100 = 146.8411$$

と計算される．T 薬については，$x_{1,i}$ が全て 0 なので

$$\hat{y}_4 = \hat{\beta}_0 + \hat{\beta}_2 x_{2,4} = 25.6236 + 1.0495 \times 0 = 25.6263$$
$$\hat{y}_5 = \hat{\beta}_0 + \hat{\beta}_2 x_{2,5} = 25.6263 + 1.0495 \times 50 = 78.0996$$
$$\hat{y}_6 = \hat{\beta}_0 + \hat{\beta}_2 x_{2,6} = 25.6264 + 1.0495 \times 100 = 130.5729$$

と計算される．

表 3.26 のデザイン行列には，いわゆるダミー変数は使われていない．ただし，第 1.8 節の SAS の Data ステップで，S 薬の場合（$x_S = 1$，$x_T = 0$），T 薬の場合（$x_S = 0$，$x_T = 1$）のように(1, 1)標示(Indicater)型のダミー変数を作成し，濃度 *dose* に対し（$x_1 = x_S \times dose$，$x_2 = x_T \times dose$）によりデザイン行列を生成している．表 3.26 のデザイン行列は，行数が少ないので Excel で濃度 *dose* データを編集したのであるが，便宜的な対応であり勧められない．基本は，（1，1）型のダミー変数（x_S，x_T）を生成し，必要なデザイン行列を内部生成することが基本である．

傾きを共通とする平行線のあてはめ(0, 1)型

表 3.27 は，S 薬と T 薬の傾きを共通とする回帰直線を得るためのデザイン行列である．薬剤のダミー変数として（0，1）型とし，S 薬の場合 $x_1=0$，T 薬の場合 $x_1=1$ を用いている．濃度が $x_2=0$ の場合，それぞれの切片は，

$$\text{S 薬の切片：}\begin{cases}\hat{y}^{(\mathrm{S})}_{x_2=0}=\hat{\beta}_0\cdot 1+\hat{\beta}_1\cdot 0+\hat{\beta}_2\cdot 0\\ \qquad\quad =28.4085\end{cases}$$

$$\text{T 薬の切片：}\begin{cases}\hat{y}^{(\mathrm{T})}_{x_2=0}=\hat{\beta}_0\cdot 1+\hat{\beta}_1\cdot 1+\hat{\beta}_2\cdot 0\\ \qquad\quad =28.4085-5.3319\\ \qquad\quad =23.0776\end{cases}$$

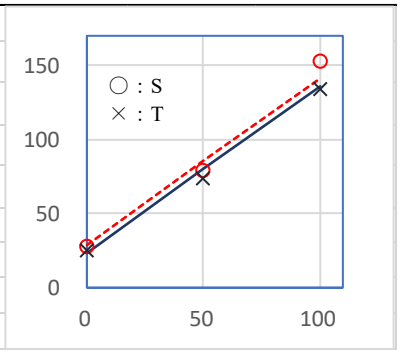
S 薬と T 薬の濃度が同じ変数 x_2 となっていることから回帰パラメータ $\hat{\beta}_2$ は共通となる．推定された 2 本の回帰式は，

$$\hat{y}^{(\mathrm{S})}=28.4085+1.1285x_2$$
$$\hat{y}^{(\mathrm{T})}=23.0766+1.1285x_2$$

のように同じ傾きを持つが，異なる切片を持つ回帰式である．

表 3.27　傾きを共通とする S 薬と T 薬の回帰直線

薬剤	濃度	y	y^		デザイン行列 $\boldsymbol{X}$ x_0	x_1	x_2	$\boldsymbol{\beta}$^		対数尤度 $\ln L_i$
S	0	28	28.4085	=	1	0	0	**28.4085**	β_0^	-2.5910
	50	79	84.8326		1	0	50	**-5.3319**	β_1^	-3.3100
	100	153	141.2568		1	0	100	**1.1285**	β_2^	-3.9098
T	0	25	23.0765		1	**1**	0			-2.6097
	50	74	79.5007		1	**1**	50			-3.2669
	100	134	135.9248		1	**1**	100			-3.3822
								$\ln L$ =		**-19.0696**



S 薬と T 薬の回帰直線の差は，切片の差の推定値 $\hat{\beta}_1$

$$\hat{\beta}_1=-5.3319$$

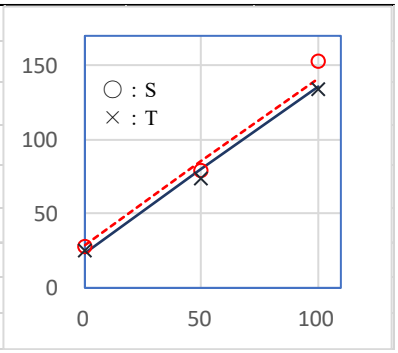
であることがわかる．

傾きを共通とする平行線のあてはめ(1, 1)型

表 3.28 は，表 3.27 と同じ切片は異なるが，共通の傾きを持つ回帰直線の推定である．ただし，ザイン行列は異なり，変数 x_0 は，S 薬の時にのみ 1 で，T 薬では 0 となっている．変数 x_1 は，逆に S 薬の時に 0 で，T 薬では 1 となっている．このような（1，1）型のデザイン行列

表 3.28　傾きを共通とする T 薬と S 薬の切片の直接推定

薬剤	濃度	y	y^		デザイン行列 $\boldsymbol{X}$ x_0	x_1	x_2	$\boldsymbol{\beta}$^		対数尤度 $\ln L_i$
S	0	28	28.4085	=	**1**	0	0	**28.4085**	β_0^	-2.5910
	50	79	84.8326		**1**	0	50	**23.0765**	β_1^	-3.3100
	100	153	141.2567		**1**	0	100	**1.1285**	β_2^	-3.9098
T	0	25	23.0765		0	**1**	0			-2.6097
	50	74	79.5006		0	**1**	50			-3.2669
	100	134	135.9247		0	**1**	100			-3.3822
									$\ln L$ =	**-19.0696**



にすることにより，T 薬と S 薬の切片を直接推定することができる．推定された 2 本の回帰式は，

$$\text{S 薬：}\begin{cases}\hat{y}^{(\mathrm{S})}=\hat{\beta}_0\cdot 1+\hat{\beta}_1\cdot 0+\hat{\beta}_2\cdot x_2\\ \quad =28.4085+\quad 1.1285x_2\end{cases}$$

$$\text{T 薬：}\begin{cases}\hat{y}^{(\mathrm{T})}=\hat{\beta}_0\cdot 0+\hat{\beta}_1\cdot 1+\hat{\beta}_2\cdot x_2\\ \quad =\quad 23.0765+1.1285x_2\end{cases}$$

のように同じ傾きであるが，異なる切片を持つ回帰式である．切片の差 d は，

$$d=\hat{\beta}_1-\hat{\beta}_0=23.0765-28.4085=-5.3320$$

として推定される．

交互作用(0, 1)型

共通の傾きを持つ回帰直線のあてはめが，統計的に支持されるのかを検討するために，表 3.27 に示した（0，1）型のダミー変数のデザイン行列に，薬剤 $\boldsymbol{X}_1$ と濃度 $\boldsymbol{X}_2$ の交互作用 $\boldsymbol{X}_3$ を

$$\text{変数 } \boldsymbol{X}_3=\boldsymbol{X}_1\boldsymbol{X}_2$$

のように積として加える．表 3.29 に交互作用列を追加した 6×4 のデザイン行列に対するポアソン回帰の解析結果を示す．パラメータ $\hat{\beta}_0=26.8140$ は，S 薬の切片であり，$\hat{\beta}_1=-2.3905$ は S 薬の切片と T 薬の切片の差であり，T 薬の切片は，

$$\hat{\beta}_0+\hat{\beta}_1=26.8140-2.3905=24.4235$$

となる．2 本の回帰式は，*dose*（濃度）を用いて

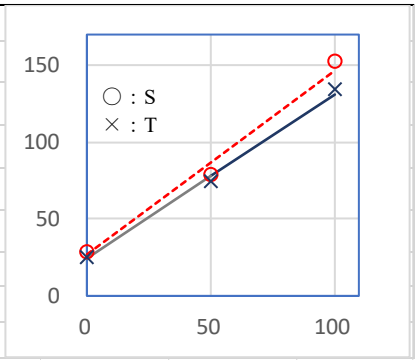
$$\hat{y}^{(\mathrm{S})}=26.8140+1.1971\cdot dose$$

$$\begin{aligned}\hat{y}^{(\mathrm{T})}&=26.8140-2.3905+(1.1971-0.1322)\cdot dose\\ &=24.4235+1.0649\cdot dose\end{aligned}$$

のように異なる切片および異なる傾きを持つ回帰式である．

表 3.29　交互作用の検討

薬剤	濃度	y	y^		デザイン行列 $\boldsymbol{X}$ x_0	x_1	x_2	x_3	$\boldsymbol{\beta}$^		対数尤度 $\ln L_i$
S	0	28	26.814	=	1	0	0	0	**26.8140**	β_0^	-2.6139
	50	79	86.667		1	0	50	0	**-2.3905**	β_1^	-3.4543
	100	153	146.519		1	0	100	0	**1.1971**	β_2^	-3.5760
T	0	25	24.424		1	**1**	0	**0**	**-0.1322**	β_3^	-2.5385
	50	74	77.667		1	**1**	50	**50**			-3.1600
	100	134	130.910		1	**1**	100	**100**			-3.4047
									$\ln L$ =		**-18.7473**



この交互作用モデルの対数尤度は，$\ln L^{(\text{交互作用})} = -18.7473$ であり，表 3.27 の主効果のみのモデルの対数尤度　$\ln L^{(\text{主効果})} = -19.0696$ との差の 2 倍

$$\chi^2_{x_3} = 2\left(\ln L^{(\text{交互作用})} - \ln L^{(\text{主効果})}\right) = 2\times(-18.7473+19.0696) = 0.6446,\quad p = 0.4220 \quad \text{N.S.}$$

が自由度 1 のカイ 2 乗分布に従うことで検定ができる．

表 3.30 に JMP によるポアソン回帰による交互作用モデルの結果を示す．(–1)*対数尤度の欄の完全の行が 18.7473 と Excel の対数用尤度に一致し，交互作用に関する項 x3 の推定値は，–0.1322 であり，尤度比検定の P 値も 0.4220 と一致することが確認できる．

表 3.30　JMP によるポアソン回帰での交互作用に対する尤度比検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度
差分	89.2672	178.5344	3
完全	18.7473		
縮小	108.0145		

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	26.8140	4.9468	29.3810	<.0001*
x1	-2.3905	6.8659	0.1212	0.7277
x2	1.1971	0.1191	94.9808	<.0001*
x3	-0.1322	0.1646	0.6447	0.4220

別々の回帰直線(1, 1)型

交互作用の検討では，薬剤と濃度についての主効果モデルに，交互作用列を加えたモデルによって行った．別々の回帰直線のパラメータの推定を直接求めることもデザイン行列の設定次第で自由にできる．表 3.31 に示したデザイン行列は，S 薬と T 薬に別々の切片と傾きを同時推定することができる．推定された 2 本の回帰式は，*dose* を用いて

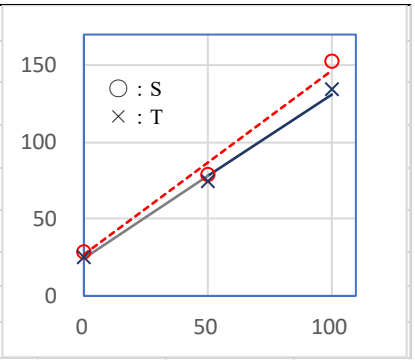
$$y^{(\text{S})} = 26.8140 + 1.1971\cdot dose$$

$$y^{(\text{T})} = 24.4235 + 1.0649\cdot dose$$

のように異なる切片および傾きを持つ回帰式である．

表 3.31　別々の回帰直線のパラメータ推定

薬剤	濃度	y	y^		デザイン行列 $\boldsymbol{X}$ x_0	x_1	x_2	x_3	$\boldsymbol{\beta}$^		対数尤度 $\ln L_i$
S	0	28	26.814	=	**1**	0	**0**	0	**26.8140**	β_0^	-2.6139
	50	79	86.667		**1**	0	**50**	0	**24.4235**	β_1^	-3.4543
	100	153	146.519		**1**	0	**100**	0	**1.1971**	β_2^	-3.5760
T	0	25	24.423		0	**1**	0	**0**	**1.0649**	β_3^	-2.5385
	50	74	77.667		0	**1**	0	**50**			-3.1600
	100	134	130.910		0	**1**	0	**100**			-3.4047
									$\ln L$ =		**-18.7473**



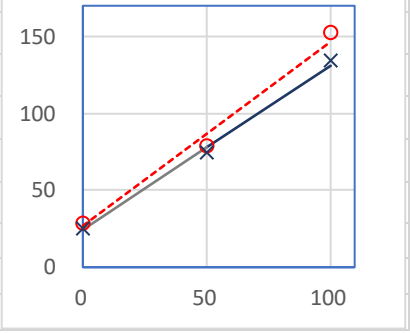
推定結果を見ると，S 薬の切片：$\hat{\beta}_0 = 26.8140$，傾き：$\hat{\beta}_2 = 1.1971$，T 薬の切片：$\hat{\beta}_1 = 24.4235$，傾き：$\hat{\beta}_2 = 1.0649$ となる．交互作用モデルと比較すると，推定されたパラメータは異なるが，推定値 $\hat{y}_i$ は完全に一致している．もちろん対数尤度も $\ln L = -18.7473$ と一致している．

別々の回帰直線(1, −1)対比型

表 3.32 に示すように薬剤に対して（1，-1）対比型ダミー変数を設定することもできる．これは，JMP によるポアソン回帰で，薬剤（S，T）を変数とした場合に相当する．

表 3.32　（1，-1）対比型を用いた別々の回帰直線のパラメータ推定

薬剤	濃度	y	y^		デザイン行列 $\boldsymbol{X}$ x_0	x_1	x_2	x_3	$\boldsymbol{\beta}$^		対数尤度 $\ln L_i$
S	0	28	26.814	=	1	1	0	0	**25.6188**	β_0^	-2.6139
	50	79	86.667		1	1	50	50	**1.1952**	β_1^	-3.4543
	100	153	146.519		1	1	100	100	**1.1310**	β_2^	-3.5760
T	0	25	24.424		1	**-1**	0	**0**	**0.0661**	β_3^	-2.5385
	50	74	77.667		1	**-1**	50	**-50**			-3.1600
	100	134	130.910		1	**-1**	100	**-100**			-3.4047
									$\ln L$ =		**-18.7473**



得られた推定値から 2 本の回帰式は，*dose* を用いて次のように推定される．

$$y^{(S)} = 25.6188 + 1.1952 + (1.1310 + 0.0661) \times dose$$
$$= 26.8140 + 1.1971 \times dose$$

$$y^{(T)} = 25.6188 - 1.1952 + (1.1310 - 0.0661) \times dose$$
$$= 24.4236 + 1.0649 \times dose$$

3.6. オフセットを含む対数リンクでの 2 本の 2 次曲線のあてはめ

第 1.11 節で取り上げた「喫煙習慣と冠動脈疾患による死亡」データを表 3.33 に再掲する［ドブソン（2008）］．第 1.11 節では，オフセットを含む 2 群の対数リンクによるポアソン回帰の導入のために，80 歳のデータを除いた結果を示した．これは，80 歳での 10 万人比での死亡数が，非喫煙者と喫煙者で逆転し，非喫煙者群の死亡数が多くなり，1 次式ではなく 2 次のポアソン回帰の検討を要する事例であった．

表 3.33　年齢階層別の喫煙習慣による冠動脈心疾患による死亡数（表 1.38 再掲）

年齢		非喫煙者（$x=0$）			喫煙者（$x=1$）		
範囲	歳	死亡	人年	10万人比	死亡	人年	10万人比
35-44	40	2	18,790	10.6	32	52,407	61.1
45-54	50	12	10,673	112.4	104	43,248	240.5
55-64	60	28	5,710	490.4	206	28,612	720.0
65-74	70	28	2,585	1083.2	186	12,663	1468.8
75-84	80	31	1,462	2120.4	102	5,317	1918.4

各年代での部分母集団の大きさが，35−44 歳代に比べ 75−84 歳代は 10 分の 1 以下となるので，10 万人比による散布図を図 3.2 に示す．左は実目盛りであり，死亡数が年齢と共に指数関数的に増加している．40 歳から 70 歳までは，喫煙者の死亡数が非喫煙者よりも多いが 80 歳で逆転している．部分母集団の数が，それぞれ（1,462，5,317）と少ないので統計的な変動の範囲内なのかも知れない．

図 3.2 右の対数目盛での散布図から，年齢が 40 歳の場合の対数死亡数が喫煙者と非喫煙者の間で最も開いていて，年齢が上がるにつれ差が縮まり 80 歳ではほとんど同じとなっている．また，年齢が上がるにつれ対数死亡数の増加は減弱している．これらのことから，単純

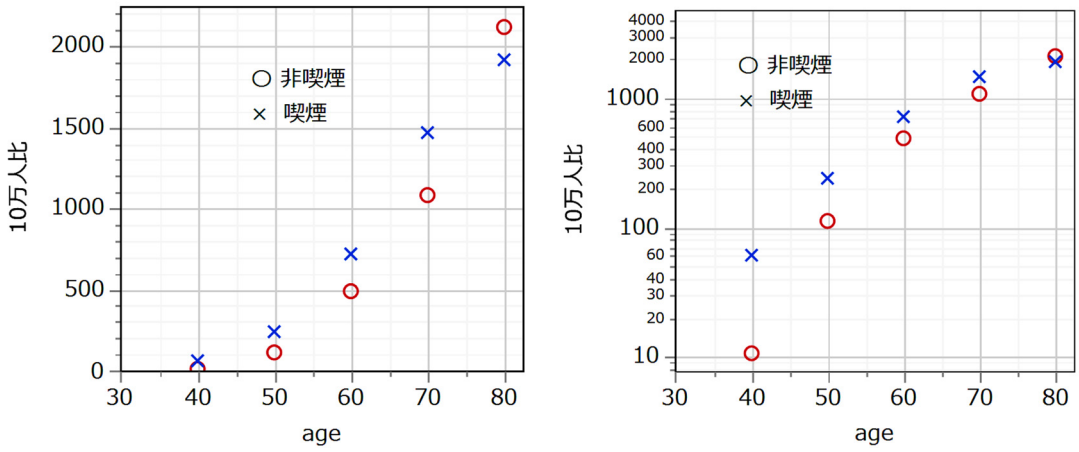


図 3.2　非喫煙・喫煙別の 10 万人比での死亡数

な対数リンクによる 2 本の直線のあてはめではなく，年齢に 2 次の項を追加する必要性が感じられる．さらに喫煙習慣と年齢の交互作用も推察される．一足飛びに 2 次式のあてはめを行う前に，最も簡単なモデルから複雑なモデルへと段階的に進めることにより，ポアソン回帰の応用について理解を深めてもらいたい．

（非喫煙･喫煙）の 2 群間比較

年齢区分を無視して，（0:非喫煙，1:喫煙）で各 5 組の死亡数 y_i とその部分母集団数 n_i が得られたとしよう．表 3.34 の Excel シートは，年齢区分を除いた 10×2 のデザイン行列でのポアソン回帰を行う．推定値を $\hat{y}_i = n_i \cdot \exp(\hat{\beta}_0 x_{0i} + \hat{\beta}_0 x_{1i})$，ポアソン確率を $P_i = \text{Poisson.dist}(y_i, \hat{y}_i, \mathit{false})$，対数尤度を $\ln L_i = \ln(P_i)$，その合計を $\ln L$ とし，$\ln L$ を最大化するように（$\hat{\beta}_0$，$\hat{\beta}_1$）をソルバーで変化させた結果である．推定されたパラメータから，10 万人比での非喫煙者の死亡数が 257.5 人，喫煙者の場合は 385.6 人と推定される．

$$\text{非喫煙者：}\ \hat{y}^{(\text{非喫煙})} = 100{,}000 \times \exp(-5.9618 + 0.5422 \times 0) = 257.5 \text{ 人}$$
$$\text{喫煙者：}\ \hat{y}^{(\text{喫煙})} = 100{,}000 \times \exp(-5.9618 + 0.5422 \times 1) = 385.6 \text{ 人}$$

表 3.34　喫煙習慣による冠動脈心疾患による死亡数

切片	喫煙	死亡	人年	10万人比	推定値	確率	対数尤度		
x_0	x_1	y	n	y'	y^	P	$\ln L_i$		最尤解
1	0	2	18,790	11	48.4	1.13E-18	-41.3229	β_0^=	**-5.9618**
1	0	12	10,673	112	27.5	4.49E-04	-7.7087	β_1^=	**0.5422**
1	0	28	5,710	490	14.7	6.58E-04	-7.3260		
1	0	28	2,585	1,083	6.7	4.75E-10	-21.4682		
1	0	31	1,462	2,120	3.8	1.99E-18	-40.7593		
1	1	32	52,407	61	232.1	3.02E-61	-139.3535		
1	1	104	43,248	240	191.5	1.44E-12	-27.2691		
1	1	206	28,612	720	126.7	2.54E-11	-24.3980		
1	1	186	12,663	1,469	56.1	1.10E-42	-96.6127		
1	1	102	5,317	1,918	23.5	5.37E-33	-74.3037		
						$\ln L$ =	**-480.5221**		

$y'_i = (y_i / n_i) \cdot 100{,}000$，　$\hat{y}_i = n_i \cdot \exp(\hat{\beta}_0 x_{0,i} + \hat{\beta}_0 x_{1,i})$，　$P_i = \text{Poisson.dist}(y_i, \hat{y}_i, \mathit{false})$，　$\ln L_i = \ln(P_i)$

2 群間の尤度比カイ 2 乗値は，切片のみをデザイン行列（10×1）とした縮小モデルでの対数尤度を

縮小モデル:	β_0^=	**-5.5144**		$\ln L$ =	**-495.0676**

を求め，完全モデル $\ln L_{\text{完全}} = -480.5521$ と縮小モデル $\ln L_{\text{縮小}} = -495.0676$ のマイナス 1 倍が，表 3.35 に示す JMP の結果に一致する．これらの差の−2 倍は，「差分」の行で示され 29.0911 となり，有意な差であると判断される．

表 3.35　JMP による喫煙習慣に関する尤度比カイ 2 乗

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	14.5456	29.0911	1	<.0001*
完全	480.5221			
縮小	495.0676			

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-5.9618	0.0995	77033.712	<.0001*
x1:smoke	0.5422	0.1072	29.0911	<.0001*

この推定結果を図 3.3 に示す. 実目盛りでは若年層での母集団数が大きいことが反映され, 10 万人比目盛り上で低い位置に推定結果が表示されているが，対数目盛上では，中ほどに推定結果が示されている.

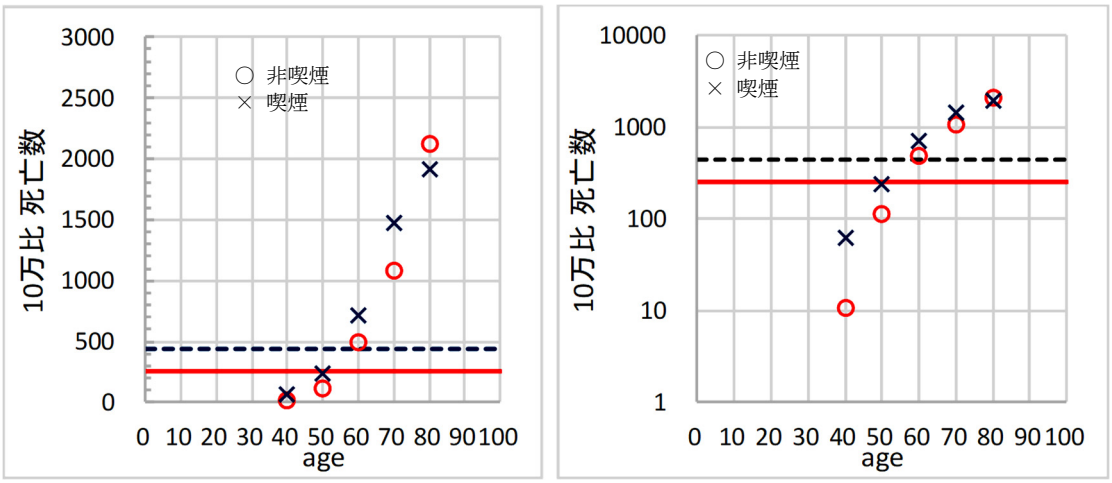


図 3.3　（非喫煙・喫煙）に対するポアソン回帰

オフセットを含む対数リンクでの 2 本の回帰直線のあてはめ

喫煙習慣に年齢を加えたモデルでは，対数死亡数に対して 2 本の直線をあてはめることを考える．その際，共通の切片を持つのか，共通の傾きを持つ平行線をあてはめるのか，別々の直線をあてはめるのかの判断が必要である．図 3.3 から，別々の直線のあてはめが適しているように思われるが，喫煙習慣により死亡数に差あるかを検討したいので，平行な直線をあてはめて，年齢に関わらず喫煙習慣の差について平均的な尤度比カイ 2 乗値で検討したい.

表 3.36 に示すように，喫煙習慣と年齢を加えた 10×3 のデザイン行列では，非喫煙群の回帰直線の切片が $\hat{\beta}_0$ となり，喫煙群の切片は $\hat{\beta}_0 + \hat{\beta}_1$ となる．傾きは $\hat{\beta}_2$ で共通で，年齢が 60 歳

の場合の 10 万人比での推定値は,

非喫煙者： $\hat{y}_{age=60}^{(非喫煙)} = 100{,}000 \times \exp(-10.6260 + 0.4064 \times 0 + 0.0836 \times 60) = 365.8$ 人

喫煙者： $\hat{y}_{age=60}^{(喫煙)} = 100{,}000 \times \exp(-10.6260 + 0.4064 \times 1 + 0.0836 \times 60) = 549.2$ 人

となる.

表 3.36　喫煙習慣別の年齢による対数リンクでの 2 本の直線のあてはめ

切片	喫煙	年齢	死亡	人年	10万人比	推定値	10万比	確率	対数尤度		
x_0	x_1	x_2:age	y	n	y'	y^	y^′	P	ln Li		最尤解
1	0	40	2	18,790	11	12.9	68.7	0.0002	-8.4925	β_0^=	**-10.6260**
1	0	50	12	10,673	112	16.9	158.6	0.0515	-2.9664	β_1^=	**0.4064**
1	0	60	28	5,710	490	20.9	365.8	0.0252	-3.6815	β_2^=	**0.0836**
1	0	70	28	2,585	1,083	21.8	843.8	0.0336	-3.3930		
1	0	80	31	1,462	2,120	28.5	1946.4	0.0640	-2.7491		
1	1	40	32	52,407	61	54.1	103.2	0.0004	-7.9464		
1	1	50	104	43,248	240	103.0	238.1	0.0389	-3.2471		
1	1	60	206	28,612	720	157.1	549.2	0.0000	-10.4973		
1	1	70	186	12,663	1,469	160.4	1266.9	0.0042	-5.4708		
1	1	80	102	5,317	1,918	155.4	2922.4	0.0000	-13.6810		
								ln L =	**-62.1250**		

$$y_i' = (y_i / n_i) \cdot 100{,}000\,, \quad \hat{y}_i = n_i \cdot \exp(\hat{\beta}_0 x_{0i} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})\,,$$
$$P_i = \text{Poisson.dist}(y_i, \hat{y}_i, \mathit{false})\,, \quad \ln L_i = \ln(P_i)$$

年齢をモデルに加えることにより，$\ln L = -62.1250$ と大きく変化する．図 3.4 に示すように年齢を加味しても喫煙習慣による死亡数に違いがあるようだが，尤度比カイ 2 乗値を求めるためには，喫煙習慣を含めない年齢のみの縮小モデルでのポアソン回帰での対数尤度が必要となる．表 3.36 の $\hat{\beta}_1$ をゼロにセットし，$\ln L$ を最大化するように，$\hat{\beta}_0$ および $\hat{\beta}_2$ を変化させると，次の結果が得られ,

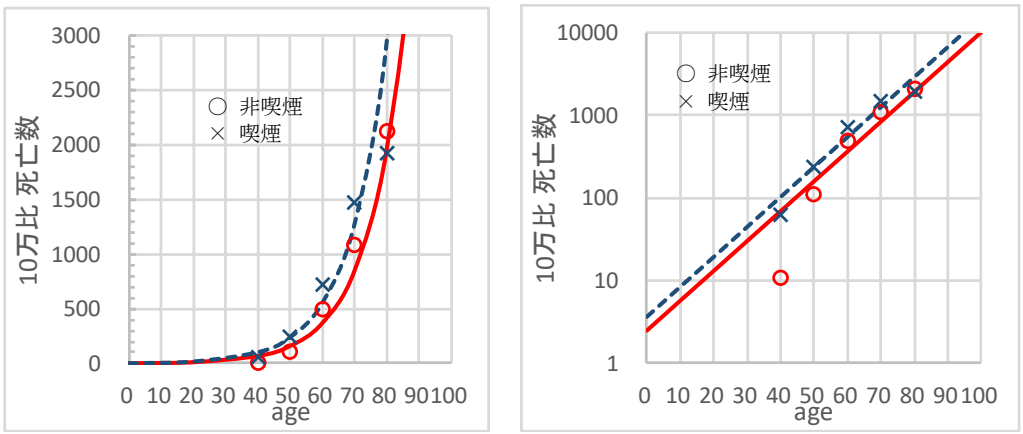


図 3.4　喫煙歴別の対数リンクでの 2 本のポアソン回帰直線

縮小モデル年齢:	$\hat{\beta}_0$=	**-10.2988**		$\ln L$ =	**-70.0397**
	$\hat{\beta}_2$=	**0.0838**			

対数尤度は，$\ln L = -70.0397$ となる．各モデルでの対数尤度，マイナス 2 倍の対数尤度，縮小モデルとの差をまとめると，

モデル	対数尤度	マイナス 2 倍	縮小モデルとの差
切片　（縮小）	-495.0676	990.1353	
切片＋煙習習慣	-480.5221	961.0441	29.0911
切片＋年齢　（縮小）	-70.0397	140.0795	
切片＋喫煙習慣＋年齢	-62.1250	124.2500	15.8294

となる．年齢を加味しない場合の喫煙歴の差についての尤度比カイ 2 乗値=29.0911 に対し，年齢を加味した場合の尤度比カイ 2 乗値=15.8294 と若干減少するが，自由度 1 のカイ 2 乗分布の 5%上側確率点は，3.84 なので，はるかに大きいカイ 2 乗値となっている．切片のみの縮小モデルに対し，年齢を加味した場合のマイナス 2 倍の対数尤度の差は，

尤度比カイ 2 乗=990.1353−140.0795=850.0558

となり，喫煙習慣に対してはるかに大きい．また，（切片＋喫煙習慣）と（切片＋喫煙習慣＋年齢）の差は，

尤度比カイ 2 乗=961.0441−124.2500=836.7941

となる，喫煙習慣が含まれているモデルについて，年齢を加味した場合のマイナス 2 倍の対数尤度の差は，少し減ってはいるが，極めて大ききく，加齢とともに死亡数が劇的に増えることを意味している．

（切片＋喫煙習慣＋年齢）モデルに対し，表 3.37 に示す JMP のポアソン回帰の出力と対比する．「モデル」の「縮小」の (-1)*対数尤度 495.0676 は，切片のみのモデルに対応し，「完全」

表 3.37　（切片＋喫煙習慣＋年齢）ポアソン回帰

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	432.9426	865.8853	2	<.0001*
完全	62.1250			
縮小	495.0676			

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-10.6258	0.2097	3777.5415	<.0001*
x1:smoke	0.4064	0.1072	15.8294	<.0001*
x2:age	0.0836	0.0029	836.7941	<.0001*

は，（切片＋喫煙習慣＋年齢）に対応し 62.1250 となる．「差分」の 865.8853 は，切片モデルに（喫煙習慣＋年齢）を加えた場合の 2 倍の差である．

表 3.37 の「パラメータ推定値」の推定値は，表 3.36 の Excel での計算結果に一致している．「標準誤差」は，Excel の計算過程でパラメータの共分散行列の計算をせずにソルバーで最尤解を直接求めたために対応できない．「x1:smoke」の「尤度比カイ 2 乗」は，15.8294 であり，すでに求めた（切片＋年齢）と（切片＋喫煙習慣＋年齢）モデルのマイナス 2 倍の対数尤度の差となっている．「x1:age」の 836.7941 は，（切片＋煙習習慣）と（切片＋喫煙習慣＋年齢）の差であることもすでに示した．

切片のみが異なる 2 本の 2 次曲線のあてはめ

図 3.4 に示した散布図から，対数目盛上では，非喫煙群の 40 歳では，回帰直線からの乖離が大きく，全体的には各点が上に凸であり，喫煙群でも同様に上に凸となっている．統計的に意味があるかを検討するためには，2 次曲線のあてはめの必要性が示唆される．ただし，2 次曲線のあてはめ，統計的に有意となっても，直線のあてはめには問題があると指摘するだけで，2 次曲線の回帰パラメータに対する意味づけは困難である．

複数の直線のあてはめの場合でも，切片が同じで傾きが異なるモデル，切片は異なるが傾きが同じモデル，別々のモデル，3 通りのモデルを第 3.5 節で示してきた．1 本の 2 次曲線の場合についてならば，迷うことはないのであるが，2 本の場合については，年齢に依存せず喫煙習慣によらず死亡数が異なるか，表 3.38 に示すように年齢の 2 乗を追加して検討する．

表 3.38　年齢について 2 乗を追加

切片	喫煙	年齢	年齢^2	死亡	人年	10万人比	推定値	10万人比	対数尤度		
x_0	x_1	x_2:age	x_3:age^2	y	n	y'	$y^{\wedge}$	$y^{\wedge}{}'$	$\ln Li$		
1	0	40	16	2	18,790	11	6.7	35.9	-3.6200	$\beta_0\hat{}=$	**-17.8583**
1	0	50	25	12	10,673	112	17.4	162.6	-3.0947	$\beta_1\hat{}=$	**0.3547**
1	0	60	36	28	5,710	490	28.5	499.4	-2.5927	$\beta_2\hat{}=$	**0.3258**
1	0	70	49	28	2,585	1,083	26.9	1040.4	-2.6105	$\beta_3\hat{}=$	**-0.1942**
1	0	80	64	31	1,462	2,120	21.5	1469.9	-4.4873		
1	1	40	16	32	52,407	61	26.8	51.2	-3.1253		
1	1	50	25	104	43,248	240	100.2	231.8	-3.3112		
1	1	60	36	206	28,612	720	203.7	712.1	-3.5959		
1	1	70	49	186	12,663	1,469	187.8	1483.4	-3.5413		
1	1	80	64	102	5,317	1,918	111.4	2095.7	-3.6430		
		$(age/10)^2$						$\ln L$ =	**-33.6218**		

縮小モデル（切片＋喫煙習慣＋年齢）の場合のマイナス 2 倍の対数尤度は，124.2500 であるのに対し，完全モデル（切片＋喫煙習慣＋年齢＋年齢^2）の場合は，67.2435 と大幅に減少

し，その差は 57.0065 であり，年齢^{2}を加えることは統計的に支持される．

モデル	対数尤度	マイナス 2 倍	縮小モデルとの差
切片＋喫煙習慣＋年齢（縮小）	-62.1250	124.2500	
切片＋喫煙習慣＋年齢＋年齢^2	-33.6218	67.2435	57.0065

図 3.5 に 2 本の 2 次曲線を描いた結果を示す．実目盛り上では，非喫煙者の 80 歳が上方に大きく外れているが，母集団の数が小さいための誤差的な変動と解釈される．しかし，図 3.5 右の対数目盛のあてはまりを見ると 40 歳代の非喫煙群でのあてはまりは悪くなっている．

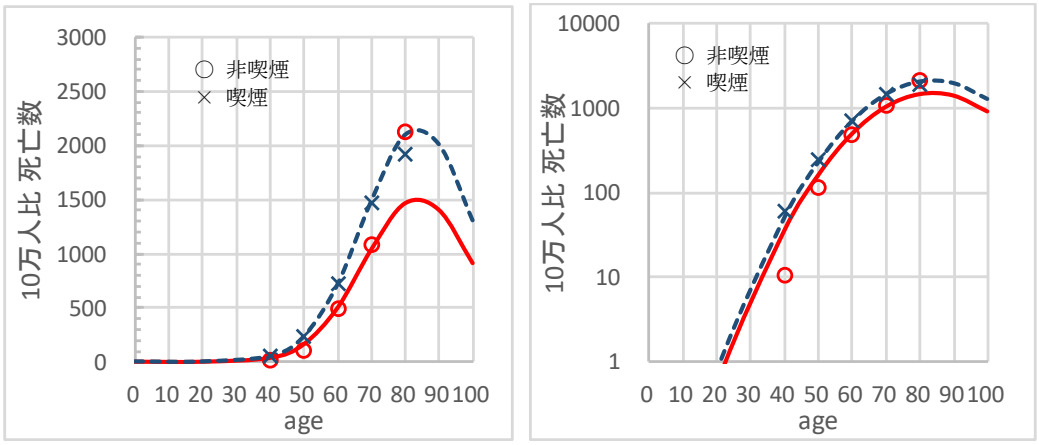


図 3.5　喫煙歴別の対数リンクでの 2 本の 2 次曲線

年齢が 60 歳の場合の 10 万人比の推定死亡数は，

非喫煙者：$\hat{y}_{age=60}^{(非喫煙)} = 100{,}000 \times \exp(-17.8583 + 0.3547 \times 0 + 0.3258 \times 60 - 0.1942 \times 6^2) = 499.4$ 人

喫煙者　　$\hat{y}_{age=60}^{(喫煙)} = 100{,}000 \times \exp(-17.8583 + 0.3547 \times 1 + 0.3258 \times 60 - 0.1942 \times 6^2) = 712.1$ 人

と推定される．

次に，（切片＋年齢＋年齢^2）を縮小モデルとして（切片＋喫煙習慣＋年齢＋年齢^2）との差により喫煙習慣の統計的性能を評価する．

縮小モデル：					
切片	$\hat{\beta}_0 =$	**-17.7907**		$\ln L =$	**-39.5501**
	$\hat{\beta}_1 =$	**-**			
年齢	$\hat{\beta}_2 =$	**0.3332**			
年齢^2	$\hat{\beta}_3 =$	**-0.2000**			

縮小モデルとの差の尤度比カイ 2 乗値は，11.8566 と統計的に有意な差となっている．

モデル	対数尤度	マイナス 2 倍	縮小モデルとの差
切片＋年齢＋年齢^2（縮小）	-39.5501	67.2435	
切片＋喫煙習慣＋年齢＋年齢^2	-33.6218	79.1002	11.8566

喫煙習慣と年齢の交互作用を含む 2 本の 2 次曲線のあてはめ

図 3.5 で，喫煙習慣によって切片が異なるモデルによって 2 次曲線のあてはめを仔細にみると，良くあてはまっているとは，言い難い．そのために，喫煙習慣と年齢の交互作用を含めたモデルにチャレンジしたくなる．ただし，厄介なのは，喫煙習慣の違いによって異なる 2 次式をあてはめる場合には，交互作用として（喫煙習慣×年齢）を加え，さらに（喫煙習慣×年齢^2）を加える必要があるかである．

喫煙習慣との交互作用を含めた場合には，統計的に有意となった場合には，年齢ごとの喫煙習慣による死亡者数の比較が必要となり，これまで示してきたような，年齢にかかわらず，喫煙習慣による死亡者を論ずることができなくなってしまう．

交互作用	モデル
なし(縮小)	切片＋喫煙習慣＋年齢＋年齢^2
1 次のみ	切片＋喫煙習慣＋年齢＋年齢^2＋(喫煙習慣×年齢)
2 次まで	切片＋喫煙習慣＋年齢＋年齢^2＋(喫煙習慣×年齢)＋(喫煙習慣×年齢^2)

このような変数が多くなると，対数尤度 $\ln L$ を最大化するための適当な初期値の設定が困難になる．与えた初期値では Excel のソルバーによって $\ln L$ を最大化できなくなることもしばしば起きる．また，パラメータ共分散行列を求めるために行列計算を行うことも極めて煩雑になるので，統計ソフト JMP でパラメータを推定し，Excel で整理し，さらにグラフ化する．表 3.39 に解析用の JMP データセットを示す．

一般化線形モデルで分布をポアソン，リンク関数を対数，オフセットを $\ln(n_i)$ とする．変数は，切片モデル，切片＋喫煙習慣,...，切片＋喫煙習慣＋・・・＋（喫煙習慣×年齢^2）のように逐次変数を増やして解析し，得られたパラメータの推定値を順次表 3.40 に示す Excel シートにコピーする．それぞれの (-1)*対数尤度もコピーし，2 倍した後に差分を計算する．この差分が追加された変数に対する尤度比カイ 2 乗値となるので，自由度 1 のカイ 2 乗分布の上側確率を Chisq.dist.RT()関数で p 値を求めている．

表 3.39　喫煙習慣と年齢の交互作用解析のための JMP データセット

	x0	x1:smoke	x2:age	x3:age^2	x4:S*age	x5:S*a^2	y	n	ln n
1	1	0	40	16	0	0	2	18,790	9.8411
2	1	0	50	25	0	0	12	10,673	9.2755
3	1	0	60	36	0	0	28	5,710	8.6500
4	1	0	70	49	0	0	28	2,585	7.8575
5	1	0	80	64	0	0	31	1,462	7.2876
6	1	1	40	16	40	16	32	52,407	10.8668
7	1	1	50	25	50	25	104	43,248	10.6747
8	1	1	60	36	60	36	206	28,612	10.2616
9	1	1	70	49	70	49	186	12,663	9.4464
10	1	1	80	64	80	64	102	5,317	8.5787

表 3.40　変数の逐次増加に対する推定値および尤度比カイ 2 乗値

変数	β_0^	β_0^～β_1^	β_0^～β_2^	β_0^～β_3^	β_0^～β_4^	β_0^～β_5^
x_0:	-5.5144	-5.9618	-10.6258	-17.8675	-19.7003	-21.4905
x_1:*smoke*		0.5422	0.4064	0.3545	2.3636	4.4053
x_2:*age*			0.0836	0.3261	0.3563	0.4136
x_3:age^2				-0.1944	-0.1977	-0.2421
x_4:$s \times age$					-0.0308	-0.0964
x_5:$s \times age^2$						0.0511
	x_0	x_1:*smoke*	x_2:*age*	x_3:age^2	x_4:$s \times age$	x_5:$s \times age^2$
(-1)対数尤度	495.0676	480.5221	62.1250	33.6217	28.3517	28.1568
2倍	990.1353	961.0441	124.2500	67.2435	56.7033	56.3136
差分		29.0911	836.7941	57.0065	10.5402	0.3898
p 値		0.0000	0.0000	0.0000	0.0012	0.5324

「差分」に対する p 値から，（喫煙習慣×年齢^2）の追加は，統計的に支持されない．最終的なモデルとして

切片＋喫煙習慣＋年齢＋年齢^2＋(喫煙習慣×年齢)

が，ポアソン回帰モデルとして有効である．ただし，前にも述べたように，喫煙習慣の統計的な検討には難点がある．

推定されたパラメータを用いて，喫煙習慣と年齢の交互作用を加えた推定曲線を図 3.6 および図 3.7 に示す．図 3.6 は，統計的には望ましいのであるが，曲線を 80 歳以上に外挿したときに動きが大きいことに難点がある．

図 3.7 は，喫煙習慣と年齢の 2 乗の交互作用迄を入れたモデルで，喫煙習慣について別々に 2 次曲線をあてはめた場合に相当する．2 次曲線のピークが，80 歳から 90 歳の間に含まれ

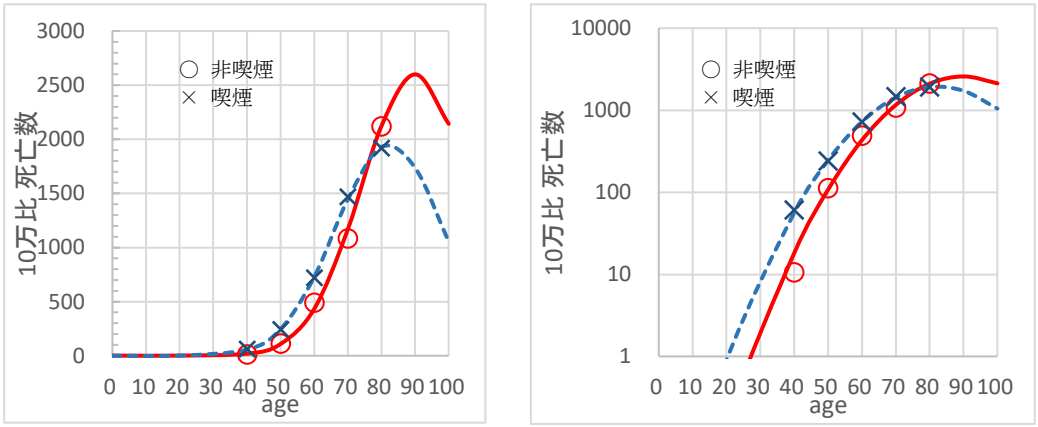


図 3.6　切片＋喫煙習慣＋年齢＋年齢^2＋(喫煙習慣×年齢)

ベストモデル

ること，それらのピークがほぼ重なり合うことから，70 歳代までは喫煙群は非喫煙群に対して死亡数が多いが，80 歳以後には死亡数の差がほとんどなくなることを意味している．これは，喫煙者でも 80 歳まで生きていたとの条件下では，それ以後に死亡確率は非喫煙者と同等であることを意味している．

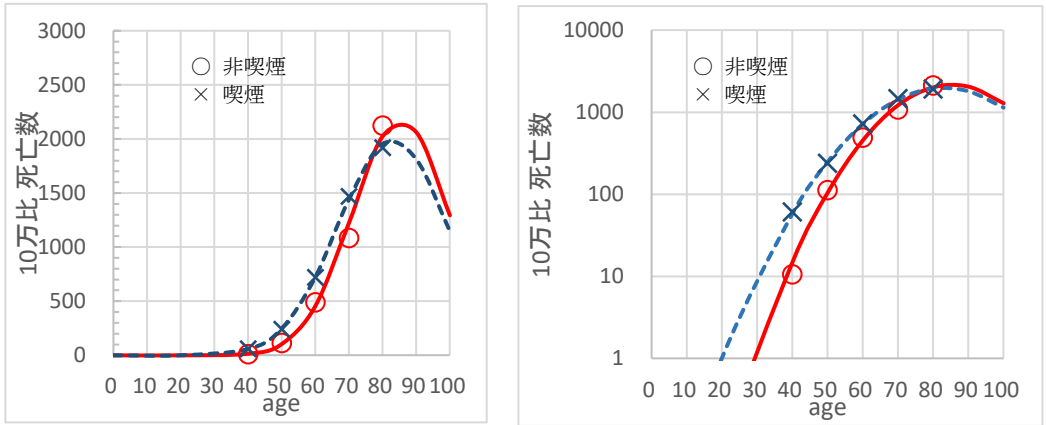


図 3.7　切片＋喫煙習慣＋年齢＋年齢^2＋(喫煙習慣×年齢)＋(喫煙習慣×年齢^2)

パラメータ過剰モデル

なお，第 12.6 節では，パラメータの共分散行列を用いて 95%信頼区間の計算およびグラフ表示を詳細に示す．

オフセットがある場合のデータで，ある疾患の死亡数を扱う場合に，年齢と共に死亡数が頭打ちになる場合に，ポアソン回帰によるモデルの探索には，本節で示したように冗長となりがちである．第 5.6 節に示すように，2 項分布を仮定したロジット解析などで，図 5.4 に示すような上限を持つシグモイド曲線のあてはめが，現実的な対応と思われる．

4. デザイン行列を用いた回帰分析入門

ポアソン分布に従うカウント・データの解析を統計ソフトで行う場合に，2 群比較でも多群比較でも，2 本の回帰直線の場合でも，何らかのデザイン行列を設定し，ポアソン回帰で解析する方法を第 3 章で例示してきた．そこで，Excel の行列関数の使い方の入門を兼ね，デザイン行列を活用した回帰分析に必要な行列計算の方法を伝統的なシグマを用いた計算方法と対比して示す．さらに，デザイン行列を用いた回帰分析から導出されるパラメータの共分散行列を活用した回帰直線の 95%信頼区間，個別データの 95%信頼区間の導出方法を示す．逆推定値に対する 95%信頼区間の推定についても伝統的な偏差平方和による計算方法に代え，パラメータの共分散行列を活用したデルタ法による近似 95%信頼区間の導出法，正確な 95%信頼区間および個別データに対応するの正確な 95%信頼区間の導出法について示す．

4.1. Excel によるデザイン行列を用いた行列計算

多くの統計解析の教科書で，説明変数を X，反応変数を Y とする回帰分析が取り上げられている．そのほとんどが，シグマを用いた偏差平方和をベースにした計算法で説明されている．そして，推定された回帰直線の 95%信頼区間は，これこれの式で与えられるとの記述に遭遇する．ポアソン回帰の場合は，どうしたらよいのだろうか．どちらの場合でも共通な方法は，あるのだろうか．幸い，デザイン行列をベースにした計算方法は，各種の回帰直線の 95%信頼区間の計算方法の基本は同じであることは，これまでも例示してきた．

ドレーパ・スミス著，中村訳（1968）「応用回帰分析」の第 1 章には，シグマを使った回帰分析，第 2 章には，同じデータについてデザイン行列を用いた計算法が丁寧に例示され，重回帰分析へ橋渡しがなされている．さらに，Draper and Smith（1998），Applied Regrettion Anarysis 3rd ed. も参考にし，シグマを使った計算との関係を示すことにより，シグマを用いた計算になれ親しんだ人達を念頭に，デザイン行列を用いた解析の利便性について示す．用いるデータは，第 1.4 節のポアソン回帰の導入で用いた人工データとし，Excel の行列関数による計算を主体にする．

シグマが出てきただけで読み飛ばしたくなるような人たちの気持ちは，その式を見ても実際のデータで自ら計算を行う意欲がわかないからだと推測する．ましてや，行列の計算式が

出てきたときに，実際のデータで計算することができないものに対し，拒絶反応が起きるのは当然のことであろう．そこで，統計ソフトの使い方，結果の見方さえ習得すれば十分だと思っている人達を念頭にして，Excel の基本的な計算機能を用い，行列計算による回帰分析を丁寧に解説する．

デザイン行列を用いた回帰式の表記

反応変数 Y_i の列ベクトルを $\boldsymbol{Y}$ とし，説明変数のデザイン行列を $\boldsymbol{X}$，推定したいパラメータの列ベクトルを $\boldsymbol{\beta}$，誤差の列ベクトルを $\boldsymbol{\varepsilon}$ とする．データは，表 1.6 の人工データである［ドブソン（2008)］．まず，Excel シートのある場所に，これと同じものを再現してもらいたい．ここに示したのは，Excel シート上の該当部分をコピーして，ワード上に「図（拡張メタファイル)」によりペーストしたものである．

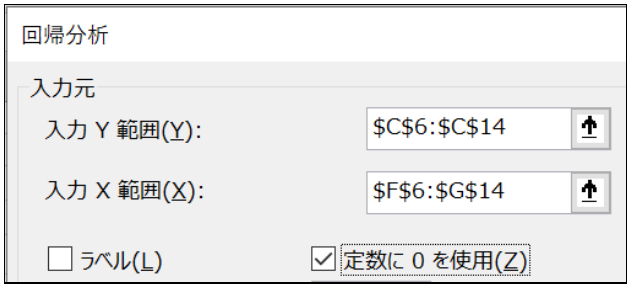
				$\boldsymbol{X}_0$	$\boldsymbol{X}_1$						
$\boldsymbol{Y}$ =	2		$\boldsymbol{X}$ =	1	-1		$\boldsymbol{\beta}$ =	β_0		$\boldsymbol{\varepsilon}$=	ε_1
	3			1	-1			β_1			ε_2
	6			1	0						ε_3
	7			1	0						ε_4
	8			1	0						ε_5
	9			1	0						ε_6
	10			1	1						ε_7
	12			1	1						ε_8
	15			1	1						ε_9
	9×1			9×2				2×1			9×1

フォントのサイズは 10 ポイント，半角英数字のフォントは Times New Roman とする．ギリシャ文字 β は「ベータ」と入力すると変換され，フォントを Times New Roman に変更してイタリックにすると「β」となる．「いぷしろん」で「ε」を生成し，「ε1」の「1」を「セル書式の設定」で「下付き」にする．このように統計で使うギリシャ文字を Excel で扱えるようになることから初めてもらいたい．「デザイン行列 $\boldsymbol{X}$ 」の「 $\boldsymbol{X}$ 」は，「X」を太文字（B）とし，さらにイタリック（ I ）としたものである．

デザイン行列 $\boldsymbol{X}$ の一般的な表記は，角括弧［・・・］あるいは括弧（・・・）で挟むのであるが，Excel シート上では表記しづらいので，太い外枠 | ・・・ | で矩形データ全体を囲む表記を用いる．デザイン行列 $\boldsymbol{X}$ は，表 1.6 の列方向の人工のデータを表 1.8 の行方向の形式で示したもので，大きさが 9 行×2 列で，9 行×1 の列ベクトル $\boldsymbol{X}_0$ と $\boldsymbol{X}_1$ を並べたものである．ベクトル $\boldsymbol{X}_0$ の値を全て 1 としているのは，回帰直線の Y 切片を推定するためである．列ベクトル $\boldsymbol{\beta}$ は，推定したいパラメータとしての切片「β_0」と傾き「β_1」である．

多くの回帰分析の統計ソフトでは，説明変数 $\boldsymbol{X}_1$ のみを与え，切片を推定するための変数 $\boldsymbol{X}_0$ の設定を必要としない．原点を通る回帰直線を求めたい場合には，切片を含まない 9 行×1 列の列ベクトル $\boldsymbol{X}_1$ のみのデザイン行列であり，統計ソフトでは，「切片を含めない」などのオプションを設定する．Excel のアドインで提供されている「データ分析ツール」の「回帰分析」では，次に示すように，

Excel
➔ データ分析
➔ 回帰分析



「定数に 0 を使用」オプションを「オン」とする．デザイン行列は，何を推定したいかにより，変幻自在であることは，第 3 章で詳細に示した．通常の回帰分析では，「定数に 0 を使用」オプションを「オフ」とし，列ベクトル $\boldsymbol{X}_1$ の範囲のみを選択する．さらに，列ベクトル $\boldsymbol{Y}$ の範囲を選択して，回帰分析を行なう．このように，Excel でも他の統計ソフトでも回帰分析では，切片を含めないことが暗黙の前提である．そのため，原点を通る回帰直線のあてはめのために「切片を含めない」とのオプションが必要となる．

列ベクトル $\boldsymbol{Y}$，デザイン行列 $\boldsymbol{X}$，列ベクトル $\boldsymbol{\beta}$，列ベクトル $\boldsymbol{\varepsilon}$ を用いた式

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

は，何を意味するのだろうか．行列 $\boldsymbol{X}$ と列ベクトル $\boldsymbol{\beta}$ の積 $\boldsymbol{X\beta}$ は，デザイン行列 $\boldsymbol{X}$ のある行と，列ベクトル $\boldsymbol{\beta}$ のセル同士を順番に掛けて足した「積和」の計算を意味する．これをデザイン行列 $\boldsymbol{X}$ の 1 行目から 9 行目まで繰り返してして，大きさが 9×1 の新たな列ベクトル $[\boldsymbol{X\beta}]$ が生成される．右辺の $\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ は，大きさが 9×1 の列ベクトル同士の足し算で，同じ行同士の足し算となり，大きさが 9×1 の列ベクトル $[\boldsymbol{X\beta}+\boldsymbol{\varepsilon}]$ となる．

	$[\boldsymbol{X\beta}]$		$\boldsymbol{\varepsilon}$		$[\boldsymbol{X\beta}+\boldsymbol{\varepsilon}]$
$\boldsymbol{X\beta}+\boldsymbol{\varepsilon}=$	$1\beta_0-1\beta_1$	+	ε_1	=	$1\beta_0-1\beta_1+\varepsilon_1$
	$1\beta_0-1\beta_1$		ε_2		$1\beta_0-1\beta_1+\varepsilon_2$
	$1\beta_0+0\beta_1$		ε_3		$1\beta_0+0\beta_1+\varepsilon_3$
	$1\beta_0+0\beta_1$		ε_4		$1\beta_0+0\beta_1+\varepsilon_4$
	$1\beta_0+0\beta_1$		ε_5		$1\beta_0+0\beta_1+\varepsilon_5$
	$1\beta_0+0\beta_1$		ε_6		$1\beta_0+0\beta_1+\varepsilon_6$
	$1\beta_0+1\beta_1$		ε_7		$1\beta_0+1\beta_1+\varepsilon_7$
	$1\beta_0+1\beta_1$		ε_8		$1\beta_0+1\beta_1+\varepsilon_8$
	$1\beta_0+1\beta_1$		ε_9		$1\beta_0+1\beta_1+\varepsilon_9$

列ベクトル $\boldsymbol{Y}$ と $\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ を等号で結んだ式

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$$

は，$\boldsymbol{Y}$ の行方向に展開して

$$\begin{aligned} Y_1 &= 1\beta_0 - 1\beta_1 + \varepsilon_1 \\ Y_2 &= 1\beta_0 - 1\beta_1 + \varepsilon_2 \\ &\vdots \\ Y_9 &= 1\beta_0 + 1\beta_1 + \varepsilon_9 \end{aligned} \qquad (4.2)$$

となる．また，添え字を使った式

$$Y_i = \beta_0 X_{0,i} + \beta_1 X_{1,i} + \varepsilon_i \quad (i = 1,\ 2,\ldots,\ 9) \qquad (4.3)$$

としても，同じ回帰モデルである．デザイン行列を使った表記は，最も簡潔な表記となっている．その意味することを Excel で表記した矩形データを連想することにより，行列計算が電卓のように身近なものとなることを期待する．

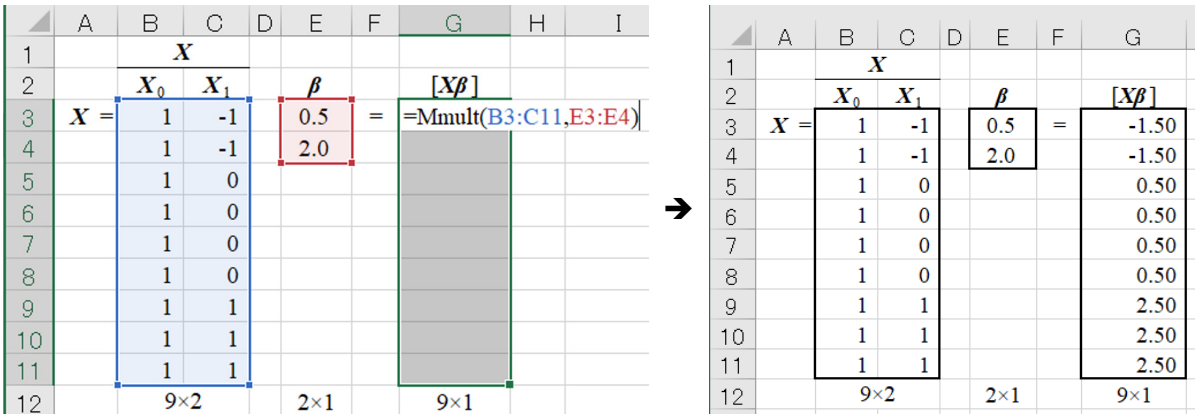
行列計算の実際

Excel シート上の列ベクトル $\boldsymbol{\beta}$ に適当な数値を入れて，実際に $\boldsymbol{X\beta}$ の計算にチャレンジしてみよう．手順は，以下に示すように，

1）[$\boldsymbol{X\beta}$] の計算結果となる 9×1 の矩形（G3:G11）を枠線で囲む．

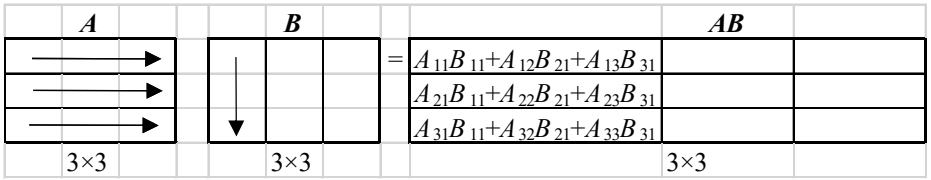
2）その 9×1 の矩形を選択し，最初の行に行列の積の関数式「=Mmult($\boldsymbol{X}$ の範囲を選択，$\boldsymbol{\beta}$ の範囲の選択)」を入力する．$\boldsymbol{X}$ の範囲には（B3:C11），$\boldsymbol{\beta}$ の範囲には（E3:E4）が自動的に設定される．

3）「コントロールキー」と「シフトキー」同時に押しながら「エンター」すると行列の積の計算が行なわれる．



デザイン行列の転置

行列の掛け算は，左側の行列 $\boldsymbol{A}$ の「**行**」と右側の行列 $\boldsymbol{B}$ の「**列**」を順番に掛けて加えた和(積和)の行列として定義されている．行列の積和の計算は，“行”方向と“列”方向であって，“列”方向と“行”の“**列・行**”ではなく，あくまで“**行・列**”の順番である．



(9 行 2 列)の $\boldsymbol{X}$ と（9 行 2 列)の $\boldsymbol{X}$ のままだと積 $\boldsymbol{XX}$ は，（2 列 vs. 9 行）と列と行の数が異なり積和の計算ができない．そこで，最初の $\boldsymbol{X}$ について行と列を入れ替え，(2 行 9 列)の $\boldsymbol{X}^T$ (転置行列)とする．行列の積 $\boldsymbol{X}^T\boldsymbol{X}$ は，内側の数が（9 列 vs. 9 行）と一致し積和の計算ができる．このように，行列の積が成り立つのは，$\boldsymbol{X}^T$ (2 行 9 列）と $\boldsymbol{X}$ (9 行 2 列）のように隣り合う行列の内側が，9 列と 9 行のように大きさが完全に一致する必要がある．行列の積の結果は，外側の（2 行 2 列）の大きさとなる．

デザイン行列の積和

行と列の入れ替えは，転置（Transpose）といい，$\boldsymbol{X}^T$ のように表記する．なお，$\boldsymbol{X}'$ あるいは ${}^t\boldsymbol{X}$ と表記する場合もあり，様々である．行列の積は，Excel の Mmult()関数を使い

$\boldsymbol{X}_0{}^T$ 行と $\boldsymbol{X}_0$ 列の 積和＝9 を $(\boldsymbol{X}^T\boldsymbol{X})$ の 1 行 1 列目へ
$\boldsymbol{X}_1{}^T$ 行と $\boldsymbol{X}_0$ 列の 積和＝1 を $(\boldsymbol{X}^T\boldsymbol{X})$ の 2 行 1 列目へ
$\boldsymbol{X}_0{}^T$ 行と $\boldsymbol{X}_1$ 列の 積和＝1 を $(\boldsymbol{X}^T\boldsymbol{X})$ の 1 行 2 列目へ
$\boldsymbol{X}_1{}^T$ 行と $\boldsymbol{X}_1$ 列の 積和＝5 を $(\boldsymbol{X}^T\boldsymbol{X})$ の 2 行 2 列目へ

	$\boldsymbol{X}^T$									$\boldsymbol{X}$			$\boldsymbol{X}^T\boldsymbol{X}$	
										$\boldsymbol{X}_0$	$\boldsymbol{X}_1$			
$\boldsymbol{X}_0{}^T$	1	1	1	1	1	1	1	1	1	1	-1	=	9	1
$\boldsymbol{X}_1{}^T$	-1	-1	0	0	0	0	1	1	1	1	-1		1	5
										1	0			
	$\boldsymbol{X}^T$:		=Transpose($\boldsymbol{X}$の範囲)							1	0		=Mmult($\boldsymbol{X}^T$の範囲, $\boldsymbol{X}$の範囲)	
			Excelの関数							1	0		Excelの関数	
										1	0			
										1	1			
										1	1			
										1	1			
				2×9						9×2			2×2	

として，(2×9) の $\boldsymbol{X}^T$ と（9×2）の $\boldsymbol{X}$ の行列の積から，(2×2) の $(\boldsymbol{X}^T\boldsymbol{X})$ 行列が生成される．なお，$\boldsymbol{X}^T\boldsymbol{X}$ ではなく $\boldsymbol{X}\boldsymbol{X}^T$ とすると，(9×2) の $\boldsymbol{X}$ と（2×9）の $\boldsymbol{X}^T$ から，(9×9) の $(\boldsymbol{X}\boldsymbol{X}^T)$ 行

列となる．このように行列の計算に際しては，行列のサイズを欄外に示しておき，内側同志のサイズが一致していることを確認することにより計算ミスが少なくなる．

シグマ流の積和の計算

行列を用いた表記は，慣れれば簡潔で見通しがいいのだが，行列計算の方法を連想しやすくなるように，シグマを用いた計算方法を示す．Excel の SumProduct()関数は，複数の配列の要素ごとの積和の計算をする便利な関数であり，次に示すように行列 $\boldsymbol{X}$ の 1 列目を $\boldsymbol{X}_0$ とし，2 列目を $\boldsymbol{X}_1$ としたときに，2×2 の積和行列の 1 列 2 行目は，

=SumProduct($\boldsymbol{X}_0$ の範囲, $\boldsymbol{X}_1$ の範囲)

として計算できる．同じ列同士の場合は，SumSq($\boldsymbol{X}_0$ の範囲) を使うこともできる．実際にデザイン行列 $\boldsymbol{X}$ の積和行列 $(\boldsymbol{X}^T\boldsymbol{X})$ は，次に示すように求めることもできるが，煩雑であり利便性に欠けるのでまったく推奨できない．

$\boldsymbol{X}_0$	$\boldsymbol{X}_1$		積和行列		（$\boldsymbol{X}^T\boldsymbol{X}$）
1	-1	＝	9	1	
1	-1		1	5	
1	0				
1	0		=SumSq($\boldsymbol{X}_0$の範囲) = 9		
1	0			=SumProduct($\boldsymbol{X}_0$の範囲, $\boldsymbol{X}_1$の範囲) = 1	
1	0		=SumProduct($\boldsymbol{X}_1$の範囲, $\boldsymbol{X}_0$の範囲) = 1		
1	1			=SumSq($\boldsymbol{X}_1$の範囲) = 5	
1	1				
1	1				
9×2					

行列の積

デザイン行列の転置は，Transpose()関数を使い，デザイン行列の掛け算は，Mmult()関数を使い $\boldsymbol{X}^T\boldsymbol{X}$ は，

$\boldsymbol{X}^T\boldsymbol{X}$ ：= Mmult(Transpose($\boldsymbol{X}$ の範囲)，$\boldsymbol{X}$ の範囲)

のように Transpose()関数を Mmult()関数の入れ子にして計算することができる．

転置したデザイン行列 $\boldsymbol{X}^T$ と $\boldsymbol{X}$ の積をシグマ記号Σで表せば，行ベクトル $\boldsymbol{X}_0{}^T$ と列ベクトル $\boldsymbol{X}_0$ の積和は n，行ベクトル $\boldsymbol{X}_0{}^T$ と列ベクトル $\boldsymbol{X}_1$ の積和は $\Sigma_i X_i$，行ベクトル $\boldsymbol{X}_1{}^T$ と列ベクトル $\boldsymbol{X}_1{}^T$ の積和は $\Sigma_i X_i^2$ となり，(2 行 2 列)の行列としてまとめて現したことになる．ここでは，$X_{1,i}$ とすべきところを略して X_i としている．$\boldsymbol{X}_0{}^T$ と $\boldsymbol{X}_0$ の積和は，行列 $\boldsymbol{X}$ の行の数 n となり，$(\boldsymbol{X}^T\boldsymbol{X})$ の 1 行 1 列目となっている．

	X^T									X			X^TX	
										X_0	X_1			
X_0^T	1	1	1	1	1	1	1	1	1	1	X_1	=	n	ΣX_i
X_1^T	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	1	X_2		ΣX_i	ΣX_i^2
										1	X_3			
										1	X_4			
										1	X_5			
										1	X_6			
										1	X_7			
										1	X_8			
										1	X_9			
					2×9					9×2			2×2	

デザイン行列 X と反応 Y との積

次に，反応変数の列ベクトル $\boldsymbol{Y}$ とデザイン行列 $\boldsymbol{X}$ の積和を示す．$\boldsymbol{X}^T\boldsymbol{Y}$ は，転置されたデザイン行列 $\boldsymbol{X}$ と列ベクトル $\boldsymbol{Y}$ の積で，$\boldsymbol{X}_0^T$ と $\boldsymbol{Y}$ についての積和は

$$1\times 2+1\times 3+\cdots+1\times 15=72$$

となり，$\boldsymbol{X}_1^T$ と $\boldsymbol{Y}$ についての積和は

$$(-1)\times 2+(-1)\times 3+\cdots+1\times 15=32$$

となる．しつこいことは承知しているが，実際に手を動かして行列計算を体験することが，最初の一歩である．

				X^T						Y		X^TY	
X_0^T	1	1	1	1	1	1	1	1	1	2	=	72	$\Sigma_i Y_i$
X_1^T	-1	-1	0	0	0	0	1	1	1	3		32	$\Sigma_i X_i Y_i$
										6			
		X^T	:	=Transpose(**X**の範囲)						7		=Mmult(X^Tの範囲, **Y**の範囲)	
				Excelの関数						8		Excelの関数	
										9			
										10			
										12			
										15			
					2×9					9×1		2×1	

シグマ記号で示せば，転置したデザイン行列 $\boldsymbol{X}^T$ と $\boldsymbol{Y}$ の積は，$\boldsymbol{X}_0^T$ と $\boldsymbol{Y}$ の積和で $\Sigma_i(X_{0,i}Y_i)$，$\boldsymbol{X}_1^T$ と $\boldsymbol{Y}$ の積和で $\Sigma_i(X_{1,i}Y_i)$ となる．

4.2. 偏差平方和ベースの回帰パラメータ推定

回帰式のパラメータ推定

求めたい回帰式を，

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i , \quad i = 1,2,...,n \tag{4.4}$$

としたときに，真の直線からの偏差ε_iの平方和は，

$$S_e = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{4.5}$$

である．通常の回帰分析においてε_iは，平均が 0，分散がσ^2の正規分布に従い，互いに独立であると仮定するのだが，実際にはどんな分布であっても，互いに独立でなくとも，最小 2 乗法での計算は可能である．このことが，最小 2 乗法による回帰分析が，ゆうずうむげに無批判的に重用されている理由である．しかし，基本中の基本であるので，丁寧に説明する．偏差平方和S_eをパラメータβ_0とβ_1で偏微分すると，式（4.6）が得られる．

$$\left.\begin{aligned} \frac{\partial S_e}{\partial \beta_0} &= -2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right) \\ \frac{\partial S_e}{\partial \beta_1} &= -2\sum_{i=1}^{n} \left[\left(Y_i - \beta_0 - \beta_1 X_i\right) X_i\right] \end{aligned}\right\} \tag{4.6}$$

偏微分に不慣れな場合には，段階的な学習が必要である．それぞれの$\varepsilon_i^{\,2}$は，

$$\begin{aligned} \varepsilon_i^{\,2} &= (Y_i - \beta_0 - \beta_1 X_i)^2 \\ &= Y_i^2 + \beta_0^{\,2} + \beta_1^{\,2} X_i^{\,2} - 2Y_i\beta_0 - 2\beta_1 X_i Y_i + 2\beta_0\beta_1 X_i \end{aligned}$$

と式を展開できるので，β_0での偏微分は，他の変数を定数と見なした微分なので，

$$\begin{aligned} \frac{\partial \varepsilon_i^{\,2}}{\partial \beta_0} &= \frac{\partial (Y_i^2 + \beta_0^{\,2} + \beta_1^{\,2} X_i^{\,2} - 2Y_i\beta_0 - 2\beta_1 X_i Y_i + 2\beta_0\beta_1 X_i)}{\partial \beta_0} \\ &= \quad 0 + 2\beta_0 + \quad 0 \quad - 2Y_i \quad - \quad 0 \quad + 2\beta_1 X_i \\ &= 2(\beta_0 - Y_i + \beta_1 X_i) \\ &= -2(Y_i - \beta_0 - \beta_1 X_i) \end{aligned}$$

のように要素に分解し，シグマで再統合すれば式（4.6）の 1 行目となる．β_1での偏微分も同様である．第 2.3 節でも示したが，偏微分を含む数学の学習をサポートするソフトを活用しつつ，合成関数の偏微分，さらに，一般化線形モデルでの対数尤度の偏微分に慣れ親しんでもらいたい．

式（4.6）を 0 と置くと，β_0 と β_1 の推定値としての $\hat{\beta}_0$ と $\hat{\beta}_1$ を求めることができる．

$$\left.\begin{aligned} -2\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)=0 \\ -2\sum_{i=1}^{n}\left[\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)X_i\right]=0 \end{aligned}\right\} \tag{4.7}$$

正規方程式

式（4.7）を $\hat{\beta}_0$ と $\hat{\beta}_1$ について解くために，両辺を –2 で割り式を展開すると，

$$\left.\begin{aligned} \sum_{i=1}^{n}Y_i-n\hat{\beta}_0-\hat{\beta}_1\sum_{i=1}^{n}X_i=0 \\ \sum_{i=1}^{n}X_iY_i-\hat{\beta}_0\sum_{i=1}^{n}X_i-\hat{\beta}_1\sum_{i=1}^{n}X_i^2=0 \end{aligned}\right\} \tag{4.8}$$

を得る．推定値 $\hat{\beta}_0$ と $\hat{\beta}_1$ が，含まれない項を右辺に移して整理すると，

$$\left.\begin{aligned} &(1)\quad n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}X_i=\sum_{i=1}^{n}Y_i \\ &(2)\quad \hat{\beta}_0\sum_{i=1}^{n}X_i+\hat{\beta}_1\sum_{i=1}^{n}X_i^2=\sum_{i=1}^{n}X_iY_i \end{aligned}\right\} \tag{4.9}$$

が得られる．これらの方程式は，正規方程式と呼ばれている．式（4.9）の左辺は，$\boldsymbol{X}^T\boldsymbol{X}$ と $\hat{\boldsymbol{\beta}}$ の積に等しく，右辺は $\boldsymbol{X}^T\boldsymbol{Y}$ と等しいので，

$\boldsymbol{X}^T\boldsymbol{X}$			$\boldsymbol{\beta}$^		$\boldsymbol{X}^T\boldsymbol{Y}$
n	ΣX_i		β_0^	=	ΣY_i
ΣX_i	$\Sigma X_i^{\,2}$		β_1^		ΣX_iY_i
2×2			2×1		2×1

が成り立ち，

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}}=\boldsymbol{X}^T\boldsymbol{Y} \tag{4.10}$$

が得られる．

正規方程式の解

正規方程式（4.9）を $\hat{\beta}_0$ と $\hat{\beta}_1$ について解くために，（1）の両辺に $1/n$ を掛け，さらに ΣX_i を掛けると次式を得る．（2）は Σ の範囲の添え字 i の表示を外してある．以後，Σ 記号の添え字 i は省略する．

$$\left.\begin{aligned} &(1)\quad \hat{\beta}_0\Sigma_iX_i+\hat{\beta}_1\frac{\left(\Sigma_iX_i\right)^2}{n}=\frac{(\Sigma_iX_i)(\Sigma_iY_i)}{n} \\ &(2)\quad \hat{\beta}_0\Sigma X_i+\hat{\beta}_1\Sigma X_i^2=\Sigma X_iY_i \end{aligned}\right\} \tag{4.11}$$

式（4.11）の（2）式からの（1）式を引いて，$\hat{\beta}_1$ について解くと

$$\left.\begin{aligned}\hat{\beta}_1 &= \frac{\Sigma X_i Y_i - \dfrac{(\Sigma X_i)(\Sigma Y_i)}{n}}{\Sigma X_i^2 - \dfrac{(\Sigma X_i)^2}{n}} \\ &= \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2} \\ &= \frac{S_{XY}}{S_{XX}}\end{aligned}\right\} \tag{4.12}$$

のように，多くのテキストで示されている結果が得られる．ここで，S_{XX} は，$\boldsymbol{X}_1$ の平均 $\bar{X}$ からの偏差の平方和の略語であり，S_{XY} は，$\boldsymbol{X}_1$ の平均 $\bar{X}$ からの偏差と $\boldsymbol{Y}$ の平均 $\bar{Y}$ からの偏差の積和の略号である．この式の変形は，$\bar{X} = (\Sigma X_i)/n$，$\bar{Y} = (\Sigma Y_i)/n$ などの関係を用い，式（4.12）の 2 行目の分子 $\Sigma (X_i - \bar{X})(Y_i - \bar{Y})$ について

$$\left.\begin{aligned}\Sigma (X_i - \bar{X})(Y_i - \bar{Y}) &= \Sigma X_i Y_i - \bar{X}\Sigma Y_i - \bar{Y}\Sigma X_i + n\bar{X}\bar{Y} \\ &= \Sigma X_i Y - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\ &= \Sigma X_i Y_i - n\bar{X}\bar{Y} \\ &= \Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{n}\end{aligned}\right\} \tag{4.13}$$

と，展開し，整理すると式（4.12）1 行目の分子となることを利用している．式（4.12）の 2 行目の分母 $\Sigma (X_i - \bar{X})^2$ については，式（4.13）の Y を X に置き換えることにより式（4.12）1 行目の分母となる．

$$\left.\begin{aligned}\Sigma (X_i - \bar{X})^2 &= \Sigma X_i^2 - 2\bar{X}\Sigma X_i + n\bar{X}^2 \\ &= \Sigma X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \Sigma X_i^2 - n\bar{X}^2 \\ &= \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\end{aligned}\right\} \tag{4.14}$$

このように，平均値を差し引いた平方和の計算は，手計算あるいは電卓の時代では，端数が出るので計算がめんどうであった．そのため，元のデータ X_i の 2 乗和の計算で済ませられ，計算量も減らすこともできるので，標準的な計算法として普及し，多くの統計の統計の教科書に引き継がれている．しかし，Excel を用いた場合には，平均値を差し引いた偏差平方和の計算の方が簡潔で扱いやすい．

正規方程式（4.9）の 1 行目は，

$$n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_i = \Sigma Y_i \tag{4.15}$$

なので，$\hat{\beta}_0$ について解くと，$\hat{\beta}_1$ を含む次式が得られる．

$$\begin{aligned}\hat{\beta}_0 &= \frac{\Sigma Y_i - \hat{\beta}_1 \Sigma X_i}{n} \\ &= \overline{Y} - \hat{\beta}_1 \overline{X}\end{aligned} \tag{4.16}$$

まとめると，回帰式の傾き $\hat{\beta}_1$ と切片 $\hat{\beta}_0$ の計算は，

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{4.17}$$

と簡潔な式となる．回帰パラメータの導出だけであれば，この簡便な式によってパラメータ $\hat{\beta}_1$ および $\hat{\beta}_0$ を計算することができ，表 4.1 に示すように Excel を用いて実に簡便にパラメータ $\hat{\beta}_1$ および $\hat{\beta}_0$ を推定することができる．ただし，説明変数が増えて 2 変数以上となった場合にも拡張は可能であるが，第 12.3 節に示すように煩雑な式の展開が求められ，3 変数以上に対しては，さらに煩雑になり，デザイン行列を用いた定式化の方が簡潔である．

偏差平方を用いたパラメータの推定の実際

表 4.1 に示す Excel による計算シートの結果を用いて，回帰パラメータは，

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{24.0000}{4.8889} = 4.9091$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 8.0000 - 4.9091 \times 0.1111 = 7.4545$$

として求められる．

表 4.1　回帰パラメータの Excel シート上での推定

i	Y	X	Y偏差	X偏差	X偏差^2	XY偏差			
1	2	-1	-6.0000	-1.1111	1.2346	6.6667		β_1^=	4.9091
2	3	-1	-5.0000	-1.1111	1.2346	5.5556		β_0^=	7.4545
3	6	0	-2.0000	-0.1111	0.0123	0.2222			
4	7	0	-1.0000	-0.1111	0.0123	0.1111			
5	8	0	0.0000	-0.1111	0.0123	0.0000			
6	9	0	1.0000	-0.1111	0.0123	-0.1111			
7	10	1	2.0000	0.8889	0.7901	1.7778			
8	12	1	4.0000	0.8889	0.7901	3.5556			
9	15	1	7.0000	0.8889	0.7901	6.2222			
	8.0000	0.1111	0.0000	0.0000	4.8889	24.0000			
	平均	平均	合計	合計	平方和	平方和			
	$\overline{Y}$	$\overline{X}$	$Y_i - \overline{Y}$	$X_i - \overline{X}$	S_{XX}	S_{XY}			

回帰分析は，この様に定式化され，ほとんどの統計の教科書で取り上げられているが，いわゆる有意差検定よりもかなり高級であり，きちっと理解し，さらなる応用のために学習し

たいと思っても難解であり，教科書に示されている計算公式の範囲内に多くの人達が留まざるを得なくなっているのが現状である．伝統的な回帰分析の解法は，更なる回帰分析を活用したいと思う人達の学習意欲をへし折るような，まさにガラスの天井のごとくである．

ガラスの天井を超えるためには，回帰分析を Excel の行列計算で実行できるようになることが最初の一歩である．とは言え，いきなりパラメータが 3 以上の場合に取り組むと，敷居が高すぎて挫折しかねない．段階的な学習としては，式（4.10）で示した正規方程式

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$$

$\boldsymbol{X}^T\boldsymbol{X}$			$\boldsymbol{\beta}$^		$\boldsymbol{X}^T\boldsymbol{Y}$
n	ΣX_i		β_0^	=	ΣY_i
ΣX_i	$\Sigma X_i^{\,2}$		β_1^		$\Sigma X_i Y_i$
2×2			2×1		2×1

に立ち戻り，学習することが望ましいのであるが，推奨できる日本語の成書は，残念ながら絶版となっているドレーパ・スミス（1986）しか見当たらない．なお，Net 書店では中古本が手に入る場合もあるが，第 3 版の原著は，Net 書店で手軽に入手できる．

そこで，「ポアソン回帰」の基礎である通常の「回帰分析」について，従来の偏差平方和を用いた解析，デザイン行列をベースにした行列計算による解析，両者の相互の関連を丁寧に示すことにした．そして，読者が偏差平方和を用いた回帰分析から，デザイン行列をベースにした回帰分析の解析法に親しみを感じてもらいたいと願っている．

4.3. デザイン行列を用いた回帰パラメータの推定

行列計算による回帰パラメータの推定

前節では，偏差平方和に基づく回帰分析のパラメータを推定する方法を示し，式（4.9）で示したシグマによる表記の正規方程式を行列によって

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$$

となることを式（4.10）で示した．

推定値 $\hat{\boldsymbol{\beta}}$ を得るために，行列計算では両辺を $(\boldsymbol{X}^T\boldsymbol{X})$ で割ることができない．そこで，逆行列の定義により $(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{I}$ が単位行列となるので，逆行列 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ を式（4.10）両辺に掛けて，

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\boldsymbol{I}\,\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (4.18)$$

を得る．2×2 の逆行列 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ ならば，$(\boldsymbol{X}^T\boldsymbol{X})$ の行列式 D を計算し，次のようにして計算することができる．実際に手軽に手計算できるのは，2×2 の場合までで，3×3 以上の場合には，Excel の行列式 Mdetarm()関数を用いても煩雑であり勧められない．

$\boldsymbol{X}^T\boldsymbol{X}$				$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$				
n	ΣX_i	-1	=	$\Sigma X_i{}^2/D$	$-\Sigma X_i/D$			
ΣX_i	$\Sigma X_i{}^2$			$-\Sigma X_i/D$	n/D			
				$D=(n\Sigma X_i{}^2)-(\Sigma X_i)^2$				
$\boldsymbol{X}^T\boldsymbol{X}$				$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$			$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	
9	1	-1	=	$5/D$	$-1/D$	=	0.1136	-0.0227
1	5			$-1/D$	$9/D$		-0.0227	0.2045
				$D=45-1=44$				
				=Mdeterm($\boldsymbol{X}^T\boldsymbol{X}$の範囲)				

実際に逆行列を前掛けすると

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$		$\boldsymbol{X}^T\boldsymbol{X}$			$(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})$	
0.1136	-0.0227	9	1	=	1.0000	0.0000
-0.0227	0.2045	1	5		0.0000	1.0000

のように単位行列になることが確認され，単位行列と $\hat{\boldsymbol{\beta}}$ の積は，

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})$		$\boldsymbol{\beta}$^		$\boldsymbol{\beta}$^
1.0000	0.0000	β^$_0$	=	β^$_0$
0.0000	1.0000	β^$_1$		β^$_1$

と元の $\hat{\boldsymbol{\beta}}$ となる．

逆行列は，Excel の Minverse()関数によって求めることができる．推定値 $\hat{\boldsymbol{\beta}}$ を求めるために必要な，$\boldsymbol{X}^T\boldsymbol{Y}$ は，第 4.1 節の「デザイン行列 $\boldsymbol{X}$ と反応 $\boldsymbol{Y}$ との積」の項の結果を用い，これらの計算結果を組み合わせて，次のような手順で推定値 $\hat{\boldsymbol{\beta}}$ が求めることができる．

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} =$$

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$			$\boldsymbol{X}^T\boldsymbol{Y}$		$\boldsymbol{\beta}$^
0.1136	-0.0227		72	=	7.4545
-0.0227	0.2045		32		4.9091
=Minverse($\boldsymbol{X}^T\boldsymbol{X}$の範囲)					= Mmult (($\boldsymbol{X}^T\boldsymbol{X}$)$^{-1}$の範囲, $\boldsymbol{X}^T\boldsymbol{Y}$の範囲)
			= Mmult (Transpose($\boldsymbol{X}$の範囲), $\boldsymbol{Y}$の範囲)		

くどいようだが，自らの手で計算しない限り，ガラスの天井を超えることはできない．

デザイン行列と偏差平方和での推定式の相違

行列計算によって推定値 $\hat{\boldsymbol{\beta}}$ が得られたのであるが，正規方程式（4.17）から，導出された $\hat{\beta}_0$ と $\hat{\beta}_1$ とは，異なる式となっているので，$(\Sigma X_i) = n\bar{X}$，$(\Sigma Y_i) = n\bar{Y}$ などの関係を用いて行列計算の式を整理すると，一致することが確認される．

$$\left.\begin{aligned}
\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} &= \begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2\end{bmatrix}^{-1}\begin{bmatrix}\Sigma Y_i \\ \Sigma X_i Y_i\end{bmatrix} \\
&= \begin{bmatrix} \dfrac{\Sigma X_i^2}{D} & \dfrac{-\Sigma X_i}{D} \\ \dfrac{-\Sigma X_i}{D} & \dfrac{n}{D}\end{bmatrix}\begin{bmatrix}\Sigma Y_i \\ \Sigma X_i Y_i\end{bmatrix} \\
&= \begin{bmatrix} \dfrac{(\Sigma X_i^2)(\Sigma Y_i) - (\Sigma X_i)(\Sigma X_i Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2} \\ \dfrac{-\Sigma X_i \Sigma Y_i + n\Sigma X_i Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}\end{bmatrix} \\
&= \begin{bmatrix} \dfrac{\bar{Y}(\Sigma X_i^2) - \bar{X}(\Sigma X_i Y_i)}{\Sigma X_i^2 - n\bar{X}^2} \\ \dfrac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2}\end{bmatrix}
\end{aligned}\right\} \qquad (4.19)$$

ここで，式（4.19）の最後の行列の 2 行目は，式（4.17）で導出された $\hat{\beta}_1$ の推定値に一致する．

$$\hat{\beta}_1 = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2} = \frac{S_{XY}}{S_{XX}} \qquad (4.20)$$

さて，式（4.17）で導出された $\hat{\beta}_0$ の推定式とは，式（4.19）の最後の行列の 1 行目は明らかに異なる．そこで，式（4.17）からスタートして，

$$
\left.
\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\
&= \frac{\bar{Y} S_{XX} - \bar{X} S_{XY}}{S_{XX}} \\
&= \frac{\left(\bar{Y} (\Sigma X_i^2) - \bar{Y} \frac{(\Sigma X_i)^2}{n} \right) - \left(\bar{X} (\Sigma X_i Y_i) - \bar{X} \frac{(\Sigma X_i)(\Sigma Y_i)}{n} \right)}{S_{XX}} \\
&= \frac{\left(\bar{Y} (\Sigma X_i^2) - n \bar{Y} \bar{X}^2 \right) - \left(\bar{X} (\Sigma X_i Y_i) - n \bar{Y} \bar{X}^2 \right)}{S_{XX}} \\
&= \frac{\bar{Y} (\Sigma X_i^2) - \bar{X} (\Sigma X_i Y_i)}{\Sigma X_i^2 - n \bar{X}^2}
\end{aligned}
\right\} \qquad (4.21)
$$

となり，式（4.19）の最後の行列の 1 行目が導出される．私にとっても見るのも嫌になる数式の変形であり，シグマによる計算と行列による回帰係数の計算結果が一致することを数式で示すことは，難儀である．実用上は，事例により数値計算の結果が一致するは，これまでの結果で明らかである．

したがって，偏差平方和ベースの回帰分析およびデザイン行列ベースの計算方法を両建てで説明することは冗長であり．気持ちとしては避けたかったのであるが，偏差平方和を用いた回帰分析を「ガラスの天井」のごとくと言い切るため，あえて両者の関係について丁寧に示した．第 12.3 節の「偏差平方和ベースの重回帰分析」と第 12.4 節の「デザイン行列ベースの重回帰分析」で，切片と他の回帰パラメータとの間の共分散が含まれていないことの功罪についてさらに言及している．

4.4. 偏差平方和ベースの回帰パラメータの分散の推定

回帰パラメータ $\hat{\beta}_1$ の分散 $Var(\hat{\beta}_1)$ は，式（4.12）の正規方程式の解を用いて，

$$\left.\begin{aligned} Var(\beta_1) &= Var\left[\frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2}\right] \\ &= Var\left[\frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2}\right] \\ &= \frac{Var(Y_i)}{\Sigma(X_i-\bar{X})^2} \\ &= \frac{\sigma^2}{\Sigma(X_i-\bar{X})^2} \end{aligned}\right\} \tag{4.22}$$

となる．分子の変形は，

$$\begin{aligned} \Sigma(X_i-\bar{X})(Y_i-\bar{Y}) &= \Sigma(X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}) \\ &= \Sigma(X_iY_i - \bar{X}Y_i) - \Sigma X_i\bar{Y} + \Sigma\bar{X}\bar{Y} \\ &= \Sigma(X_iY_i - \bar{X}Y_i) - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\ &= \Sigma(X_i-\bar{X})Y_i \end{aligned}$$

を用いて，$\Sigma(X_i-\bar{X})$ を Y_i に関してコンスタント化するためである．$\hat{\beta}_0$ の分散 $Var(\hat{\beta}_0)$ も，正規方程式の解を用いて，$\bar{Y}$ と $\hat{\beta}_1$ が無相関であり，$\bar{X}$ はコンスタントなので，

$$\left.\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{Y}-\hat{\beta}_1\bar{X}) \\ &= Var(\bar{Y}) + \bar{X}^2 Var(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\bar{X}^2}{\Sigma(X_i-\bar{X})^2} \\ &= \frac{\sigma^2\Sigma X_i^{\,2}}{n\Sigma(X_i-\bar{X})^2} \end{aligned}\right\} \tag{4.23}$$

となる．共分散 $Cov(\hat{\beta}_0, \hat{\beta}_1)$ は，

$$\left.\begin{aligned} Cov(\hat{\beta}_0, \beta_1) &= Cov(\bar{Y}-\hat{\beta}_1\bar{X}, \hat{\beta}_1) \\ &= -\bar{X}\,Var(\hat{\beta}_1) \\ &= \frac{-\bar{X}\sigma^2}{\Sigma(X_i-\bar{X})^2} \end{aligned}\right\} \tag{4.24}$$

となる．

回帰パラメータ $\hat{\boldsymbol{\beta}}$ の共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，これらの計算式から

$$\left.\begin{aligned}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) &= \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix} \\ &= \begin{bmatrix} \dfrac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} & \dfrac{-\bar{X}}{\Sigma(X_i-\bar{X})^2} \\ \dfrac{-\bar{X}}{\Sigma(X_i-\bar{X})^2} & \dfrac{1}{\Sigma(X_i-\bar{X})^2} \end{bmatrix}\sigma^2 \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2 \end{aligned}\right\} \tag{4.25}$$

となる．分散共分散 $Var(\hat{\boldsymbol{\beta}})$ は，デザイン行列 $\boldsymbol{X}^T$ と $\boldsymbol{X}$ の積の逆行列 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ に誤差分散 σ^2 を掛けた結果に一致する．ただし，誤差分散 σ^2 は，未知なので，推定誤差 ε_i の平方和を自由度 $(n-2)$ で割った $\hat{\sigma}^2$ を用いて計算する．

$$\left.\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2} \\ &= \frac{\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{n-2}\end{aligned}\right\} \tag{4.26}$$

これまで，$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ については，次の式を示してきたのであるが，式 (4.25) と異なるので，式の変形を行う．$(\boldsymbol{X}^T\boldsymbol{X})$ の行列式は，

$\boldsymbol{X}^T\boldsymbol{X}$				$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	
n	ΣX_i	-1	=	$\Sigma X_i{}^2/D$	$-\Sigma X_i/D$
ΣX_i	$\Sigma X_i{}^2$			$-\Sigma X_i/D$	n/D
				$D=(n\Sigma X_i^2)-(\Sigma X_i)^2$	

としてきた．ただし，D は，

$$\left.\begin{aligned}D &= n\Sigma X_i^2-(\Sigma X_i)^2 \\ &= n(\Sigma X_i^2-n\bar{X}^2) \\ &= n\Sigma(X_i-\bar{X})^2\end{aligned}\right\} \tag{4.27}$$

と変形できるので，

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \dfrac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} & \dfrac{-\Sigma X_i}{n\Sigma(X_i-\bar{X})^2} \\ \dfrac{-\Sigma X_i}{n\Sigma(X_i-\bar{X})^2} & \dfrac{n}{n\Sigma(X_i-\bar{X})^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} & \dfrac{-\bar{X}}{\Sigma(X_i-\bar{X})^2} \\ \dfrac{-\bar{X}}{\Sigma(X_i-\bar{X})^2} & \dfrac{1}{\Sigma(X_i-\bar{X})^2} \end{bmatrix} \tag{4.28}$$

と一致することが確かめられる．このように，従来のシグマを用いたパラメータの分散の計算式とデザイン行列を用いた計算結果が同じであることが示された．したがって，従来の計算式ではなく，デザイン行列ベースの計算手順で簡単に得れるパラメータの共分散行列を基本とすることは，ガラスの天井を超えるための必須の知識である．

4.5. デザイン行列を用いた回帰分析の実際

パラメータの推定

表 4.2 にこれまで示してきた計算方法を総合し，デザイン行列を用いた回帰分析を例示する．全ての計算過程を 1 枚の Excel シートで示そうとしたために，やや見づらくなっているので，丁寧な解説を付け加える．なお，Excel の分析ツールの回帰分析を活用した後述する表 4.6 と比較してもらいたい．

表 4.2　デザイン行列を用いた回帰分析

i	$\boldsymbol{X}$		$\boldsymbol{Y}$	$\boldsymbol{Y}$^	$\boldsymbol{\varepsilon}=\boldsymbol{Y}-\boldsymbol{Y}$^
1	1	-1	2	2.5455	-0.5455
2	1	-1	3	2.5455	0.4545
3	1	0	6	7.4545	-1.4545
4	1	0	7	7.4545	-0.4545
5	1	0	8	7.4545	0.5455
6	1	0	9	7.4545	1.5455
7	1	1	10	12.3636	-2.3636
8	1	1	12	12.3636	-0.3636
9	1	1	15	12.3636	2.6364
	9.00	1.00	0.1136	-0.0227	72.0000
	1.00	5.00	-0.0227	0.2045	32.0000
	$\boldsymbol{X}^T\boldsymbol{X}$		$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$		$\boldsymbol{X}^T\boldsymbol{Y}$

項	推定値	分散	*SE*	*t* 値	*p* 値
β_0^=	7.4545	0.2952	0.5433	13.72	0.0000
β_1^=	4.9091	0.5313	0.7289	6.73	0.0003
	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$		sqrt(分散)		T.Dist.2T(*t*, 7)

$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$=	18.1818	共分散	0.2952	-0.0590
σ^2=	2.5974	行列	-0.0590	0.5313
			$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma$^2	

デザイン行列 $\boldsymbol{X}$ に対し， $\boldsymbol{X}^T\boldsymbol{X}$ の結果が表の下段に 2×2 の矩形内に Excel の Mmult()関数および Transpose()関数を用いて

$$= \text{Mmult}(\text{Transpose}(\boldsymbol{X}\text{の範囲}), \boldsymbol{X}\text{の範囲})) = \begin{array}{|cc|}\hline 9.00 & 1.00 \\ 1.00 & 5.00 \\ \hline\end{array}$$

と計算されている．その横に $\boldsymbol{X}^T\boldsymbol{X}$ の逆行列 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ が，Minverse()関数を用いて

$$= \text{Minverse}(\boldsymbol{X}^T\boldsymbol{X}\text{の範囲}) = \begin{array}{|cc|}\hline 0.1136 & -0.0227 \\ -0.0227 & 0.2045 \\ \hline\end{array}$$

となり，デザイン行列 $\boldsymbol{X}$ の転置行列列 $\boldsymbol{X}^T$ と列ベクトル $\boldsymbol{Y}$ との積 $\boldsymbol{X}^T\boldsymbol{Y}$ が

$$= \text{Mmult}(\text{Transpose}(\boldsymbol{X}\text{の範囲}), \boldsymbol{Y}\text{の範囲}) = \begin{array}{|c|}\hline 72.0000 \\ 32.0000 \\ \hline\end{array}$$

として計算されている．回帰パラメータの推定値 $\hat{\boldsymbol{\beta}}$ は，表の中段の「推定値」の欄に

$$= \text{Mmult}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\text{の範囲}, \boldsymbol{X}^T\boldsymbol{Y}\text{の範囲}) = \begin{array}{|c|}\hline 7.4545 \\ 4.9091 \\ \hline\end{array}$$

と $\hat{\beta}_0 = 7.4545$ ， $\hat{\beta}_1 = 4.9091$ として計算されている．誤差分散は，回帰の推定値 $\hat{\boldsymbol{Y}}$ を

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \text{Mmult}(\boldsymbol{X}\text{の範囲}, \hat{\boldsymbol{\beta}}\text{の範囲})$$

で求め，誤差ベクトル $\boldsymbol{\varepsilon}$ が列ベクトル $\boldsymbol{Y}$ と推定ベクトル $\hat{\boldsymbol{Y}}$ の差

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{Y}\text{の範囲} - \hat{\boldsymbol{Y}}\text{の範囲})$$

として求められている．誤差平方和 S_e を

$$S_e = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \text{Mmult}(\text{Transpose}(\hat{\boldsymbol{\varepsilon}}\text{の範囲}), \hat{\boldsymbol{\varepsilon}}\text{の範囲})$$

により計算し，データ数 n からパラメータの数 2 を引いた自由度で割った平均平方が誤差分散 $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-2} = \frac{18.1818}{9-2} = 2.5974$$

として計算されている．

デザイン行列を用いた回帰分析の最大の利点は，回帰パラメータ $\hat{\beta}_0$ と $\hat{\beta}_1$ の分散および共分散が 2×2 の行列として得られることである．式（4.25）からパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，

$$\begin{aligned}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) &= \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2\end{aligned} \qquad (4.29)$$

であることを示した．デザイン行列の積和の逆行列 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ および誤差分散 $\hat{\sigma}^2$ は，

0.1136	-0.0227
-0.0227	0.2045

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=$	18.1818
$\sigma\text{^}2=$	2.5974

として計算されているので，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，

$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) =$

0.1136	-0.0227
-0.0227	0.2045

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

2.5974
$\sigma\text{^}2$

$=$

0.2952	-0.0590
-0.0590	0.5313

$\boldsymbol{\Sigma}(\boldsymbol{\beta}\text{^}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma\text{^}2$

と計算される．分散 $Var(\hat{\beta}_0) = 0.2952$ ， $Var(\hat{\beta}_1) = 0.5313$ は， $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$ の対角要素となっているので，次のように推定値の分散が得られる．

項	推定値	分散
$\beta_0\text{^}=$	7.4545	0.2952
$\beta_1\text{^}=$	4.9091	0.5313

更に標準誤差 SE を分散の平方根 sqrt()関数で求め，推定値/SE で t 値を計算し，t 分布の両側確率の p 値を T.dist.2T()関数

$$p\text{ 値} = \text{T.dist.2T}(t\text{値}, (9-2))$$

で計算し，回帰パラメータについての推定および t 検定が次のように行える．

項	推定値	分散	SE	t 値	p 値
β_0^=	7.4545	0.2952	0.5433	13.72	0.0000
β_1^=	4.9091	0.5313	0.7289	6.73	0.0003
$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$			sqrt(分散)		T.Dist.2T(t, 7)

行列計算による回帰パラメータの推定は，表 4.1 に示した偏差平方和をベースにした推定よりも複雑ではあるが，Excel シート 1 枚の中に，回帰分析の基本的な結果が網羅されている．行列計算による回帰パラメータの推定の良さは，複数の変数を対象にした多項式回帰，共分散分析などに拡張した場合でも，同じ計算手順が適用できることにある．

分散分析表

分散分析表が，回帰分析の全体的な評価をするために使われている．分散分析表は，各種の偏差平方和をベースに構成されている．表 4.2 では，反応 Y_i に対して，推定された回帰パラメータ $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ との偏差 $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ を計算し，その平方和を求めた．この偏差平方和を，誤差平方和 S_e といい，推定された回帰直線からのズレ（誤差平方和）の大きさを表している．

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (4.30)$$

反応 Y_i 自体についての統計量として，平均と分散が代表的であるが，表 4.3 に示すように平均 $\overline{Y}$ からの偏差を，$d_i = Y_i - \overline{Y}$ としたときの偏差平方和を 全体の平方和の意味で S_T とする．

$$S_T = \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2 \qquad (4.31)$$

回帰直線の推定値は，$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ であり，平均 $\overline{Y}$ からの偏差 $R_i = \hat{Y}_i - \overline{Y}$ の平方和が回帰の平方和 S_R となる．

$$S_R = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}_i)^2 \qquad (4.32)$$

表 4.3　各種の平方和の計算

i	$\boldsymbol{X}$		Y_i	$\overline{Y}$	$Y_i - \overline{Y}$	Y_i^	Y_i^$-\overline{Y}$	$Y_i - Y$^			
1	1	-1	2	8.00	-6.00	2.55	-5.45	-0.55		β_0^=	7.4545
2	1	-1	3	8.00	-5.00	2.55	-5.45	0.45		β_1^=	4.9091
3	1	0	6	8.00	-2.00	7.45	-0.55	-1.45			
4	1	0	7	8.00	-1.00	7.45	-0.55	-0.45			
5	1	0	8	8.00	0.00	7.45	-0.55	0.55			
6	1	0	9	8.00	1.00	7.45	-0.55	1.55			
7	1	1	10	8.00	2.00	12.36	4.36	-2.36			
8	1	1	12	8.00	4.00	12.36	4.36	-0.36			
9	1	1	15	8.00	7.00	12.36	4.36	2.64			
				8.00	136.00		117.82	18.18			
				$\overline{Y}$	S_T		S_R	S_e			
自由度			9	1	9−1=8	2	2−1=1	9-2=7		2	

このように，元の反応Y_iについての平方和には，

$$\left.\begin{aligned} S_T &= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 \\ &= S_e+S_R \end{aligned}\right\} \qquad (4.33)$$

との関係があり，平方和の分解とも言われており，最小 2 乗法においては重要な公式になっている（本節の末尾に式の証明を示す）．実際に，これらの平方和を計算した結果を表 4.4 に示す．

回帰パラメータの推定には，デザイン行列を使うことを勧めるが，分散分析表の作成には，各種の平方和の定義に基づいた計算方法が，理解しやすい．平方和の計算は，SumSq()関数の使用が効率的である．

$$S_T=\sum_{i=1}^{9}(Y_i-\bar{Y})^2=\mathrm{SumSq}(\boldsymbol{Y}\text{の範囲}-\bar{Y})=136.00\,, \qquad df=9-1=8$$

$$\hat{\boldsymbol{Y}}=\mathrm{Mmult}(\boldsymbol{X}\text{の範囲},\boldsymbol{\beta}\text{の範囲})\,, \qquad df=2$$

$$S_R=\sum_{i=1}^{9}(\hat{Y}_i-\bar{Y})^2=\mathrm{SumSq}(\hat{\boldsymbol{Y}}\text{の範囲}-\bar{Y})=117.82\,, \qquad df=2-1=1$$

$$S_e=\sum_{i=1}^{9}(\hat{Y}_i-Y_i)^2=\mathrm{SumSq}(\hat{\boldsymbol{Y}}\text{の範囲}-\boldsymbol{Y}\text{の範囲})=18.18\,, \qquad df=9-2=7$$

計算結果を，表 4.4 の分散分析表にまとめる．計算原理を習得したの後は，表 4.6 に示すように Excel の分析ツールの回帰分析によっても同じ結果が得られるので，自ら計算することにこだわる理由はない．

表 4.4　回帰に対する分散分析表

要因		平方和	自由度	平均平方	F 値	p 値
回帰	S_R	117.8182	2-1=1	117.8182	45.3600	0.0003
誤差	S_e	18.1818	9-2=7	2.5974		
全体	S_T	136.0000	9-1=8			

反応$\boldsymbol{Y}$の自由度は，データ数nであるが，全体の平方和S_Tの自由度は，計算のために反応$\boldsymbol{Y}$から求めた$\bar{Y}$を使っているので，自由度が 1 つ分減少して$df_T=n-1$となる．誤差平方和S_eは，回帰の推定値$\hat{Y}_i$の計算のために$\hat{\beta}_0$と$\hat{\beta}_1$を用いているので自由度が 2 つ減り$df_e=n-2$となる．回帰の平方和S_Rは，自由度が 2 の$\hat{Y}_i$に対し，自由度 1 の$\bar{Y}$の差の平方和なので 1 つ減り$df_R=1$となる．平方和と同様に自由度の推定にも

$$\begin{aligned} df_T &= df_R+df_e \\ &= 1+(n-2)=n-1 \end{aligned} \qquad (4.34)$$

が成り立つ．分散分析における自由度については，各種の便宜的な説明が行われているが，偏差平方和の定義式に立ち返ることにより，自由度の本質的な理解となる．

F 値は，回帰の平均平方 117.82 を誤差の平均平方 2.60 で割って 45.36 が計算されている．この F 値は，分子の自由度 1，分母の自由度 7 の F 分布の上側確率で，F.dist.RT()関数を使い，=F.dist.RT(45.36, 1, 7)＝45.36 が計算されている．これらの Excel の確率分布の計算になれることも大切である．

分散分析表は，最小 2 乗法による分析に際して，結果を概観するために有益である．しかし，最尤法によるポアソン回帰では，対数尤度を用いた要約となり，偏差平方和をベースにした分散分析表の適用ができないが，基本的な考え方は同様であり，第 11.3 節の表 11.7 に「通常の回帰とポアソン回帰の対比」として対応関係を示してあるので，参照されたい．

パラメータの共分散行列の活用

Excel のデータ分析ツールの回帰分析により，パラメータの推定などが手軽にできることから Excel の行列関数を使った回帰分析をする意義はないと思われるかもしれない．多くの Excel ユーザの悩みは，回帰直線の 95%信頼区間および 95%予測区間（個別データの 95%信頼区間）を散布図上に描きたいと思っても手軽に解決する機能が見いだせないことにある．

デザイン行列を用いた行列計算による回帰分析の計算過程で，表 4.2 に示したようにパラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ の計算が含まれており，この行列を使って，パラメータの標準誤差 SE を計算している．パラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ があれば，95%信頼区間を散布図上に描くのは容易である．表 4.6 に示すように Excel のデータ分析ツールの回帰分析で出力される分散分析表の誤差分散 $\hat{\sigma}^2$ が得られるので，切片を含めた説明変数をデザイン行列 $\boldsymbol{X}$ とし，行列計算で $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ を求め

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$$

により，パラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ が容易に得られる．

回帰直線の 95%信頼区間

回帰直線の 95%信頼区間を求めるためには，回帰の推定値 $\hat{Y}_i$ の分散が必要となる．合成分散の一般的な公式により，

$$\begin{aligned} Var(\hat{Y}_i) &= Var(\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= Var(\hat{\beta}_0) + 2Cov(\hat{\beta}_0, \hat{\beta}_1)X_i + Var(\hat{\beta}_1)X_i^2 \end{aligned} \tag{4.35}$$

により，計算する．任意の行ベクトルを

$$\boldsymbol{x} = [1 \quad x]$$

とすれば，次の 2 次形式による計算方法で，

$$\left.\begin{aligned} Var(\hat{y}) &= \boldsymbol{x}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2]\boldsymbol{x}^T \\ &= \begin{bmatrix}1 & x\end{bmatrix}\begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & Var(\hat{\beta}_1)\end{bmatrix}\begin{bmatrix}1 \\ x\end{bmatrix} \\ &= \begin{bmatrix}1 & x\end{bmatrix}\begin{bmatrix} Var(\hat{\beta}_0)+Cov(\hat{\beta}_0,\hat{\beta}_1)x \\ Cov(\hat{\beta}_0,\hat{\beta}_1)+Var(\hat{\beta}_1)x\end{bmatrix} \\ &= Var(\hat{\beta}_0)+2Cov(\hat{\beta}_0,\hat{\beta}_1)x+Var(\hat{\beta}_1)x^2 \end{aligned}\right\} \quad (4.36)$$

合成分散の一般的な公式に一致する．

Excel できれいな 95%信頼区間の滑かな曲線を描くためには，表 4.5 に示すように，描きたい X 軸の範囲内で，適当な間隔の x を設定し，推定値を $\hat{y}=\boldsymbol{x}\hat{\boldsymbol{\beta}}$，分散 $Var(\hat{y})$ を

$$Var(\hat{y}) = \boldsymbol{x}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2]\boldsymbol{x}^T \quad (4.37)$$

で計算する．表 1.10 のポアソン回帰での 95%信頼区間と比べると，X の小さい方の幅が広がり，大きい方の幅が狭くなっている．

表 4.5　回帰直線の 95%信頼区間

切片	x	y^	$Var(y$^$)$	$L95\%$	$U95\%$	個別 $L95\%$	個別 $U95\%$
1	-2	-2.36	2.66	-6.22	1.49	-7.78	3.06
1	-1.8	-1.38	2.23	-4.91	2.15	-6.58	3.81
1	-1.6	-0.40	1.84	-3.61	2.81	-5.38	4.58
1	-1.4	0.58	1.50	-2.32	3.48	-4.21	5.37
1	-1.2	1.56	1.20	-1.03	4.16	-3.05	6.17
1	-1	2.55	0.94	0.25	4.84	-1.90	7.00
1	-0.8	3.53	0.73	1.51	5.55	-0.79	7.84
1	-0.6	4.51	0.56	2.74	6.27	0.31	8.71
1	-0.4	5.49	0.43	3.95	7.04	1.38	9.60
1	-0.2	6.47	0.34	5.09	7.85	2.42	10.53
1	0	7.45	0.30	6.17	8.74	3.43	11.48
:							
1	1.8	16.29	1.80	13.11	19.47	11.33	21.25
1	2	17.27	2.18	13.78	20.77	12.10	22.44

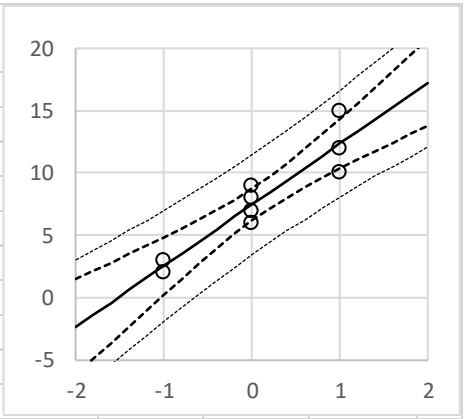


表 4.5 の 1 行目の $\boldsymbol{x}=\begin{bmatrix}1 & -2\end{bmatrix}$ に対する推定値 $\hat{y}$ は，

$$\begin{aligned} \hat{y}_{(x=-2)} &= \boldsymbol{x}\hat{\boldsymbol{\beta}} \\ &= \begin{bmatrix}1 & -2\end{bmatrix}\begin{bmatrix}7.4545 \\ 4.9091\end{bmatrix} \\ &= -2.3636 \end{aligned}$$

となり，分散は，

$$
\begin{aligned}
Var(\hat{y}_{(x=-2)}) &= \boldsymbol{x}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2]\boldsymbol{x}^T \\
&= \begin{bmatrix}1 & -2\end{bmatrix}\begin{bmatrix}0.2952 & -0.0590 \\ -0.0590 & 0.5313\end{bmatrix}\begin{bmatrix}1 \\ -2\end{bmatrix} \\
&= 2.6564
\end{aligned}
$$

95%信頼区間は，

$$
\begin{aligned}
L95\% &= \hat{y}_{(x=-2)} - t_{0.05}(9-2)\sqrt{Var(\hat{y}_{(x=-2)})} \\
&= -2.3636 - 2.3646 \times \sqrt{2.6564} \\
&= -6.2176
\end{aligned}
$$

$$
\begin{aligned}
U95\% &= -2.3636 + 2.3646 \times \sqrt{2.6564} \\
&= 1.4904
\end{aligned}
$$

として計算されている．

個別データの 95%信頼区間は，回帰直線の分散 $Var(\hat{Y}_i)$ に 1 個分のデータの分散 $\hat{\sigma}^2 = 2.5974$ を加えた分散を使う．

$$
\begin{aligned}
\text{個別}L95\% &= \hat{y} - t_{0.05}(9-2)\sqrt{Var(\hat{y}) + \hat{\sigma}^2} \\
&= -2.3636 - 2.3646 \times \sqrt{2.6564 + 2.5974} \\
&= -7.7837
\end{aligned}
$$

$$
\begin{aligned}
\text{個別}L95\% &= -2.3636 + 2.3646 \times \sqrt{2.6564 + 2.5974} \\
&= 3.0564
\end{aligned}
$$

第 1 行目で作成した計算式は，行方向にフィルハンドルでコピーすることにより計算される．この結果を Excel 散布図にまとめた結果が示されている．Excel 散布図は，きめ細かな設定ができる優れものである．

伝統的な方法

多くの教科書で回帰直線の 95%信頼区間についての記述は，ほとんどが，以下の形式で示されている（過度な標準化となっている）．ある x_0 について

$$
\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{4.38}
$$

から，$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ なので，$\hat{\beta}_0$ について解くと，$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ となり，これを式（4.38）に代入して

$$
\begin{aligned}
\hat{Y} &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 \\
&= \bar{y} + \hat{\beta}_1 (x_0 - \bar{x})
\end{aligned} \tag{4.39}
$$

が得られる．$\hat{Y}$ の分散 $Var(\hat{Y})$ は，$\bar{y}$ と $\hat{\beta}_1$ の共分散が 0 なので，式 (4.22) の $Var(\hat{\beta}_1)$ を用いて，次式で与えられる．

$$\left.\begin{aligned} Var(\hat{Y}) &= Var[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})] \\ &= Var(\bar{y}) + (x_0 - \bar{x})^2 Var(\hat{\beta}_1) \\ &= \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]\hat{\sigma}^2 \end{aligned}\right\} \quad (4.40)$$

また，個別データに対する分散 $Ver(\hat{Y}_{個別})$ は，

$$\left.\begin{aligned} Ver(\hat{Y}_{個別}) &= Ver(\hat{Y}) + \hat{\sigma}^2 \\ &= \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]\hat{\sigma}^2 \end{aligned}\right\} \quad (4.41)$$

で与えられる．単回帰分析の場合には，これらのシグマでの式を用いればよいのであるが，最尤法によるポアソン回帰の場合には，そもそも偏差平方和の計算をしないので，これらの式は使えない．回帰分析でも，変数が増えた場合には，これらの式の拡張が必要であり，第 12.3 節の「偏差平方和ベースの重回帰分析」で詳しく説明しているので参考にしてもらいたい．

このように，多くの教科書で定式化されている単回帰についてシグマを用いた計算公式は，単回帰分析でのみに対するものであり，他の問題に対して応用することができない．このような状況は，再々繰り返すが，多くの読者に対して応用力を封じ込めるような「ガラスの天井」のごとくである．Excel の「データ分析」の「回帰分析」，回帰分析のための LinEst()関数は，重回帰分析もサポートする優れものではあるが，95%信頼区間の計算についてはサポートされていない．

例えば 2 次式 $y = \beta_0 + \beta_1 x + \beta_2 x^2$ をあてはめ，その 95%信頼区間を描きたいとしても，従来の計算方法では，解決の糸口はつかみ難い．行列計算の場合ならば，変数の数が増えてもパラメータの共分散行列は，常に $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$ であり，変数の数によらず 95%信頼区間の計算方法は同じである．

第 2 章でも示したように，最尤法によるポアソン回帰の場合であっても，パラメータの分散共分散行列（共分散行列，と分散を省略する場合もある）を用いた 95%信頼区間の計算の考え方は，通常の回帰分析の場合と全く同じで，汎用的である．

また，第 12.5 節では，デザイン行列をベースにした行列計算により，2 次曲線の 95%信頼区間を Excel の散布図で作成する手順を詳細に示している．単に 2 次式の回帰パラメータを出すだけならば，Excel の「回帰分析」を使えばいいのだが，2 次曲線の 95%信頼区間の計算はどのようにしたらよいのだろうか．

東京大学教養学部統計学教室編（1992），「自然科学の統計学」の「第 2 章　線形モデルと最小二乗法」には，2 次式についての正規方程式が展開され，パラメータの推定値が求められ，2 次曲線が例示されている．しかし，95%信頼区間の計算式の例示は見いだせない．もちろん，最小二乗推定量の分散の一般式としての記述はあるものの，残念ながら事例としては，単回帰分析の伝統的な計算方法が示され，95%信頼区間の計算式も偏差平方和を用いた伝統的な記述となっている．そこで，第 12.5 節でパラメータの共分散行列の活用事例として 2 次式の 95%信頼区間および予測区間を推定しグラフ化する方法を示す．

現実的な対応

Excel の「回帰分析」は手軽に使える優れものであるので，パラメータの共分散分行列の計算を付け加えることにより，95%信頼区間の計算を自在にできるようになる．表 4.6 に示すよ

表 4.6　分析ツールの回帰分析およびパラメータの共分散行列の計算

i	X		Y		分散分析表（分析ツールの回帰分析）						
1	1	-1	2			自由度	変動	分散	分散比	有意 F	
2	1	-1	3		回帰	1	117.8182	117.8182	45.3600	0.0003	
3	1	0	6		残差	7	18.1818	2.5974			
4	1	0	7		合計	8	136.0000				
5	1	0	8								
6	1	0	9			係数	標準誤差	t	P-値	下限 95%	上限 95%
7	1	1	10		切片	7.4545	0.5433	13.7212	0.0000	6.1699	8.7392
8	1	1	12		X 値 1	4.9091	0.7289	6.7350	0.0003	3.1855	6.6327
9	1	1	15								
	9.00	1.00			0.1136	-0.0227		0.2952	-0.0590	$t_{0.05}(7)$=	2.3646
	1.00	5.00			-0.0227	0.2045		-0.0590	0.5313		
	X^TX				$(X^TX)^{-1}$			$\Sigma(\beta\hat{})=(X^TX)^{-1}\sigma\hat{}^2$			

うに，Excel の「回帰分析」を使い，分散分析表から残差分散 $\hat{\sigma}^2 = 2.5974$ を，回帰係数の表から $\hat{\beta}_0 = 7.4545$，$\hat{\beta}_1 = 4.9091$ を得る．パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，デザイン行列 $\boldsymbol{X}$ を用い

$$\boldsymbol{X}^T\boldsymbol{X} = \text{Mmult(Transpose(}\boldsymbol{X}\text{の範囲), }\boldsymbol{X}\text{の範囲)}$$

$$= \begin{array}{|cc|}\hline 9.00 & 1.00 \\ 1.00 & 5.00 \\ \hline \end{array}$$

$$\boldsymbol{X}^T\boldsymbol{X}$$

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$$

$$= \text{Minverse}(\boldsymbol{X}^T\boldsymbol{X}\text{の範囲}) \times \hat{\sigma}^2$$

$$= \begin{array}{|cc|}\hline 0.1136 & -0.0227 \\ -0.0227 & 0.2045 \\ \hline \end{array} \times \boxed{2.5974}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$=$

0.2952	-0.0590
-0.0590	0.5313

$\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma\hat{}^2$

によって計算できる．これらを用いて，表 4.5 で示した回帰直線の 95%信頼区間の計算が可能となる．

表 4.5 に示した Excel によるグラフ作成の手順は，すでに示したが，ここでは，できる限り表 4.7 に示すように行列計算を使った計算方法で示す．もちろん，個別データに対しての計算式をフィルハンドルでコピーしても同じ結果が得られるのであるが，データ数が変化したときなど，行列計算による計算式の方が，変更に際して見通しがよく，操作性にも優れている．

表 4.7　Excel の行列計算機能を用いた上側 95%信頼区間の一括計算

LINEST　=Q5:Q13+T.INV.2T(0.05,7)*SQRT(R5:R13)

	N	O	P	Q	R	S	T	U	V
4		切片	x	$y\hat{}$	$Var(y\hat{})$	$L95\%$	$U95\%$	個別$L95\%$	個別$U95\%$
5		1	-2.0	-2.3636	2.6564	-6.2176	=Q5:Q13+T	-7.7837	3.0564
6		1	-1.5	0.0909	1.6677	-2.9627	3.1445	-4.7925	4.9743
7		1	-1.0	2.5455	0.9445	0.2474	4.8435	-1.9048	6.9957
8		1	-0.5	5.0000	0.4870	3.3498	6.6502	0.8471	9.1529
9		1	0.0	7.4545	0.2952	6.1699	8.7392	3.4329	11.4762
10		1	0.5	9.9091	0.3689	8.4728	11.3454	5.8365	13.9817
11		1	1.0	12.3636	0.7084	10.3734	14.3538	8.0643	16.6630
12		1	1.5	14.8182	1.3135	12.1082	17.5282	10.1419	19.4944
13		1	2.0	17.2727	2.1842	13.7781	20.7674	12.1020	22.4434

手順 1）X 軸の範囲を 0.5 の増分で(–2～+2)とし，切片 1 も含めデザイン行列 $\boldsymbol{X}$ とする．

手順 2）推定値 $\hat{\boldsymbol{y}}$ を次式で計算する．

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$$
$$=\mathrm{Mmultt}(\boldsymbol{X}\text{の範囲},\hat{\boldsymbol{\beta}}\text{の範囲})$$

手順 3）推定値 $\hat{y}_i$ の分散を計算し，フィルハンドルで計算式をコピーする．

$$Var(\hat{y}_i)=\boldsymbol{x}_i\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i^T$$
$$=\mathrm{Mmult}(\mathrm{Mmult}(\boldsymbol{x}_i\text{の範囲},\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{の範囲}),\mathrm{Transpose}(\boldsymbol{x}_i\text{の範囲}))$$

手順 4）回帰の推定値の 95%信頼区間を一括で計算する

$$\boldsymbol{L}95\%=\hat{\boldsymbol{y}}-t_{0.05}(9-2)\sqrt{Var(\hat{\boldsymbol{y}})}$$
$$=\hat{\boldsymbol{y}}\text{の範囲}-\mathrm{Tinv.2T}(0.05,9-2)\times\mathrm{Sqrt}(Var(\hat{\boldsymbol{y}})\text{の範囲})$$

$$\boldsymbol{U}95\%=\hat{\boldsymbol{y}}+t_{0.05}(9-2)\sqrt{Var(\hat{\boldsymbol{y}})}$$
$$=\hat{\boldsymbol{y}}\text{の範囲}+\mathrm{Tinv.2T}(0.05,9-2)\times\mathrm{Sqrt}(Var(\hat{\boldsymbol{y}})\text{の範囲})$$

手順 5）個別の回帰の推定値の 95%信頼区間を一括で計算する

$$\text{個別}\boldsymbol{L}95\% = \hat{\boldsymbol{y}} - t_{0.05}(9-2)\sqrt{Var(\hat{\boldsymbol{y}}) + \hat{\sigma}^2}$$
$$= \hat{\boldsymbol{y}}\text{の範囲} - \text{Tinv.2T}(0.05, 9-2) \times \text{Sqrt}(Var(\hat{\boldsymbol{y}})\text{の範囲}+\hat{\sigma}^2)$$

$$\text{個別}\boldsymbol{U}95\% = \hat{\boldsymbol{y}} + t_{0.05}(9-2)\sqrt{Var(\hat{\boldsymbol{y}}) + \hat{\sigma}^2}$$
$$= \hat{\boldsymbol{y}}\text{の範囲} + \text{Tinv.2T}(0.05, 9-2) \times \text{Sqrt}(Var(\hat{\boldsymbol{y}})\text{の範囲}+\hat{\sigma}^2)$$

手順 6） Excel グラフの「データの選択」機能を使い，推定値 $\hat{\boldsymbol{y}}$，95%信頼区間 $L95\%$，$U95\%$，個別$L95\%$，個別$U95\%$ を追加し，「データ系列の書式設定」で適当な線グラフとする．

平方和の分解に対する補足

式（4.33）で，全体の平方和 S_T が，回帰の平方和 S_R と誤差平方和 S_e の和に分解できるとしたが，式の展開を省略して結論だけを示したので補足をする．式（4.33）では，第 3 項があり，これがゼロとなることを示さなかった．

$$\left.\begin{aligned} S_T &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}[(Y_i - \hat{Y}_i) + (\hat{Y} - \bar{Y})]^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}[(Y_i - \hat{Y}_i)(\hat{Y} - \bar{Y})] \\ &= S_e + S_R + 2\sum_{i=1}^{n}[(Y_i - \hat{Y}_i)(\hat{Y} - \bar{Y})] \\ &= S_e + S_R \end{aligned}\right\} \quad (4.42)$$

第 3 項は，次のように展開して 0 となる．

$$\left.\begin{aligned} \sum_{i=1}^{n}[(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})] &= \sum_{i=1}^{n}\hat{\varepsilon}_i\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}\hat{\varepsilon}_i \\ &= \sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= \hat{\beta}_0\sum_{i=1}^{n}\hat{\varepsilon}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{\varepsilon}_i X_i \\ &= 0 \end{aligned}\right\} \quad (4.43)$$

これは，式（4.7）の正規方程式を ε_i で置き換えた次の式が，0 となることを用いている．

$$\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n}\hat{\varepsilon}_i = 0 \quad (4.44)$$

$$\sum_{i=1}^{n}\left[\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)X_i\right] = \sum_{i=1}^{n}\left[\left(Y_i - \hat{Y}_i\right)X_i\right] = \sum_{i=1}^{n}\hat{\varepsilon}_i X_i = 0 \quad (4.45)$$

4.6. 逆推定値に対する各種の 95%信頼区間の推定

図 4.1 に示すように逆推定は，検量線に対して未知検体の濃度を知りたい場合に，未知検体から反応 y_0 が得られた場合の濃度 $\hat{x}_0$ を推定する問題である．検量線が直線の場合に

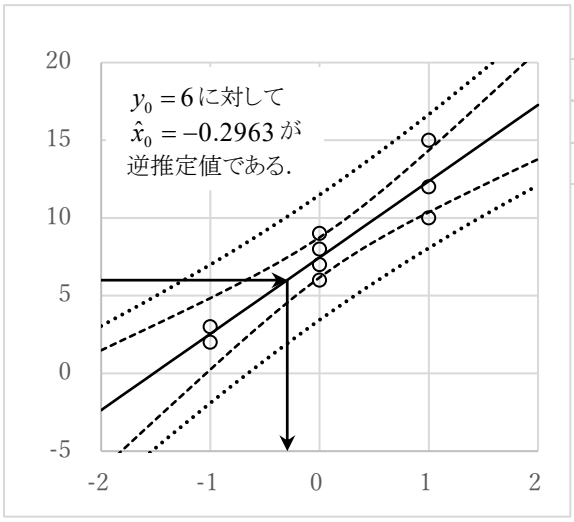
$$y_0 = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_0$$

の関係から，

$$\begin{aligned} \hat{x}_0 &= \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} \\ &= \frac{6.0 - 7.4545}{7.9091} = -0.2963 \end{aligned} \tag{4.46}$$

と，容易に濃度 $\hat{x}_0$ を得ることができる．難しいのは，$\hat{x}_0$ の 95%信頼区間の算出である．これは，推定された逆推定値 $\hat{x}_0$ が回帰パラメータの比で表されており，一般的な線形式に対する合成分散の計算公式が使えないからである．

項	推定値
β_0^=	7.4545
β_1^=	4.9091



共分散	0.2952	-0.0590
行列	-0.0590	0.5313
	$\Sigma(\beta$^$)=(X^TX)^{-1}\sigma$^^{2}	

図 4.1　逆推定の例示

逆推定値の 95%信頼区間の求め方について，参考になるのは，竹内（1979），「数理統計学，第 29 章 IV 回帰直線自体についての推論」である．JMP を用いた逆推定については，芳賀（2010），「医薬品開発のための統計学，第 3 部 非線形モデル，第 1 章 （4）JMP による逆推定の解析」が詳しい．パラメータの共分散分散行列を活用については，Collett（2003），「Modeling Binary Data 2nd. ed., 4.2.1 Approximate standard error of an estimated effective dose」が参考になる．逆推定の 95%信頼区間に関して，高橋（2013a），「応用回帰分析 I －各種の重み付き回帰における逆推定－」も，第 9 章で文献に対する「文献に対する批判的吟味」を行っている．高橋（2013b），「回帰分析・再入門 －統計ソフトが対応していない生物統計の各種の課題を Excel

でサクサク解こう」は，「基礎セミナー じっくり勉強すれば身につく統計入門」シリーズの第 7 回目の資料集で，本章でのデザイン行列を用いた回帰分析についてスライドを用いて説明をしている．

デルタ法による近似 95%信頼区間

任意の合成分散式に対しては，デルタ法によって合成分散を求めることができる．そのために，求めたい $\hat{x}_0$ の式（4.46）に対して，$\hat{\beta}_0$ と $\hat{\beta}_1$ で偏微分し，

$$d_0 = \frac{\partial \hat{x}_0}{\partial \hat{\beta}_0} = \frac{-1}{\hat{\beta}_1} \tag{4.47}$$

$$d_1 = \frac{\partial \hat{x}_0}{\partial \hat{\beta}_1} = \frac{-(y_0 - \hat{\beta}_0)}{\hat{\beta}_1^2} \tag{4.48}$$

これらの偏微分式を行ベクトル

$$\boldsymbol{d} = [d_0 \; d_1] \tag{4.49}$$

としたとき，$\hat{x}_0$ の分散 $Var(\hat{x}_0)$ は，

$$\begin{aligned} Var(\hat{x}_0) &= \boldsymbol{d}\,\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{d}^T \\ &= [d_0 \; d_1] \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_0) \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \end{bmatrix} \end{aligned} \tag{4.50}$$

の 2 次形式の計算で求めることができる．

表 4.8　逆推定の近似 95%信頼区間

y_0	x_0^	d_0	d_1	$Var(X_0$^$)$	$L95\%$	$U95\%$	個別$L95$	個別$U95$
2	-1.1111	-0.2037	0.2263	0.0449	-1.6122	-0.6100	?	?
3	-0.9074	-0.2037	0.1848	0.0348	-1.3488	-0.4660		
6	-0.2963	-0.2037	0.0604	0.0156	**-0.5920**	**-0.0006**		
7	-0.0926	-0.2037	0.0189	0.0129	-0.3611	0.1759		
8	0.1111	-0.2037	-0.0226	0.0120	-0.1477	0.3699		
9	0.3148	-0.2037	-0.0641	0.0129	0.0463	0.5833		
10	0.5185	-0.2037	-0.1056	0.0156	0.2228	0.8142		
12	0.9259	-0.2037	-0.1886	0.0266	0.5402	1.3117		
15	1.5370	-0.2037	-0.3131	0.0568	0.9735	2.1006		

表 4.8 の 1 行目の $y_0 = 2$ に対する逆推定値 $\hat{x}_0$ は，表 4.2 の $\hat{\beta}_0$ と $\hat{\beta}_1$ の推定値を用いて

$$\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \frac{2 - 7.4545}{4.9091} = -1.1111$$

であり，$\hat{x}_0$ に対する $\hat{\beta}_0$ と $\hat{\beta}_1$ で偏微分式は，

$$d_0 = \frac{-1}{\hat{\beta}_1} = \frac{-1}{4.9091} = -0.2037$$

$$d_1 = \frac{-(y_0 - \hat{\beta}_0)}{\hat{\beta}_1^2} = \frac{-(2 - 7.4545)}{4.9091^2} = 0.2263$$

$$\boldsymbol{d} = [-0.2037 \;\; 0.2263]$$

なので，分散$Var(\hat{x}_0)$は，

$$\begin{aligned} Var(\hat{x}_0) &= \boldsymbol{d}\,\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{d}^T \\ &= \begin{bmatrix} -0.2037 & 0.2263 \end{bmatrix} \begin{bmatrix} 0.2952 & -0.0590 \\ -0.0590 & 0.5313 \end{bmatrix} \begin{bmatrix} -0.2037 \\ 0.2263 \end{bmatrix} \\ &= 0.0449 \end{aligned}$$

と計算されている．下側 95%点と上側 95%点は，

$$\begin{aligned} L95\% &= \hat{x}_0 - t_{0.05}(9-2)\sqrt{Var(\hat{x}_0)} \\ &= -1.1111 - 2.3646 \times \sqrt{0.0449} = -1.6122 \\ U95\% &= \hat{x}_0 + t_{0.05}(9-2)\sqrt{Var(\hat{x}_0)} \\ &= -1.1111 + 2.3646 \times \sqrt{0.0449} = -0.6100 \end{aligned}$$

で計算されている．第 1 行目で作成した計算式は，行方向にフィルハンドルでコピーすることにより計算される．なお，この方法では，個別データの 95%信頼区間の計算ができない．

逆推定値に対する正確な 95%信頼区間

逆推定値に対する正確な 95%信頼区間の算出方法には，いくつかの方法があるので，図 4.1 で示した回帰直線の推定値 $\hat{y}$ の 95%信頼区間を活用する方法を示す．ある $y_0 = 6$ に対する逆推定値は，

$$\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \frac{6 - 7.4545}{4.9091} = 0.2693$$

と推定される．図 4.2 に示すように，ある $y_0 = 6$ の水平線 2 つの 95%信頼区間の交点に着目する．水平線と 95%信頼曲線の上側の交点の X 軸 $\hat{x}_{L95}$ が，$\hat{x}_0 = -0.269$ の X 軸方向の下側 95%点となるが，このままでは推定できない．何らかの探索的な方法が必要となる．

交点の推定値 $\hat{x}_{L95}$ に対する回帰の推定値 $\hat{y}_{L95}$ は，未知の $\hat{x}_{L95}$ を用いて

$$\hat{y}_{L95} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{L95} \tag{4.51}$$

である．この $\hat{y}_{L95}$ に対する Y 軸方向の上側 95%点は，$y_0 = 6$ と等しいので，次の等式が成り立つ．

$$y_0 = \hat{y}_{L95} + t_{0.05}\sqrt{Var(\hat{y}_{L95})} \tag{4.52}$$

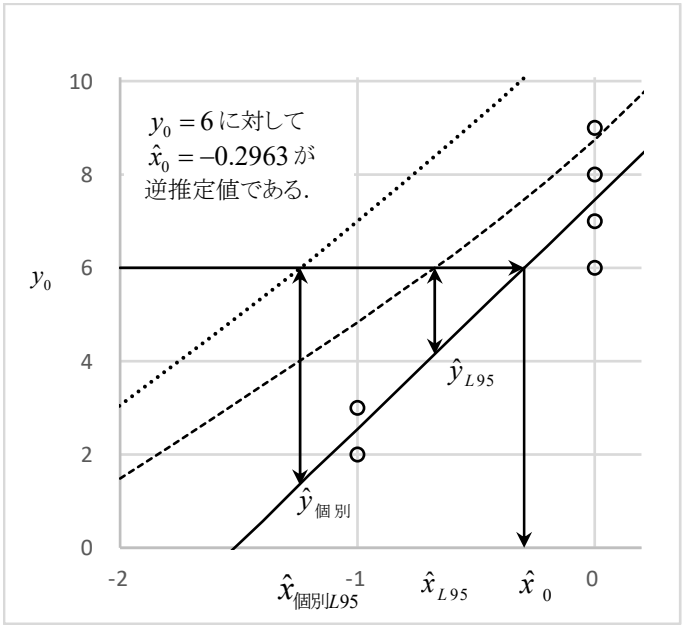


図 4.2　逆推定の 95%信頼区間の算出の例示

推定したいのは，$\hat{x}_{L95}$ なので，$\hat{y}_{L95}$ を $\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{L95}$

$$y_0 = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{L95} + t_{0.05} \sqrt{Var(\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{L95})} \tag{4.53}$$

に置き換える．この式を $\hat{x}_{L95}$ について解くことにより，逆推定値 $\hat{x}_{L95}$ の下側 95%点が推定できる．上の式を右辺の $\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{L95}$ を左辺に移し，右辺の $t_{0.05}$ を左辺の分母とし，両辺を平方すると，次式が得られる．

$$\left(\frac{y_0 - \hat{\beta}_0 - \hat{\beta}_1 \hat{x}_{L95}}{t_{0.05}} \right)^2 = Var(\hat{\beta}_0) + 2Cov(\hat{\beta}_0, \hat{\beta}_1) \hat{x}_{L95} + Var(\hat{\beta}_1) \hat{x}_{L95}^2 \tag{4.54}$$

右辺を左辺に移して $\hat{x}_{L95}$ について整理すると，

$$\left[Var(\hat{\beta}_0) - \frac{(y_0 - \hat{\beta}_0)^2}{t_{0.05}^2} \right] + \left[2Cov(\hat{\beta}_0, \hat{\beta}_1) + \frac{2(y_0 - \hat{\beta}_0)\hat{\beta}_1}{t_{0.05}^2} \right] \hat{x}_{L95} + \left[Var(\hat{\beta}_1) - \frac{\hat{\beta}_1^2}{t_{0.05}^2} \right] \hat{x}_{L95}^2 = 0 \tag{4.55}$$

が得られる．この複雑な式は，幸い $\hat{x}_{L95}$ に関する 2 次式に

$$a + b\hat{x}_{L95} + c\hat{x}_{L95}^2 = 0 \tag{4.56}$$

なるので，2 次式の解の公式

$$\hat{x}_{L95} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c} \tag{4.57}$$

により，$\hat{x}_{L95}$ を求めることができる．解は 2 つあるが，小さい方が $\hat{x}_{L95}$ となり，大きい方が上側 95%点 $\hat{x}_{U95}$ となる．

個別データの正確な 95%信頼区間

個別の上側 95%曲線と $y_0=6$ を通る水平線との交点は，$\hat{x}_{個別L95}$ に対する回帰の推定値

$$\hat{y}_{個別L95}=\hat{\beta}_0+\hat{\beta}_1\hat{x}_{個別L95} \tag{4.58}$$

の上側 95%点でもある．個別の上側 95%点は，y_0 と等しいので，次の等式が成り立つ．

$$y_0=\hat{y}_{L95}+t_{0.05}\sqrt{Var(\hat{y}_{個別L95})+\hat{\sigma}^2} \tag{4.59}$$

推定したいのは，$\hat{x}_{個別L95}$ なので，

$$y_0=\hat{\beta}_0+\hat{\beta}_1\hat{x}_{個別L95}+t_{0.05}\sqrt{Var(\hat{\beta}_0+\hat{\beta}_1\hat{x}_{個別L95})+\hat{\sigma}^2} \tag{4.60}$$

と置き換える．式（4.59）を $\hat{x}_{個別L95}$ について解くと式（4.55）と同様な結果となるが，2 次式のパラメータ a について

$$個別a'=a+\hat{\sigma}^2$$

となるのみで，他は同様である．

これらから，表 4.9 に示すように，与えられた y_0 に対する逆推定値の 95%信頼区間，および，個別データの 95%信頼区間を推定することができる．

表 4.9　逆推定の正確な 2 種類の 95%信頼区間

	2次式のパラメータ			逆推定	95%信頼区間		個別	個別95% CL	
y_0	a	b	c	$x^{\wedge}_0$	$x^{\wedge}_{L95}$	$x^{\wedge}_{U95}$	a'	個別L95	個別U95
2	-5.0258	-9.6959	-3.7787	-1.1111	-1.8450	-0.7209	-2.4284	-2.2846	-0.2813
3	-3.2536	-7.9399	-3.7787	-0.9074	-1.5433	-0.5579	-0.6562	-2.0150	-0.0862
6	-0.0832	-2.6721	-3.7787	-0.2963	**-0.6745**	**-0.0327**	2.5142	**-1.2426**	**0.5354**
7	0.2582	-0.9162	-3.7787	-0.0926	-0.4094	0.1669	2.8556	-0.9990	0.7565
8	0.2419	0.8397	-3.7787	0.1111	-0.1652	0.3875	2.8394	-0.7628	0.9850
9	-0.1320	2.5956	-3.7787	0.3148	0.0553	0.6316	2.4654	-0.5343	1.2212
10	-0.8636	4.3516	-3.7787	0.5185	0.2549	0.8967	1.7338	-0.3132	1.4648
12	-3.4000	7.8634	-3.7787	0.9259	0.6129	1.4681	-0.8026	0.1076	1.9733
15	-9.8872	13.1312	-3.7787	1.5370	1.1031	2.3719	-7.2898	0.6936	2.7815

表 4.9 の 1 行目の $y_0=2$ の逆推定値 $\hat{x}_0$ は，表 4.2 の $\hat{\beta}_0$ と $\hat{\beta}_1$ の推定値を用いて

$$\hat{x}_0=\frac{y_0-\hat{\beta}_0}{\hat{\beta}_1}=\frac{2-7.4545}{4.9091}=-1.1111$$

表 4.2：

$\beta_0^{\wedge}=$	7.4545	$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=$	18.1818	共分散	0.2952	-0.0590
$\beta_1^{\wedge}=$	4.9091	$\sigma^{\wedge 2}=$	2.5974	行列	-0.0590	0.5313
$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$		$t_{0.05}=$	2.3646		$\boldsymbol{\Sigma}(\boldsymbol{\beta}^{\wedge})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{\wedge 2}$	

である．2 次式のパラメータ a，b，c は，

$$a = Var(\hat{\beta}_0) - \frac{(y_0 - \hat{\beta}_0)^2}{t_{0.05}^2} = 0.2952 - \frac{(2 - 7.4545)^2}{2.3646^2} = -5.0258$$

$$b = 2Cov(\hat{\beta}_0, \hat{\beta}_1) + \frac{2(y_0 - \hat{\beta}_0)\hat{\beta}_1}{t_{0.05}^2} = 2 \times (-0.0590) + \frac{2 \times (2 - 7.4545) \times 4.9091}{2.3646^2} = -9.6959$$

$$c = Var(\hat{\beta}_1) - \frac{\hat{\beta}_1^2}{t_{0.05}^2} = 0.5313 - \frac{7.4545^2}{2.3646^2} = -3.7778$$

となる．下側 95%点 $\hat{x}_{L95}$ は，

$$\hat{x}_{L95} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c}$$

$$= \frac{-(-9.6959) \pm \sqrt{(-9.6959)^2 - 4 \times (-5.0258) \times (-3.7787)}}{2 \times (-3.7778)} = (-1.8450, -0.7209)$$

となる．個別の下側 95%点は，

$$個別a' = a + \hat{\sigma}^2 = -5.0258 + 2.5974 = -2.4284$$

なので，

$$\hat{x}_{個別L95} = \frac{-b \pm \sqrt{b^2 - 4a'c}}{2c}$$

$$= \frac{-(-9.6958) \pm \sqrt{(-9.6959)^2 - 4 \times (-2.4284) \times (-3.7787)}}{2 \times (-3.7787)} = (-2.2846, -0.2813)$$

と計算される．図 4.2 の逆推定の 95%信頼区間の算出の例示で用いた $\hat{x}_{L95}$ に関する逆推定値は，表 4.9 から該当部分を抜粋した結果を表 4.10 に示すように，$\hat{x}_0 = -0.296$，$\hat{x}_{L95} = -0.675$，$\hat{x}_{個別L95} = -1.243$ である．

表 4.10　近似および正確な逆推定値の比較

		逆推定	95%信頼区間		個別95% CL	
	y_0	$x^\wedge_0$	$x^\wedge_{L95}$	$x^\wedge_{U95}$	個別L95	個別U95
正確	6	-0.2963	-0.6745	-0.0327	-1.2426	0.5354
近似	"	"	-0.5920	-0.0006	?	?

Excel ソルバーを用いた逆推定の正確な 95%信頼区間

Excel ソルバーには，目標値を最大化または最小化する以外に，設定した推定値となるように変数セルを変更する機能がある．この機能を使うことにより，逆推定値 y_0 の水平線と 95%信頼区間の曲線が交わる点を目標値に設定することができる．Excel のソルバーで $\hat{x}_{L95}$ または $\hat{x}_{U95}$ を変化させて

$$\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{L95}, \quad \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{U95} \ \rightarrow \ y_0$$

y_0 となるような推定値 $\hat{x}_{L95}$ または $\hat{x}_{U95}$ を得る．表 4.2 に示した回帰分析の結果から，パラメータ $\hat{\boldsymbol{\beta}}$ の推定値と共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を表 4.11 再掲する．ある x， $x=0$ とした場合の y の推定値は，

$$\hat{y}_{(x=0)} = 7.4545 + 4.9091 \times 0 = 7.4545$$

となり，その分散 $Var(\hat{y}_{(x=0)})$ は，合成分散の 2 次形式により，

$$Var(\hat{y}_{(x=0)}) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.2952 & -0.0590 \\ -0.0590 & 0.5313 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 0.2952$$

と推定されるので，95%信頼区間の上限 $U95\%$ は，

$$U95\% = 7.4545 + 2.3646 \times \sqrt{0.2952} = 8.7392$$

が計算されている．この結果が，表 4.11 の「推定値」の行に示されている．

「推定値」の行を次の行にコピー&ペーストし，「回帰下限」とする．Excel ソルバーの目標値として「回帰下限」の「$U95\%$」欄の 8.7392 をセットし，さらに「指定値」として「6」を入力する．「変数セルの変更」に $x=$ 0 の位置としてセットする．ソルバーを実行すると，「回帰下限」が $y_0 = 6$ となるような $x_0 = -0.6745$ を推定してくれる．これが，正確な逆推定値における 95%信頼区間の下限である．

表 4.11　Excel ソルバーによる逆推定の実際

		β_0^=	7.4545	共分散	0.2952	-0.0590	σ^^{2}=	2.5974	
		β_1^=	4.9091	行列	-0.0590	0.5313	$t_{0.05}$=	2.3646	
		切片	x	y^	$Var(Y$^$)$	$L95\%$	$U95\%$	個別 $L95\%$	個別 $U95\%$
推定値		1	**0**	7.4545	0.2952	6.1699	**8.7392**	3.4329	11.4762
回帰下限		1	**-0.6745**	4.1433	0.6165		**6.0000**	：目標値(6 にセット)	
回帰上限		1	**-0.0327**	7.2943	0.2996	**6.0000**			
個別下限		1	**-1.2426**	1.3545	1.2622				**6.0000**
個別上限		1	**0.5354**	10.0831	0.3843			**6.0000**	

次いで，$L95\%$ について同様の手順を繰り返すことにより，$y_0 = 6$ に対する 95%信頼区間として（–0.6745，–0.0327）が推定される．更に個別の $U95\%$ と $L95\%$ については，（–1.2426，–0.5345）と推定される．これは，表 4.10 に示した指定結果に一致する．

4.7. JMP による回帰分析と逆推定

これまで，Excel を用いて回帰分析のパラメータの共分散行列を活用し，各種の 95%信頼区間について示してきた．統計ソフト JMP の「二変量のあてはめ」によって回帰直線の 95%信頼区間のきれいなグラフを手軽に作成できるのであるが，内部でどのような計算式が使われているかを出力させることが可能であり，大変興味深い．また，JMP の「モデルのあてはめ」を使うことにより，逆推定も手軽に行うことができる．

「二変量の関係」による回帰分析

表 4.12 に JMP の「二変量の関係」による回帰分析の結果を示す．「分散分析」の結果は，表 4.4 で示した結果に相当し，「パラメータの推定値」は，表 4.2 に相当する．さらに，回帰直線の 95%信頼区間，および，個別データの 95%信頼区間を計算式付きで JMP ファイルへ出力することもできる．

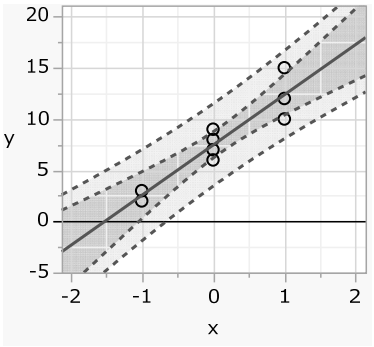
表 4.12　JMP の「二変量の関係」による回帰分析

分散分析

要因	自由度	平方和	平均平方	F値
モデル	1	117.8182	117.8182	45.3600
誤差	7	18.1818	2.5974	p値(Prob>F)
全体(修正済み)	8	136.0000		0.0003*

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	7.4545	0.5433	13.72	<.0001*
x	4.9091	0.7289	6.73	0.0003*



回帰直線の 95%信頼区間の計算式

表 4.13 の「下側 95%」および「上側 95%」の欄は，計算式を含めた出力になっており，内部での計算式を確認できるようになっている．回帰の 95%信頼区間の下限については，表 4.14 に示すような計算式を表示させることができる．

計算式内の VecQuadraticc()関数は，デザイン行列の計算 $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ 結果に対し，$\boldsymbol{X}$ の行ベクトル $\boldsymbol{x}_i$ の 2 次形式の計算を行う関数であり，さらに，誤差分散 $\hat{\sigma}^2 = 2.5974$ を掛け，$\hat{Y}_i$ の分散 $Var(\hat{Y}_i)$ の平方根が計算されている．$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ に $\hat{\sigma}^2$ を掛けた結果は，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ となっており，また，推定値 $\hat{Y}_i$ の分散の計算にも $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を挟んだ 2 次形式になっていて，伝統的な偏差平方和を用いた計算ではないことが確認できる．

表 4.13　JMP の「二変量の関係」による信頼区間の出力

	x	y	予測式 y	下側95%	上側95%	L95% 個別	U95% 個別
1	-1	2	2.52	0.25	4.84	-1.90	7.00
2	-1	3	2.52	0.25	4.84	-1.90	7.00
3	0	6	7.45	6.17	8.74	3.43	11.48
4	0	7	7.45	6.17	8.74	3.43	11.48
5	0	8	7.45	6.17	8.74	3.43	11.48
6	0	9	7.45	6.17	8.74	3.43	11.48
7	1	10	12.39	10.37	14.35	8.06	16.66
8	1	12	12.39	10.37	14.35	8.06	16.66
9	1	15	12.39	10.37	14.35	8.06	16.66
10	-2	•	-2.42	-6.22	1.49	-7.78	3.06
11	-1.5	•	0.05	-2.96	3.14	-4.79	4.97
12	-0.5	•	4.98	3.35	6.65	0.85	9.15
13	0.5	•	9.92	8.47	11.35	5.84	13.98
14	1.5	•	14.85	12.11	17.53	10.14	19.49
15	2	•	17.32	13.78	20.77	12.10	22.44

注）10 行目～15 行目は，グラフ表示の際のため，x を与えて予測値などの計算をさせている．

表 4.14　JMP の「二変量の関係」による上側 95%の計算式

$$\left(7.4545454545 + 4.9090909091 \bullet x\right)$$

$$-2.3646242516 \bullet \sqrt{\text{Vec Quadratic}\left(\begin{bmatrix} 0.114 & -0.023 \\ -0.023 & 0.205 \end{bmatrix}, [1] \,||\, x\right) \bullet 2.5974025974}$$

$$\text{VecQuadraticc}() \bullet \hat{\sigma}^2 = \begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} 0.1136 & -0.0227 \\ -0.0227 & 0.2045 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} * \begin{bmatrix} 2.5974 \end{bmatrix}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} \qquad \sigma\text{^}2$$

回帰の 95%信頼区間の下限の計算式は，

$$L95\% = \hat{Y} - t_{0.05}(7)\sqrt{Var(\hat{Y})}$$

であることが，読み取れる．

個別データの 95%信頼区間は，「$L95\%$ 個別」の欄の計算式から，表 4.15 に示すように 2 次形式の関数が使われていることがわかる．

$$L95\%\text{個別} = \hat{Y} - t_{0.05}(7)\sqrt{Var(\hat{Y}) + \hat{\sigma}^2}$$

$$\sqrt{Var(\hat{Y}) + \hat{\sigma}^2} = \sqrt{\text{Vec Quadratic}((\boldsymbol{X}^T\boldsymbol{X})^{-1}, [1 \;\; x])\hat{\sigma}^2 + \hat{\sigma}^2}$$

表 4.15　JMP の「二変量の関係」による個別データの下側 95%の計算式

$$\left(7.4545454545 + 4.9090909091 \bullet x\right) - 2.3646242516 \bullet \sqrt{\text{Vec Quadratic}\left(\begin{bmatrix} 0.114 & -0.023 \\ -0.023 & 0.205 \end{bmatrix}, \begin{bmatrix} 1 \end{bmatrix} \| x\right) \bullet 2.5974025974 + 2.5974025974}$$

これらのことから，JMP での回帰分析は，デザイン行列ベースの計算となっていることを垣間見ることができる．

「モデルのあてはめ」による逆推定

逆推定を行い，Excel での計算結果との相互検証を行う．「二変量の関係」での回帰分析は，信頼区間の標示などで豊富な機能があるが，逆推定には対応していない．「モデルのあてはめ」による回帰分析を用いて表 4.16 に示すように，$x_0 = 6$ に対する逆推定を行ない，表 4.17 に結果を示す．

表 4.16　JMP の「逆推定」の設定

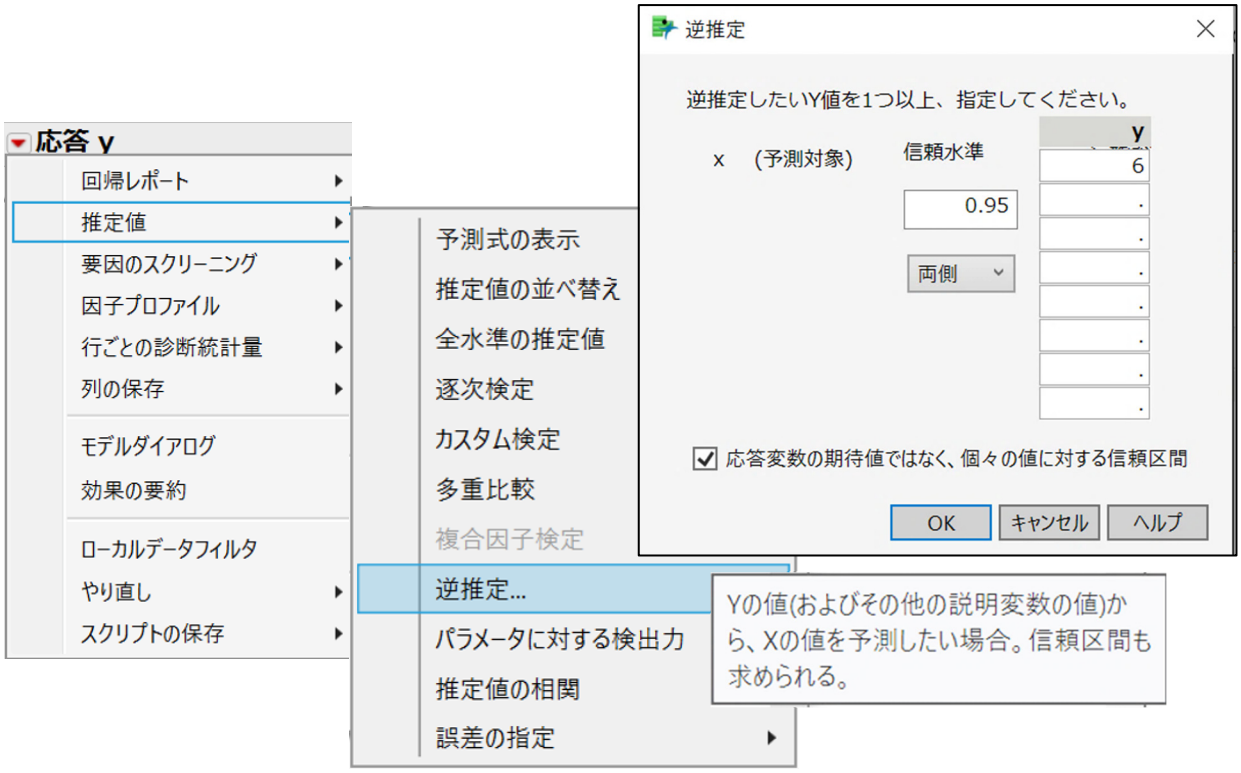
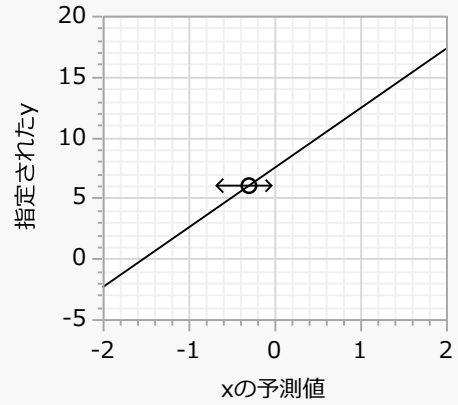


表 4.17　JMP の「モデルのあてはめ」による逆推定

逆推定

指定されたy	xの予測値	下側95%	上側95%
6.0000	-0.2963	-0.6745	-0.0327

応答の期待値に対する信頼区間



逆推定

指定されたy	xの予測値	下側95%	上側95%
6.0000	-0.2963	-1.2426	0.5354

個別の応答に対する信頼区間

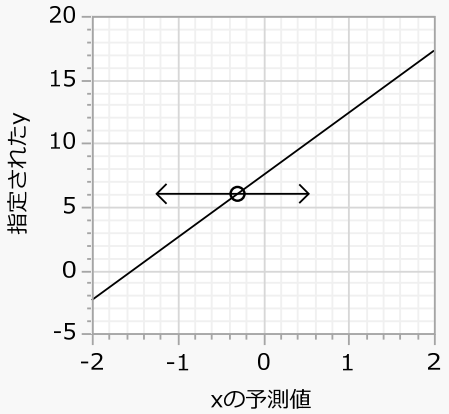


表 4.12 の「分散分析」および「パラメータの推定値」と全く同じ結果が得られることを確認し，$y_0 = 6$ についての逆推定を行う．なお，この結果は，表 4.10 と完全に一致する．

表 4.10 抜粋

	逆推定	95%信頼区間		個別95% CL	
y_0	$x^{\wedge}{}_0$	$x^{\wedge}{}_{L95}$	$x^{\wedge}{}_{U95}$	個別L95	個別U95
6	-0.2963	-0.6745	-0.0327	-1.2426	0.5354

通常の回帰分析でのパラメータの共分散行列を用いた逆推定の計算方法は，ポアソン回帰，ワイブル回帰のみならず，ロジスティック回帰による 10%あるいは 50%有効量の推定なども同様に適用できる汎用的な方法である．

非線形回帰を用いた逆推定値の 95%信頼区間の直接推定

回帰直線を推定し，得られたパラメータの推定値と共分散行列を用いて，ある y_0 に対する逆推定値 $\hat{x}_0$ を推定する方法を示してきた．通常の回帰分析は，次式

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{4.61}$$

で与えられる．ある y_0 に対する逆推定値 $\hat{x}_0$ を通る回帰式は，

$$Y_i - y_0 = \beta_1 (X_i - \hat{x}_0) + \varepsilon_i \tag{4.62}$$

で与えられる．

ある y_0 を観測値 Y_i の平均値とし，$\hat{x}_0$ を X_i の平均とすると式（4.62）は，重心を通る回帰式となり，パラメータは傾き β_1 のみとなる．切片 β_0 は，

$$\beta_0 = \bar{Y} + \beta_1 \bar{X} \tag{4.63}$$

で計算できる．ただし，ある y_0 に対する逆推定値 $\hat{x}_0$ は，回帰パラメータが推定されないと求めることができない．式（4.62）を

$$Y_i = y_0 + \beta_1 (X_i - \hat{x}_0) + \varepsilon_i \tag{4.64}$$

のように，ある y_0 を右辺に移項した式とする．この式を用いて，β_1 と $\hat{x}_0$ を同時に推定し，それらの分散も推定したい．

式（4.64）パラメータに関して偏微分すると，

$$\frac{\partial Y_i}{\partial \beta_1} = (X_i - \hat{x}_0), \qquad \frac{\partial Y_i}{\partial \hat{x}_0} = -\beta_1 \tag{4.65}$$

となり，互いに他のパラメータが残り，線形式ではなくなり通常の回帰分析で解くことができない．そのために，偏微分式を新たな変数とした反復回帰が必要となる［ドレーパ・スミス（1968），第 10 章 非線形推定序説］.

式（4.64）に対して JMP の非線形回帰を使えば，表 4.17 右に示した回帰の逆推定 $\hat{x}_0$ の正確な 95%信頼区間を直接推定することができる．詳しくは，芳賀（2010），医薬品開発のための統計解析，第 3 部，大和田（2010）線形モデルと非線形モデルの基本的な考え方－逆推定の解析，標準誤差と信頼限界－，中西（2016）じっくり勉強すれば身につく統計入門 第 12 回－非線形回帰を用いた逆推定の基礎－を参照のこと．

5. 反復重み付き最尤法によるポアソン回帰

ポアソン回帰を含む一般化線形モデルに対する最尤法には，2 種類の解法がある．伝統的には，重み付き回帰の反復によって最尤解を求める方法である．他方は，第 2 章で導入したニュートン・ラフソン法による最尤解を求める方法である．手計算の時代から反復重み付き回帰によって，2 値反応データから 50 パーセント致死量 LD_{50} を求める方法とし，プロビット法が知られている．コンピュータの性能向上により，2 階の偏微分行列を使うニュートン・ラフソン法による計算も可能となってきた．どちらでもパラメータの推定値は一致するが，パラメータの共分散行列には，若干の相違がある．反復重み付き最尤法は，パラメータ数が増えても対応が容易であり，Excel での計算には適している．

5.1. 反復重み付きポアソン回帰

第 1.4 節の人工データでポアソン回帰のパラメータ推定に対し，説明なしに反復重み付き回帰による最尤法を用いた．ポアソン回帰は，説明変数 X_i が大きくなるにつれ推定値 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ の分散が推定値 $\hat{Y}_i$ 比例して大きくなり，観測値 Y_i と回帰の推定値 $\hat{Y}_i$ との偏差 $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ に等分散性を仮定することができない場合に用いられる．

回帰パラメータの推定に際して観測値が，ポアソン分布に従う場合には，最小 2 乗法による回帰分析ではなく，ポアソン分布の確率関数を用いた最尤法を適用する必要がある．第 2 章で導入したポアソン分布の確率関数を用いる最尤法は，計算機の性能が低い時代には，容易ではなかった．そのために，偏差 $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ の平方を求める際に $\hat{Y}_i$ の分散で割って基準化した平方和を用いる重み付き回帰が考案された．

ただし，重み付き回帰には，厄介な問題が内在する．これは，基準化のために，何らかの方法で得られた回帰の推定値 $\hat{Y}_i^{(0)}$ の分散を用い，重み付き回帰を実施して新たな推定値 $\hat{Y}_i^{(1)}$ と分散を得ると $\hat{Y}_i^{(0)}$ の分散との間にわずかなブレが生じて一致しなくなる．そのため，さらに，推定値 $\hat{Y}_i^{(1)}$ の分散を用いて重み付き回帰を行い，推定値 $\hat{Y}_i^{(2)}$ 得る．といった反復を繰り返して，回帰パラメータのブレがなくなるまで，重み付き回帰を反復する必要が生ずる．

第 1.9 節で示した植物の種子数のデータの解析は，説明変数 X の増加に伴い反応 Y が指数関数的に増加する場合に対し，ポアソン分布の確率関数を用いたポアソン回帰を取り上げ，第 2.5 節で理論的側面を示した．このような現象に対し，反応 Y を対数変換し直線化して通常の回帰分析を行えばとの誘惑にかられる．しかし，厄介な問題が発生する．元のカウント・データにゼロがあると対数変換できない．元のデータが対数正規分布に従っていれば，対数変換後は正規分布になることは，良く知られている．しかし，元のデータがポアソン分布に従っていた場合に，対数変換後の分布は，どのような分布になるのかは知られていない．

これらの問題を解決するための方法とし，元データを対数変換するのではなく，仮の対数推定値 $\ln(\hat{Y}_i)$ に対し，後述する式（5.20）から，

$$Z_i = \ln(\hat{Y}_i) + \frac{\hat{Y}_i - \exp[\ln(\hat{Y}_i)]}{\exp[\ln(\hat{Y}_i)]}$$

のような巧妙な変換を施し，Z_i について反復重み付き回帰を適用する．$Y_i = 0$，$\ln(\hat{Y}_i) = -0.1$ の場合であれば，

$$\begin{aligned} Z_i &= \ln(\hat{Y}_i) + \frac{\hat{Y}_i - \exp[\ln(\hat{Y}_i)]}{\exp[\ln(\hat{Y}_i)]} \\ &= -0.1 + \frac{0 - 0.9048}{0.9048} = -1.1000 \end{aligned}$$

として重み付き回帰の対象とする「データ」とするのである．元のスケールでは

$$\exp(Z_i) = \exp(-1.10) = 0.3329$$

となり，元のゼロの値にげたをはかせ，対数変換したことになる．これは，ポアソン回帰だけではなく一般化線形モデルとしてのロジスティック回帰，プロビット回帰でも同様な変換が行なわれている．詳しくは，高橋（2017），「一般化線形モデルを Excel で極め活用する－プロビット法・ロジット法・補 2 重対数法－」を参照のこと

第 1.5 節，第 1.10 節，第 1.11 節では，ある地域で発生した死亡者数などのカウント・データの解析に際し，部分母集団の人数が得られる場合のデータを示した．年齢などで層別した場合にも，部分母集団の人数が得られるような場合に，その部分集団の人数に対し対数を取ってオフセットとしたポアソン回帰を示した．オフセットがある場合に対するポアソン回帰は，部分母集団の 1 人あたりのイベント発生数（発生率）をベースにしており，本章では，対数リンクでオフセットを考慮した反復重み付き回帰による最尤法による計算方法も示す．

5.2. 重み付き回帰の基礎

重み付き回帰では，偏差平方和の計算に際して分散$Var(\hat{Y}_i)$の逆数を掛けることで分散に関して基準化している．ポアソン分布は，平均と分散が等しいことから，得られた推定値$\hat{Y}_i$と分散が$Var(\hat{Y}_i)=\hat{Y}_i$と同じになる．さて，推定値$\hat{Y}_i$を得るためには，何らかの回帰分析が必要となるのだが，最初は「重み」となる$\hat{Y}_i$が求まっていないので重み付き回帰は実施できない．そのために通常の回帰分析を行い，仮の推定値 $\hat{Y}_i^{(0)}=\hat{\beta}_0^{(0)}+\hat{\beta}_1^{(0)}X_i$を求め，$\hat{Y}_i^{(0)}$の逆数を重みとした回帰を行う．

正規方程式

誤差がポアソン分布に従う回帰式を，

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i,\quad i=1,2,...,n,\quad \varepsilon_i\sim Poisson(Y_i) \tag{5.1}$$

または，

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim Poisson(\boldsymbol{Y}) \tag{5.2}$$

とする．推定されたパラメータを用いた $\hat{Y}_i^{(0)}=\hat{\beta}_0^{(0)}+\hat{\beta}_1^{(0)}X_i$ の推定値$\hat{Y}_i^{(0)}$に対して重みを分散の逆数$\hat{w}_i=1/\hat{Y}_i^{(0)}$としたときに，$\hat{Y}_i^{(0)}$からの偏差$\varepsilon_i$の重み付き平方和は，

$$Q=\sum_{i=1}^{n}\hat{w}_i\varepsilon_i^2=\sum_{i=1}^{n}\hat{w}_i(Y_i-\beta_0-\beta_1X_i)^2 \tag{5.3}$$

となる．式（5.3）をβ_0とβ_1で偏微分すると次式が得られる．

$$\left.\begin{aligned}\frac{\partial Q}{\partial\beta_0}&=-2\sum_{i=1}^{n}\hat{w}_i(Y_i-\beta_0-\beta_1X_i)\\ \frac{\partial Q}{\partial\beta_1}&=-2\sum_{i=1}^{n}\hat{w}_i(Y_i-\beta_0-\beta_1X_i)X_i\end{aligned}\right\} \tag{5.4}$$

式（5.4）を 0 と置くと，β_0とβ_1の推定値としての$\hat{\beta}_0$と$\hat{\beta}_1$が得られる．これらの偏微分した式を整理すると，

$$\left.\begin{aligned}\sum_{i=1}^{n}\hat{w}_iY_i-\hat{\beta}_0\sum_{i=1}^{n}\hat{w}_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{w}_iX_i&=0\\ \sum_{i=1}^{n}\hat{w}_iX_iY_i-\hat{\beta}_0\sum_{i=1}^{n}\hat{w}_iX_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{w}_iX_i^2&=0\end{aligned}\right\} \tag{5.5}$$

となる．式(5.5)で，$\hat{\beta}_0$と$\hat{\beta}_1$が含まれない項を右辺に移して整理すると，次の正規方程式

$$\left.\begin{aligned}\hat{\beta}_0\sum_{i=1}^{n}\hat{w}_i+\hat{\beta}_1\sum_{i=1}^{n}\hat{w}_iX_i&=\sum_{i=1}^{n}\hat{w}_iY_i\\ \hat{\beta}_0\sum_{i=1}^{n}\hat{w}_iX_i+\hat{\beta}_1\sum_{i=1}^{n}\hat{w}_iX_i^2&=\sum_{i=1}^{n}\hat{w}_iX_iY_i\end{aligned}\right\} \tag{5.6}$$

が得られる．なお，これらの式の展開について理解が不十分と思われた場合には，第 4 章を先に学習してもらいたい．

重みを含む行列計算

デザイン行列を用いた計算に際し，重み $\hat{w}_i$ をベクトル化した場合だと行列のサイズが合わないのでデザイン行列の積 $\boldsymbol{X}^T\boldsymbol{X}$ に組み込めない．そこで，対角要素を $\hat{W}_{ii}=\hat{w}_i$，それ以外は 0 とする $n\times n$ の行列 $\boldsymbol{W}$ を定義する．これを，$\boldsymbol{X}^T\boldsymbol{X}$ の中に入れ込み

$$\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X} \tag{5.7}$$

とすると，行列のそれぞれの内側のサイズが一致する．表 1.6 の人工データのデータ数は $n=9$ なので，9×9 の行列として例示する．重み行列 $\hat{\boldsymbol{W}}$ は，9×9 の行列であり，$\boldsymbol{X}^T\boldsymbol{X}$ の中に入れ込むと，内側のそれぞれの行列のサイズが一致する．

	$\boldsymbol{X}^T$										$\boldsymbol{W}\hat{}$									$\boldsymbol{X}$	
																				$\boldsymbol{X}_0$	$\boldsymbol{X}_1$
$\boldsymbol{X}_0^T$	1	1	1	1	1	1	1	1	1		$w_1\hat{}$	0	0	0	0	0	0	0	0	1	X_1
$\boldsymbol{X}_1^T$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9		0	$w_2\hat{}$	0	0	0	0	0	0	0	1	X_2
											0	0	$w_3\hat{}$	0	0	0	0	0	0	1	X_3
											0	0	0	$w_4\hat{}$	0	0	0	0	0	1	X_4
											0	0	0	0	$w_5\hat{}$	0	0	0	0	1	X_5
											0	0	0	0	0	$w_6\hat{}$	0	0	0	1	X_6
											0	0	0	0	0	0	$w_7\hat{}$	0	0	1	X_7
											0	0	0	0	0	0	0	$w_8\hat{}$	0	1	X_8
											0	0	0	0	0	0	0	0	$w_9\hat{}$	1	X_9
				2×9											9×9					9×2	

$\boldsymbol{X}^T$ と $\hat{\boldsymbol{W}}$ の積は，2×9 の行列となり，さらに $\boldsymbol{X}$ との積は，2×2 の行列となり，正規方程式（5.6）の左辺と同様の形式となる．

					$\boldsymbol{X}^T\boldsymbol{W}\hat{}$					$\boldsymbol{X}$			
										$\boldsymbol{X}_0$	$\boldsymbol{X}_1$		$\boldsymbol{X}^T\boldsymbol{W}\hat{}\boldsymbol{X}$
$\boldsymbol{X}_0^T\boldsymbol{w}\hat{}$	$w_1\hat{}$	$w_2\hat{}$	$w_3\hat{}$	$w_4\hat{}$	$w_5\hat{}$	$w_6\hat{}$	$w_7\hat{}$	$w_8\hat{}$	$w_9\hat{}$	1	X_1	=	$\Sigma w_i\hat{}$ $\Sigma w_i\hat{}X_i$
$\boldsymbol{X}_1^T\boldsymbol{w}\hat{}$	$w_1\hat{}X_1$	$w_2\hat{}X_2$	$w_3\hat{}X_3$	$w_4\hat{}X_4$	$w_5\hat{}X_5$	$w_6\hat{}X_6$	$w_7\hat{}X_7$	$w_8\hat{}X_8$	$w_9\hat{}X_9$	1	X_2		$\Sigma w_i\hat{}X_i$ $\Sigma w_i\hat{}X_i^2$
										1	X_3		
										1	X_4		
										1	X_5		
										1	X_6		
										1	X_7		
										1	X_8		
										1	X_9		
					2×9					9×2			2×2

2×2 の $\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ 行列に回帰の推定値 $\hat{\boldsymbol{\beta}}$ を掛けシグマで表記すると，

$\boldsymbol{X}^T\boldsymbol{W}\hat{}\boldsymbol{X}$			$\boldsymbol{\beta}\hat{}$		
$\Sigma w_i\hat{}$	$\Sigma w_i\hat{}X_i$		$\beta_0\hat{}$	=	$\hat{\beta}_0\Sigma\hat{w}_i + \hat{\beta}_1\Sigma\hat{w}_iX_i$
$\Sigma w_i\hat{}X_i$	$\Sigma w_i\hat{}X_i^2$		$\beta_1\hat{}$		$\hat{\beta}_0\Sigma\hat{w}_iX_i + \hat{\beta}_1\Sigma\hat{w}_iX_i^2$

となり，正規方程式（5.6）の左辺に等しくなる．

$\boldsymbol{X}^T\hat{\boldsymbol{W}}$ は既に計算しているので，$\boldsymbol{X}^T\hat{\boldsymbol{W}}$ と $\boldsymbol{Y}$ との積は，2×1 のベクトルとなり，回帰パラメータの推定値正規方程式（5.6）の右辺に等しくなる．

				$\boldsymbol{X}^T\boldsymbol{W}\hat{}$						$\boldsymbol{Y}$		$\boldsymbol{X}^T\boldsymbol{W}\hat{}\boldsymbol{Y}$
$\boldsymbol{X}_0{}^T\boldsymbol{w}\hat{}$	$w_1\hat{}$	$w_2\hat{}$	$w_3\hat{}$	$w_4\hat{}$	$w_5\hat{}$	$w_6\hat{}$	$w_7\hat{}$	$w_8\hat{}$	$w_9\hat{}$	Y_1	=	$\Sigma w_i\hat{}X_i$
$\boldsymbol{X}_1{}^T\boldsymbol{w}\hat{}$	$w_1\hat{}X_1$	$w_2\hat{}X_2$	$w_3\hat{}X_3$	$w_4\hat{}X_4$	$w_5\hat{}X_5$	$w_6\hat{}X_6$	$w_7\hat{}X_7$	$w_8\hat{}X_8$	$w_9\hat{}X_9$	Y_2		$\Sigma w_i\hat{}X_iY_i$
										Y_3		
										Y_4		
										Y_5		
										Y_6		
										Y_7		
										Y_8		
										Y_9		
					2×9					9×1		2×1

デザイン行列を用いた計算結果が，正規方程式（5.6）の左辺と右辺に等しいことから，

$$(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Y} \tag{5.8}$$

$\boldsymbol{X}^T\boldsymbol{W}\hat{}\boldsymbol{X}$		$\boldsymbol{\beta}\hat{}$		$\boldsymbol{X}^T\boldsymbol{W}\hat{}\boldsymbol{Y}$
$\Sigma w_i\hat{}$	$\Sigma w_i\hat{}X_i$	$\beta_0\hat{}$	=	$\Sigma w_i\hat{}Y_i$
$\Sigma w_i\hat{}X_i$	$\Sigma w_i\hat{}X_i^2$	$\beta_1\hat{}$		$\Sigma w_i\hat{}X_iY_i$

となる．逆行列 $(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ を両辺にかけて

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Y} \tag{5.9}$$

が得られる．重みが全て 1 の場合は，$\boldsymbol{W}$ が消えて，通常の回帰式

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{5.10}$$

となる．

5.3. 恒等リンクの場合のポアソン回帰

初期パラメータの推定

第 1.4 節でドブソン（2008）の人工データについて詳細を示さず，反復重み付き回帰を用いた結果を示した．第 5.2 節で示した Excel で重み付き回帰のために必要な行列計算の方法を用いて，反復計算の詳細を示す．表 5.1 に示すように，重みなしの回帰パラメータ $\hat{\boldsymbol{\beta}}^{(0)}$ を推定し，推定値 $\hat{\boldsymbol{Y}}^{(0)}$ を求め，重みを計算する．これは，前節の表 4.2 と同様の計算結果で，重みの計算が追加されている．なお，第 1.4 節では，ここに示した回帰係数を初期値としては用いずに，ドブソン（2008）で示されている $\hat{\boldsymbol{\beta}}^{(0)}=[7.0 \quad 5.0]^T$ を使っている．

表 5.1　反復(0)の回帰

	デザイン行列			回帰					
	$\boldsymbol{X}$			推定値$^{(0)}$	重み				
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	$\boldsymbol{Y}$^	$\boldsymbol{w}$^$=1/\boldsymbol{Y}$^				
1	1	-1	2	2.5455	0.3929	$\boldsymbol{X}^T\boldsymbol{X}=$	9.0000	1.0000	
2	1	-1	3	2.5455	0.3929		1.0000	5.0000	
3	1	0	6	7.4545	0.1341				
4	1	0	7	7.4545	0.1341	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}=$	0.1136	-0.0227	
5	1	0	8	7.4545	0.1341		-0.0227	0.2045	
6	1	0	9	7.4545	0.1341				
7	1	1	10	12.3636	0.0809	$\boldsymbol{X}^T\boldsymbol{Y}=$	72.0000	β_0^$=$	**7.4545**
8	1	1	12	12.3636	0.0809		32.0000	β_1^$=$	**4.9091**
9	1	1	15	12.3636	0.0809				$\boldsymbol{\beta}^{(0)}$^
				$\boldsymbol{X\beta}^{(0)}$^					

推定値 $\hat{\boldsymbol{\beta}}^{(0)}$ を次式で求め，

7.4545	=	0.1136	-0.0227	72.0000
4.9091		-0.0227	0.2045	32.0000
$\boldsymbol{\beta}^{(0)}$^		$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$		$\boldsymbol{X}^T\boldsymbol{Y}$

推定値 $\hat{\boldsymbol{Y}}^{(0)}=\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)}$ を得る．さらに，重みを列ベクトル $\hat{\boldsymbol{w}}^{(0)}=1/\hat{\boldsymbol{Y}}^{(0)}$（この演算子「/」は，セル同士の除算）として計算する．重みは，$x_{1,1}=-1$ の場合に $\hat{w}_1=1/2.5455=0.3929$，$x_{1,3}=0$ の場合に $\hat{w}_3=1/7.4545=0.1341$，$x_{1,7}=1$ の場合に $\hat{w}_7=1/12.3636=0.0809$ と小さくなっている．

重み付き回帰

表 5.1 では，重みを列ベクトル $\hat{\boldsymbol{w}}$ として計算しているので，これを対角要素とするマトリックス $\hat{\boldsymbol{W}}$ とする必要がある．ただし，Excel の行列関数に列ベクトルを対角行列とする行列関数がないので，シート上に 9×9 の行列の枠を作成し，その対角要素に重みを手作業で代入する必要がある．この方法は，データ数 n の変化に対し煩雑な操作を必要とするので，代替の計算手段が必要となる．

$\boldsymbol{w}$^		=Diag($\boldsymbol{w}$^) Excelにはこの様な対角化の関数がない								
0.3929	➔	**0.3929**	0	0	0	0	0	0	0	0
0.3929		0	**0.3929**	0	0	0	0	0	0	0
0.1341		0	0	**0.1341**	0	0	0	0	0	0
0.1341		0	0	0	**0.1341**	0	0	0	0	0
0.1341		0	0	0	0	**0.1341**	0	0	0	0
0.1341		0	0	0	0	0	**0.1341**	0	0	0
0.0809		0	0	0	0	0	0	**0.0809**	0	0
0.0809		0	0	0	0	0	0	0	**0.0809**	0
0.0809		0	0	0	0	0	0	0	0	**0.0809**

重みを使う行列計算は，回帰式の

$$\hat{\boldsymbol{\beta}}^{(1)} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Y}$$

の $\boldsymbol{X}^T\hat{\boldsymbol{W}}$ の項である．この結果と同値となる代替の行列計算として，重みを列ベクトル $\hat{\boldsymbol{w}}$ としたときに，デザイン行列と重みベクトルとのセル同士の積の演算子「*」を用いた $\boldsymbol{X}*\hat{\boldsymbol{w}}$

$$\boldsymbol{X}^T\hat{\boldsymbol{W}} = (\boldsymbol{X}*\hat{\boldsymbol{w}})^T$$

を転置することにより実現できる．実際には，次のような結果となる．

$\boldsymbol{X}$			$\boldsymbol{w}$		$\boldsymbol{X}*\boldsymbol{w}$	
$\boldsymbol{X}_0$	$\boldsymbol{X}_1$					
1	X_1	*	w_1^	=	w_1^	w_1^ X_1
1	X_2		w_2^		w_2^	w_2^ X_2
1	X_3		w_3^		w_3^	w_3^ X_3
1	X_4		w_4^		w_4^	w_4^ X_4
1	X_5		w_5^		w_5^	w_5^ X_5
1	X_6		w_6^		w_6^	w_6^ X_6
1	X_7		w_7^		w_7^	w_7^ X_7
1	X_8		w_8^		w_8^	w_8^ X_8
1	X_9		w_9^		w_9^	w_9^ X_9
9×2			9×1		9×2	

転置

$\boldsymbol{X}^T\boldsymbol{W} = (\boldsymbol{X}*\boldsymbol{w})^T$

w_1^	w_2^	w_3^	w_4^	w_5^	w_6^	w_7^	w_8^	w_9^
w_1^X_1	w_2^X_2	w_3^X_3	w_4^X_4	w_5^X_5	w_6^X_6	w_7^X_7	w_8^X_8	w_9^X_9
				2×9				

表 5.2 に，セル同士の積「*」の演算子を用いた重み付き回帰の結果を示す．表の左側は，反復(0)の回帰であり，恒等「リンク関数」は $\boldsymbol{Z} = \boldsymbol{Y}$ なので，

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(1)} &= (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Z} \\ &= [(\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1}(\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{Z}\end{aligned}$$

により，重み付き回帰のパラメータを推定する．表の下段には，行列計算の結果が表示されている．回帰パラメータ $\hat{\boldsymbol{\beta}}^{(1)}$ は，Excel の Mmult()関数を用いて

$$\hat{\boldsymbol{\beta}}^{(1)} = \text{Mmult}(\ ((\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{X})^{-1}\text{の範囲},\ (\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{Z}\text{の範囲}) = \begin{array}{|c|}\hline \mathbf{7.4517} \\ \mathbf{4.9351} \\ \hline \end{array}$$

$\boldsymbol{\beta}^{(1)}$^

が計算されている．

表 5.2　反復が(1)の重み付き回帰

	デザイン行列			回帰		リンク	重み付回帰	推定値
	$\boldsymbol{X}$			推定値	重み	関数	推定値	差
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	$\boldsymbol{Y}$^	$\boldsymbol{w}$^= 1/$\boldsymbol{Y}$^	$\boldsymbol{Z}=\boldsymbol{Y}$	$\boldsymbol{Z}$^	$\boldsymbol{Y}$^ - $\boldsymbol{Z}$^
1	1	-1	2	2.5455	0.3929	2.00	2.5165	0.0289
2	1	-1	3	2.5455	0.3929	3.00	2.5165	0.0289
3	1	0	6	7.4545	0.1341	6.00	7.4517	0.0029
4	1	0	7	7.4545	0.1341	7.00	7.4517	0.0029
5	1	0	8	7.4545	0.1341	8.00	7.4517	0.0029
6	1	0	9	7.4545	0.1341	9.00	7.4517	0.0029
7	1	1	10	12.3636	0.0809	10.00	12.3868	-0.0232
8	1	1	12	12.3636	0.0809	12.00	12.3868	-0.0232
9	1	1	15	12.3636	0.0809	15.00	12.3868	-0.0232
		β_0^=	7.4545	$\boldsymbol{X\beta}^{(0)}$^	β_0^=	**7.4517**	$\boldsymbol{X\beta}^{(1)}$^	**0.1389**
		β_1^=	4.9091		β_1^=	**4.9351**		絶対値の和
			$\boldsymbol{\beta}^{(0)}$^			$\boldsymbol{\beta}^{(1)}$^		
				1.5649	-0.5431	0.7824	0.4132	8.9813
				-0.5431	1.0284	0.4132	1.1906	1.0284
				$(\boldsymbol{X}^*\boldsymbol{w}^\wedge)^T\boldsymbol{X}$		$[(\boldsymbol{X}^*\boldsymbol{w}^\wedge)^T\boldsymbol{X}]^{-1}$		$(\boldsymbol{X}^*\boldsymbol{w}^\wedge)^T\boldsymbol{Z}$

さらに，$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(1)}$ を計算し，次いで推定値間の差 $(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Z}})$ から，その絶対値の和を求め，乖離の度合を絶対偏差の和で評価すると，

0.1389
絶対値の和

が得られる．推定されたパラメータ間の差は，小数点以下 3 桁目で異なっている．

7.4517	−	7.4545	=	-0.0029
4.9351		4.9091		0.0261
$\boldsymbol{\beta}^{(1)}$^		$\boldsymbol{\beta}^{(0)}$^		差

反復重み付き回帰（2）

表 5.2 で示した通常の回帰分析に引き続き，重み付き回帰で得られたパラメータは，初期パラメータに対し，推定値の差の絶対値の和は 0.1389 とかなり大きい．表 5.3 に示すように，第 2 回目の重み付き回帰を繰り返してみる．反復が(1)の重み付き回帰で推定されたパラメータをコピーし，$\hat{\boldsymbol{\beta}}^{(1)}$ には計算式が入っているので値のみをペースト

第 1 反復：	β_0^=	7.4545	$\boldsymbol{X\beta}^{(0)}$^	β_0^=	**7.4517**	$\boldsymbol{Z\beta}^{(1)}$^	**0.1389**
	β_1^=	4.9091		β_1^=	**4.9351**		絶対値の和
		$\boldsymbol{\beta}^{(0)}$^			$\boldsymbol{\beta}^{(1)}$^		

すると，

第 2 反復：								
	β_0^=	7.4517	$X\beta^{(1)}$^		β_0^=	**7.4516**	$Z\beta^{(2)}$^	**0.0008**
	β_1^=	4.9351			β_1^=	**4.9353**		絶対値の和
		$\beta^{(1)}$^				$\beta^{(2)}$^		

$\hat{\boldsymbol{\beta}}^{(2)}$ が自動的に計算される．表 5.3 に示すように絶対値の和は，0.0008 とかなり小さくなり，パラメータも少数点以下 3 桁まで一致している．

表 5.3　反復が (2) の重み付き回帰

	デザイン行列			回帰		リンク	重み付回帰	推定値
	X			推定値	重み	関数	推定値	差
i	X_0	X_1	Y	Y^	w^=1/Y^	$Z=Y$	Z^	Y^ - Z^
1	1	-1	2	2.5165	0.3974	2.00	2.5163	0.0002
2	1	-1	3	2.5165	0.3974	3.00	2.5163	0.0002
3	1	0	6	7.4517	0.1342	6.00	7.4516	0.0000
:								
9	1	1	15	12.3868	0.0807	15.00	12.3869	-0.0001
		β_0^=	7.4517	$X\beta^{(1)}$^	β_0^=	**7.4516**	$X\beta^{(2)}$^	**0.0008**
		β_1^=	4.9351		β_1^=	**4.9353**		絶対値の和
			$\beta^{(1)}$^			$\beta^{(2)}$^		
				1.5737	-0.5526	0.7817	0.4165	8.9999
				-0.5526	1.0369	0.4165	1.1863	1.0002
				$(X*w^)^T X$		$[(X*w^)^T X]^{-1}$		$(X*w^)^T Z$

反復重み付き回帰（3）

表 5.4 に示すように，反復が (3) の重み付き回帰の結果は，絶対値の和が，0.0000 となり，$\hat{\boldsymbol{\beta}}^{(2)}$ と $\hat{\boldsymbol{\beta}}^{(3)}$ は，少数点以下 4 桁まで等しくなったので，解が求まったとみなす．反復を繰り返せば更に精度の高い推定値を得ることができる．

表 5.4　反復が (3) の重み付き回帰

	デザイン行列			回帰		リンク	重み付回帰	推定値
	X			推定値	重み	関数	推定値	差
i	X_0	X_1	Y	Y^	w^=1/Y^	$Z=Y$	Z^	Y^ - Z^
1	1	-1	2	2.5163	0.3974	2.00	2.5163	0.0000
2	1	-1	3	2.5163	0.3974	3.00	2.5163	0.0000
3	1	0	6	7.4516	0.1342	6.00	7.4516	0.0000
:								
9	1	1	15	12.3869	0.0807	15.00	12.3869	0.0000
		β_0^=	7.4516	$X\beta^{(2)}$^	β_0^=	**7.4516**	$X\beta^{(3)}$^	**0.0000**
		β_1^=	4.9353		β_1^=	**4.9353**		絶対値の和
			$\beta^{(2)}$^			$\beta^{(3)}$^		
				1.5738	-0.5526	0.7817	0.4166	9.0000
				-0.5526	1.0370	0.4166	1.1863	1.0000
				$(X*w)^T X$		$[(X*w)^T X]^{-1}$		$(X*w)^T Z$

なお，反復(0)として表 5.1 に示した回帰分析を行うことは，煩わしい．ここに示した手作業の場合には，最初の初期値として $\hat{\boldsymbol{Y}}$ が 0 にならないような初期値 $\hat{\boldsymbol{\beta}}^{(0)}=[\ 2.0\quad 1.0\ ]^T$ などを与えても，反復計算が 1 回ほど多くなるだけで表 5.4 と同様な解を求めることができる．

これまでに示した結果は，第 1.4 節に対する補足説明となっているので，結果の解釈，共分散行列を用いた 95%信頼区間の作図などを，合わせて見てほしい．なお，第 1.4 節では，$\hat{\boldsymbol{\beta}}^{(0)}=[\ 7.0\quad 5.0\ ]^T$ と異なる初期値を与えたが，収束結果は当然ながら一致している．

回帰パラメータについてのワルド検定

反復重み付き回帰によるポアソン回帰では，パラメータに関する共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ が，反復計算の過程で $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=[(\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1}$ として

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=\begin{array}{|cc|}\hline 0.7817 & 0.4166\\ 0.4166 & 1.1863\\ \hline\end{array}\quad [(\boldsymbol{X}*\boldsymbol{w}\hat{})^T\boldsymbol{X}]^{-1} \tag{5.11}$$

求まっている．パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の対角要素が，$\hat{\beta}_0$ と $\hat{\beta}_1$ の分散なので，それぞれの分散の平方根で標準誤差 SE を計算し，

$$\text{Wald カイ 2 乗}=\left(\frac{\text{推定値}}{SE}\right)^2 \tag{5.12}$$

により Wald カイ 2 乗検定統計量を求められる．これが，自由度 1 のカイ 2 乗分布に従うことから，p 値を求めることができる．表 5.5 に示したように，$\hat{\beta}_1$ の分散は，1.1863 なので，

$$\chi^2_{\hat{\beta}_1}=\left(\frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}}\right)^2=\left(\frac{4.9353}{\sqrt{1.1892}}\right)^2=20.5320$$

となる．

表 5.5　反復重み付き回帰による Wald 検定

項	推定値	分散	標準誤差	Waldカイ2乗	p 値
X_0	7.4516	0.7817	0.8841	71.0357	0.0000
X_1	4.9353	1.1863	1.0892	20.5320	0.0000

第 2.4 節では，反復重付き回帰ではなく，ニュートン・ラフソンによる解析法を示し，計算の過程で求められている負のヘッセ行列の逆数 $(-\boldsymbol{H})^{-1}$ がパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ とな

ることを示した．第 2.4 節では，尤度比カイ 2 乗検定の結果を示したが，Wald カイ 2 乗検定を行うことも可能である．表 2.15 に示したパラメータの共分散行列 $(-\boldsymbol{H})^{-1}$ を次に示す．

$$(-\boldsymbol{H})^{-1} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) =$$

負の逆行列 $(-H)^{-1}$	
0.7817	0.4160
0.4160	1.1915

(5.13)

共分散行列といっても式（5.13）と式（5.11）で示したのでは，若干の差異がある．その結果として，表 5.6 に示した共分散行列 $(-\boldsymbol{H})^{-1}$ の対角要素を分散とした Wald 検定結果も表 5.5 のカイ 2 乗値と若干の差異が生ずる．

表 5.6　負のヘッセ行列の逆行列 $(-\boldsymbol{H})^{-1}$ を用いた Wald 検定

項	推定値	分散	標準誤差	Waldカイ2乗	*p* 値
X_0	7.4516	0.7817	**0.8842**	71.0300	0.0000
X_1	4.9353	1.1915	**1.0915**	20.4428	0.0000

尤度比検定

第 2.4 節の表 2.17 で示した JMP によるポアソン回帰は，尤度比検定を標準的に出力するので，Wald 検定とは若干の差異が生ずる．表 2.17 を表 5.7 として再掲する．表 5.5 に示した反復重み付きによる $\hat{\beta}_1$ のカイ 2 乗値が 20.5320 であるのに対し，尤度比カイ 2 乗の結果は 16.5260 とかなり異なる．

表 5.7　JMP によるポアソン回帰の結果（表 2.17 再掲）

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	7.4516	0.8842	71.0299	<.0001*
x	4.9353	1.0915	16.5260	<.0001*

推定値の共分散

共分散

	切片	x
切片	0.7817	0.4160
x	0.4160	1.1915

正規分布を仮定した最小 2 乗法による回帰分析では，各種の統計ソフトによって結果が異なることはなくなったが，一般化線形モデルとしてのポアソン回帰では，SAS の GENMODE プロシジャの解析手法と JMP の一般線形モデルに対する最尤法の解析方法が異なるために，若干の差異が生ずる．従って，解析結果を公表する際には，使用した統計ソフトが用いている最尤法の計算方法，さらに，尤度比検定なのかワルド検定なのかを明示する必要がある．

5.4. 対数リンクでのポアソン回帰

第 1.5 節で取り上げた冠動脈心疾患の死亡者数は，オフセットがある対数リンクの事例であるが，オフセットを無視して，表 5.8 に示すようにオフセットなしの対数リンクの事例として取り上げる［ドブソン（2008)］.

対数リンク

リンク関数が「恒等」の場合に対し「対数リンク」の場合は，ポアソン回帰の特徴的な反復重み付き回帰となる．死亡者数 y_i が，指数関数的に増加するので，

$$Y_i = \exp(\beta_0 + \beta_1 X_i) + \varepsilon_i, \quad \varepsilon_i \sim Poisson(\hat{y}_i) \tag{5.14}$$

をあてはめている．この式に対して両辺に対数を取ると誤差 ε_i が和の形で入っているために線形化できない.

$$\ln Y_i = \ln[\exp(\beta_0 + \beta_1 X_i) + \varepsilon_i] \tag{5.15}$$

そこで，回帰係数を推定値で表記することにより，

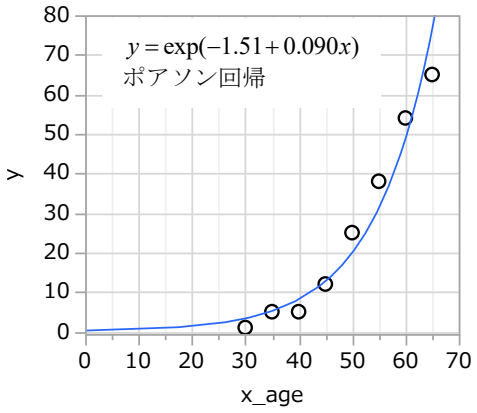
$$\ln \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{5.16}$$

と，線形化でき，反復重み付き回帰を行なう．このために指数ではなく，対数リンクと言われている.

さて，元のデータ y_i は，カウント・データなので $y_i = 0$ も発生する．ゼロについて対数はマイナスの無限大となり，計算不能となってデータに含められない．どうしたら良いのだろうか．便宜的には，全データに 1 を加えて対数を取ることも考えられるし，ゼロの場合に 0.5 に置き換えることも考えられる.

表 5.8　オーストラリアのある地方の冠動脈心疾患の死亡者数

	年齢層	死亡者数
No.	X	Y
1	30	1
2	35	5
3	40	5
4	45	12
5	50	25
6	55	38
7	60	54
8	65	65



一般化線形モデルは，ポアソン回帰の対数リンクのみならず，この様なリンク関数によっては，変換不能となるような場合でも，元の反応データのまま適切に対応できる計算アルゴリズムとなっている．表 5.9 に示すように年齢層が 30 歳代の死亡者数が $Y_1 = 0$ 人であると仮定した場合に，便宜的に $Y_1 = 0.5$ 人とおいて，対数変換すると，

$$\ln Y_1 = \ln(0.5) = -0.6931$$

を得る．他のデータについても対数を取り，回帰係数

$$\ln \hat{Y}_i = -3.4664 + 0.1257 X_{1,i}$$

を求めることは可能であるが，便宜的な対応と言わざるを得ない．

表 5.9　対数変換できない場合の便宜的な対応

	デザイン行列 $\boldsymbol{X}$		死亡者数	対数変換値	対数予測値				
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	ln$\boldsymbol{Y}$	ln$\boldsymbol{Y}$^				
1	1	30	**0.5**	**-0.6931**	**0.3043**	$\boldsymbol{X}^T\boldsymbol{X}$=	8.0	380.0	
2	1	35	5	1.6094	0.9327		380.0	19100.0	
3	1	40	5	1.6094	1.5611				
4	1	45	12	2.4849	2.1896	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$=	2.2738	-0.0452	
5	1	50	25	3.2189	2.8180		-0.0452	0.0010	
6	1	55	38	3.6376	3.4465				$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\ln\boldsymbol{Y})$
7	1	60	54	3.9890	4.0749	$\boldsymbol{X}^T\ln\boldsymbol{Y}$=	20.0305	β_0^ =	**-3.4664**
8	1	65	65	4.1744	4.7033		1083.42	β_1^ =	**0.1257**

一般化線形モデルでの反復重み付き回帰のアルゴリズムでは，何らかの最初の初期パラメータを得た後に，便宜的な対応ではなく，元のスケールのデータを変換する際に「補正式」を用いている．

対数リンクの場合，$Y_i = 0$ としたときの場合について例示する．補正式は，一般化線形モデルの公式により，

$$Z_i = \ln \hat{Y}_i + \frac{Y_i - \exp(\ln \hat{Y}_i)}{\exp(\ln \hat{Y}_i)} \qquad (5.17)$$

であり，何らかの初期パラメータによって，$\hat{Y}_i > 0$ が推定された場合に，補正値 Z_i は，$Y_i = 0$ の場合でも計算不能とならない．表 5.9 で推定された $\hat{\beta}_0 = -3.4664$，$\hat{\beta}_1 = 0.1257$ の場合に，推定値は，

$$\begin{aligned} \ln \hat{Y}_1 &= -3.4664 + 0.1257 X_{1,1} \\ &= -3.4664 + 0.1257 \times 30 = 0.3043 \end{aligned}$$

となり，

$$
\begin{aligned}
Z_i &= \ln \hat{Y}_1 + \frac{Y_1 - \exp(\ln \hat{Y}_1)}{\exp(\ln \hat{Y}_1)} \\
&= 0.3043 + \frac{0 - 0.3043}{0.3043} \\
&= -0.6957
\end{aligned}
$$

これを元のスケールに直すと

$$
Y_i' = \exp(-0.6957) = 0.4987
$$

おおよそ，0.5 程度の値となる．他の Z_i，$i = 2,3,...,9$ についても同じ補正式で調整して，この調整値 Z_i について重み付き回帰分析を行う．

式（5.17）は，リンク関数が「対数」の場合の一般化線形モデルの公式から与えられるとしたのであるが，他のリンク関数の場合を含む一般公式を用いて導出する．対数リンク式を

$$
\eta_i = \ln \hat{Y}_i \tag{5.18}
$$

とし，元の推定値を

$$
\left.
\begin{aligned}
\mu_i &= \exp(\ln \hat{Y}_i) \\
\ln \mu_i &= \ln \hat{Y}_i
\end{aligned}
\right\} \tag{5.19}
$$

としたときに，一般線形モデルの公式により，補正した対数 Z_i は，

$$
\left.
\begin{aligned}
Z_i &= \eta_i + (Y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \\
&= \eta_i + (Y_i - \mu_i)\left(\frac{\partial \ln(\mu_i)}{\partial \mu_i}\right) \\
&= \eta_i + \frac{(Y_i - \mu_i)}{\mu_i} \\
&= \ln \hat{Y}_i + \frac{Y_i - \exp(\ln \hat{Y}_i)}{\exp(\ln \hat{Y}_i)}
\end{aligned}
\right\} \tag{5.20}
$$

となる．重みは，

$$
\left.
\begin{aligned}
\hat{w}_i &= \frac{1}{Var(\hat{Y}_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \\
&= \frac{1}{Var(\hat{Y}_i)\left(\frac{\partial \ln \mu_i}{\partial \mu_i}\right)^2} \\
&= \frac{1}{\frac{\mu_i}{\mu_i^2}} = \mu_i = \exp(\eta_i) = \exp(\ln \hat{Y}_i)
\end{aligned}
\right\} \tag{5.21}
$$

となる．この公式は，リンク関数が「対数」のみならず，「プロビット」，「ロジット」，「補 2 重対数」などでも使われる．

重み付き回帰

さて，表 5.10 に示すように $y_1 = 0.5$ を $y_1 = 1$ と元のデータに戻し，回帰係数を推定し直すと初期値が，

$$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \ln(\boldsymbol{Y}) = \begin{bmatrix} -2.8310 \\ 0.1141 \end{bmatrix}$$

として得られる．この推定値を用いて反復重み付き回帰を行う．得られた初期値 $\hat{\boldsymbol{\beta}}^{(0)}$ を $\hat{\boldsymbol{\beta}}^{(m-1)}$ の欄に値のみペーストすると，反復 1 の結果が $\hat{\boldsymbol{\beta}}^{(m)}$ の欄に求められている．なお，初期パラメータ $\hat{\boldsymbol{\beta}}^{(0)}$ の推定に際し，$Y_i = 0$ がある場合には，欠測値として扱ってもなんら差し支えない．

表 5.10　＜反復 1＞冠動脈心疾患の死亡者数（対数リンク）

	デザイン行列		死亡	回帰		補正リンク	重付回帰	推定値
	$\boldsymbol{X}$		者数	推定値	重み	関数	推定値	差
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	$\ln\boldsymbol{Y}$^	$\boldsymbol{w}$^=$\boldsymbol{Y}$^	$\boldsymbol{Z}$	$\boldsymbol{Z}$^	$\ln\boldsymbol{Y}$^ - $\boldsymbol{Z}$^
1	1	30	1	0.5931	1.8096	0.1457	1.2472	-0.6541
2	1	35	5	1.1638	3.2019	1.7253	1.6903	-0.5266
3	1	40	5	1.7344	5.6657	1.6169	2.1335	-0.3991
4	1	45	12	2.3051	10.0253	2.5021	2.5767	-0.2716
5	1	50	25	2.8758	17.7394	3.2851	3.0199	-0.1441
6	1	55	38	3.4465	31.3893	3.6571	3.4631	-0.0166
7	1	60	54	4.0171	55.5423	3.9894	3.9063	0.1109
8	1	65	65	4.5878	98.2800	4.2492	4.3495	0.2384
β_0^=	-2.8310	β_0^=	**-2.8310**		β_0^=	**-1.4120**	$\boldsymbol{X\beta}^{(m)}$^	**2.3613**
β_1^=	0.1141	β_1^=	**0.1141**		β_1^=	**0.0886**	絶対値の和	
$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \ln(\boldsymbol{Y})$			$\boldsymbol{\beta}^{(m-1)}$^			$\boldsymbol{\beta}^{(m)}$^		
				223.7	13178.2	0.2734	-0.0046	852.29
				13178.2	789404.0	-0.0046	0.0001	51363.34
				$(\boldsymbol{X}*\boldsymbol{w}$^$)^T\boldsymbol{X}$		$[(\boldsymbol{X}*\boldsymbol{w}$^$)^T\boldsymbol{X}]^{-1}$		$(\boldsymbol{X}*\boldsymbol{w}$^$)^T\boldsymbol{Z}$

Excel シートに埋め込まれている計算式は，次の通りである．

1）　$\ln\hat{\boldsymbol{Y}} = \mathrm{Mmult}(\boldsymbol{X}$の範囲, $\hat{\boldsymbol{\beta}}^{(m-1=0)}$の範囲）

2）　$\hat{\boldsymbol{w}} = \exp(\ln\hat{\boldsymbol{Y}}$の範囲$)) = \hat{\boldsymbol{Y}}$

3）　$\boldsymbol{Z} = \ln\hat{\boldsymbol{Y}} + (\boldsymbol{Y} - \exp(\ln\hat{\boldsymbol{Y}})) / \exp(\ln\hat{\boldsymbol{Y}})$　：　以下「**の範囲**」は省略

4）　$(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X} = \mathrm{Mmult}(\mathrm{Transpose}(\boldsymbol{X} * \hat{\boldsymbol{w}}), \boldsymbol{X})$

5）　$[(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1} = \mathrm{Minvers}((\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X})$

6）　$(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{Z} = \mathrm{Mmult}(\mathrm{Transpose}(\boldsymbol{X} * \hat{\boldsymbol{w}}), \boldsymbol{Z})$

7）　$\hat{\boldsymbol{\beta}}^{(1)} = [(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1}(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{Z} = \mathrm{Mmult}([(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1}, (\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{Z})$

8）　$\hat{\boldsymbol{Z}} = \mathrm{Mmult}(\boldsymbol{X}, \hat{\boldsymbol{\beta}}^{(m=1)})$

9）　$\ln\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Z}} = \ln\hat{\boldsymbol{Y}}$の範囲 $- \hat{\boldsymbol{Z}}$の範囲

10）　絶対値の和=Sum(Abs($\ln\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Z}}$)）

反復重み付き回帰（2）および（3）

第 1 反復では，「$\ln \hat{\boldsymbol{Y}} - \hat{\boldsymbol{Z}}$ の絶対値の和」が，2.3613 と大きいので，第 2 反復を行う．第 1 反復で得られたパラメータの推定値

$\hat{\boldsymbol{\beta}}^{(1)}=$	
-1.4120	
0.0886	
$\boldsymbol{\beta}^{(m)\wedge}$	

を

-2.8310	
0.1141	
$\boldsymbol{\beta}^{(m-1)\wedge}$	

に値のみをペースト

すると第 2 反復の結果は

$(m=2)$:	
-1.5053	
0.0898	
$\boldsymbol{\beta}^{(m)\wedge}$	

が得られるが，「$\ln \hat{\boldsymbol{Y}} - \hat{\boldsymbol{Z}}$ の絶対値の和」が，0.2866 と，まだ大きいので，（m=2）の結果を，$\hat{\boldsymbol{\beta}}^{(m-1)}$ に値のみをペーストして第 3 反復を行った結果を表 5.11 に示す．

表 5.11　＜反復 3＞冠動脈心疾患の死亡者数（対数リンク）

	デザイン行列		死亡	回帰		補正リンク	重付回帰	推定値
	$\boldsymbol{X}$		者数	推定値	重み	関数	推定値	差
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	$\ln\boldsymbol{Y}\wedge$	$\boldsymbol{w}\wedge=\boldsymbol{Y}\wedge$	$\boldsymbol{Z}$	$\boldsymbol{Z}\wedge$	$\ln\boldsymbol{Y}\wedge - \boldsymbol{Z}\wedge$
1	1	30	1	1.1901	3.2875	0.4943	1.1888	0.0014
2	1	35	5	1.6394	5.1520	1.6099	1.6382	0.0012
3	1	40	5	2.0886	8.0737	1.7079	2.0876	0.0010
:								
8	1	65	65	4.3348	76.3105	4.1866	4.3348	0.0001
$\beta_0\wedge=$	-2.8310	$\beta_0\wedge=$	**-1.5053**		$\beta_0\wedge=$	**-1.5078**	$\boldsymbol{X\beta}^{(m)\wedge}$	**0.0058**
$\beta_1\wedge=$	0.1141	$\beta_1\wedge=$	**0.0898**		$\beta_1\wedge=$	**0.0899**		絶対値の和
$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \ln(\boldsymbol{Y})$			$\boldsymbol{\beta}^{(m-1)\wedge}$			$\boldsymbol{\beta}^{(m)\wedge}$		
				205.1	11753.5	0.2176	-0.0037	747.26
				11753.5	689087.2	-0.0037	0.0001	44216.90
				$(\boldsymbol{X}*\boldsymbol{w}\wedge)^T\boldsymbol{X}$		$[(\boldsymbol{X}*\boldsymbol{w}\wedge)^T\boldsymbol{X}]^{-1}$		$(\boldsymbol{X}*\boldsymbol{w}\wedge)^T\boldsymbol{Z}$

$\ln \hat{\boldsymbol{Y}} - \hat{\boldsymbol{Z}}$ の絶対値の和が，0.0058 とかなり小さくなる．第 3 反復は，まだ収束はしていないが，次の反復で収束する直前の結果である．したがって，

$(m=3)$:	
$\beta_0\wedge=$	**-1.5078**
$\beta_1\wedge=$	**0.0899**
	$\boldsymbol{\beta}^{(m)\wedge}$

が，オフセットなしの対数リンクで推定されたパラメータになる．パラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta}\wedge)$ は，

共分散行列：$\boldsymbol{\Sigma}(\boldsymbol{\beta}\wedge)=$	
0.2176	-0.0037
-0.0037	0.0001
$[(\boldsymbol{X}*\boldsymbol{w}\wedge)^T\boldsymbol{X}]^{-1}$	

となる．

95%信頼区間

パラメータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})$を用いて，回帰直線の 95%信頼区間を求める．それぞれの i ごとに，$\boldsymbol{x}_i = [x_{0,i} \quad x_{1,i}]$ としたとき $\hat{Z}_i$ の分散は，次の 2 次形式で

$$Var(\hat{Z}_i) = \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T$$

求められる．$\hat{Z}_i$ の 95%信頼区間は，

$$(L95\%, U95\%) = \hat{Z}_i \pm 1.96\sqrt{Var(\hat{Z}_i)}$$

であり，表 5.12 に計算した結果を示す．元のスケールは，対数の 95%信頼区間について指数を取って計算したものである．

表 5.12　対数リンクの場合の 95%信頼区間

	$\boldsymbol{X}$		対 数						元のスケール		
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Z}$	$\boldsymbol{Z}\hat{}$	$Var(\boldsymbol{Z}\hat{})$	$L95\%$	$U95\%$	$\boldsymbol{Y}$	$\boldsymbol{Y}\hat{}$	$L95\%$	$U95\%$
1	1	30	0.4943	**1.1888**	**0.0532**	**0.7367**	**1.6408**	1	**3.2830**	**2.0891**	**5.1592**
2	1	35	1.6099	1.6382	0.0371	1.2606	2.0158	5	5.1458	3.5274	7.5067
3	1	40	1.7079	2.0876	0.0243	1.7821	2.3931	5	8.0656	5.9426	10.9470
:											
8	1	65	4.1866	4.3348	0.0087	4.1519	4.5176	65	76.3060	63.5559	91.6140

パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いた計算の実例を，$i=1$ の場合について示す．まず，対数についての $Var(\hat{Z}_1)$ は，2 次形式の計算法で，0.0532 が得られる．

$Var(Z_1\hat{})$=	1	30	0.2178	-0.0037	1	= 0.0532
	$\boldsymbol{x}_1$		-0.0037	0.0001	30	
	$\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{}) = [(\boldsymbol{X}*\boldsymbol{w}\hat{})^T \boldsymbol{X}]^{-1}$				$\boldsymbol{x}_1^T$	

信頼区間の計算は，

$$\hat{Z}_1 \pm 1.96\sqrt{Var(\hat{Z}_1)} = 1.1888 \pm 1.96\sqrt{0.0532} = (0.7367, 1.6408)$$

で得られる．指数を取って元のスケールでは，

$$\hat{Y}_1 = \exp(\hat{Z}_1) = \exp(1.1888) = 3.2830$$
$$L95\% = \exp(0.7367) = 2.0891$$
$$U95\% = \exp(1.6408) = 5.1592$$

となる．

これらの計算結果より，図 5.1 に対数軸と元のスケールでの 95%信頼区間を示す．なお，この図は Excel ではなく，JMP で対数リンクでのポアソン回帰を行い，推定値および 95%信頼区間を JMP ファイルに書き出して Excel の結果と一致することを確認の上,「重ね合わせプロット」で作成したものである．

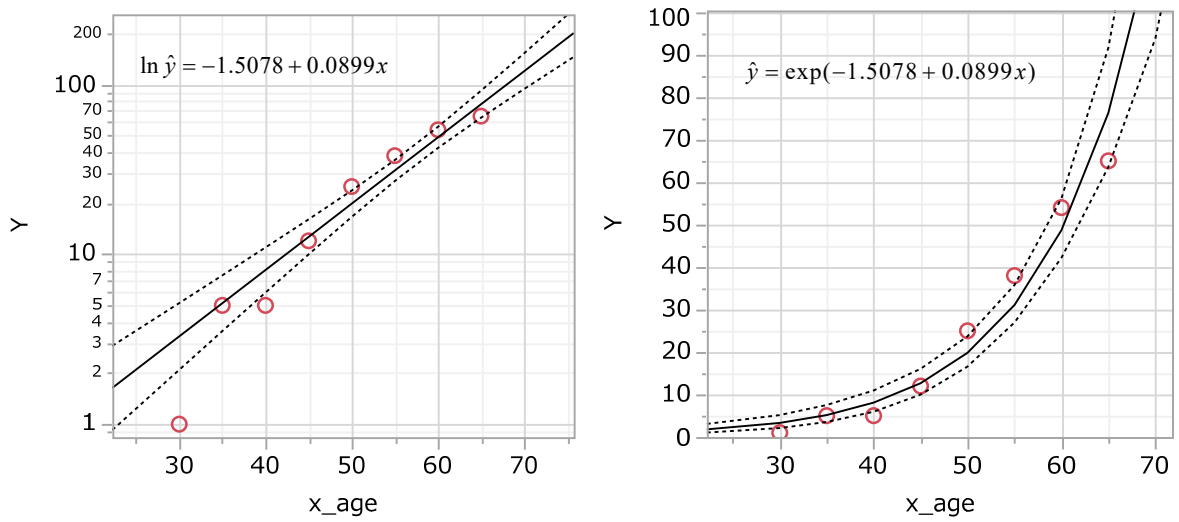


図 5.1　冠動脈心疾患の死亡者数に対するポアソン回帰曲線と 95%信頼区間

この解析に用いたデータに 0 が含まれていなかったので，図 5.1 左の対数目盛上の観測データのプロットには，元の $Y_1=1$， $Y_2=5,\ldots,$ $Y_8=65$ が使われている．データに 0 が含まれている場合には, 対数変換すると $-\infty$ となり標示することができない. 反復重み付き回帰では, 元の Y_i を使うのではなく，調整された対数変換値 Z_i に対して重み付き回帰を行っている．表 5.11 に示されているように，元の $Y_1=1$ に対し， $Z_1=0.4943$ なので，指数を取ると $Y_1'=\exp(Z_1)=1.6394$ となることを認識しておくことが必要である．もちろん他のすべての Y_i に対し $Y_2'=\exp(Z_2)=\exp(1.6099)=5.0022$ のように調整されている．

2 次式のあてはめ

図 5.1 左の対数リンクによる回帰直線に対して，元データのプロットは上に凸の曲線状となり，直線のあてはめには疑問が残る．これは，年齢が高くなれば，人口が減り冠動脈心疾患による死亡者数も相対的に減るであろうし，10 万人当たりの死亡数も全死亡の一部でもあり，上限があり頭打ちになることも考えられる．そこで，65 歳までのデータから頭打ちが確認できるのか検討する．表 5.13 に示すように年齢について 2 乗の項を追加し，直線のあてはめが妥当かを検討する．

反復重み付き回帰の利便性は，説明変数の数を増やしても，行列のサイズが大きくなるだけで，Excel の計算シートの本体に大きな変更がない．表 5.13 に示すように，デザイン行列 $\boldsymbol{X}$ のサイズを 8×2 から 8×3 に変更し，パラメータ $\boldsymbol{\beta}$ のサイズズを 2×1 から 3×1 に変更するが，最初の $\boldsymbol{\beta}^{(0)}$ の計算式 $\boldsymbol{\beta}^{(0)}=(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{X}^T\ln(\boldsymbol{Y})$ に変更はない．ただし，Excel では，行列の範囲をフィルハンドルによって拡大する必要がある．他の計算式も同様に行列の範囲を変更しなければならないので，煩雑さはいなめないが機械的な作業レベルで対応できる．

表 5.13　冠動脈心疾患の死亡者数に対する 2 次式のあてはめ（対数リンク）

	デザイン行列			死亡	回帰		補正リンク	重付回帰	推定値
	$\boldsymbol{X}$			者数	推定値	重み	関数	推定値	差
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$X_2=\boldsymbol{X}_1{}^{\wedge 2}$	$\boldsymbol{Y}$	$\ln\boldsymbol{Y}^{\wedge}$	$\boldsymbol{w}^{\wedge}=\boldsymbol{Y}^{\wedge}$	$\boldsymbol{Z}$	$\boldsymbol{Z}^{\wedge}$	$\ln\boldsymbol{Y}^{\wedge}-\boldsymbol{Z}^{\wedge}$
1	1	30	900	1	0.2151	1.2400	0.0216	0.2151	0.0000
2	1	35	1225	5	1.1259	3.0829	1.7477	1.1259	0.0000
3	1	40	1600	5	1.9224	6.8372	1.6537	1.9224	0.0000
4	1	45	2025	12	2.6046	13.5262	2.4918	2.6046	0.0000
5	1	50	2500	25	3.1726	23.8702	3.2200	3.1726	0.0000
6	1	55	3025	38	3.6264	37.5769	3.6376	3.6264	0.0000
7	1	60	3600	54	3.9659	52.7675	3.9893	3.9659	0.0000
8	1	65	4225	65	4.1912	66.0991	4.1745	4.1912	0.0000
$\beta_0{}^{\wedge}=$	-7.6076		$\beta_0{}^{\wedge}=$	**-7.6486**		$\beta_0{}^{\wedge}=$	**-7.6486**	$\boldsymbol{X\beta}^{(m)\wedge}$	**0.0000**
$\beta_1{}^{\wedge}=$	0.3277		$\beta_1{}^{\wedge}=$	**0.3307**		$\beta_1{}^{\wedge}=$	**0.3307**	絶対値の和	
$\beta_2{}^{\wedge}=$	-0.0022		$\beta_2{}^{\wedge}=$	**-0.0023**		$\beta_2{}^{\wedge}=$	**-0.0023**		
$\hat{\boldsymbol{\beta}}^{(0)}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\ln(\boldsymbol{Y})$				$\boldsymbol{\beta}^{(m-1)\wedge}$			$\boldsymbol{\beta}^{(m)\wedge}$		
			2.05E+02	1.18E+04	6.86E+05	6.9415	-0.2619	0.0024	7.50E+02
			1.18E+04	6.86E+05	4.06E+07	-0.2619	0.0100	-0.0001	4.41E+04
			6.86E+05	4.06E+07	2.44E+09	0.0024	-0.0001	8.73E-07	2.62E+06
			$(\boldsymbol{X}^*\boldsymbol{w}^{\wedge})^T\boldsymbol{X}$			$\boldsymbol{\Sigma}(\boldsymbol{\beta}^{\wedge})=[(\boldsymbol{X}^*\boldsymbol{w}^{\wedge})^T\boldsymbol{X}]^{-1}$			$(\boldsymbol{X}^*\boldsymbol{w}^{\wedge})^T\boldsymbol{Z}$

表 5.13 で求めたパラメータ $\boldsymbol{\beta}^{(m)}$，パラメータの共分散行列の対角要素から $\boldsymbol{\beta}^{(m)}$ の分散を表 5.14 に取り出し，検定統計量の計算を追加する．

表 5.14　冠動脈心疾患の死亡者数に対する Wald 検定の結果

項	推定値	分散	標準誤差	Waldカイ2乗	p 値
X_0	-7.6486	6.9415	2.6347	8.4277	0.0037
X_1	0.3307	0.0100	0.1000	10.9341	0.0009
$X_2=X_1{}^{\wedge 2}$	-0.0023	8.73E-07	9.34E-04	5.9802	0.0145

Wald 検定の結果で，2 次の項が p=0.0145 と有意であることから，65 歳までのデータで頭打ちが統計的に確認されたことになる．そこで，表 5.12 と同様に表 5.15 に 95%信頼区間の計算を行い，図 5.2 にポアソン回帰の 2 次曲線を示す．図 5.1 に比べて，あてはまりは良くなり，対数死亡者数目盛りの場合に場合に年齢が増えるにつれ徐々に頭打ちとなることが観察される．実目盛りでは，70 歳がピークとなり，減少に転じている．

ただし，対数目盛上での 2 次式のあてはめは便宜的な方法ではあり，得られた 2 次式のパラメータについて意味づけすることは生産的でないが，少なくとも直線のあてはめには難点があると理解できることに意義がある．

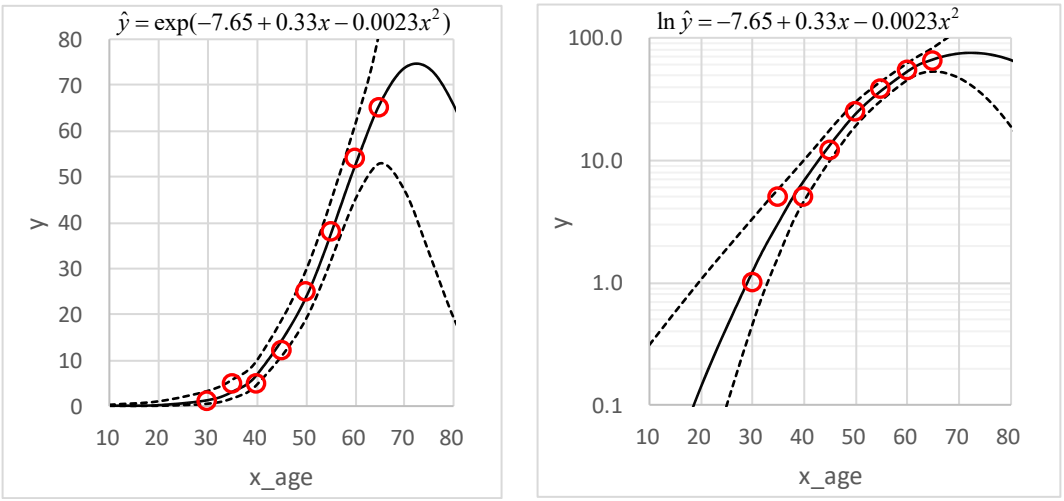


図 5.2　冠動脈心疾患の死亡者数に対するポアソン 2 次回帰と 95%信頼区間

表 5.15　2 次式の 95%信頼区間

		$\boldsymbol{X}$		―――― 対数 ――――				―― 元のスケール ――		
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{Z}$^	$Var(\boldsymbol{Z}$^$)$	L95%	U95%	$\boldsymbol{Y}$^	L95%	U95%
1	1	10	100	-4.5704	**3.0094**	-7.9705	-1.1702	0.010	0.000	0.310
2	1	20	400	-1.9491	1.0481	-3.9557	0.0575	0.142	0.019	1.059
3	1	30	900	0.2151	0.2553	-0.7752	1.2054	1.240	0.461	3.338
4	1	35	1225	1.1259	0.1036	0.4951	1.7567	3.083	1.641	5.793
5	1	40	1600	1.9224	0.0380	1.5402	2.3045	6.837	4.666	10.020
6	1	50	2500	3.1726	0.0130	2.9495	3.3958	23.870	19.096	29.838
7	1	60	3600	3.9659	0.0063	3.8101	4.1217	52.768	45.154	61.664
8	1	65	4225	4.1912	0.0129	3.9687	4.4136	66.099	52.914	82.570
9	1	70	4900	4.3022	0.0538	3.8474	4.7569	73.860	46.873	116.384
10	1	75	5625	4.2989	0.1663	3.4995	5.0983	73.621	33.100	163.748
11	1	80	6400	4.1814	0.4008	2.9407	5.4222	65.460	18.928	226.382
12	1	85	7225	3.9497	0.8204	2.1744	5.7250	51.920	8.797	306.446

対数リンクの 2 次式の分散も表 5.12 の 1 次式の場合と同様に，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いた 2 次形式で求めることができる．計算の実例を，$i=1$の場合について示す．まず，対数についての$Var(\hat{Z}_1)$は，2 次形式の計算法で，3.0094 が得られる．この分散を用いて 95%信頼区間の計算が行える．

$Var(Z_1$^$)$=	1	10	100	6.9415	-0.2619	0.0024	1	=	3.0094
		$\boldsymbol{x}_1$		-0.2619	0.0100	-0.0001	10		
				0.0024	-0.0001	8.73E-07	100		
				$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)=[(\boldsymbol{X}*\boldsymbol{w}$^$)^T\boldsymbol{X}]^{-1}$			$\boldsymbol{x}_1^T$		

5.5. 対数リンクでオフセットがある場合のポアソン回帰

第 5.4 節では，冠動脈心疾患の死亡者数についてオフセットを無視して対数リンクによるポアソン回帰を例示した．元のデータである第 1.5 節のデータには，部分母集団の人数が得られているので，表 5.16 にように死亡率を求めることができる［ドブソン（2008)］．このデータについては，第 2.6 節で，ポアソン確率を用いたニュートン・ラフソン法による最尤法の解析法でも取り上げており，ここでは，反復重み付き回帰による解析法について示す．

表 5.16　冠動脈心疾患の死亡率（表 1.11，表 2.21 再掲）

	年齢層	死亡者数	母集団	死亡率	1万比
No.	X	Y	人数 n	%	人
1	30	1	17,742	0.0056	0.56
2	35	5	16,554	0.0302	3.02
3	40	5	16,059	0.0311	3.11
4	45	12	13,083	0.0917	9.17
5	50	25	10,784	0.2318	23.18
6	55	38	9,645	0.3940	39.40
7	60	54	10,706	0.5044	50.44
8	65	65	9,933	0.6544	65.44

このように死亡率が求められるデータに対しては，一般化線形モデルで，分布を 2 項分布とし，リンク関数を（ロジット or プロビット or 補 2 重対数）とした解析が適しているようにも思われる．しかし，死亡率が最も高い年齢 65 歳階層でも 1.0 パーセントにも届かないデータに，0～100 パーセントの範囲を規定するシグモイド曲線のあてはめるような統計モデルを適用することは，適切と言えるのであろうか．

全ての死亡を対象にした場合には，0～100 パーセントの死亡率を想定することは理にかなっているように思われる．しかし，冠動脈心疾患での高年齢での死亡率はおおよそ 15%程度［厚労省・2018 年，人口動態統計，第 6 表 性別にみた死因順位（第 10 位まで）別 死亡数・死亡率（人口 10 万対)・構成割合］であるので，0～100 パーセントを仮定した方法の適用は，違和感が付きまとう．リンク関数を対数としたポアソン回帰が望ましいように思われるが，分母があるカウント・データに対してどのように適用するのだろうか．

オフセットを含めたポアソン回帰

死亡人数 Y_i に対して部分母集団の人数 n_i を加えた指数曲線をあてはめるモデル

$$Y_i = n_i \exp(\beta_0 + \beta_0 X_i) + \varepsilon_i , \quad \varepsilon_i \sim Poisson(Y_i) \tag{5.22}$$

について，誤差項を含まない推定式として両辺に対数を取ると，

$$\ln \hat{Y}_i = \ln n_i + \hat{\beta}_0 + \hat{\beta}_0 X_i \tag{5.23}$$

右辺にオフセット $\ln n_i$ を含む式が得られる．オフセット $\ln n_i$ を左辺に移動して整理すると

$$\ln \frac{\hat{Y}_i}{n_i} = \hat{\beta}_0 + \hat{\beta}_0 X_i \tag{5.24}$$

が得られる．これは，死亡率の対数について，回帰直線をあてはめることに相当する．もちろん，単純に死亡率の対数に対し，通常の回帰分析を行なう問題ではない．誤差の分布はポアソン分布を仮定しなければならない．式（5.23）で考えれば，$\ln n_i + \hat{\beta}_0$ がそれぞれの年齢層 i についての切片となっている．このことから，傾きは共通で，切片に年齢層ごとにオフセットを持つ回帰直線を求める問題と解される．ただし，推定する切片は，対数死亡率に対する切片 β_0 であり，それぞれの年齢層 i について別々に推定しているわけではなく，単にオフセットしているだけである．なお，オフセットとは，ある基準からのズレの意味で，$\hat{\beta}_0$ を基準として $\ln n_i$ 分ずらし切片とすることを意味する．オフセットを考慮した回帰直線のあてはめについては，第 2.6 節の図 2.3 に例示してある．

補正値

反復重み付き回帰で，オフセットがある対数リンク式を

$$\eta_i = \ln\left(\frac{\hat{Y}_i}{n_i}\right)$$

とし，元の推定値を

$$\mu_i = n_i \exp\left[\ln\left(\frac{\hat{Y}_i}{n_i}\right)\right]$$

$$\ln\left(\frac{\mu_i}{n_i}\right) = \ln\left(\frac{\hat{Y}_i}{n_i}\right)$$

としたときに，一般線化形モデルの公式により，対する補正値 Z_i は

$$\left.\begin{aligned} Z_i &= \eta_i + (Y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \\ &= \eta_i + (Y_i - \mu_i)\left(\frac{\partial \ln(\mu_i)}{\partial \mu_i}\right) \\ &= \eta_i + \frac{(Y_i - \mu_i)}{\mu_i} \\ &= \ln\left(\frac{\hat{Y}_i}{n_i}\right) + \frac{Y_i - n_i \exp[\ln(\hat{Y}_i / n_i)]}{n_i \exp[\ln(\hat{Y}_i / n_i)]} \end{aligned}\right\} \tag{5.25}$$

となる．式（5.22）の誤差分布の分散は，$Var(Y_i)=\mu_i$ なので，計算に用いる重み w_i は，

$$\left.\begin{aligned} w_i &= \frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \\ &= \frac{1}{\mu_i\left(\frac{\partial \ln \mu_i}{\partial \mu_i}\right)^2} \\ &= \frac{1}{\frac{\mu_i}{\mu_i^2}} = \mu_i = n_i \exp\left[\ln\left(\frac{\hat{Y}_i}{n_i}\right)\right] \end{aligned}\right\} \qquad (5.26)$$

となる．

反復計算

さて，元のデータに戻って，初期値は，表 5.17 から

$$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \ln\left(\frac{\boldsymbol{Y}}{\boldsymbol{n}}\right) = \begin{bmatrix} -13.1819 \\ 0.1331 \end{bmatrix}$$

ただし，$\ln\left(\frac{\boldsymbol{Y}}{\boldsymbol{n}}\right)$ は，セル同士の除算結果に対し対数を取ったベクトル

として求められる．

表 5.17　オフセットを考慮した初期値の計算

	デザイン行列		死亡	母集団	死亡率の				
	$\boldsymbol{X}$		者数	人数	対数				
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	$\boldsymbol{n}$	$\ln(\boldsymbol{Y}/\boldsymbol{n})$				
1	1	30	1	17,742	-9.7837	$\boldsymbol{X}^T\boldsymbol{X}=$	8.00	380.00	
2	1	35	5	16,554	-8.1049		380.00	19100.00	
3	1	40	5	16,059	-8.0746	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}=$	2.2738	-0.0452	
4	1	45	12	13,083	-6.9942		-0.0452	0.0010	
5	1	50	25	10,784	-6.0669				$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\ln\boldsymbol{Y}/\boldsymbol{n})$
6	1	55	38	9,645	-5.5366	$\boldsymbol{X}^T\ln(\boldsymbol{Y}/\boldsymbol{n})=$	-54.8797	β_0^=	**-13.1819**
7	1	60	54	10,706	-5.2896		-2467.04	β_1^=	**0.1331**
8	1	65	65	9,933	-5.0292				$\boldsymbol{\beta}^{(0)}$

この初期値を用いて表 5.18 に示すように反復重み付き回帰を行う．この結果は，3 反復目の結果で，まだ収束はしていないが，次の反復で収束する直前の結果である．なお，計算の手順は，第 5.4 節の対数リンクの場合の $\ln(\boldsymbol{Y})$ を $\ln(\boldsymbol{Y}/\boldsymbol{n})$ に変更しただけで，他の手順は全く同じである．

表 5.18　＜第 3 反復＞冠動脈心疾患の死亡者数（対数リンク，オフセットあり）

	デザイン行列		死亡	母集団	死亡率の	回帰		リンク	重付回帰	推定値
	$\boldsymbol{X}$		者数	人数	対数	推定値	重み	関数	推定値	差
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{Y}$	$\boldsymbol{n}$	$\ln(\boldsymbol{Y}/\boldsymbol{n})$	$\ln\boldsymbol{Y}\hat{}$	$\boldsymbol{w}=\boldsymbol{Y}\hat{}$	$\boldsymbol{Z}$	$\boldsymbol{Z}\hat{}$	$\ln\boldsymbol{Y}\hat{}-\boldsymbol{Z}\hat{}$
1	1	30	1	17,742	-9.7837	-8.4919	3.6393	-9.2171	-8.4947	0.0028
2	1	35	5	16,554	-8.1049	-7.9701	5.7220	-8.0962	-7.9725	0.0024
3	1	40	5	16,059	-8.0746	-7.4483	9.3537	-7.9137	-7.4503	0.0020
4	1	45	12	13,083	-6.9942	-6.9264	12.8408	-6.9919	-6.9281	0.0017
5	1	50	25	10,784	-6.0669	-6.4046	17.8356	-6.0029	-6.4059	0.0013
6	1	55	38	9,645	-5.5366	-5.8828	26.8802	-5.4691	-5.8837	0.0009
7	1	60	54	10,706	-5.2896	-5.3610	50.2783	-5.2870	-5.3615	0.0006
8	1	65	65	9,933	-5.0292	-4.8392	78.6061	-5.0123	-4.8394	0.0002
		$\beta_0\hat{}=$	-13.1819	$\beta_0\hat{}=$	**-11.6228**		$\beta_0\hat{}=$	**-11.6278**	$\boldsymbol{X\beta}^{(m)\hat{}}$	**0.0119**
		$\beta_1\hat{}=$	0.1331	$\beta_1\hat{}=$	**0.1044**		$\beta_1\hat{}=$	**0.1044**		絶対値の和
		$\hat{\boldsymbol{\beta}}^{(0)}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\ln(\boldsymbol{Y}/\boldsymbol{n})$			$\boldsymbol{\beta}^{(m-1)\hat{}}$			$\boldsymbol{\beta}^{(m)\hat{}}$		
						205.2	11757.7	0.2049	-0.0035	-1157.6
						11757.7	690267.6	-0.0035	0.0001	-64626.5
						$(\boldsymbol{X}^*\boldsymbol{w})^T\boldsymbol{X}$		$[(\boldsymbol{X}^*\boldsymbol{w})^T\boldsymbol{X}]^{-1}$		$(\boldsymbol{X}^*\boldsymbol{w})^T\boldsymbol{Z}$

オフセットを用いた推定

表 5.18 の推定結果から，対数発現率に対する回帰式

$$\ln\frac{\hat{Y}_i}{n_i}=-11.6278+0.1044X_{1,i}$$

が得られる．そこで，表 5.19 に示すように，年齢階層別の人数の推定値に換算しよう．年齢が 30 歳代の場合の死亡率の推定値は，

$$\hat{Z}_1=-11.6278+0.1044\times 30=-8.4947$$

となる．これは，死亡率

$$p_1=\frac{1}{17,742}=0.000056$$

に対する推定値

$$\hat{\pi}_1=\exp(-8.4947)=0.00020$$

である．年齢階層別の人数の推定値は，

$$\begin{aligned}\hat{Y}_1&=n_1\exp(\hat{\beta}_0+\hat{\beta}_0X_{1,1})\\&=17,742\exp(-8.4947)=3.63\end{aligned}$$

となる．それぞれの階層での分母が異なるので，オフセットされた推定値が適切であるが，全体として解釈するためには適さない．そのために，分母を区切りの良い人数に揃えて結果を示すことにする．分母の人数は，9,647～17,742 人なので，10,000 人あたりに換算する．年齢階層 30 歳の得られた死亡数は，1 人なので，

$$Y_1' = \frac{10,000}{n_i} Y_1 = \frac{10,000}{17,742} \times 1 = 0.56$$

0.56 人となり，10,000 人当たりの推定値は，

$$\hat{Y}_1' = 10,000 \times \exp(-8.4947) = 2.05$$

と，2.05 人となる．この様に換算することにより，人口 10,000 人あたりの死亡数となり，指数曲線での図が作成可能となる．オフセット付きの人数では，各層間の直接的な比較は不向きであり，またグラフ表示にもなじまない

表 5.19　1 万人当たりの冠動脈心疾患の死亡者数の推定

	デザイン行列		死亡	母集団	回帰	人数	10,000人	10,000人
	X		者数	人数	推定値	推定値	あたりの	あたりの
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	***Y***	***n***	***Z***^	***Y***^	人数	推定人数
1	1	30	1	17,742	-8.4947	3.63	0.56	2.05
2	1	35	5	16,554	-7.9725	5.71	3.02	3.45
3	1	40	5	16,059	-7.4503	9.33	3.11	5.81
4	1	45	12	13,083	-6.9281	12.82	9.17	9.80
5	1	50	25	10,784	-6.4059	17.81	23.18	16.52
6	1	55	38	9,645	-5.8837	26.86	39.40	27.84
7	1	60	54	10,706	-5.3615	50.25	50.44	46.94
8	1	65	65	9,933	-4.8394	78.59	65.44	79.12

図 5.3 に 1 万人比の死亡者数に対して，対数目盛と実目盛りでのポアソン回帰の結果を示す．前節と同様に，各年齢のプロット点が推定値に対して上に凸であり，前節と同様に 2 次式のあてはめが必要となる．

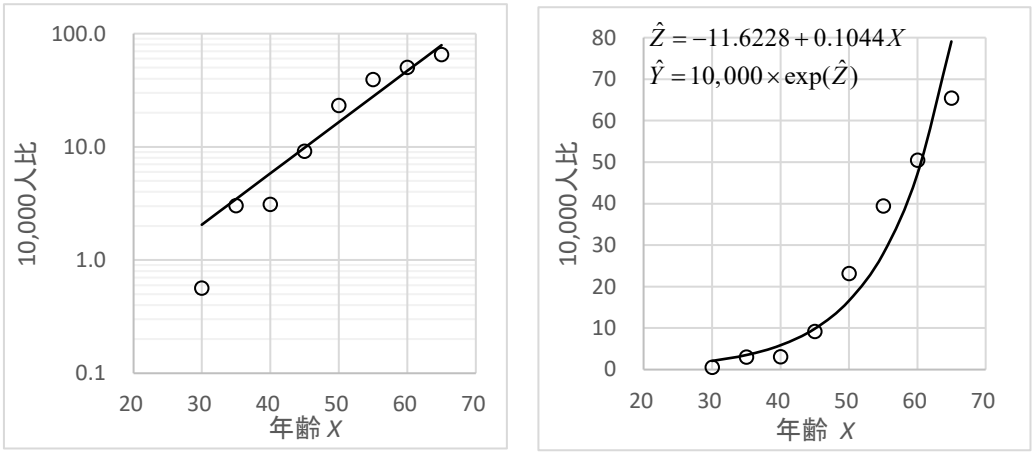


図 5.3　冠動脈心疾患の人口 10,000 人当たりの死亡者数にポアソン回帰

5.6. 2 項分布を仮定した（プロビット・補 2 重対数・ロジット）解析

オフセットがある対数リンクのポアソン回帰については，前節で解析事例を示した．オフセットがある場合には，観測されたカウント・データに対して部分母集団のサイズが分かっている場合は，ポアソン分布を仮定しなくても，2 項分布を仮定した（0, 1）反応に対するロジスティック回帰などの適用が妥当とも思われる．前節では，「死亡率が最も高い年齢 65 歳階層でも 1.0 パーセントにも届かないデータに，0～100 パーセントの範囲を仮定するシグモイド曲線のあてはめを行う統計モデルを適用することは，適切と言えるのであろうか.」と述べた.

ポアソン回帰も一般化線形モデルの枠組みであるので，分布を 2 項分布とし，リンク関数を（プロビット or 補 2 重対数 or ロジット）による一般化線形モデル解析を適用した場合に結果をどのように解釈するのか，ポアソン回帰の結果と対比して示す．一般化線形モデルでのリンク関数をロジットとした場合には，独立した統計手法として確立しているロジスティック回帰と同じ結果を得るので，「プロビット」の場合を最初に解説する．

プロビット

2 項分布に従う出現率 $P_i = Y_i / n_i$ について，説明変数を X_i とした場合に，標準正規分布の分布関数をシグモイド曲線として

$$P_i = \Phi(\beta_0 + \beta_1 X_i) + \varepsilon_i, \quad \varepsilon_i \sim Binomial(Y_i, n_i; \pi_i) \tag{5.27}$$

ただし，Φ は，標準正規分布

をあてはめたい．ここで，

$$\beta_0 = \frac{-\mu}{\sigma}, \quad \beta_1 = \frac{1}{\sigma}$$

で置き換えると，

$$\begin{aligned}\Phi(\beta_0 + \beta_1 X_i) &= \Phi\left(\frac{-\mu}{\sigma} + \frac{1}{\sigma} X_i\right) \\ &= \Phi\left(\frac{X_i - \mu}{\sigma}\right)\end{aligned}$$

となり，線形式のパラメータから，標準正規分布への変換パラメータは，

$$\mu = \frac{-\beta_0}{\beta_1}, \quad \sigma = \frac{1}{\beta_1}$$

として与えられる.

リンク関数の「プロビット」は，標準正規分布の逆関数 Φ^{-1} を用いて

$$Z_i = \Phi^{-1}(\hat{P}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (5.28)$$

ただし，Φ^{-1} ：逆標準正規分布

で線形化し，反復重み付き回帰で解くので，線形化した場合の分布を気にする必要がなくなる．この方法は，生物統計の分野で 50%致死量 LD_{50} を求めるための方法として Finney（1971），Probit Analisis 3rd ed. および Finney（1978），Statistical Metod in Biological Assay 3rd ed. により知られている方法でもあり，詳細は，高橋（2017）を参照のこと．

ここでは，プロビット変換した式（5.28）ではなく，式（5.27）によりシグモイド曲線の直接あてはめを行う．表 5.20 に示すように冠動脈心疾患の死亡率に対して，年齢が 30 歳（$X_{1,1}=30$）正規分布の下側確率を

$$\begin{aligned}\hat{\pi}_1 &= \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_{1,1}) \\ &= \Phi(-4.6119 + 0.0337 \times 30) \\ &= \text{Norm.dist}(-3.5999,\ 0,\ 1,\ \textit{true}) = 0.000159\end{aligned}$$

で求めている．この予測死亡率 π_1 に対し，2 項分布の確率を「尤度」として

$$\begin{aligned}L_1 &= \text{Binom.dist}(Y_1, n_1, \hat{\pi}_1, \textit{false}) \\ &= \text{Binom.dist}(1,\ 17742,\ 0.000159,\ \textit{false}) \\ &= 0.1676\end{aligned}$$

で求める．それの対数尤度の和は，

$$\begin{aligned}\ln L &= \sum_{i=1}^{8} \ln L_i \\ &= \ln(0.1676) + \ln(0.1755) + \cdots + \ln(0.0176) \\ &= -1.7861 - 1.7402 - \cdots - 4.0383 \\ &= -23.2990\end{aligned}$$

表 5.20　冠動脈心疾患の死亡率に対する正規分布曲線のあてはめ

	デザイン行列		死亡	母集団	死亡率		シグモイド			組合せ
	X		者数	人数	***Y/n***	***Xβ^***	標準正規	二項分布		無
i	X_0	X_1	***Y***	***n***	***P***	***Z^***	***π^***	尤度 L_i	$\ln L_i$	$\ln L_i'$
1	1	30	1	17,742	0.0001	-3.5999	**0.000159**	**0.1676**	-1.7861	-11.5698
2	1	35	5	16,554	0.0003	-3.4312	0.000300	0.1755	-1.7402	-45.5240
3	1	40	5	16,059	0.0003	-3.2625	0.000552	0.0644	-2.7426	-46.3746
4	1	45	12	13,083	0.0009	-3.0939	0.000988	0.1106	-2.2017	-95.9583
5	1	50	25	10,784	0.0023	-2.9252	0.001721	0.0290	-3.5392	-177.6532
6	1	55	38	9,645	0.0039	-2.7566	0.002921	0.0137	-4.2875	-249.8657
7	1	60	54	10,706	0.0050	-2.5879	0.004828	0.0516	-2.9635	-339.5517
8	1	65	65	9,933	0.0065	-2.4192	0.007777	0.0176	-4.0383	-392.7210
					β_0^=	**-4.6119**			**-23.2990**	-1,359.2183
					β_1^=	**0.0337**		対数尤度	$\ln L$	$\ln L'$

となっている．これは，ソルバーで $\ln L$ が最大になるように $\hat{\beta}_0$ と $\hat{\beta}_1$ を変化させた結果である．ここに示した対数尤度は，2 項分布の確率の対数の和であるが，統計ソフトでは，2 項確率の組み合わせ計算を除いている場合があるので， $\ln L'$

$$\begin{aligned}\ln L' &= \sum_{i=1}^{8}\left[y_i \ln \hat{\pi}_i + (n_i - y_i)\ln(1-\hat{\pi}_i)\right] \\ &= -11.57 - 45.52 - \cdots - 392.72 \\ &= -1{,}359.22\end{aligned}$$

の計算も付け足してある．この $\ln L' = -1{,}359.22$ は， 表 5.21 に示す JMP によるプロビット解析の「モデル」の欄の「完全」，「(−1)*対数尤度」の欄の「1359.2183」に一致する．

JMP によるプロビット解析は，データを死亡の（0：あり，1：なし）を行方向に展開し，それぞれの人数を設定し，一般線形モデルで次のように分布とリンク関数を設定する．

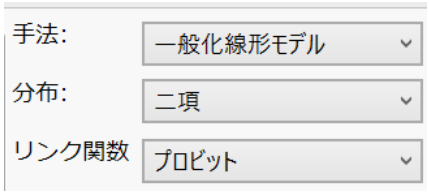
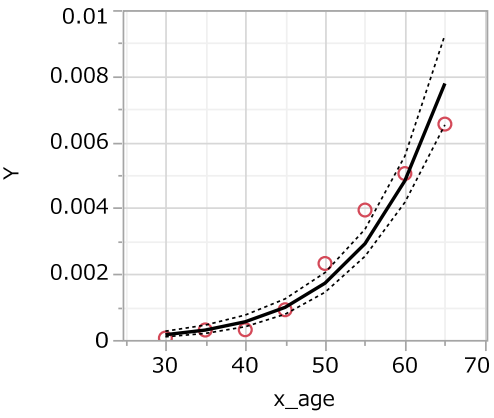


表 5.21 に結果を示す．Excel で求めた結果と一致することが確かめられる．95%信頼区間の図は，推定値を JMP ファイルに別途書き出して，重ね合わせプロットで作成したものである．

表 5.21　JMP によるプロビット解析

モデル	(-1)*対数尤度	尤度比カイ2乗
差分	123.5484	247.0968
完全	1359.2183	
縮小	1482.7667	

項	推定値	標準誤差	尤度比カイ2乗
切片	-4.6119	0.1402	2580.1056
x_age	0.0337	0.0025	247.0968



標準正規分布に変換するためのパラメータは，

$$\hat{\mu}_{NOR} = \frac{-\hat{\beta}_0}{\hat{\beta}_1} = \frac{-(-4.6119)}{0.0337} = 136.72\,, \qquad \hat{\sigma}_{NOR} = \frac{1}{\hat{\beta}_1} = \frac{1}{0.0337} = 29.64$$

なので，形式的に平均が 136.72 歳，標準偏差が 29.64 歳とした正規分布をあてはめていることになる．

補 2 重対数

2 項分布に従う出現率 $P_i = Y_i / n_i$ について，説明変数を X_i とした場合に，最小極値分布をシグモイド曲線として

$$P_i = 1 - \exp\left[-\exp(\beta_0 + \beta_1 X_{1,i})\right] + \varepsilon_i, \quad \varepsilon_i \sim Binomial(Y_i, n_i; \pi_i) \tag{5.29}$$

用いる．式（5.29）の推定値にした回帰式について解くと

$$\exp\left[-\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i})\right] = 1 - \hat{P}_i$$

$$\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i}) = -\ln(1 - \hat{P}_i)$$

$$Z_i = \ln\left[-\ln(1 - \hat{P}_i)\right] = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_{1,i} \tag{5.30}$$

となり，標準的な関数で変換式が表せる．式（5.28）に示したプロビット法と同様に，リンク関数が「補 2 重対数」の形式になっている．

ここでは，補 2 重対数変換した式（5.30）ではなく，式（5.29）によるシグモイド曲線の直接あてはめを行う．表 5.22 に対数尤度 $\ln L_i$ が計算できるような適当な $\hat{\beta}_0$ と $\hat{\beta}_1$ を設定し，Excel のソルバーで，対数尤度 $\ln L$ を最大化するように $\hat{\beta}_0$ と $\hat{\beta}_1$ を変化させた結果を示す．

表 5.22　冠動脈心疾患の死亡率に対する最小極値分布のあてはめ

	デザイン行列		死亡	母集団	死亡率		シグモイド			組合せ
	X		者数	人数	***Y/n***	***Xβ^***	最小極値	二項分布		無
i	X_0	X_1	***Y***	***n***	***P***	***Z^***	***π^***	尤度 L_i	$\ln L_i$	$\ln L_i'$
1	1	30	1	17,742	0.0001	-8.4956	0.000204	0.0966	-2.3376	-12.1213
2	1	35	5	16,554	0.0003	-7.9729	0.000345	0.1677	-1.7857	-45.5695
3	1	40	5	16,059	0.0003	-7.4501	0.000581	0.0522	-2.9536	-46.5856
4	1	45	12	13,083	0.0009	-6.9273	0.000980	0.1114	-2.1949	-95.9515
5	1	50	25	10,784	0.0023	-6.4045	0.001653	0.0220	-3.8157	-177.9297
6	1	55	38	9,645	0.0039	-5.8817	0.002786	0.0084	-4.7837	-250.3619
7	1	60	54	10,706	0.0050	-5.3590	0.004695	0.0474	-3.0488	-339.6370
8	1	65	65	9,933	0.0065	-4.8362	0.007906	0.0142	-4.2522	-392.9349
					β_0^=	**-11.6323**			**-25.1721**	-1,361.0914
					β_1^=	**0.1046**			対数尤度 $\ln L$	$\ln L'$

プロビットの場合とほぼ同様なパラメータの推定値

$$\hat{\beta}_0 = -11.6323, \quad \hat{\beta}_1 = 0.1046$$

が得られている．

$$\hat{\mu}_{SEV} = \frac{-\hat{\beta}_0}{\hat{\beta}_1} = \frac{-(-11.6323)}{0.1046} = 111.25, \quad \hat{\sigma}_{SEV} = \frac{1}{\hat{\beta}_1} = \frac{1}{0.1046} = 9.56$$

従って，位置パラメータが 111.25 歳，形状パラメータが 9.56 歳の最小極値分布をあてはめたことになる．

JMP による補 2 重対数による解析結果を表 5.23 に示す．Excel での結果に一致することが確認される．

表 5.23　JMP による補 2 重対数解析

モデル	(-1)*対数尤度	尤度比カイ2乗
差分	121.6753	243.3506
完全	1361.0914	
縮小	1482.7667	

項	推定値	標準誤差	尤度比カイ2乗
切片	-11.6323	0.4531	1501.6613
x_age	0.1046	0.0078	243.3506

ロジット

2 項分布に従う出現率 $P_i = Y_i / n_i$ について，説明変数を X_i とした場合に，ロジスティック分布をシグモイド曲線として

$$P_i = \frac{\exp(\beta_0 + \beta_1 X_{1,i})}{1+\exp(\beta_0 + \beta_1 X_{1,i})} + \varepsilon_i , \quad \varepsilon_i \sim Binomial(Y_i, n_i; \pi_i) \tag{5.31}$$

用いる．リンク関数の「ロジット」は，式（5.31）を推定式にした回帰式について解くと

$$\hat{P}_i + \hat{P}_i \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i}) = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i})$$

$$\hat{P}_i = (1-\hat{P}_i)\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i})$$

$$Z_i = \ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} \tag{5.32}$$

となり，標準的な関数で表わされることから広く使われている．

表 5.24　冠動脈心疾患の死亡率に対するロジスティック曲線のあてはめ

	デザイン行列		死亡	母集団	死亡率	ロジスティック				組合せ
	X		者数	人数	Y/n	$X\beta$^	曲線	二項分布		無
i	X_0	X_1	Y	n	p	Z^	π^	尤度 L_i	$\ln L_i$	$\ln L_i'$
1	1	30	1	17,742	0.0001	-8.4978	0.000204	0.0971	-2.3317	-12.1154
2	1	35	5	16,554	0.0003	-7.9741	0.000344	0.1679	-1.7846	-45.5685
3	1	40	5	16,059	0.0003	-7.4505	0.000581	0.0523	-2.9505	-46.5825
4	1	45	12	13,083	0.0009	-6.9269	0.000980	0.1114	-2.1949	-95.9515
5	1	50	25	10,784	0.0023	-6.4032	0.001653	0.0221	-3.8124	-177.9265
6	1	55	38	9,645	0.0039	-5.8796	0.002788	0.0084	-4.7755	-250.3537
7	1	60	54	10,706	0.0050	-5.3560	0.004698	0.0475	-3.0464	-339.6346
8	1	65	65	9,933	0.0065	-4.8323	0.007905	0.0143	-4.2504	-392.9331
					β_0^=	**-11.6395**			**-25.1465**	-1,361.0658
					β_1^=	**0.1047**			対数尤度 $\ln L$	$\ln L'$

ここでは，ロジット変換した 式（5.32）ではなく，式（5.31）によるシグモイド曲線の直接あてはめを行う．表 5.24 に対数尤度 $\ln L_i$ が計算できるような適当な $\hat{\beta}_0$ と $\hat{\beta}_1$ を設定し，Excel のソルバーで，対数尤度 $\ln L$ を最大化するように $\hat{\beta}_0$ と $\hat{\beta}_1$ を変化させた結果を示す．

補 2 重対数の場合と同様なパラメータの推定値

$$\hat{\beta}_0 = -11.6395\,, \qquad \hat{\beta}_1 = 0.1047$$

が得られている．

$$\hat{\mu}_{LGT} = \frac{-\hat{\beta}_0}{\hat{\beta}_1} = \frac{-(-11.6325)}{0.1046} = 111.14\,, \qquad \hat{\sigma}_{LGT} = \frac{1}{\hat{\beta}_1} = \frac{1}{0.1047} = 9.55$$

従って，位置パラメータが 111.14 歳，形状パラメータが 9.55 歳のロジスティック分布をあてはめたことになる．

JMP によるロジットによる解析結果を
表 5.26 に示す．Excel での結果に一致することが確認される．

表 5.25　JMP によるロジット解析

モデル	(-1)*対数尤度	尤度比カイ2乗	項	推定値	標準誤差	尤度比カイ2乗
差分	121.7009	243.4019	切片	-11.6395	0.4538	1499.0600
完全	1361.0658		x_age	0.1047	0.0078	243.4019
縮小	1482.7667					

ポアソン・プロビット・補 2 重対数・ロジット

対数リンクでオフセットがある場合のポアソン回帰に対比して，2 値データに対する誤差分布を 2 項分布，リンク関数を（プロビット・補 2 重対数・ロジット）とした解析結果を示してきた．いずれの方法でも同様の結果が得られるのであるが，推定されたパラメータを相互に比較する．
表 5.26 に比較結果を示すが，プロビットを除いて，他の方法はほとんど同様な値となっていてることをあらためて認識させられる．

（プロビット・補 2 重対数・ロジット）は，どれも 0〜1 の範囲のシグモイド曲線としての分布関数のあてはめであり，単調増加を前提にしている．第 5.4 節で取り上げたように，反応に頭打ちがあるような場合に，2 次式に対するポアソン回帰を行った．シグモイド曲線のあてはめを前提にする解析に対して，シグモイド曲線を否定するような 2 次式の適用は受け入れがたい．したがって，稀な現象であるが母集団のサイズが分かっているような場合については，オフセット付きの対数リンクのポアソン回帰が標準的な方法として薦められる．

なお，補 2 重対数の最小極値分布は，左に長く裾を引く分布であり，その位置パラメータは，累積分布の 0.623 パーセント点となり，正規分布とロジスティック分布の 0.50 パーセント点と異なるが，ここでの事例は，発現率が 1%未満であり，3 種の方法にほとんど違いがない．

表 5.26　JMP によるポアソン・プロビット・補 2 重対数・ロジット解析

	ポアソン	プロビット	補2重対数	ロジット
β_0^	-11.6278	-4.6119	-11.6323	-11.6395
β_1^	0.1044	0.0337	0.1046	0.1047

ポアソン回帰のみならず，(0，1) データに対する一般化線形モデルに対して，統計ソフトの結果と Excel のソルバーを用いた結果を確認することは，統計モデルに対する知識を確実なものにすることが期待される．

上限があるシグモイド曲線のあてはめ

シグモイド曲線のあてはめの基本は，反応が 0%から 100%の範囲である．ただし，現実の用量反応関係では，低用量の場合に自然反応などがあり 0%に収束しない場合も良く知られている．また，特定の疾患の死亡率では，本質的に 100%になりえない場合もあり，シグモイド曲線の上限を得られたデータから推測することが現実的である．第 2.6 節では，「死亡率の上限を新たな変数とするロジスティック回帰」について例示したので参考にしてもらいたい．同様な方法で，上限をパラメータ U_{limit} として加えたプロビット曲線

$$P_i' = U_{limit}\Phi(\beta_0 + \beta_1 X_i) + \varepsilon_i, \quad \varepsilon_i \sim Binomial(r_i, n_i; \pi_i) \tag{5.33}$$

によって推定された結果を図 5.4 に示す．このようなシグモイド曲線が求められる条件は，低年齢層のデータおよび高年齢層の死亡率に頭打ち現象がみられる場合である．推定結果は，上限が 0.8966 パーセント，平均が 57.85 歳，標準偏差が 11.09 歳のシグモイド曲線（累積正規分布）があてはめられる．

$$\hat{\pi}'_{NOR} = 0.008966 \times \Phi(-5.2153 + 0.0901x)$$

$$\hat{\mu}'_{NOR} = \frac{-\hat{\beta}_0}{\hat{\beta}_1} = \frac{-(-5.2153)}{0.0901} = 57.85, \quad \hat{\sigma}'_{NOR} = \frac{1}{\hat{\beta}_1} = \frac{1}{0.0901} = 11.09$$

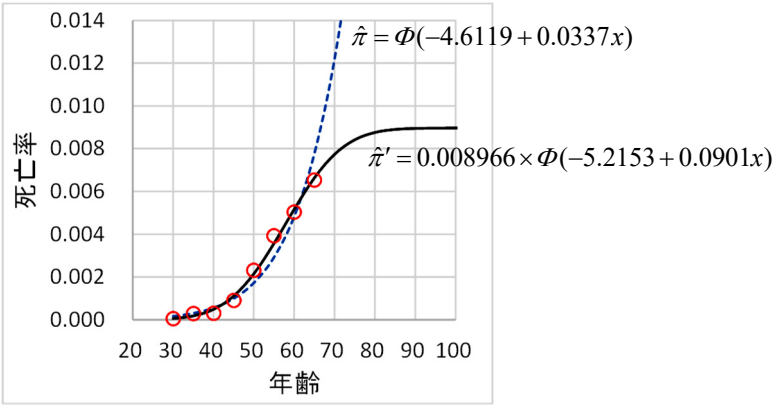


図 5.4　上限を持つプロビット曲線のあてはめ

6. 過分散・ゼロ過剰への対処

第 1.13 節で取り上げた, 雌に連結する雄のサテライト数は, 平均が 2.9191 匹, 分散が 9.9120 匹であり，分散/平均の比が 3.40 と 1 よりもかなり大きな過分散であり，ポアソン分布のあてはめは否定的であった．そのために，過分散を考慮するガンマ・ポアソン分布（負の 2 項分布）のあてはめを JMP で行ったが，ゼロとなるサテライト数が多いこともあり，ゼロ過剰（Zero-Inflated）モデルの適用の検討も示唆した．また，第 1.12 節で取り上げた「医院への通院回数」のように，稀な現象ではないカウント・データは，過分散となりがちであることが知られている．このように過分散によりポアソン分布があてはめに難点があるカウント・データに対し，分散の大きさを調整できるガンマ・ポアソン分布（負の 2 項分布）のあてはめが望まれる．そこで，ガンマ・ポアソン分布の特徴について示し，ゼロ過剰ポアソン分布，および，ゼロ過剰ガンマ・ポアソン分布についても検討を加える．さらに，これらの分布を誤差分布としたガンマ・ポアソン回帰，および，ゼロ過剰ガンマ・ポアソン回帰を適用した結果についても示す．

6.1. 負の 2 項分布

成功数を固定

負の 2 項分布は，成功の確率を π としたときに，一定の成功数 k が得られるまでの試行を行なった場合の失敗数 Y の分布であり，次式で与えられる．

$$NegBinomial(Y;\, k,\pi) = \left[\binom{Y+k-1}{k-1} \pi^{k-1} (1-\pi)^{Y} \right] \pi \qquad (6.1)$$

簡単な事例で説明しよう．成功の確率を $\pi = 0.65$，成功数を $k = 3$ と固定する．成功する場合を○，失敗する場合を×とした時に，3 回連続して成功する場合は，

1	2	3 回目
○	○	○

となり，失敗の数は $Y = 0$ なので，その確率は，

$$NegBinomial(0;\, k,\pi) = \left[\binom{0+3-1}{3-1} \pi^{3-1} (1-\pi)^{0} \right] \pi = \binom{2}{2} 0.65^{3} = 0.2746$$

である．最後の第 3 回目の試行は，常に成功○なので，第 2 回目までの試行に対する確率が主体となる．

次に 4 回目の施行が成功○となる場合を考えよう. 第 1 回目から第 3 回目まで 1 回は失敗, 2 回は成功の場合は, 4 回目で成功○なので成功数は $k=3$ となり, 次の 3 通りの場合である.

1	2	3	4 回目
×	○	○	○
○	×	○	○
○	○	×	○

失敗数は $Y=1$，成功数は $k=3$ なので，この確率は,

$$NegBinomial(1;\ k,\pi)=\left[\binom{1+3-1}{3-1}\pi^{3-1}(1-\pi)^{1}\right]\pi=\binom{3}{2}0.65^{3}\times 0.35^{1}=3\times 0.0961=0.2884$$

である，引き続き 5 回目が成功○となる場合を考えよう．第 1 回目から第 4 回目までに 2 回失敗し，5 回目で成功する場合は，次の 6 通りであり,

1	2	3	4	5 回目
×	×	○	○	○
×	○	×	○	○
×	○	○	×	○
○	×	×	○	○
○	×	○	×	○
○	○	×	×	○

その確率は,

$$NegBinomial(2;k,\pi)=\left[\binom{2+3-1}{3-1}\pi^{3-1}(1-\pi)^{2}\right]\pi=\binom{4}{2}0.65^{3}\times 0.35^{2}=6\times 0.0336=0.2018$$

である．更に 6 回目で成功，7 回目で成功と全体の成功の数を $k=3$ と固定し，順次計算して確率が 0.0000 となるまで計算した結果を表 6.1 に示す．2 項分布と負の 2 項分布を対比すると，2 項分布は，N 回目の試行中 $Y=0,1,...,N$ 回の成功する確率であり，負の 2 項分布は，最後の試行が成功するとした場合に，試行回数は制限せずに $k-1$ 回が成功する確率に，最後の試行が成功する確率を掛ける.

分布	試行回数	成功の確率	
2 項分布	N 回の試行中	Y 回の成功する確率	
負の 2 項分布	$Y+k-1$ 回の試行中	$k-1$ 回の成功する確率	k 回目は成功

負の 2 項分布は，成功数 k を固定し失敗数 Y を変化させることに対応している．組み合わせ数は，Excel の $\text{Combin}(Y+k-1,k-1)$ 関数で計算し，（成功の確率の k 乗）×（失敗の確率の

Y 乗）を求め，その積から負の 2 項分布の確率を計算している．なお，Excel には，負の 2 項分布の関数 NegBinom.dist()

$$NegBinomial(Y;k,\pi) = \text{NegBinom.dist}(Y,k,\pi,\ false)$$

を使って直接計算することもできる．

交通事故の件数

ある地域で交通事故の発生の（あり，なし）の日を想定する．失敗を事故が起きた日とする．事故が起きなかった日を成功とし，観測を始めて事故が起きなかった日が k=3 日となるまでに事故が起きた日の数を失敗数 Y とする．事故が起きなかった日（成功）が 3 日となった翌日から再び観測を始める．事故が起きない日（成功）の確率を π=0.65 とした場合の失敗数 Y （事故が起きた日数）は，負の 2 項分布に従うと想定される．

表 6.1 に，負の 2 項分布に従うとしとしたときの失敗数 Y の確率の計算結果を示す．試行数を $(Y+k-1)$ としたときの成功数が $(k-1)$ となる組み合わせ数を Excel の Combin(Y+k–1, k–1) で計算し，成功が k 日かつ事故が起きた日数 Y の確率 $\pi^{k-1}(1-\pi)^{Y}\pi$ を計算し，それらの積を計算している．組合せ数と確率の積から失敗数 Y に対する負の 2 項分布の確率が計算されている．もちろん, NegBinom.dist()関数で計算した結果と一致する．このような観察を 1,000 回間続けたとした場合の失敗 Y の件数 n_i を計算している．事故が 3 日間連続して起きなかった（$Y_1=0$）のは 275 件，1 日だけ事故が起きた（$Y_2=1$）のは 289 件で，最も事故が多発した（$Y_{10}=9$）場合には 1 件である．

表 6.1　負の 2 項分布（k=3，π=0.65）に対する事故件数の例

	試行数	成功数	失敗数	負の二項分布			1,000回中
i	Y+k-1	k-1	Y	組合せ	$\pi^{k-1}(1-\pi)^{Y}\pi$	確率 P^{NB}	件数 n
1	2	2	0	1	0.274625	0.274625	275
2	3	2	1	3	0.096119	0.288356	289
3	4	2	2	6	0.033642	0.201849	202
4	5	2	3	10	0.011775	0.117745	118
5	6	2	4	15	0.004121	0.061816	62
6	7	2	5	21	0.001442	0.030290	30
7	8	2	6	28	0.000505	0.014135	14
8	9	2	7	36	0.000177	0.006361	6
9	10	2	8	45	0.000062	0.002783	3
10	11	2	9	55	0.000022	0.001190	1
11	12	2	10	66	0.000008	0.000500	0
12	13	2	11	78	0.000003	0.000207	0
13	14	2	12	91	0.000001	0.000084	0
14	15	2	13	105	0.000000	0.000034	0
		k= 3		**π= 0.65**		0.999977	1,000

負の 2 項分布 vs. ポアソン分布

表 6.2 に示すように，ある地域での事故がない日（成功）が 3 日となるまでの事故の発生件数の統計が得られたして，この分布はポアソン分布に従うと見なすことができるのだろうか．事故件数の平均 $\overline{Y}$ は，

$$\overline{Y} = \frac{\sum_{i=1}^{10} n_i Y_i}{N} = \frac{1,604}{1,000} = 1.6040$$

となり，分散 $Var(Y)$ は，

$$Var(Y) = \frac{\sum_{i=1}^{10} n_i (Y_i - \overline{Y})^2}{N-1} = \frac{2399.1840}{999} = 2.4016$$

であり，過分散（分散/平均＝1.4972）である．ポアソン分布を仮定した場合の

$$P_i = \text{Poisonn.dist}(Y_i, \mu = 1.6040, false)$$

の確率に 1,000 回を掛けた事故件数との差は，大きなプラスマイナスがあり，あてはまっているとは言い難い．事故件数の分布がポアソン分布と見なせない場合には，ある地域で特異的に事故が発生しやすい交差点がいくつかあるような場合，複数のポアソン分布が混合している場合にも過分散が起きやすい．

表 6.2　事故件数の人工データに対するポアソン分布のあてはめ

	事故日数	発生件数	ポアソン分布		件数
i	Y	n	確率 P	件数	差
1	0	275	0.2011	201.1	**-73.9**
2	1	289	0.3225	322.5	33.5
3	2	202	0.2587	258.7	56.7
4	3	118	0.1383	138.3	20.3
5	4	62	0.0555	55.5	**-6.5**
6	5	30	0.0178	17.8	**-12.2**
7	6	14	0.0048	4.8	**-9.2**
8	7	6	0.0011	1.1	**-4.9**
9	8	3	0.0002	0.2	**-2.8**
10	9	1	0.0000	0.0	**-1.0**
11	10	0	0.0000	0.0	**0.0**
	全体	1,000	1.6040	2.4016	1.4972
		N	平均	分散	比

Excel には，平均と分散の計算に Avarege()関数と Ver.s()関数があるが，度数 n_i を含む場合の関数はないので，平均=SumProduct(***Y*** の範囲, ***n*** の範囲)/*N*，分散=SumProduct((***Y*** の範囲 － 平均)^2 , ***n*** の範囲) / (*N*–1) で直接計算することができる．

図 6.1 に負の 2 項分布を仮定して生成した事故件数データに対し，ポアソン分布をあてはめた場合を示す．負の 2 項分布から生成されたデータから，平均が 1.6040，分散が 2.4061 であり，ポアソン分布は，平均が 1.6040，分散が同じく 1.6040 なので，ポアソン分布は，負の 2 項分布に対して，相対的に尖った形状となっている．

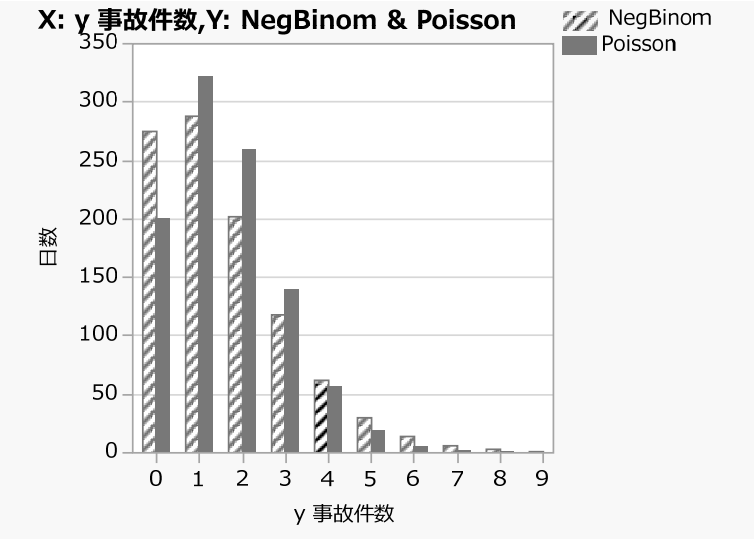


図 6.1　負の 2 項分布に従う事故件数データへのポアソン分布のあてはめ

負の 2 項分布のパラメータ推定

実際に観察されたデータで負の 2 項分布をあてはめてパラメータを推定したいが，どうしたら良いのだろうか．ポアソン分布の位置パラメータ μ は，算術平均と等しいが，負の 2 項分布のパラメータである出現確率 π，成功数 k を与えられた度数分布から推定したい．ある初期値を与えて負の 2 項分布の確率 P_i を計算し，対数尤度

$$\ln L^{NB} = \sum_{i=1}^{10} n_i (\ln \hat{P}_i^{NB}) = \sum_{i=1}^{10} n_i (\ln(\text{NegBinom.dist}(Y_i, \hat{k}, \hat{\pi}, \mathit{false})))$$

を求めて，Excel のソルバーで，$\ln L^{NB}$ を最大にするように出現確率 $\hat{\pi}$，成功数 $\hat{k}$ を変化させれば，最尤解が求められると思うかもしれない．残念ながら，成功数 $\hat{k}$ は整数なので，NegBinom.dist() 関数を使った場合には，ソルバーで変化させ最尤解を得ることができない．

負の 2 項分布は，階乗の計算を含むので，これをガンマ関数で置き換え，

$$\left.\begin{aligned}\left[\binom{Y+k-1}{k-1}\pi^{k-1}(1-\pi)^{Y}\right]\pi &= \frac{(Y+k-1)!}{[(Y+k-1)-(k-1)]!\,(k-1)!}\pi^{k}(1-\pi)^{Y} \\ &= \frac{\Gamma(Y+k)}{\Gamma(Y+1)\,\Gamma(k)}\pi^{k}(1-\pi)^{Y}\end{aligned}\right\} \qquad (6.2)$$

実数でも計算できるようにする．成功数 k が整数であれば，式（6.2）は，負の 2 項分布に一致するが，成功数 k が実数の場合に離散分布の“負の 2 項分布”と言って良いのだろうか．失敗数 Y も実数で与えることも可能であり，連続分布となっている．

事故件数のデータにガンマ関数を用いた“負の 2 項分布”のパラメータを推定するための Excel シートを表 6.3 に示す．“成功数” $\hat{k} = 5.0$ 回，成功の確率 $\hat{\pi} = 0.50$ を初期値にし，ソルバーで対数尤度 $\ln L^{NB'}$ を最大化した結果を示す．最尤解として $\hat{k} = 3.1516$，$\hat{\pi} = 0.6627$ が得られ

ている．人工的に作成した事故件数のデータのパラメータは$k=3$，$\pi=0.65$であったので，完全には一致しないが，“負の 2 項分布”のパラメータ推定を，Excel のソルバーを用いた最尤法により推定できる．なお，総件数Nを増やすことにより，パラメータの推定値は，$k=3$，$\pi=0.65$に漸近する．

表 6.3　ガンマ関数を用いた“負の 2 項分布”のパラメータ推定

	事故日数	件数	試行数	成功数	ガンマ関数・負の二項分布			対数尤度
i	Y	n	$Y+k-1$	$k-1$	組合せ	$\pi^k(1-\pi)^Y$	$P^\wedge_i{}^{NB'}$	$\ln L_i$
1	0	275	2.1516	2.1516	1.0000	0.273457	0.2735	-356.57
2	1	289	3.1516	2.1516	3.1516	0.092233	0.2907	-357.07
3	2	202	4.1516	2.1516	6.5421	0.031109	0.2035	-321.58
4	3	118	5.1516	2.1516	11.2342	0.010493	0.1179	-252.30
5	4	62	6.1516	2.1516	17.2771	0.003539	0.0611	-173.26
6	5	30	7.1516	2.1516	24.7118	0.001194	0.0295	-105.70
7	6	14	8.1516	2.1516	33.5735	0.000403	0.0135	-60.25
8	7	6	9.1516	2.1516	43.8931	0.000136	0.0060	-30.74
9	8	3	10.1516	2.1516	55.6982	0.000046	0.0026	-17.91
10	9	1	11.1516	2.1516	69.0138	0.000015	0.0011	-6.84
11	10	0	12.1516	2.1516	83.8629	0.000005	0.0004	0.00
		最尤解	k^=	**3.1516**	π^=	**0.6627**	$\ln L^{NB'}=$	**-1682.23**
		初期値	k^=	**5.0000**	π^=	**0.5000**	$\ln L^{NB'}=$	**-2602.52**

注）なぜ，一定の成功数kが得られるまでの試行を行なった場合の失敗数Yの分布を負の 2 項分布と言うのだろうか．箕谷（2010），統計分布ハンドブック 増補版，p608-10 に「負の 2 項展開に現れる 2 項係数が，負の 2 項分布と同じであることから，この分布は負の 2 項分布とよばれる．」ことが示されている．Excel での組合せ数の計算のガンマ関数は，Gamma()を用いている．対数尤度$\ln L_i = n_i \ln(P_i)$で求め，それらを足し合わせて全体の対数尤度$\ln L$を計算している．

計算シートの$i=1$の行は，

$$Y_1+\hat{k}-1=0+3.1516-1=2.1516$$

$$\text{組合せ}=\frac{\text{Gamma}\left(Y_1+\hat{k}\right)}{\text{Gamma}\left(Y_1+1\right)\times\text{Gamma}\left(\hat{k}\right)}=\frac{2.3106}{1\times 2.3106}=1.0000$$

$$\hat{\pi}^{\hat{k}}(1-\hat{\pi})^{Y_1}=0.6627^{3.1516}\times(1-0.6627)^0=0.273457$$

$$\hat{P}_1^{NB'}=1.0000\times 0.273457=0.2735$$

$$\ln L_1^{NB'}=n_1\ln(\hat{P}_1)=n_1\ln(0.2735)=275\times(-1.2966)=-356.57$$

として計算されている．対数尤度$\ln L_{NB}$は，

$$\ln L^{NB'}=\sum_{i=1}^{10}\ln L_i=-356.57-357.07-\cdots-6.84=-1682.23$$

である．なお，負の 2 項分布のパラメータ推定ができたとしても，成功数$\hat{k}=3.1516$回，成功の確率$\hat{\pi}=0.6627$と解釈はできても，ポアソン分布のパラメータと不整合であり，扱いにくい．

6.2. ガンマ・ポアソン分布

位置および形状パラメータに変換

Agresti（2013），Categorical Data Analysis 3rd. の section 14.4 の Negative Binomial Regression に，ガンマ分布とポアソン分布を融合した場合に，“負の 2 項分布”となることが示されている．“負の 2 項分布”は，式（6.2）で示したように，

$$NegBinom(Y;k,\pi)=\frac{\Gamma(Y+k)}{\Gamma(Y+1)\Gamma(k)}\pi^{k}(1-\pi)^{Y} \tag{6.3}$$

となる．負の 2 項分布のパラメータは，π と k であり，ポアソン分布のパラメータと全く異なるので，パラメータが平均 μ（位置パラメータ）と過分散 σ（形状パラメータ）となるように変換したい．負の 2 項分布の成功の確率 π を平均 μ と k で

$$\pi=\frac{k}{\mu+k} \tag{6.4}$$

のように置き換えると

$$GammaPoisson(Y;\mu,k)=\frac{\Gamma(Y+k)}{\Gamma(Y+1)\Gamma(k)}\left(\frac{k}{\mu+k}\right)^{k}\left(1-\frac{k}{\mu+k}\right)^{Y} \tag{6.5}$$

となる．さらに，k を $1/\sigma$ で置き換え，整理すると

$$\left.\begin{aligned}GammaPoisson(Y;\mu,\sigma)&=\frac{\Gamma(Y+1/\sigma)}{\Gamma(Y+1)\Gamma(1/\sigma)}\left(\frac{1/\sigma}{\mu+1/\sigma}\right)^{1/\sigma}\left(1-\frac{1/\sigma}{\mu+1/\sigma}\right)^{Y}\\&=\frac{\Gamma(Y+1/\sigma)}{\Gamma(Y+1)\Gamma(1/\sigma)}\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}\left(1-\frac{1}{1+\sigma\mu}\right)^{Y}\\&=\frac{\Gamma(Y+1/\sigma)}{\Gamma(Y+1)\Gamma(1/\sigma)}\cdot\frac{(\mu\sigma)^{Y}}{(1+\mu\sigma)^{Y+1/\sigma}}\end{aligned}\right\} \tag{6.6}$$

が得られる．前節の表 6.1 で，成功の数を $k=3$，成功の確率 $\pi=0.65$ とする負の 2 項分布の確率を計算し，1,000 回の観察で事故件数の分布を例示した．ガンマ・ポアソン分布のパラメータは，式（6.4）を平均 μ（位置パラメータ）について解くと，

$$\begin{aligned}\mu&=\frac{(1-\pi)k}{\pi}\\&=\frac{(1-0.65)\times 3}{0.65}=1.6154\end{aligned} \tag{6.7}$$

と負の 2 項分布の期待値になり，過分散 σ（形状パラメータ）は，

$$\sigma=\frac{1}{k}=\frac{1}{3}=0.3333 \tag{6.8}$$

となる．

負の二項分布の分散は，

$$Var(Y)=\frac{(1-\pi)k}{\pi^2} \tag{6.9}$$

なので，式（6.4）の $\pi=k/(\mu+k)$ を代入し，整理し，さらに k を $k=1/\sigma$ 置き換えると

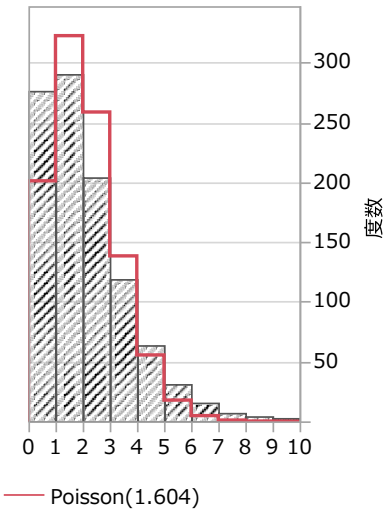
$$\left.\begin{aligned} Var(Y)&=\frac{(1-\pi)k}{\pi^2}\\ &=\mu\cdot\frac{\mu+k}{k}\\ &=\mu(1+\mu\sigma)\end{aligned}\right\} \tag{6.10}$$

が得られる．ガンマ・ポアソン分布の分散は，期待値 μ を含んでいる．したがって，過分散の形状パラメータ σ は，いわゆる分散の平方根に相当しているわけではないが，期待値 μ が推定されれば，分散も推定できるような「形状パラメータ」と理解される．

ガンマ・ポアソン分布のパラメータ推定

JMP の「一変量の分布」を用いて，事故件数データにポアソン分布およびガンマ・ポアソン分布の当てはめた結果を表 6.4 に示す．ポアソン分布のあてはめは，Pearson の適合度検定の結果からも，$p<0.0001$ と全く支持されない．当然ながらガンマ・ポアソン回帰は良くあてはまっていて $p=0.5631$ と棄却されない．

表 6.4　事故の人工データに対するポアソン分布のあてはめ



Poisson分布のあてはめ

パラメータ推定値

種類	パラメータ	推定値
尺度	λ	1.6040

指標	
(-2)*対数尤度	3452.1855
AICc	3454.1895
BIC	3459.0932

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
1495.75062	<.0001*

ガンマPoissonのあてはめ

パラメータ推定値

種類	パラメータ	推定値
位置	λ	1.6040
過分散	σ	1.5089

指標	
(-2)*対数尤度	3364.4503
AICc	3368.4623
BIC	3378.2658

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
991.255338	0.5631

推定されたパラメータは，$\hat{\lambda}=\hat{\mu}=1.6040$ と式（6.7）の期待値 1.6154 とほぼ同様の推定値が得られている．推定された過分散パラメータは，$\hat{\sigma}=1.5089$ と式（6.8）での $\sigma=1/k=0.3333$ と全く異なる．これは，JMP（Ver. 14）の「一変量のあてはめ」でガンマ・ポアソン分布のあてはめた場合は，過分散 σ' として次式

$$\sigma'=1+\mu\sigma=1+1.6154\times 0.3333=1.5385$$

が使われている．このσ'は，過分散がないポアソン分布の場合に，$\sigma'=1$となり，さらに$\mu\sigma$を加えて平均値と過分散のスケールを合わせるための対応と解される．表 6.3 で推定された負の 2 項分布のパラメータ$\hat{\pi}=0.6627$および$\hat{k}=3.1513$を使い，

$$\hat{\mu}=\frac{(1-\hat{\pi})\hat{k}}{\hat{\pi}}=\frac{(1-0.6627)\times 3.1516}{0.6627}=1.6040$$

$$\hat{\sigma}=\frac{1}{\hat{k}}=\frac{1}{3.1516}=0.3173$$

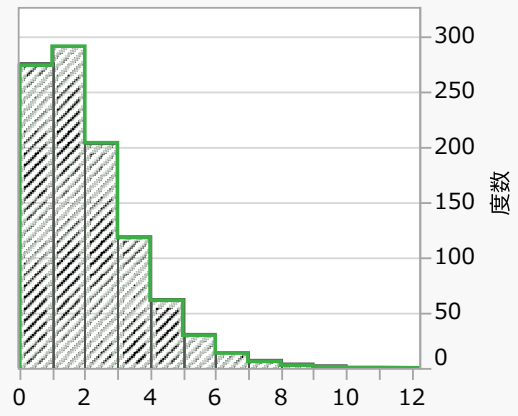
が得られる．表 6.4 で示した$\hat{\mu}$に一致し，$\hat{\mu}$と$\hat{\sigma}$から，

$$\hat{\sigma}'=1+\hat{\mu}\hat{\sigma}=1+1.6040\times 0.3173=1.5089$$

が推定され，表 6.4 の「過分散σ」に一致する．「指標」の欄の (–2)*対数尤度=3364.4503 は，表 6.3 で求めた対数尤度$\ln L=-1682.23$をマイナス 2 倍した結果に一致する．

2019 年にリリースされた JMP の Ver. 15 では，離散分布の選択画面から「ガンマ Poisson 分布のあてはめ」が消え，「負の 2 項分布のあてはめ」に置き換わっている．推定されるパラメータは，「平均 λ」は同じであるが，「過分散σ」に代え「ばらつきσ」となり，$\hat{\sigma}=0.3173$が推定されている．したがって，「負の 2 項分布」の通常のパラメータ（k，π）が推定されているわけではない．

表 6.5　事故の人工データに対する負の 2 項分布のあてはめ



負の二項分布のあてはめ

パラメータ		推定値
平均	λ	1.6040
ばらつき	σ	0.3173
指標		
(-2)*対数尤度		3364.4503
AICc		3368.4623
BIC		3378.2658

適合度検定

	X2	自由度	Prob>X2
Pearsonのカイ2乗	0.1085365	6	1.0000

JMP（Ver 15.1）での負の 2 項分布のあてはめ．
適合度検定で自由度が 6 となっているが，理由は調査中．

過分散パラメータを変化させた場合の形状

ガンマ・ポアソン分布の過分散パラメータσが 0 に漸近した場合に，ポアソン分布となることが知られているので，実際に確認する．負の 2 項分布の場合は，成功数kは，1 以上の整数であるが，ガンマ・ポアソン分布の場合は実数となるので，組合せ数の計算をガンマ関数

に置き換えて，負の 2 項分布をガンマ・ポアソン分布として成功数 k を実数化して，いくつかの k について分布の形状について検討する．

ガンマ・ポアソン分布の計算で，σが小さくなると $\Gamma(1/\sigma)$ が増大するために計算不能となるので，Excel の対数ガンマ関数 Gammaln()を使って，計算する必要がある．

$$\left.\begin{aligned} GammaPoisson(Y;\sigma,\mu) &= \frac{\Gamma(Y+1/\sigma)}{\Gamma(Y+1)\Gamma(1/\sigma)}\cdot\frac{(\mu\sigma)^{Y}}{(1+\mu\sigma)^{Y+1/\sigma}} \\ &= \exp\left[\ln\Gamma(Y+1/\sigma)-\ln\Gamma(Y+1)-\ln\Gamma((1/\sigma)+Y\ln(\mu\sigma)-(Y+1/\sigma)\ln(1+\mu\sigma)\right] \end{aligned}\right\} \quad (6.11)$$

表 6.6 および図 6.2 に示すように，$\sigma=5$ の場合は，$P_{y=0}=0.6433$ のように大きな確率で，$\mu=1.6154$ を確保するために，Y の大きい方に長く裾を引いている．小さな $\sigma=0.0002$ に対しては，ポアソン分布との差はなくなっている．

表 6.6　過分散σを変化させた場合のガンマ・ポアソン分布の確率（$\mu=1.6154$）

	ガンマ･ポアソン 過分散 σ						ポアソン	σ=0.0002
Y	**5.0**	2.0	1.0	0.3333	0.1000	**0.0002**	P	との差
0	**0.6433**	0.4862	0.3824	0.2746	0.2237	**0.1989**	**0.1988**	0.0001
1	0.1145	0.1856	0.2362	0.2884	0.3111	0.3211	0.3212	0.0000
2	0.0611	0.1063	0.1459	0.2018	0.2380	0.2594	0.2594	0.0000
3	0.0399	0.0677	0.0901	0.1177	0.1324	0.1397	0.1397	0.0000
4	0.0284	0.0452	0.0556	0.0618	0.0598	0.0564	0.0564	0.0000
5	0.0212	0.0311	0.0344	0.0303	0.0233	0.0182	0.0182	0.0000
6	0.0164	0.0217	0.0212	0.0141	0.0081	0.0049	0.0049	0.0000
7	0.0129	0.0154	0.0131	0.0064	0.0026	0.0011	0.0011	0.0000
8	0.0103	0.0110	0.0081	0.0028	0.0008	0.0002	0.0002	0.0000
9	0.0084	0.0080	0.0050	0.0012	0.0002	0.0000	0.0000	0.0000
10	0.0069	0.0058	0.0031	0.0005	0.0001	0.0000	0.0000	0.0000
11	0.0057	0.0042	0.0019	0.0002	0.0000	0.0000	0.0000	0.0000
12	0.0047	0.0031	0.0012	0.0001	0.0000	0.0000	0.0000	0.0000
13	0.0039	0.0023	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000
平均 μ=	**1.6154**	1.6154	1.6154	1.6154	1.6154	1.6154	1.6154	
成功確率 π=	0.7558	0.5532	0.3824	0.1711	0.0583	**0.0001**		
成功数　k=	0.20	0.50	1.00	3.00	10.00	5000.00	—	
JMP: $1+\mu\sigma$=	9.0770	4.2308	2.6154	1.5385	1.1615	1.0003	1.0000	

$GammaPoisson(y=0;\ \sigma=5,\ \mu=1.6154)=0.6433$ に対する計算は，

=Exp(Gammaln(0+1/5) - Gammaln(0+1) - Gammaln(1/5) +0×ln((1.6154×5) – (0+1/5)×ln((1+1.6154×5))

=Exp(1.5241+0.0000 - 1.5241+0.0000 - 0.4411) =Exp(-0.4411)= 0.6433　で求められている．

図 6.2 に示すように過分散パラメータの $\sigma=0.10$ の場合に●印が，ほぼポアソン分布に一致している．過分散パラメータ $\sigma=0.10$ の場合（＊印）は，ほぼポアソン分布に重なっているが，$\sigma=0.33$（□印）になるとポアソン分布に比べ $Y=0$ の確率が上方に乖離し，$Y=1,2,3$ の場合は，下方に乖離し，$Y\geq 4$ の場合は，図では識別しづらいが，●印から上方に乖離しているこ

とが表 6.6 からわかる．このような上下への乖離は，平均を $\mu = 1.6154$ と同じとするために，反応を 0 とする割合が極端に大きくなり，それを補正するために右に長く裾を引く分布となる．

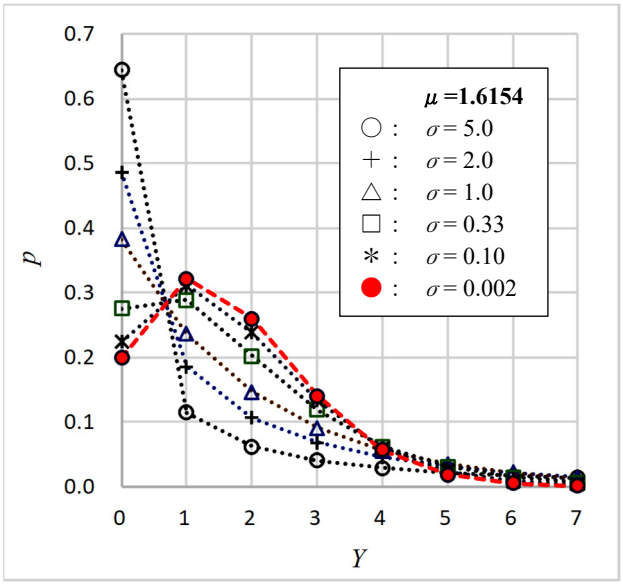


図 6.2　平均 $\mu = 1.6154$ に対するガンマ・ポアソン分布の形状の比較

図 6.3 に平均を $\mu = 4.0$ とした場合のガンマ・ポアソン分布の形状を示す．形状パラメータ $\sigma = 0.33$ までは，ポアソン分布と同様に一山型であるが，$\sigma \geq 1.0$ 以上では，指数分布的な片流れ型となっている．

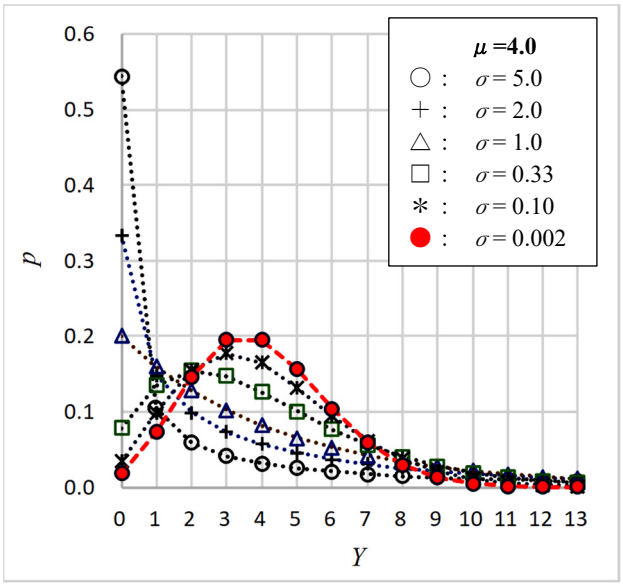


図 6.3　平均 $\mu = 4.0$ に対するガンマ・ポアソン分布の形状の比較

なお，負の 2 項分布とガンマ・ポアソン分布の数学的な解説については，岩崎（2010），「カウントデータの統計解析」の「第 6 章　負の二項分布」に詳しく示されている．

6.3. 過分散の事例

過分散の事例とし，第 1.12 節の図 1.14 および表 1.43 で「医院への通院回数」について JMP の「一変量の分布」でガンマ・ポアソン分布をあてはめた結果を示した［Cameron and Trivedi（1998)］．ここでは，Excel のソルバーを用いた最尤法によるあてはめを示す．表 6.7 で示すように通院回数の合計は，5,190 回であり，平均は，SumProduct()関数を用いて

$$\text{平均}=\frac{\text{SumProduct}(\boldsymbol{n}\text{の範囲},\boldsymbol{Y}\text{の範囲})}{\text{Sum}(\boldsymbol{n}\text{の範囲})}=\frac{1{,}566}{5{,}190}=0.3017$$

として求められる．分散は，

$$\text{分散}=\frac{\text{SumProduct}[\boldsymbol{n}\text{の範囲},(\boldsymbol{Y}\text{の範囲}-\text{平均})\wedge 2]}{\text{Sum}(\boldsymbol{n}\text{の範囲})-1}=\frac{3{,}305.48}{5{,}190-1}=0.6370$$

として求められる．その比は，$0.6370/0.3017=2.1112$ と過分散となっている．

表 6.7　医院への通院回数に対するポアソン分布およびガンマ･ポアソン分布のあてはめ

		度数	構成比		ポアソン分布			ガンマ･ポアソン分布		構成比
i	Y	n	p		$P\wedge^{P}$	$\ln L_i$		$P\wedge^{GP}$	$n\ln L_i^{GP}$	との差
1	0	4,141	0.7979		0.7395	-1249.5		**0.8011**	**-918.3**	0.0032
2	1	782	0.1507		0.2231	-1173.0		0.1344	-1569.7	-0.0163
3	2	174	0.0335		0.0337	-590.1		0.0411	-555.3	0.0076
4	3	30	0.0058		0.0034	-170.6		0.0145	-127.1	0.0087
5	4	24	0.0046		0.0003	-198.5		0.0054	-125.2	0.0008
6	5	9	0.0017		0.0000	-99.7		0.0021	-55.4	0.0004
7	6	12	0.0023		0.0000	-168.8		0.0008	-85.0	-0.0015
8	7	12	0.0023		0.0000	-206.6		0.0003	-95.8	-0.0020
9	8	5	0.0010		0.0000	-102.5		0.0001	-44.4	-0.0008
10	9	1	0.0002		0.0000	-23.9		0.0001	-9.8	-0.0001
	合計	5,190			$\ln L^{P}=$	**-3983.2**		$\ln L^{GP}=$	**-3586.0**	
	平均=	**0.3017**	$\mu\wedge=$		**0.3017**	$\mu\wedge=$		**0.3017**		
	分散=	**0.6370**				$\sigma\wedge=$		**2.6487**		
	分散/平均=	**2.1112**				$\sigma'\wedge=$		**1.7992**	$\hat{\sigma}'=1+\hat{\mu}\hat{\sigma}$	

表 6.7 のポアソン分布のあてはめは，適当な $\hat{\mu}$ を設定し，対数尤度 $\ln L$ を Excel のソルバーで最大化した結果を示す．もちろん $\hat{\mu}=0.3017$ は，算術平均に一致している．ガンマ・ポアソン分布は，式（6.6）を用いて適当な $\hat{\mu}$ と $\hat{\sigma}$ を初期値を設定し，分布の確率 P_i^{GP} を計算し，対数尤度 $\ln L_i^{GP}=n_i\ln P_i^{GP}$ の和としての対数尤度 $\ln L^{GP}$ を Excel のソルバーで最大化した結果を示す．

ガンマ・ポアソン分布の確率は，$Y=0$ の場合

$$\begin{aligned} P_1^{GP} &= \frac{\Gamma(Y+1/\sigma)}{\Gamma(Y+1)\Gamma(1/\sigma)} \cdot \frac{(\mu\sigma)^Y}{(1+\mu\sigma)^{Y+1/\sigma}} \\ &= \frac{\Gamma(0+1/2.6487)}{\Gamma(0+1)\Gamma(1/2.6487)} \cdot \frac{(0.3017\times 2.26487)^0}{(1+0.3017\times 2.6487)^{0+1/2.6487}} \\ &= 1\times\frac{1}{1.2483} = 0.8011 \end{aligned}$$

として計算される．その対数尤度 $\ln L_1^{GP}$ は，

$$\begin{aligned} n_1 L_1^{GP} &= n_1 \ln P_1^{GP} \\ &= 4{,}141\times\ln(0.8011) \\ &= -918.3 \end{aligned}$$

である．全体の対数尤度 $\ln L_{GP}$ は，

$$\begin{aligned} \ln L^{GP} &= \sum_{i=1}^{10}(n_i \ln L_i^{GP}) \\ &= -918.3 - 1569.7 - \cdots - 9.8 \\ &= -3586.0 \end{aligned}$$

となっている．ポアソン分布の場合の対数尤度 $\ln L_P$ との差 $\ln L_{diff}$ は，

$$\begin{aligned} \ln L_{diff} &= \ln L^{GP} - \ln L^P \\ &= -3586.0 - (-3983.2) = 397.2 \end{aligned}$$

と桁違いに大きい．差の 2 倍がおおよそ自由度 1 のカイに乗分布に従うとことから，分布のあてはめの比較の目安として使うとすれば，明らかにガンマ・ポアソン分布のあてはめが良いと判断される．JMP の「一変量の分布」の結果は省略するが，表 6.4 に示したと同様に「(–2)*対数尤度」の出力があり，一致する．

ポアソン分布があてはまるかの目安に分散と平均の比をこれまで用いてきた．医院への通院回数のデータでは，分散/平均=2.1122 であり，ガンマ・ポアソン分布の過分散パラメータでは，$\hat{\sigma}=2.6487$ とやや大きい．JMP での過分散 $\hat{\sigma}'=1+\hat{\mu}\hat{\sigma}=1.7992$ であり，過分散が $\hat{\sigma}=0.0$ の場合に $\hat{\sigma}'=1+\hat{\mu}\hat{\sigma}=1.0$ となることから，分散と平均の比と比較する場合と同様な特性となり，使い勝手が良い．

図 6.4 に示すように，ポアソン分布の場合には，通院回数が 0 回の場合に低めな推定となり，1 回の場合に高めの推定となっているのに対し，ガンマ・ポアソン分布の場合には，良くあてはまっていることが読み取れる．

「医院への通院回数」は，カウント・データではあるが，ポアソン分布を仮定する稀な現象とは言い難くなっている．「医院へ通院」する回数は，病気の種類あるいは治療の内容によって一定の期間あたりの通院回数が異なる幾つかの集団から構成されていると想定すれば，必然的に過分散となりがちである．単純に分布のあてはめとするならば，過分散パラメータを持つガンマ・ポアソン分布があてはめられると言っていいのかも知れない．

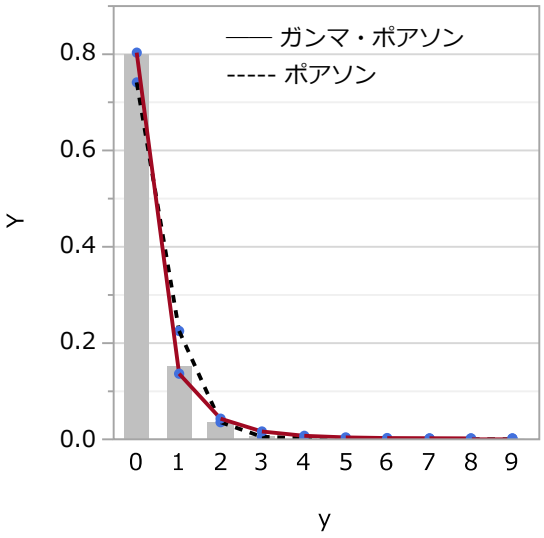


図 6.4　ガンマ・ポアソン分布とポアソン分布のあてはめ

このような得られたデータから，過分散であることが判明した場合には，数院回数に影響を与えるような未知の因子が存在している判断し，探索的な解析が必要と思うべきである．これについては，第 7 章を参照してもらいたい．

6.4. ゼロ過剰ポアソン分布のあてはめ

ゼロの反応が過大となっているの場合に，ゼロとなる件数を割り引いた残りでポアソン分布をあてはめるゼロ過剰（Zero-Inflated）モデルが，Cameron and Trivedi（1998）に示されている．SAS の GENMOD プロシジャでは，ゼロ過剰ポアソン回帰が使えるようになっているが，どのような性質なのか Excel を用いて検討する．なお，JMP の Ver. 15.1 からゼロ過剰ポアソン分布のあてはめがサポートされている．

反応が 0 の場合にポアソン分布のあてはめから除外する確率を ω とする．反応 0 を含む残りのデータに対して，ポアソン分布の確率の和が 1 ではなく $(1-\omega)$ となるようにパラメータ ω および μ を推定する．ゼロ過剰ポアソン分布の確率関数は，次のように定義される．

$$\left.\begin{aligned} f(Y;\omega,\mu) &= \omega + (1-\omega)Poisson(Y;\ \mu), \quad Y=0 \\ f(Y;\omega,\mu) &= (1-\omega)Poisson(Y;\ \mu), \quad Y=1,2,\ldots \end{aligned}\right\} \qquad (6.12)$$

このモデルは，Y が 0 となる部分集団，Y が 0 以上となる部分集団が混在しているような混合分布を想定することになる．

過分散データの事例として，第 1.13 節の「雌のカブトガニに結合する雄のサテライト数」のデータを引き続き用いる［アグレスティ（2003）］．表 6.8 に式（6.12）を用いたゼロ過剰ポアソン分布，式（6.6）を用いたガンマ・ポアソン分布のあてはめ結果を示す．表 6.8 に示すように，ゼロ過剰ポアソン分布のあてはめは，適当な初期値（$\hat{\mu}=3.0000$，$\hat{\omega}=0.3000$）をセットし，ソルバーで対数尤度を最大化するような $\hat{\mu}$ と $\hat{\omega}$ を変化させ，$\hat{\mu}=4.4990$，$\hat{\omega}=0.3512$ を得る．

「確率 ω」の欄は，$Y=0$ に対する $\hat{\omega}$ の推定結果であり，「確率 P」の欄は，平均が 4.4990 のポアソン分布に $(1-\hat{\omega})$ を掛けた結果である．従って，「確率 P」の合計は，$(1-\hat{\omega})=0.6488$ となる．それぞれの Y_i に対する確率 P_i と尤度 $\ln L_i$ は，

$$Y_1=0\ :\ \left\{\begin{aligned} P_1 &= \hat{\omega}+(1-\hat{\omega})\times \text{Poisson.dist}(Y_1=0,\ \hat{\mu}=4.4990,\ false) \\ &= 0.3512+(1-0.3512)\times 0.0111 \\ &= 0.3512+0.007 \\ \ln L_1 &= 62\times\ln(0.3512+0.0072) \\ &= -63.6217 \end{aligned}\right.$$

$$Y_2=1\ :\ \left\{\begin{aligned} p_2 &= (1-\hat{\omega})\times \text{Poisson.dist}(Y=1,\ \hat{\mu}=4.4990,\ false) \\ &= (1-0.3512)\times 0.0500=0.0325 \\ \ln L_2 &= 16\times\ln(0.0325) \\ &= -54.8430 \end{aligned}\right.$$

：

全体の対数尤度 $\ln L^{ZP}$ は，

$$\begin{aligned}\ln L^{ZP} &= \sum_i \ln L_i^{ZP} \\ &= -63.6217 - 54.8430 - \cdots - 10.2731 \\ &= -381.6146\end{aligned}$$

として計算されている．

表 6.8　ゼロ過剰ポアソン分布およびガンマ・ポアソン分布のあてはめ

		初期値	μ^=	**3.0000**	$\ln L^{ZP}$	μ^=	**3.0000**	$\ln L^{GP}$	
			ω^=	**0.3000**	**-415.6585**	σ^=	**1.0000**	**-385.1084**	
		推定結果	μ^=	**4.4990**	$\ln L^{ZP}$	μ^=	**2.9191**	$\ln L^{GP}$	
			ω^=	**0.3512**	**-381.6146**	σ^=	**1.3197**	**-383.7046**	
	サテライト		**ゼロ過剰 ポアソン**				**ガンマ・ポアソン**		
i	Y	n	確率 ω^	確率 P^	$\ln L_i^{ZP}$	n^	確率 P^	$\ln L_i^{GP}$	n^
1	0	62	**0.3512**	**0.0072**	**-63.6217**	62.0	0.3021	-74.2038	52.3
2	1	16		**0.0325**	**-54.8430**	5.6	0.1818	-27.2802	31.4
3	2	9		0.0730	-23.5529	12.6	0.1268	-18.5841	21.9
4	3	19		0.1095	-42.0234	18.9	0.0926	-45.2178	16.0
5	4	19		0.1232	-39.7899	21.3	0.0690	-50.7897	11.9
6	5	15		0.1108	-32.9970	19.2	0.0522	-44.3038	9.0
7	6	13		0.0831	-32.3403	14.4	0.0397	-41.9326	6.9
8	7	4		0.0534	-11.7191	9.2	0.0305	-13.9664	5.3
9	8	6		0.0300	-21.0323	5.2	0.0234	-22.5187	4.1
10	9	3		0.0150	-12.5963	2.6	0.0181	-12.0336	3.1
11	10	3		0.0068	-14.9925	1.2	0.0140	-12.7995	2.4
12	11	1		0.0028	-5.8915	0.5	0.0109	-4.5196	1.9
13	12	1		0.0010	-6.8726	0.2	0.0085	-4.7707	1.5
14	13	0		0.0004	0.0000	0.1	0.0066	0.0000	1.1
15	14	1		0.0001	-9.0689	0.0	0.0052	-5.2686	0.9
16	15	1		0.0000	-10.2731	0.0	0.0040	-5.5156	0.7
	計	173	0.3512	0.6488		173.0	0.9854		170.5
			1.0000						

表 6.8 の右側のガンマ・ポアソン分布のあてはめは，式（6.6）により確率を計算し，それぞれの対数尤度から，全体の対数尤度が計算されている．もちろん，初期値（$\hat{\mu}=3.0000$，$\hat{\sigma}=1.0000$）からソルバーで，全体の対数尤度 $\ln L^{GP}$ を最大化した結果である．

確率　$$\hat{P}_i^{GP} = \frac{\Gamma(Y_i + 1/\hat{\sigma})}{\Gamma(Y_i+1)\Gamma(1/\hat{\sigma})} \cdot \frac{(\hat{\mu}_i\hat{\sigma})^{Y_i}}{(1+\hat{\mu}_i\hat{\sigma})^{Y_i+1/\hat{\sigma}}}$$

対数尤度　$$\ln L_i^{GP} = n_i \ln(\hat{P}_i^{GP})$$

$$\begin{aligned}\ln L^{GP} &= \sum_{i=1}^{16} \ln L_i^{GP} \\ &= -74.2038 - 27.2802 - \cdots - 5.5156 \\ &= -383.7046\end{aligned}$$

図 6.5 にデータ数 $N=173$ のそれぞれの分布の確率 Y を掛けて推定件数 $\hat{n}_i = N\hat{P}_i$ を算出し，元のツガイ数 n_i のヒストグラムの上に上書きした結果を示す．ゼロ過剰ポアソン分布のあてはまりが，ガンマ・ポアソン分布よりもサテライト数が 1 および 2 に対して推定値が追従している．

対数尤度の比較でも，ゼロ過剰ポアソン分布の $\ln L_{ZP}$ とガンマ・ポアソン分布の $\ln L_{GP}$ の比較では，

$$\begin{aligned}\ln L_{diff} &= \ln L^{ZP} - \ln L^{GP} \\ &= -381.6146 - (-383.7046) \\ &= 2.0900\end{aligned}$$

と，わずかではあるが，ゼロ過剰ポアソン分布のあてはめがほんの少し良い．

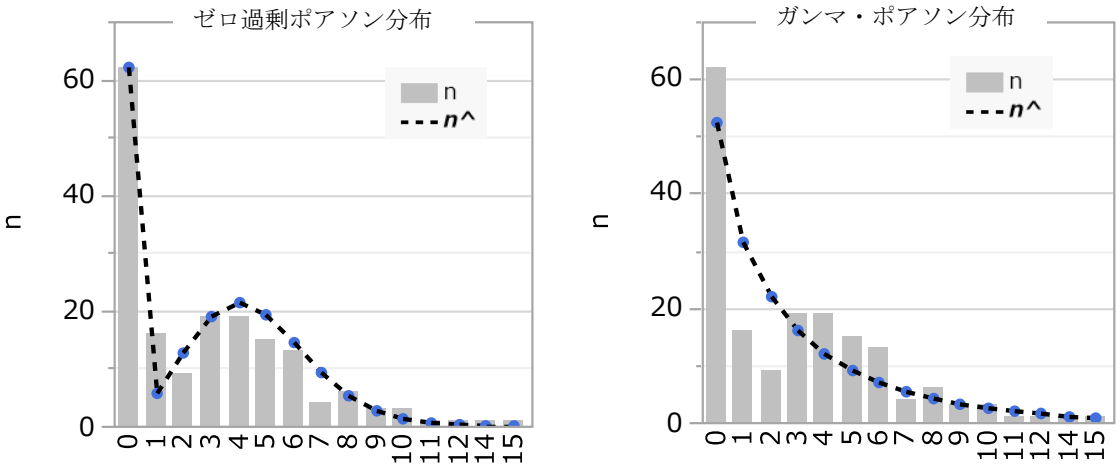


図 6.5　Zero-Ifuratied ポアソン分布およびガンマ・ポアソン分布のあてはめ

JMP（Ver 15.1）から，ゼロ過剰ポアソン分布のあてはめができるようになったので，表 6.9 に結果を示す．推定値は，平均 $\lambda : \hat{\mu} = 4.4990$，ゼロ強調 $\pi : \hat{\omega} = 0.3512$ と，表 6.8 の結果に一致することが確認される．

表 6.9　JMP によるゼロ過剰ポアソン分布のあてはめ

ゼロ強調 Poisson分布のあてはめ					
パラメータ		推定値	標準誤差	下側95%	上側95%
平均	λ	4.4990	0.2055	4.1074	4.9129
ゼロ強調	π	0.3512	0.0369	0.2813	0.4252
指標					
(-2)*対数尤度		763.2292			
AICc		767.2998			
BIC		773.5357			

雄の結合がない雌の数に対して，ポアソン分布のあてはめから除外する雌の割合は $\hat{\omega}=0.3512$ と推定されているので，

$$N\hat{\omega}=173\times 0.3512=60.75$$

と，62 匹中の大部分となる．このことからも，雄が雌に結合しやすい場合と，結合しにくい何らかの理由があると推論される．図 6.6 に示すように，サテライト数がゼロを除いた場合に，平均=4.5494，分散=8.0134 と分散が平均の 2 倍弱であり，過分散となっており，適合度の検定でもポアソン分布のあてはめは支持されない．

図 6.6　ゼロを除いた場合のポアソン分布のあてはめ

さらなる探索的な解析については，第 7.2 節で取り上げているが，高橋（2019a），「最尤法による探索的ポアソン回帰」にもまとめられている．

6.5. ゼロ過剰ガンマ・ポアソン分布のあてはめ

ゼロの反応が過大となっている場合に，ゼロとなる件数を割り引いた残りにガンマ・ポアソン分布をあてはめる．全体の数を N としたときに，反応が 0 の場合にガンマ・ポアソン分布のあてはめをしない割合を ω とする．残りの $(1-\omega)$ にガンマ・ポアソン分布をあてはめる．反応が 0 の場合の確率は，ω にガンマ・ポアソン分布の確率に $(1-\omega)$ を掛けたものを加える．

$$\left.\begin{aligned} f(Y;\omega,\mu,\sigma) &= \omega + (1-\omega)GammaPoisson(Y;\mu,\sigma), \quad Y=0 \\ f(Y;\omega,\mu,\sigma) &= (1-\omega)GammaPoisson(Y;\mu,\sigma), \quad Y=1,2,\ldots \end{aligned}\right\} \quad (6.13)$$

$$GammaPoisson(Y;\mu,\sigma) = \frac{\Gamma(Y+1/\sigma)}{\Gamma(Y+1)\Gamma(1/\sigma)}\frac{(\mu\sigma)^{Y}}{(1+\mu\sigma)^{Y+1/\sigma}}$$

このモデルは，ゼロ過剰ポアソン分布と同様に，Y が 0 となる部分集団，Y が 0 以上となる部分集団が混在しているような混合分布を想定することになる．表 6.10 にゼロ過剰ガンマ・ポアソン分布をあてはめた結果を示す．

表 6.10　最尤法によるゼロ過剰ガンマ・ポアソン分布のあてはめ

	初期値	**4.0000**	μ^=	**4.3287**		
		1.0000	σ^=	**0.2242**	$\ln L^{ZGP}$	
		0.4000	ω^=	**0.3256**	**-369.3516**	
	サテライト数	ツガイ数	ゼロ過剰ガンマ・ポアソン			
i	Y	n	確率 ω^	確率 P^	$\ln L_i^{ZGP}$	n^
1	0	62	0.3256	0.0327	-63.6218	62.0
2	1	16		0.0719	-42.1183	12.4
3	2	9		0.0967	-21.0263	16.7
4	3	19		0.1025	-43.2710	17.7
5	4	19		0.0942	-44.8846	16.3
6	5	15		0.0785	-38.1695	13.6
7	6	13		0.0610	-36.3678	10.5
8	7	4		0.0449	-12.4163	7.8
9	8	6		0.0317	-20.7172	5.5
10	9	3		0.0216	-11.5073	3.7
11	10	3		0.0143	-12.7406	2.5
12	11	1		0.0093	-4.6816	1.6
13	12	1		0.0059	-5.1365	1.0
14	13	0		0.0037	0.0000	0.6
15	14	1		0.0023	-6.0961	0.4
16	15	1		0.0014	-6.5967	0.2
	計	173	0.3256	0.6724		172.7
			0.9980			

ゼロ過剰ポアソン分布をあてはめた場合に比べ，$\hat{\omega}=0.3256$ とやや小さ目になり，$Y_1=0$ に対する推定値 $N\hat{\omega}$ は，

$$N\hat{\omega}=173\times 0.3256=56.34$$

となる．残りの $173\times(1-0.3256)=116.66$ ツガイ数に対するガンマ・ポアソン分布のあてはめは，平均が $\hat{\mu}=4.3287$，過分散が $\hat{\sigma}=0.2243$ となり，全体的にあてはまりがよくなっている．

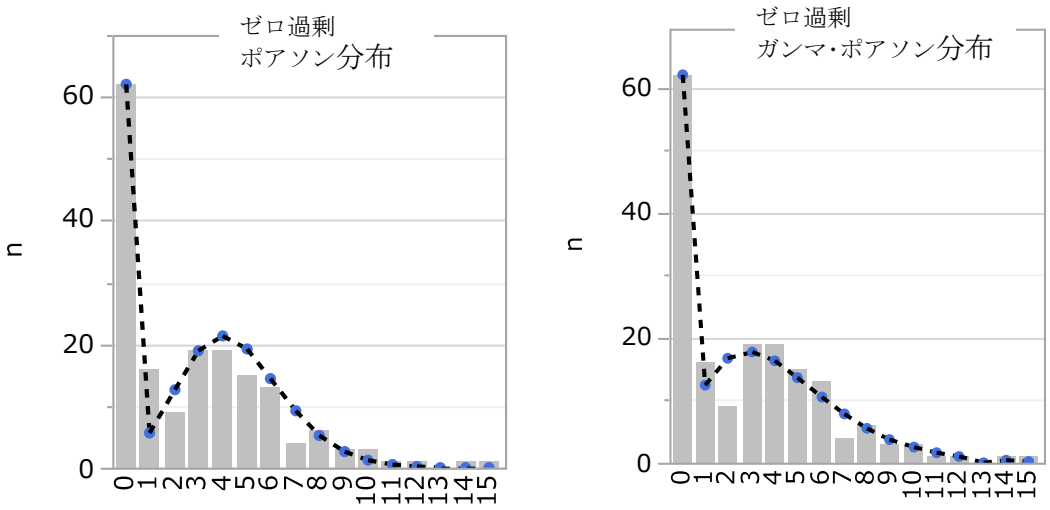


図 6.7　ゼロ過剰（ポアソン分布およびガンマ・ポアソン分布）のあてはめ

これまでに検討した結果を，対数尤度の大きさによってモデルのあてはめについて評価する．基本は，ポアソン分布のあてはめで，第 1.13 節の表 1.46 に示したように，JMP の「一変量の分布」によりサテライト数の平均は 2.9191，分散は 9.9120 であり，その比は 3.40 と過分散になっている．ポアソン分布をあてはめ，棒グラフ上に上書きした結果を見ても，誤差分布にポアソン分布を仮定することは絶望的である．もちろん，適合度の検定でもポアソン分布であるとは言えない．

過分散となる場合は，何らかの条件によりサテライト数の平均が大きく異なる部分集団の集まりが複数存在すると推測される．甲羅の色によってサテライト数の平均が大きく異なり，それに伴って過分散が解消されるのであろうか．あるいは，甲羅の幅，あるいは，体重を幾つかに区分したときにサテライト数の平均が大きく異なり，過分散が解消されるのだろうか．検討すべき事項である．ポアソン分布のマイナス 2 倍の対数尤度は，表 1.46 から，$-2\ln L^P=988.0893$ であり，ガンマ・ポアソン分布の場合は，表 6.8 から，$\ln L_{GP}=-383.7046$ である．ゼロ過剰ポアソン分布のあてはめは，$\ln L^{ZP}=-381.6146$ であり，ゼロ過剰ガンマ・ポアソン分布のあてはめは，表 6.10 から $\ln L^{ZGP}=-369.3516$ とある．これらを表 6.11 にまとめ直す．

表 6.11　4 種のモデルのマイナス 2 倍の対数尤度の比較

				ポアソン vs.	ゼロ過剰
ゼロ過剰	分布タイプ	$\ln L$	$(-2)\ln L$	ガンマ・ポアソン	なし・あり
なし	ポアソン分布（*P*）	-494.0447	**988.0893**	基準	基準
	ガンマ・ポアソン分布（*GP*）	-383.7046	767.4092	220.6801	
あり	ゼロ過剰 ポアソン分布（*ZP*）	-381.6146	**763.2292**	基準	224.8602
	ゼロ過剰 ガンマ・ポアソン分布（*ZGP*）	-369.3516	738.7032	24.5260	28.7061

マイナス 2 倍の対数尤度の差について，自由度 1 のカイ 2 乗分布の 5 パーセント点 3.84 をおおよその目安として用いる．ポアソン分布 vs. ガンマ・ポアソン分布のマイナス 2 倍の対数尤度の差は，220.6801 であり，圧倒的にガンマ・ポアソン分布のあてはめが支持される．

ゼロ過剰ポアソン分布 vs. ガンマ・ポアソン分布の差は，24.5260 あり，ガンマ・ポアソン分布のあてはめがかなり改善している．ポアソン分布でゼロ過剰にした場合には，マイナス 2 倍の対数尤度の差は，224.8602 とゼロ過剰にした場合が圧倒的に支持される．ガンマ・ポアソン分布でゼロ過剰にした場合には，マイナス 2 倍の対数尤度の差は，28.7061 とゼロ過剰にした場合にあてはめがかなり改善される．

この結果から，ゼロ過剰ガンマ・ポアソン分布が，最もあてはまりがいいとの結論である．ただし，この分布を使った回帰分析を実施することは，可能ではあるが，ゼロの割合が，説明変数の値により変化するような場合に，第 6.6 節および第 6.7 節で取り上げるが，ゼロがないにも関わらず確率的には存在するというような矛盾が内在し，結果の解釈に困難さを感ずる．

JMP（Ver 15.1）から，ゼロ強調負の 2 項分布（ゼロ過剰ガンマ・ポアソン分布）のあてはめができるようになったので，表 6.12 に結果を示す．推定値は，（平均 $\lambda:\hat{\mu}=4.3287$，ばらつき $\hat{\sigma}=0.2242$，ゼロ強調 $\pi:\hat{\omega}=0.3257$）と，表 6.10 に一致することが確認される．注）$\pi:\hat{\omega}$ は少数点以下 4 桁で 0.0001 の食い違いがあるが，正確には，$\pi=0.325650$，$\hat{\omega}=0.325649$ あり，少数点以下 6 桁での 0.000001 の違いである．

表 6.12　JMP によるゼロ強調・負の 2 項分布のあてはめ

ゼロ強調　負の二項分布のあてはめ

パラメータ		**推定値**	**標準誤差**	**下側95%**	**上側95%**
平均	λ	4.3287	0.2958	3.7497	4.9289
ばらつき	σ	0.2242	0.0783	0.1039	0.4317
ゼロ強調	π	0.3257	0.0402	0.2464	0.4050

指標	
(-2)*対数尤度	738.7032
AICc	744.8452
BIC	754.1630

6.6. ガンマ・ポアソン回帰

第 1.13 節では，雌のカブトカニに連結する雄のサテライト数のデータにポアソン分布をあてはめ，過分散データの場合にどのような不具合が起きるかを示した．サテライト数を反応変数，甲羅の幅を説明変数とする対数リンクによるポアソン回帰を行い，過分散があるデータの場合には，個別データの 95%信頼区間が極端に過小評価になることを示した．

ポアソン回帰 vs. ガンマ・ポアソン回帰

第 1.13 節の表 1.50 で，ガンマ・ポアソン分布のあてはめを行なった結果，見た目にはあてはまりが悪いと思われた．ただし，適合度検定ではあてはめは否定されなかった．この様に過分散データに対して，ガンマ・ポアソン分布を誤差分布とする回帰分析を行った場合，どのような結果になるのであろうか．あるいは，どのような注意点があるのだろうか．そこで，Excel による対数リンクによるガンマ・ポアソン回帰を行い，ポアソン回帰の場合と比較検討することにした．表 6.13 に結果を示す．

表 6.13　甲羅の幅を説明変数とするガンマ・ポアソン回帰

			β^$_0$=	**-3.3048**		β^$_0$=	**-4.0523**	
			β^$_1$=	**0.1640**	$\ln L^P$	β^$_1$=	**0.1921**	$\ln L^{GP}$
					-461.59	過分散 σ^=	**1.1055**	**-375.65**
		サテ	—— ポアソン回帰 ——			— ガンマ・ポアソン回帰 —		
	幅	ライト	対数	確率	対数尤度	対数	確率	対数尤度
No	x	Y	リンク y^	P^	$\ln L_i^P$	リンク y^	P^	$\ln L_i^{GP}$
1	28.3	8	3.8103	0.0244	-3.7132	3.9874	0.0324	-3.4291
2	22.5	0	1.4715	0.2296	-1.4715	1.3089	0.4451	-0.8094
3	26.0	9	2.6128	0.0011	-6.7709	2.5636	0.0148	-4.2130
4	24.8	0	2.1459	0.1170	-2.1459	2.0358	0.3443	-1.0663
5	26.0	4	2.6128	0.1424	-1.9492	2.5636	0.0721	-2.6302
:								
171	28.0	0	3.6274	0.0266	-3.6274	3.7642	0.2266	-1.4846
172	27.0	0	3.0786	0.0460	-3.0786	3.1064	0.2600	-1.3472
173	24.5	0	2.0429	0.1297	-2.0429	1.9218	0.3568	-1.0306

ポアソン回帰の場合は，

対数リンク　$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$

確率　$\hat{P}_i^P = \text{Poisson.dist}(Y_i, \hat{y}_i, \mathit{false})$

対数尤度　$\ln L_i^P = \ln(\hat{P}_i^P)$

$\ln L$　$\ln L^P = \sum_{i=1}^{173} \ln L_i^P$

とし，$\hat{\beta}_0$ と $\hat{\beta}_1$ に $\ln L_i$ が計算されるような初期値を設定し，ソルバーで，$\ln L_i$ の和である $\ln L^P$ が最大になるように $\hat{\beta}_0$ と $\hat{\beta}_1$ を変化させた結果である．推定された $\hat{\beta}_0 = -3.3048$ および $\hat{\beta}_1 = 0.1640$ が最尤解として求まっている．これは，表 1.47 に示した JMP による結果と一致している．

次に，ガンマ・ポアソン回帰の場合は，式（6.6）のガンマ・ポアソン分布関数を用いて，

対数リンク	$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$
確率	$\hat{P}_i^{GP} = \dfrac{\Gamma(Y_i + 1/\hat{\sigma})}{\Gamma(Y_i+1)\Gamma(1/\hat{\sigma})} \cdot \dfrac{(\hat{y}_i\hat{\sigma})^{Y_i}}{(1+\hat{y}_i\hat{\sigma})^{Y_i+1/\hat{\sigma}}}$
対数尤度	$\ln L_i^{GP} = \ln(\hat{P}_i^{GP})$
$\ln L$	$\ln L^{GP} = \sum_{i=1}^{173} \ln L_i^{GP}$

とし，$\hat{\beta}_0$ と $\hat{\beta}_1$ に $\ln L_i$ が計算されるような初期値を入れ，ソルバーで $\ln L^{GP}$ が最大になるように $\hat{\beta}_0$，$\hat{\beta}_1$ および $\hat{\sigma}$ を変化させた結果である．推定された $\hat{\beta}_0 = -4.0523$，$\hat{\beta}_1 = 0.1921$ および $\hat{\sigma} = 1.1055$ が最尤解として求まっている．

ポアソン回帰の対数尤度 $\ln L^P = -461.58$ に比べ，ガンマ・ポアソン回帰の対数尤度は，$\ln L_{GP} = -375.65$ と大きくなり，対数尤度の差は，

$$\begin{aligned} \ln L_{diff} &= \ln L^{GP} - \ln L^P \\ &= -375.65 - (-461.58) = 85.94 \end{aligned}$$

と大きく，ガンマ・ポアソン回帰のあてはめが支持される．また，ガンマ・ポアソン回帰場合に，切片 $\hat{\beta}_0$ は小さくなり，傾きは大きくなっている．図 6.8 の甲羅の幅とサテライト数の散布図上に，対数リンクの回帰直線（指数曲線）を上書きした結果を示す．

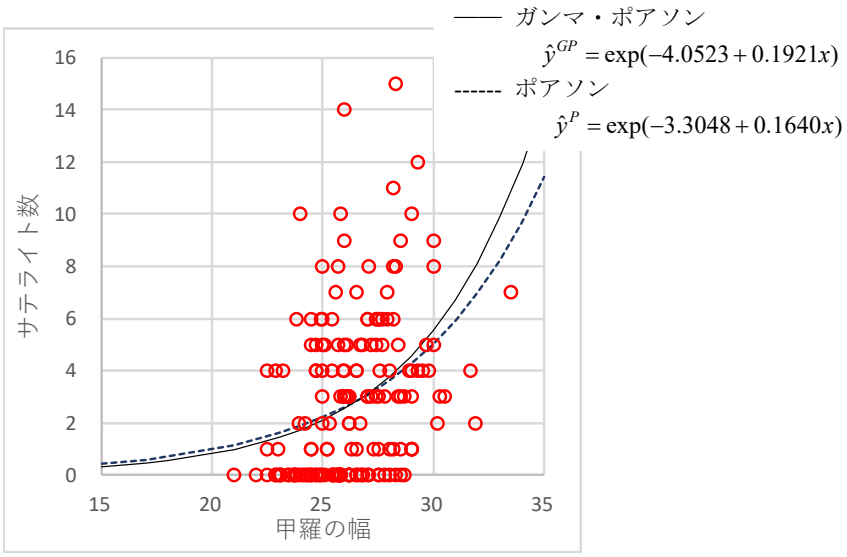


図 6.8　対数リンクでのガンマ・ポアソン回帰

甲羅の幅に対するポアソン回帰およびガンマ・ポアソン回帰の指数曲線の推定値を表 6.14 に示す．ポアソン回帰の切片は，

$$\hat{y}^{P}_{(x=0)} = \exp(-3.3047 + 0.1640 \times 0) = 0.0367$$

であり，ガンマ・ポアソン回帰の切片は，

$$\hat{y}^{GP}_{(x=0)} = \exp(-4.0524 + 0.1921 \times 0) = 0.0174$$

と低めに推定されている．

表 6.14　ポアソン回帰およびガンマ・ポアソン回帰の推定値

	ポアソン	ガンマ･ポアソン	
x	y^{P}^	y^{GP}^	$1+y^{GP}$^σ^
0	0.0367	0.0174	1.0192
10	0.1893	0.1186	1.1311
20	0.9764	0.8098	1.8952
25	2.2175	2.1156	3.3388
30	5.0359	5.5271	7.1102
35	11.4366	14.4401	16.9636

甲羅の幅 $x = 25$ の場合には，ポアソン回帰の場合は，

$$\hat{y}^{P}_{(x=25)} = \exp(-3.3047 + 0.1640 \times 25) = 2.2175$$

であり，ガンマ・ポアソン回帰の場合は，

$$\hat{y}^{GP}_{(x=25)} = \exp(-4.0524 + 0.1921 \times 25) = 2.1156$$

と低めであるが，$x = 30$ の場合には，ガンマ・ポアソン回帰の方が 14.4401 と高めとなっている．

ポアソン回帰の分散は，回帰の推定値の伸びと同値で大きくなる．ガンマ・ポアソン回帰の場合に過分散 σ は，一定で変化しないので，式（6.10）の分散の式から最初の $\hat{y}^{GP}$ を除いた $\hat{\sigma}'$（JMP での過分散の式）

$$\hat{\sigma}' = 1 + \hat{y}^{GP}\hat{\sigma}$$

を用いて計算した結果を示した．

甲羅の幅 x に対するガンマ・ポアソン分布のあてはめ

ガンマ・ポアソン回帰は，甲羅の幅 x の変化に対し，どのような分布をあてはめているのだろうか．表 6.15 に $x_i = (20, 25, 30)$ の場合の推定値 $\hat{y}^{P}_{x_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ に対するポアソン分布の確率

$$\hat{P}^{P}_{x_i,j} = \text{Poisson.dist}(Y_{x_i,j}, \hat{y}^{P}_{x_i}, false)$$

を計算した結果を示す．同様にガンマ・ポアソン回帰の推定値 $\hat{y}^{GP}_{x_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ に対するガンマ・ポアソン分布の確率

$$\hat{P}_{x,j}^{GP} = \frac{\Gamma(Y_{x,j}+1/\hat{\sigma})}{\Gamma(Y_{x,j}+1)\Gamma(1/\hat{\sigma})} \cdot \frac{(\hat{y}_x^{GP}\hat{\sigma})^{Y_{x,j}}}{(1+\hat{y}_x^{GP}\hat{\sigma})^{Y_{x,j}+1/\hat{\sigma}}}$$

を計算した結果も示す. 甲羅の幅 $x=25$ の場合では, ポアソン分布の $Y=0$ 場合は確率が 0.1089 であるのに対し，ガンマ・ポアソン分布の確率は 0.3360 とかなり上回っている.

表 6.15　ポアソン回帰およびガンマ・ポアソン回帰の誤差の確率分布

		$\beta\hat{}_0$=	**-3.3047**		$\beta\hat{}_0$=	**-4.0524**	
		$\beta\hat{}_1$=	**0.1640**		$\beta\hat{}_1$=	**0.1921**	
					過分散 $\sigma\hat{}$=	**1.1055**	
		——— ポアソン回帰 ———			—— ガンマ・ポアソン回帰 ——		
	x_i=	20	25	30	20	25	30
	$y\hat{}_{xi}$=	0.9764	2.2175	5.0359	0.8098	2.1156	5.5271
j	Y	ポアソン　確率			ガンマ・ポアソン　確率		
1	0	0.3767	0.1089	0.0065	0.5608	0.3360	0.1696
2	1	0.3678	0.2414	0.0327	0.2396	0.2129	0.1318
3	2	0.1796	0.2677	0.0824	0.1078	0.1420	0.1079
4	3	0.0584	0.1979	0.1384	0.0493	0.0963	0.0898
5	4	0.0143	0.1097	0.1742	0.0227	0.0659	0.0753
6	5	0.0028	0.0486	0.1754	0.0105	0.0453	0.0635
7	6	0.0005	0.0180	0.1473	0.0049	0.0312	0.0537
8	7	0.0001	0.0057	0.1059	0.0023	0.0216	0.0455
9	8	0.0000	0.0016	0.0667	0.0011	0.0149	0.0386
10	9	0.0000	0.0004	0.0373	0.0005	0.0103	0.0328
11	10	0.0000	0.0001	0.0188	0.0002	0.0072	0.0280
12	11	0.0000	0.0000	0.0086	0.0001	0.0050	0.0238
13	12	0.0000	0.0000	0.0036	0.0001	0.0035	0.0203
14	13	0.0000	0.0000	0.0014	0.0000	0.0024	0.0173
15	14	0.0000	0.0000	0.0005	0.0000	0.0017	0.0148
16	15	0.0000	0.0000	0.0002	0.0000	0.0012	0.0126
17	16	0.0000	0.0000	0.0001	0.0000	0.0008	0.0108

表 6.15 に示したポアソン回帰およびガンマ・ポアソン回帰の甲羅の幅 $x=(20, 25, 30)$ に対するサテライト数の分布の確率を図 6.9 に示す．甲羅の幅が 20 cm の場合は，片流れ的であ

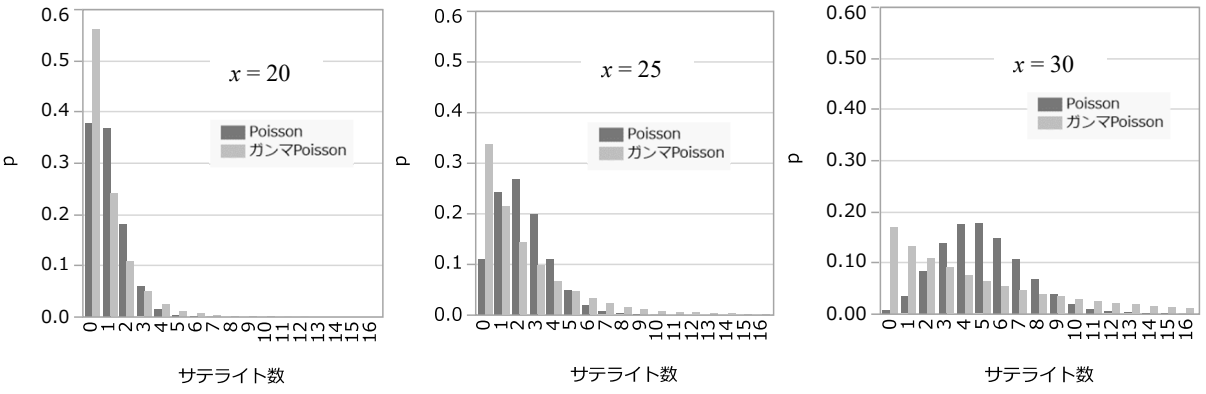


図 6.9　対数リンクでの確率分布

るのが，甲羅の幅が大きくなるにつれて，ポアソン回帰の場合は，一山型であるが，ガンマ・ポアソン回帰の場合は，片流れが続いている．

対数リンクによるガンマ・ポアソン回帰の指数曲線は，図 6.8 に示したのであるが，$x=(20, 25, 30)$ に対して表 6.15 で計算したガンマ・ポアソン分布の確率を重ね書きした結果を図 6.10 に示す．

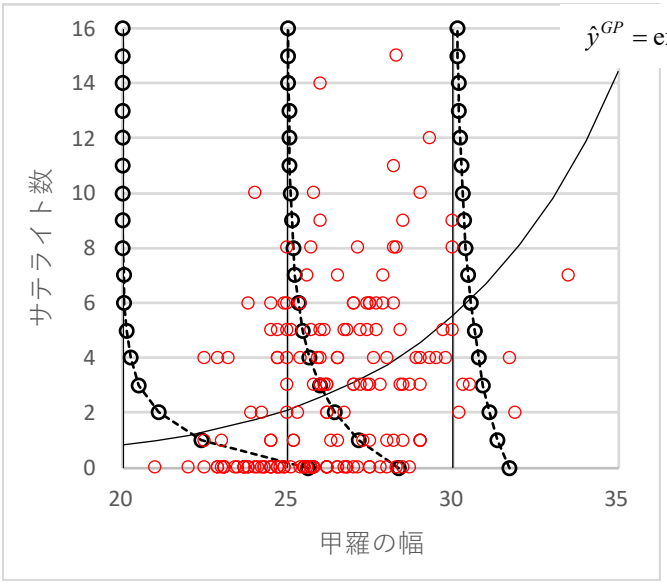


図 6.10　対数リンクでのガンマ・ポアソン回帰における確率分布

甲羅の幅 $x=25$ の場合にサテライト数 Y に対する出現確率は，○印の分布を適切に反映していると判断される．ただし，甲羅の幅 $x=30$ では，サテライト数 $Y=0$ が観察されないにも関わらず出現確率が高止まりしている．このことは，甲羅の幅に関わらずサテライト数の分布が，一律にガンマ・ポアソン分布に従っていると仮定することが妥当ではないとの判断になる．

6.7. ゼロ過剰ガンマ・ポアソン回帰

第 6.6 節で示したようにガンマ・ポアソン回帰に引き続き，ゼロ過剰ガンマ・ポアソン回帰の振る舞いについて示す．第 6.3 節で示したように稀な現象ではないカウント・データでは，ゼロ過剰ガンマ・ポアソンのあてはめが支持されていた．雌のカブトカニに連結する雄のサテライト数のデータに対し，甲羅の幅を説明変数とする対数リンクによる，ゼロ過剰ガンマ・ポアソン回帰を適用し，その振る舞いについて検討する．

ガンマ・ポアソン回帰 vs. ゼロ過剰ガンマ・ポアソン回帰

表 6.16 にゼロ過剰ガンマ・ポアソン回帰の結果を示す．対数リンクは，

対数リンク　　$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$

で計算し，サテライト数 y_i がゼロの場合は，

ゼロ対応　　$Y_i = 0$ ：　$\hat{P}_{0,i}^{ZGP} = \hat{\omega}$

$Y_i \neq 0$ ：　$\hat{P}_{0,i}^{ZGP} = 0$

とし，全てのサテライト数 y_i について

確率　　$$\hat{P}_i^{ZGP} = (1-\hat{\omega}) \cdot \frac{\Gamma(Y_i + 1/\hat{\sigma})}{\Gamma(Y_i+1)\Gamma(1/\hat{\sigma})} \cdot \frac{(\hat{y}_i\hat{\sigma})^{Y_i}}{(1+\hat{y}_i\hat{\sigma})^{Y_i+1/\hat{\sigma}}}$$

対数尤度　　$$\ln L_i^{ZGP} = \ln(\hat{P}_{0,i}^{ZGP} + \hat{P}_i^{ZGP})$$

$\ln L$　　$$\ln L^{ZGP} = \sum_{i=1}^{173} \ln L_i^{ZGP}$$

表 6.16　甲羅の幅を説明変数とするゼロ過剰ガンマ・ポアソン回帰

							ω=	**0.3099**	
			β^$_0$=	**-4.0523**			β^$_0$=	**-0.3638**	
			β^$_1$=	**0.1921**	$\ln L^{GP}$		β^$_1$=	**0.0671**	$\ln L^{ZGP}$
			過分散 σ^=	**1.1055**	**-375.65**		過分散 σ^=	**0.2442**	**-367.49**
		サテ	ガンマ・ポアソン回帰			ゼロ過剰ガンマ・ポアソン回帰			
	幅	ライト	対数	確率	対数尤度	対数	ゼロ	確率	対数尤度
No	x	Y	リンク y^	P	$\ln L_i^{GP}$	リンク y^	P_0	P	$\ln L_i^{ZGP}$
1	28.3	8	3.9874	0.0324	-3.4291	4.6389	0.0000	0.0363	-3.3168
2	22.5	0	1.3089	0.4451	-0.8094	3.1438	0.3099	0.0669	-0.9758
3	26.0	9	2.5636	0.0148	-4.2130	3.9757	0.0000	0.0182	-4.0083
4	24.8	0	2.0358	0.3443	-1.0663	3.6682	0.3099	0.0503	-1.0210
5	26.0	4	2.5636	0.0721	-2.6302	3.9757	0.0000	0.0949	-2.3547
：									
171	28.0	0	3.7642	0.2266	-1.4846	4.5465	0.3099	0.0324	-1.0719
172	27.0	0	3.1064	0.2600	-1.3472	4.2515	0.3099	0.0374	-1.0575
173	24.5	0	1.9219	0.3568	-1.0306	3.5952	0.3099	0.0523	-1.0155

とし，（$\hat{\omega}$，$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\sigma}$）に $\ln L_i^{ZGP}$ が計算されるような初期値を入れ，ソルバーで $\ln L^{ZGP}$ が最大になるように（$\hat{\omega}$，$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\sigma}$）を変化させた結果である．推定された $\hat{\omega}=0.3099$ $\hat{\beta}_0=-0.3638$，$\hat{\beta}_1=0.0671$ および $\hat{\sigma}=0.2442$ が最尤解として求まっている．

ガンマ・ポアソン回帰の対数尤度 $\ln L^{GP}=-375.65$ に比べ，ゼロ過剰ガンマ・ポアソン回帰の対数尤度は，$\ln L^{ZGP}=-367.49$ と大きくなり，対数尤度の差は，

$$\begin{aligned}\ln L_{diff} &= \ln L^{ZGP} - \ln L^{GP} \\ &= -367.49-(-375.65)=8.15\end{aligned}$$

と大きく，ゼロ過剰ガンマ・ポアソン回帰のあてはめが支持される．また，ガンマ・ポアソン回帰の場合に，切片 $\hat{\beta}_0$ は大きくなり，傾き $\hat{\beta}_1$ は小さく，過分散 $\hat{\sigma}=0.2442$ とかなり小さくなっている．図 6.11 に甲羅の幅とサテライト数の散布図上に，対数リンクの回帰直線（指数曲線）を上書きした結果を示す．

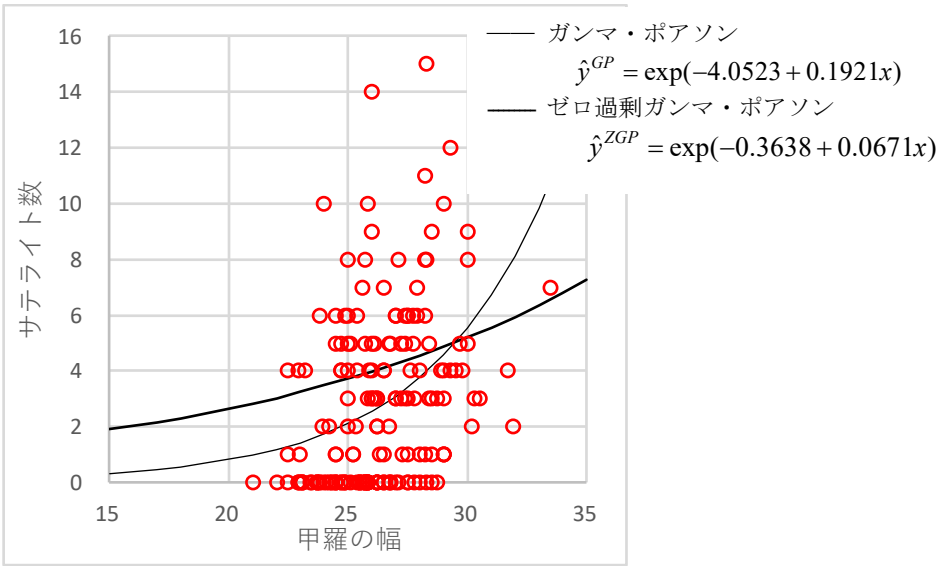


図 6.11　対数リンクでのゼロ過剰ガンマ・ポアソン回帰

甲羅の幅 x に対するガンマ・ポアソン回帰およびゼロ過剰ガンマ・ポアソン回帰の指数曲線を次式で推定する．ガンマ・ポアソン回帰の切片は，

$$\hat{y}_{(x=0)}^{GP}=\exp(-4.0523+0.1921\times 0)=0.0174$$

であり，セロ過剰ガンマ・ポアソン回帰の切片は，

$$\hat{y}_{(x=0)}^{ZGP}=\exp(-0.3638+0.0671\times 0)=0.6950$$

と高めに推定されている．

甲羅の幅 $x=25$ の場合には，ガンマ・ポアソン回帰の場合は，

$$\hat{y}^{GP}_{(x=25)} = \exp(-4.0523 + 0.1921 \times 25) = 2.1156$$

であり，ゼロ過剰ガンマ・ポアソン回帰の場合は，

$$\hat{y}^{ZGP}_{(x=25)} = \exp(-0.3638 + 0.0671 \times 25) = 3.7178$$

と低めであるが，$x=30$ の場合には，ゼロ過剰ガンマ・ポアソン回帰の方が低めと逆転している．

甲羅の幅 x に対するゼロ過剰ガンマ・ポアソン分布のあてはめ

ゼロ過剰ガンマ・ポアソン回帰は，甲羅の幅 x の変化に対し，どのような分布をあてはめているのか図 6.12 に $x_i = (20, 25, 30)$ に対応するゼロ過剰ガンマ・ポアソン回帰の推定値 $\hat{y}_{x_i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ に対するゼロ過剰ガンマ・ポアソン分布の確率 $\hat{P}$ を Y_i，$i = 0, 1, ..., 16$ について

$$y_0 = 0 \ : \quad \hat{P}^{ZGP}_{x_i,j} = \hat{\omega} + (1-\hat{\omega}) \cdot \frac{\Gamma(Y_{x_i,j} + 1/\hat{\sigma})}{\Gamma(Y_{x_i,j}+1)\Gamma(1/\hat{\sigma})} \cdot \frac{(\hat{y}^{ZGP}_{x_i}\hat{\sigma})^{Y_{x_i,j}}}{(1+\hat{y}^{ZGP}_{x_i}\hat{\sigma})^{Y_{x_i,j}+1/\hat{\sigma}}}$$

$$y_0 \neq 0 \ : \quad \hat{P}^{ZGP}_{x_i,j} = (1-\hat{\omega}) \cdot \frac{\Gamma(Y_{x_i,j} + 1/\hat{\sigma})}{\Gamma(Y_{x_i,j}+1)\Gamma(1/\hat{\sigma})} \cdot \frac{(\hat{y}^{ZGP}_{x_i}\hat{\sigma})^{Y_{x_i,j}}}{(1+\hat{y}^{ZGP}_{x_i}\hat{\sigma})^{Y_{x_i,j}+1/\hat{\sigma}}}$$

計算し（表 6.17），図 6.12 にプロットとした結果を示す．

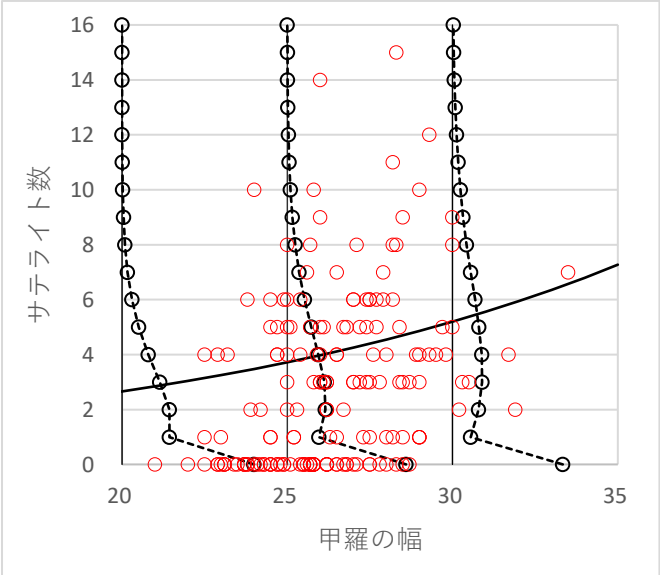


図 6.12　対数リンクでのゼロ過剰ガンマ・ポアソン回帰における確率分布

甲羅の幅 $x=25$ の場合にサテライト数 y に対する出現確率は，○印の分布を適切に反映していると判断される．ただし，甲羅の幅 $x=30$ では，サテライト数 $Y=0$ がまったく観察されないにも関わらず出現確率が高止まりしている．このことは，甲羅の幅に関わらずサテライト数の分布が，一律にゼロ過剰ガンマ・ポアソン分布に従っていると仮定することが妥当ではないとの判断になる．

表 6.17　ガンマ・ポアソン回帰およびゼロ過剰ガンマ・ポアソン回帰の誤差の確率分布

					ω=	**0.3099**	
		$\beta\hat{}_0$=	**-4.0524**		$\beta\hat{}_0$=	**-0.3638**	
		$\beta\hat{}_1$=	**0.1921**		$\beta\hat{}_1$=	**0.0671**	
		過分散 $\sigma\hat{}$=	**1.1055**		過分散 $\sigma\hat{}$=	**0.2442**	
		ガンマ・ポアソン回帰			**ゼロ過剰 ガンマ・ポアソン回帰**		
	x_i=	20	25	30	20	25	30
	$y\hat{}_{xi}$=	0.8098	2.1156	5.5271	2.6585	3.7178	5.1992
j	Y	ガンマ・ポアソン 確率			ゼロ過剰 ガンマ・ポアソン確率		
1	0	0.5608	0.3360	0.1696	0.3989	0.3589	0.3340
2	1	0.2396	0.2129	0.1318	0.1434	0.0954	0.0551
3	2	0.1078	0.1420	0.1079	0.1438	0.1157	0.0785
4	3	0.0493	0.0963	0.0898	0.1150	0.1119	0.0893
5	4	0.0227	0.0659	0.0753	0.0803	0.0944	0.0886
6	5	0.0105	0.0453	0.0635	0.0512	0.0727	0.0802
7	6	0.0049	0.0312	0.0537	0.0305	0.0525	0.0680
8	7	0.0023	0.0216	0.0455	0.0173	0.0360	0.0549
9	8	0.0011	0.0149	0.0386	0.0095	0.0238	0.0426
10	9	0.0005	0.0103	0.0328	0.0050	0.0152	0.0320
11	10	0.0002	0.0072	0.0280	0.0026	0.0095	0.0234
12	11	0.0001	0.0050	0.0238	0.0013	0.0058	0.0168
13	12	0.0001	0.0035	0.0203	0.0006	0.0035	0.0118
14	13	0.0000	0.0024	0.0173	0.0003	0.0020	0.0082
15	14	0.0000	0.0017	0.0148	0.0002	0.0012	0.0056
16	15	0.0000	0.0012	0.0126	0.0001	0.0007	0.0038
17	16	0.0000	0.0008	0.0108	0.0000	0.0004	0.0025

7. 過分散がある場合の探索的ポアソン回帰

第 7 章の最初の事例は，第 1.7 節の細菌を用いた 2×2 の実験結果を用い，要因配置型のカウント・データに対する探索的ポアソン回帰のアプローチの基本が示されている．第 2 の事例は，第 1.13 節のカブトガニの観察データについての事例で，2 変量ポアソン回帰，2 元配置型ポアソン回帰，さらに共分散分析型ポアソン回帰を扱っている．第 3 の事例は，第 1 章で取り上げなかった事例で，2 群比較において，ポアソン分布が仮定できない社会調査に関するカウント・データに対し，ゼロ過剰（Zero-Inflated）ポアソン分布などを扱っている．

7.1. ネズミチフス菌のコロニー数の事例

稀に起きる現象の観察から得られるカウント・データには，ポアソン分布が良くあてはまるが，稀とは言い難い事象のカウント・データの場合には，過分散となりがちでポアソン分布のあてはまりが悪くなることを第 6 章で示した．稀に起きる現象を対象にした実験的研究において，ポアソン分布があてはまることを期待していたが，過分散となってしまった場合を想定しよう．このような場合には，反応が大きめになるような複数の要因が内在していることが疑われる．

異なる実験条件データの併合

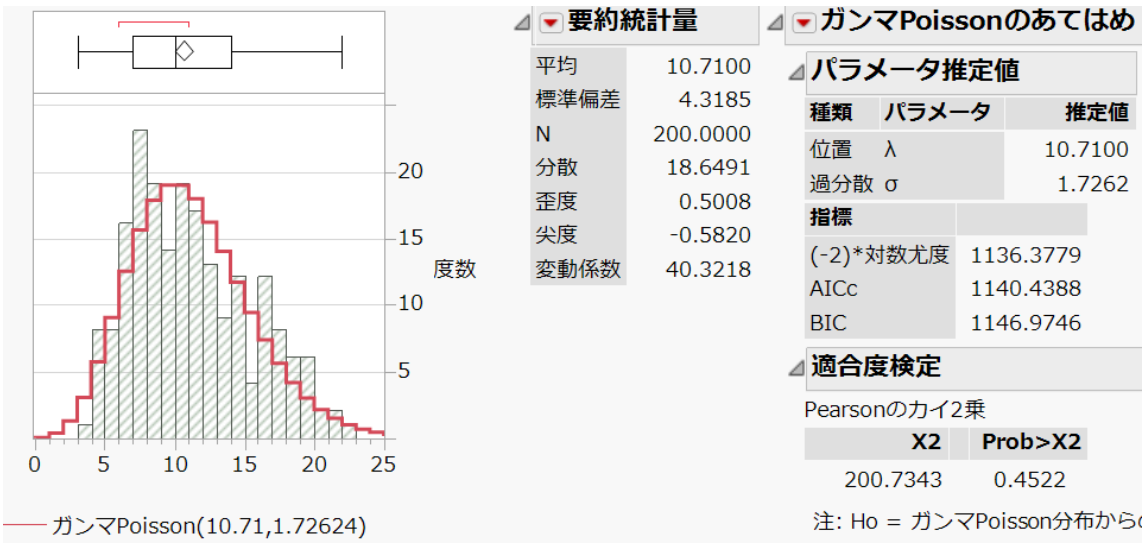
第 1.7 節で，ネズミチフス菌を用いた実験データの事例を示し，それぞれの条件下でポアソン分布があてはめることを示した［吉村ら（1992)］．実験は，2 つの要因に対する陰性対照群の特質を吟味するための実験的研究データであった．探索的なデータ解析を行う事例として表 7.1 に示すように実験は，あらかじめ設定した実験条件が結果におよぼす影響がないことを前提とし，200 枚のシャーレ全体の結果を得たとする．

表 7.1　ネズミチフス菌株に関するコロニー数（表 1.22：再掲）

		コロニー数																				
溶媒	活性化	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	計
A_1	B_1	0	1	0	0	2	1	2	3	5	3	2	3	3	6	6	4	5	1	2	1	50
	B_2	1	5	4	10	7	6	5	6	3	1	1	1	0	0	0	0	0	0	0	0	50
A_2	B_1	0	0	1	1	2	2	3	4	7	7	5	5	1	6	2	2	1	1	0	0	50
	B_2	0	2	3	5	12	10	4	6	2	2	1	3	0	0	0	0	0	0	0	0	50
	計	**1**	**8**	**8**	**16**	**23**	**19**	**14**	**19**	**17**	**13**	**9**	**12**	**4**	**12**	**8**	**6**	**6**	**2**	**2**	**1**	**200**

全 200 枚のシャーレ上のコロニー数についてのヒストグラム，要約統計量，ガンマ・ポアソン分布のあてはめ結果を表 7.2 に示す．平均=10.7100，分散=18.6491 とその比は 1.74 と過分散である．表には示していないが，ポアソン分布をあてはめた場合の適合度のカイ 2 乗値は，346.5154 であり，自由度 N-1=199 のカイ 2 乗に従うとして結果は $p < 0.0001$ でありポアソン分布のあてはめは棄却される．過分散を考慮したガンマ・ポアソン分布のあてはめに対する適合度のカイ 2 乗値は，200.7343， $p = 0.4522$ であり，あてはめは棄却されない．

表 7.2　全 200 シャーレ上のコロニー数に対するガンマ・ポアソン分布のあてはめ



説明変数ごとの層別

探索的な解析の第 1 歩は，得られている説明変数と反応変数間の関係の把握である．図 7.1 に示した A:溶媒および B:活性化で層別したヒストグラムを見ると，A:溶媒では，A_1 も A_2 も

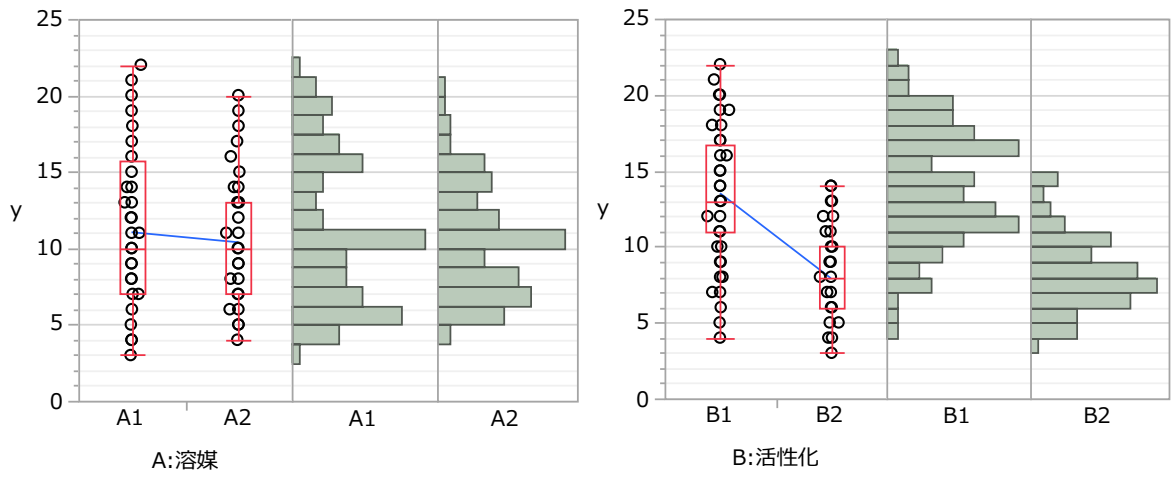


図 7.1　説明変数ごとの層別ヒストグラム

全体の分布と同様であり過分散は解消していない．B:活性化では，B_1 と B_2 の 2 つの平均値が異なる分布となり，過分散が解消することが期待される．

実際に分散/平均の比を計算した結果を表 7.3 に示す．B:活性化は（B_1：1.14，B_2：0.79）と全体の 1.74 に比べて縮小している．A:溶媒では（A_1：2.18，A_2：1.27）と過分散は解消していない．過分散の主な原因は，B:活性化であり，これを考慮すればピュアなポアソン分布と見なすことができそうである，ただし，図 7.1 から，実験条件 B_1 に 2 つの山があり，更なる探索が必要とも思われれる．

表 7.3　説明変数ごとの分散/平均の比

要因	水準	N	平均	分散	比
A:溶媒	A_1	100	11.0400	24.0388	2.18
	A_2	100	10.3800	13.2279	1.27
B:活性化	B_1	100	13.5100	15.3635	1.14
	B_2	100	7.9100	6.2847	0.79
	全体	200	10.7100	18.6491	1.74

説明変数の組み合わせによる層別

A:溶媒および B:活性化の各 2 水準を組み合わせて図 7.2 に示すように層別ヒストグラムを作成する．A:溶媒の各水準を B:活性化の水準で分割すると過分散が解消さる．

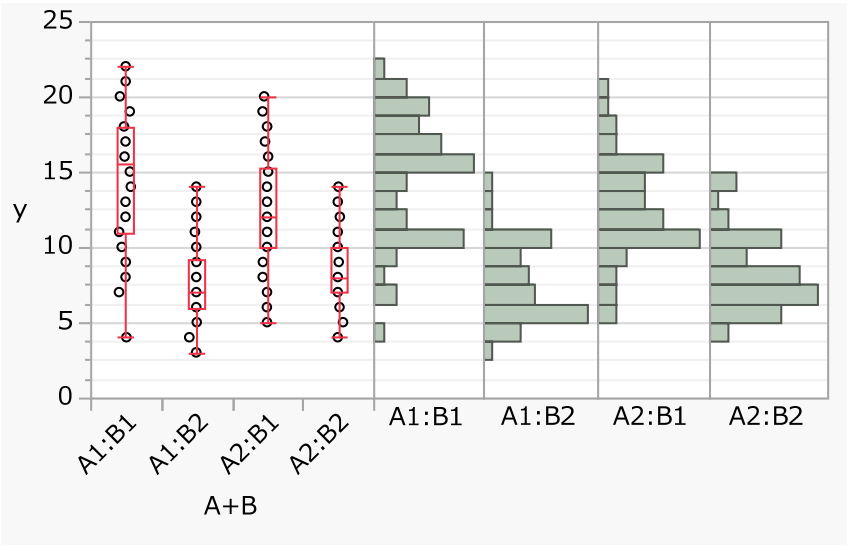


図 7.2　説明変数の組み合わせによる層別ヒストグラム

表 7.4 に示すように分散と平均の比が 1 を超えているのは，A_1:B_1 であり．平均=14.5400 に対して，分散=17.2331 やや過分散傾向であり，偶然とは思えない 2 つの山があり，まだ同定

できない原因が隠されているかもしれない．他の組み合わせについては，分散が平均より小さめなので過分散が解消されており，A:溶媒と B:活性化の 2 つの要因に対しポアソン分布を仮定した 2×2 の要因配置型の解析が可能となる．これについては，第 3.4～3.5 節に各種のデザイン行列を用いた解析方法と結果の見方についてすでに例示した．

表 7.4　説明変数の組み合わせによる分散/平均の比

A:溶媒	B:活性化	N	平均	分散	比
A_1	B_1	50	14.5400	17.2331	1.19
	B_2	50	7.5400	6.3351	0.84
A_2	B_1	50	12.4800	11.6424	0.93
	B_2	50	8.2800	6.0833	0.73
	全体	200	10.7100	18.6491	1.74

適合度のカイ 2 乗検定

カウント・データなので，ポアソン分布があてはまることを期待したとしても，本節に示すように異なる発生原因が内在する場合には，過分散となりやすいことを例示した．ポアソン分布とガンマ・ポアソン分布に対する適合度のカイ 2 乗検定について Excel で計算した結果を表 7.5 に示す．コロニー数 y_i，$i=1,2,...,20$ に対し，その頻度 n_i が示されている．平均は 10.7100 なので，y_i に対するポアソン分布の確率は，

$$P_i=\frac{\hat{\mu}^{y_i}e^{-\hat{\mu}}}{y_i!}=\text{Poisson.dist}(y_i,\hat{\mu},\mathit{false})$$

によって計算されている．

ポアソン分布に対する適合度のカイ 2 乗検定は，第 1.7 節の表 1.22 で引用した吉村ら（1992）で使われていたカウント・データ y_i をそのまま使う方法を示した．ポアソン分布のあてはめの場合は，$Var(y_i)=\hat{\mu}$ なので平方和を $\hat{\mu}$ で割った

$$\chi^2=\sum_{i=1}^{20}\frac{n_i(y_i-\hat{\mu})^2}{\hat{\mu}}$$
$$=5.5503+33.6314+\cdots+11.9014=346.5154$$

によって計算されている．この χ^2 は，自由度 $N-1=199$ のカイ 2 乗分布に従うことから

$$p=\text{Chisq.dist.rt}(346.5154,199)=4.49\times10^{-10}$$

がえられ，ポアソン分布のあてはめは支持されない．

ガンマ・ポアソン分布は，式（6.6）から

$$P_i^{GP}=\frac{\Gamma(y_i+1/\hat{\sigma})}{\Gamma(y_i+1)\Gamma(1/\hat{\sigma})}\frac{(\hat{\mu}\hat{\sigma})^{y_i}}{(1+\hat{\mu}\hat{\sigma})^{y_i+1/\hat{\sigma}}}$$

表 7.5　ポアソン分布およびガンマ・ポアソン分布のあてはめ

	コロニー			ポアソン分布				ガンマ･ポアソン分布			
	数	頻度		確率	適合度	推定値		確率	適合度	推定値	対数尤度
i	y	n		$P^{Poisson}$	カイ2乗	n^		P^{GP}	カイ2乗	n^	$-2\ln L_i^{GP}$
1	3	1		0.0046	5.5503	0.9		0.0154	3.2153	3.1	8.3489
2	4	8		0.0122	33.6314	2.4		0.0287	19.4825	5.7	56.8052
3	5	8		0.0262	24.3541	5.2		0.0453	14.1082	9.1	49.5127
4	6	16		0.0468	33.1415	9.4		0.0627	19.1987	12.5	88.6115
5	7	23		0.0716	29.5588	14.3		0.0782	17.1232	15.6	117.2278
6	8	19		0.0958	13.0287	19.2		0.0894	7.5475	17.9	91.7399
7	9	14		0.1140	3.8224	22.8		0.0951	2.2143	19.0	65.8787
8	10	19		0.1221	0.8943	24.4		0.0950	0.5181	19.0	89.4426
9	11	17		0.1189	0.1335	23.8		0.0899	0.0773	18.0	81.8977
10	12	13		0.1061	2.0199	21.2		0.0812	1.1701	16.2	65.2901
11	13	9		0.0874	4.4068	17.5		0.0703	2.5528	14.1	47.7989
12	14	12		0.0669	12.1278	13.4		0.0586	7.0256	11.7	68.0938
13	15	4		0.0478	6.8736	9.6		0.0472	3.9818	9.4	24.4206
14	16	12		0.0320	31.3547	6.4		0.0369	18.1636	7.4	79.1579
15	17	8		0.0201	29.5530	4.0		0.0281	17.1199	5.6	57.1437
16	18	6		0.0120	29.7726	2.4		0.0209	17.2471	4.2	46.4384
17	19	6		0.0068	38.5009	1.4		0.0151	22.3033	3.0	50.2957
18	20	2		0.0036	16.1165	0.7		0.0107	9.3362	2.1	18.1359
19	21	2		0.0018	19.7729	0.4		0.0075	11.4543	1.5	19.5848
20	22	1		0.0009	11.9014	0.2		0.0051	6.8944	1.0	10.5532
計	20	200	μ^=	10.7100	**346.5154**	$=\chi^2$	μ^=	**10.7100**	**200.7343**	$=\chi^2$	**1136.3779**
区分数		N			199	$=df$	σ^=	**0.0678**	199	$=df$	$\ln L^{GP}$
					0.0000	$=p$	σ'^=	1.7262	0.4522	$=p$	
	$Var(y)=\mu$^=			10.7100	$Var(y)=\mu$^(μ^σ^+1))=			18.4880			

である，ガンマ・ポアソン分布のパラメータは，Excel のソルバーで対数尤度を最大化（マイナス 2 倍の場合には最小化）するように $\hat{\mu}$ と $\hat{\sigma}$ を変化させて推定する．表 7.2 の JMP で求められた過分散パラメータ $\hat{\sigma}'=1.7262$ は，$\sigma'=1+\mu\sigma$ の関係から $\hat{\sigma}$ を

$$\hat{\sigma}=\frac{\hat{\sigma}'-1}{\hat{\mu}}=\frac{1.7262-1}{10.7100}=0.0678$$

と換算することができる．適合度のカイ 2 乗検定は，式（6.10）で示したガンマ・ポアソン分布の分散

$$Var(\hat{y})=\hat{\mu}(1+\hat{\mu}\hat{\sigma})$$
$$=10.7100\times(1+10.7100\times0.0678)=18.4880$$

を用いて，

$$\chi^2_{GP}=\sum_{i=1}^{20}\frac{n_i(y_i-\hat{\mu})^2}{\hat{\mu}(1+\hat{\mu}\hat{\sigma})}$$
$$=3.2153+19.4825+\cdots+6.8944=200.7343$$

となる．χ^2_{GP} は，自由度 $N-1=198$ のカイ 2 乗分布に従うことから

$$p=\text{Chisq.dist.rt}(200.7343,198)=0.4522$$

がえられ，ガンマ・ポアソン分布のあてはめは棄却されない．

このように，統計ソフトの出力結果を鵜呑みにすることなく，Excel を用いて再現できるかを確認することは，統計モデルの理論の理解をより確実なものとし，統計ソフトが対応していない課題に対し，応用できる力の根源となる．

JMP のバージョン 14 をベースにした解析例を示してきたのであるが，バージョン 15 が新たに提供されたので，使い始めると結果がまったく異なる事態に直面する．ポアソン分布のあてはめに対する適合度の検定は，表 7.6 に示すようにバージョン 15 では，カイ 2 乗=75.1451 であるのに対し，表 7.7 に示すようにバージョン 14 では，カイ 2 乗=346.5154 とまったく異なる．少なくともバージョン 14 についての適合度の検定は表 7.5 に示すように Excel で再現できるが，バージョン 15 の自由度 12 となっていることも不可解である．新しい理論に基づいているのかも知れないが，調査中である．

表 7.6　JMP バージョン 15 のポアソン分布の適合度の検定

要約統計量

平均	10.71
標準偏差	4.3184657
平均の標準誤差	0.3053616
平均の上側95%	11.31216
平均の下側95%	10.10784
N	200

Poisson分布のあてはめ

パラメータ		推定値	標準誤差	下側95%	上側95%
平均	λ	10.71	0.2314087	10.262827	11.169978

指標	
(-2)*対数尤度	1172.297
AICc	1174.3172
BIC	1177.5954

適合度検定

	X2	自由度	Prob>X2
Pearsonのカイ2乗	75.145111	12	<.0001*

表 7.7　JMP バージョン 14 のポアソン分布の適合度の検定

要約統計量

平均	10.71
標準偏差	4.3184657
平均の標準誤差	0.3053616
平均の上側95%	11.31216
平均の下側95%	10.10784
N	200

Poisson分布のあてはめ

パラメータ推定値

種類	パラメータ	推定値	下側95%信頼限界	上側95%信頼限界
尺度	λ	10.71	10.262827	11.169978

指標	
(-2)*対数尤度	1172.297
AICc	1174.3172
BIC	1177.5954

適合度検定

Pearsonのカイ2乗

X2	Prob>X2
346.515406	<.0001*

注: Ho = Poisson分布からのデータ。 p値が小さい場合はHoを棄却。

7.2. カブトガニのサテライト数に対する探索的解析

得られたカウント・データが過分散である場合，第 7.1 節では，名義尺度の 2 つの説明変数で層別することにより，分散と平均の比が小さくなることを示した．第 1.13 節の表 1.45 に示したデータには，説明変数として順序尺度（甲羅の色，後体部の棘の状態）の 2 変数，連続尺度（甲羅の幅，体重）の 2 変数も含まれている［アグレスティ（2003)］．これらの説明変数が，反応変数であるサテライト数にどのような関わり合いがあるのか探索的解析を行う．なお，この探索的解析については，高橋（2019a)「最尤法による探索的ポアソン回帰」で詳細に示している．

甲羅の色・後体部の棘

最初の一歩は，説明変数が名義あるいは順序尺度の場合には，それらのカテゴリ（水準）で層別してサテライト数の分布の形状を観察し，過分散の程度を把握することである．JMP の「二変量の関係」でサテライト数を説明変数で層別した結果を図 7.3 に示す．

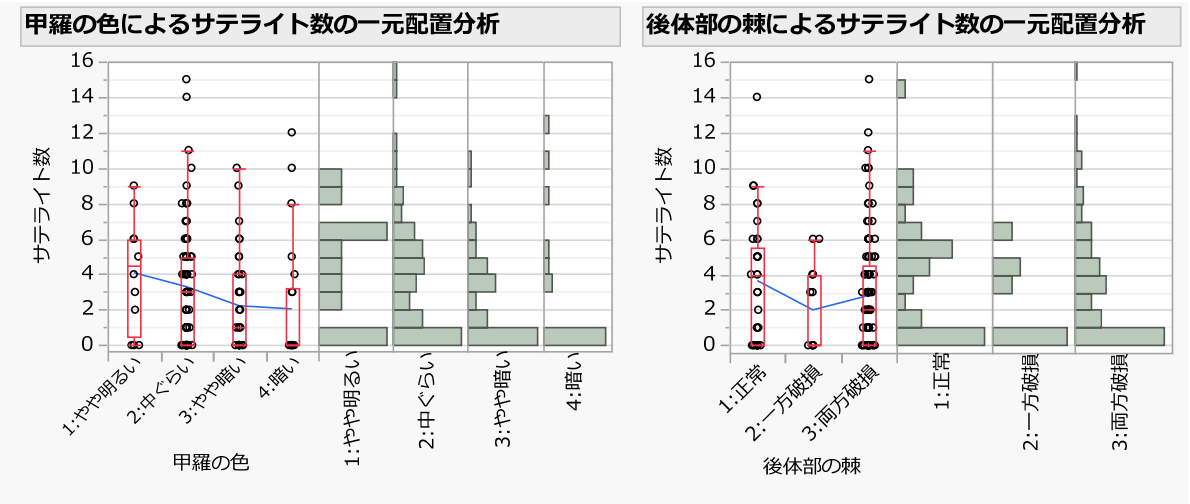


図 7.3　甲羅の色および後体部の棘の状態とサテライト数の関連

雌の甲羅の色については，暗くなるに従いゼロ・カウントの割合が増えサテライト数の平均値が減少傾向であることが読み取れる．雌の後体部の棘の状態については，正常の場合には，サテライト数の 5 匹あたりに山があり，雄が連結する割合が多いようであるが，サテライト数の平均値は同程度である．

表 7.8 に甲羅の色と後体部の棘の状態を組み合わせた場合のサテライト数につての N，平均，分散，および，分散と平均の比を示す．甲羅の色が暗くなるにつれて後体部の棘は，正常

から一方破損，さらに両方破損へ移行するが，ある程度のサテライト数が観察されている場合には，分散と平均の比が 2 以上であり過分散が解消する様子はない．

表 7.8　甲羅の色別 後体部の棘別 の分散/平均の比

甲羅の色	棘の状態	N	平均	分散	分散/平均
1:やや明るい	**1:正常**	**9**	**4.44**	**10.53**	**2.37**
	2:一方破損	2	4.50	4.50	1.00
	3:両方破損	1	0.00	-	-
2:中ぐらい	**1:正常**	**24**	**3.29**	**12.13**	**3.68**
	2:一方破損	8	1.75	6.21	3.55
	3:両方破損	63	3.49	10.03	2.87
3:やや暗い	**1:正常**	**3**	**5.33**	**10.33**	**1.94**
	2:一方破損	4	1.75	4.25	2.43
	3:両方破損	37	2.03	6.25	3.08
4:暗い	**1:正常**	**1**	**0.00**	-	-
	2:一方破損	1	0.00	-	-
	3:両方破損	20	2.25	13.99	6.22
	全体	173	2.92	9.91	3.40

甲羅の幅・体重

雌の甲羅の幅とサテライト数の関連については，第 1.13 節で対数リンクによるポアソン回帰の結果を示した．第 6.6 節および第 6.7 節では，甲羅の幅とサテライト数の関連についてガンマ・ポアソン回帰およびゼロ過剰ガンマ・ポアソン回帰の結果を示した．図 7.4 に示すように，甲羅の幅が小さい場合，および，体重が軽い場合にはサテライト数がゼロの場合が多く，甲羅の幅あるいは体重が大きくなるに従い，サテライト数が急激に増えている．

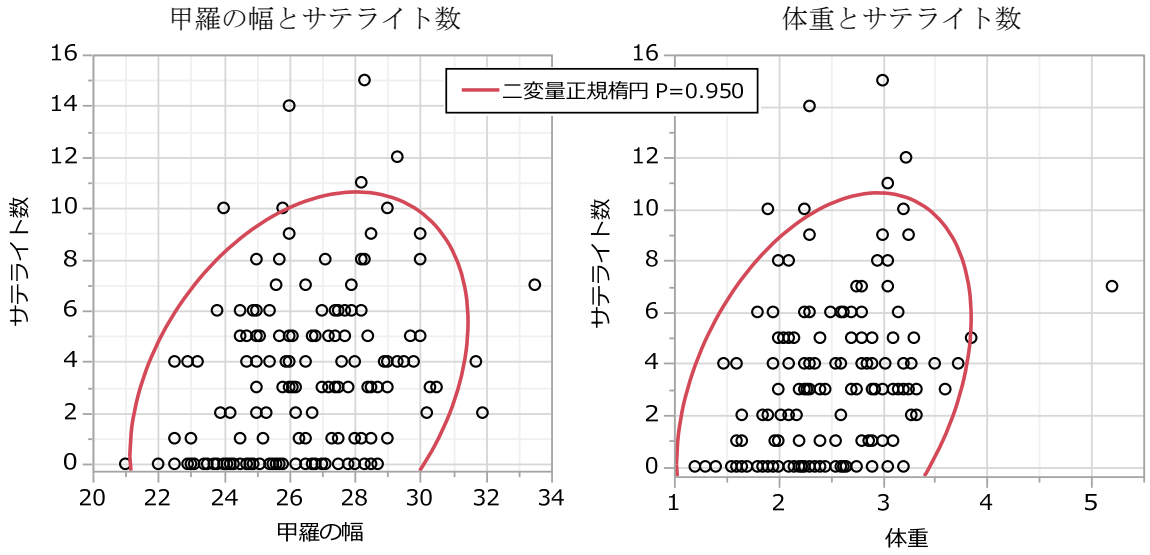


図 7.4　甲羅の幅および体重の増加に伴うとサテライト数の変化

探索的な解析を行う際には，きめ細かなデータの吟味が大切である．サテライト数が甲羅の幅に対して指数関数的に増加するという仮定をしたのだが，散布図を仔細に見れば，幅が 29 cm を超えると，ゼロ・カウントもなくなり，逆にサテライト数は減少傾向となる．体重とサテライト数の関係も，3.0 kg を超えるとサテライト数も同様に減少傾向となる．

甲羅の幅を Y 軸，体重を X 軸，サテライト数をラベルとした散布図を図 7.5 に示す．相関が高いことは自明である．観察すべきことは，サテライト数に対する甲羅の幅と体重の相互関係を読み解くことにある．まず，甲羅の幅を固定して左から右に水平方向にサテライト数の変化を追と増加傾向が読み取れる．次に体重を固定して下から上に垂直方向にサテライト数の変化を追うと少なくなったり多くなったりし，はっきりした傾向が読み取れない．左下から右上は，明らかにサテライト数が増加傾向となってる．

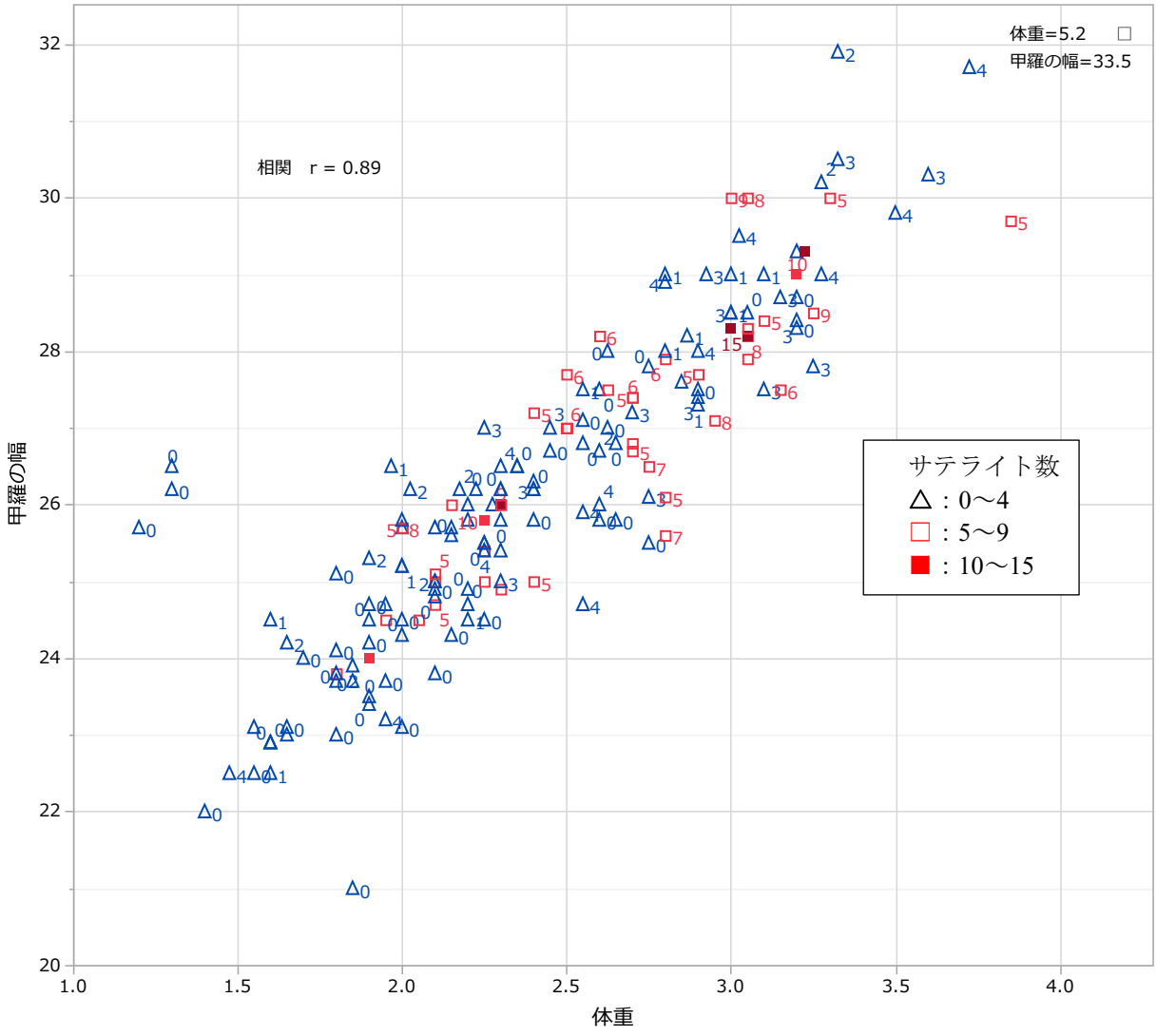


図 7.5　甲羅の幅および体重の散布図にサテライト数と

ポアソン重回帰

過分散を承知で，対数リンクによるポアソン重回帰を行い，甲羅の幅か体重か，どちらがサテライト数との関連が高いか検討する．表 7.9 に示すように，甲羅の幅の推定値は，0.0461，体重の推定値は，0.4470 であり，尤度比検定の結果は，体重のみが有意な差である．

表 7.9　対数リンクによるポアソン重回帰

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-1.2952	0.8989	2.0691	0.1503
甲羅の幅	0.0461	0.0467	0.9658	0.3257
体重	0.4470	0.1586	7.9780	0.0047*

共分散

	切片	甲羅の幅	体重
切片	0.8080	-0.0412	0.1156
甲羅の幅	-0.0412	0.0022	-0.0067
体重	0.1156	-0.0067	0.0252

この結果は，甲羅の幅が増加した場合にサテライト数が増大することを否定しているのではなく，図 7.5 で示したように，体重が同じ場合に，甲羅の幅が大きくなる Y 軸方向にサテライト数の増加が見いだされにくいことを反映している．従って，体重が増えれば，甲羅の幅も広くなり，サテライト数も増えるが，体重を固定した場合に甲羅の幅が広くなってもサテライト数はさほど増えないと解釈される．図 7.6 は，JMP による対数リンクでのポアソン 2 変量回帰に引き続き「予測プロファイル」の機能を用い，体重を（2，3，4 kg）と変化させた場合に，甲羅の幅がサテライト数に及ぼす影響を図示したものである．甲羅の幅は体重の増

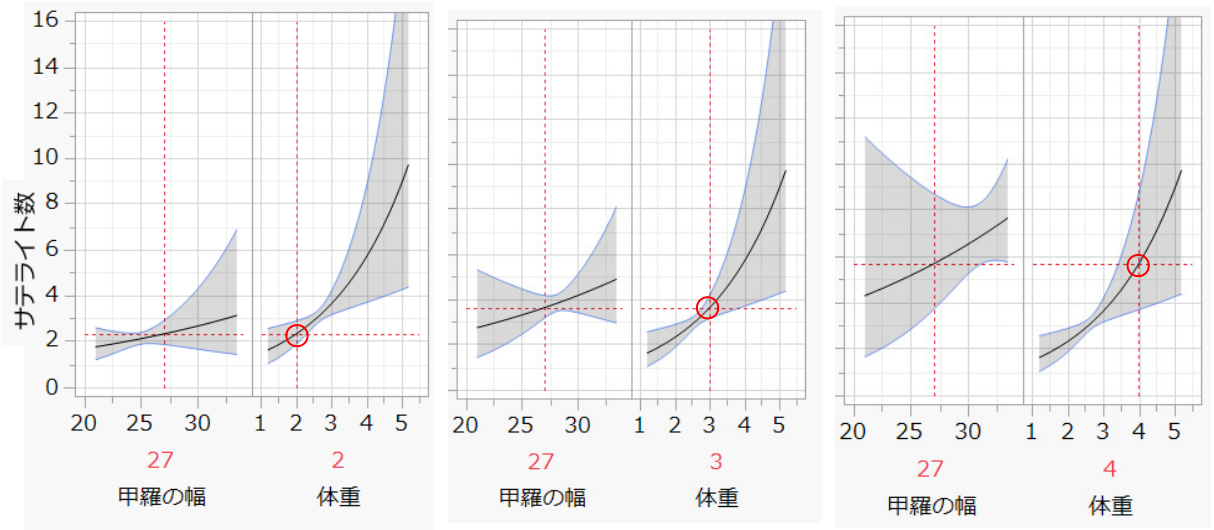


図 7.6　体重を変化させた場合の甲羅の幅とサテライト数との関連

加に伴いサテライト数も全体的には増加しているが，95%信頼区間の表示から，傾きがマイナスになる可能性も読み取れ，このことが表 7.9 に示したように甲羅の幅が p=0.3257 であることに対応する．

JMP の予測プロファイルは，ポアソン重回帰に限らず通常の重回帰の場合でも推定結果の可視化に役立つ．これは，1 変数の場合の回帰の推定値に対して 95%信頼区間あるいは 95%予測区間を図示することは一般的であるが，2 変数の場合にはどうしたら良いのだろうか．そこで，2 つの変数の一方を固定し他方を変化させた場合，回帰の推定値および 95%信頼区間を示すことにより可視化する方法であり，JMP では「予測プロファイル」と称している．

Excel による量的変数に対する予測プロファイル

単回帰分析において回帰直線の 95%信頼区間を散布図上に重ね書きすることは，結果の解釈をするために一般的に行なわれている．ほとんどの回帰分析の入門書には，95%信頼区間の描画のための式が示されている．ポアソン回帰でも同様に 95%信頼区間の描画のための式を第 1.4 節で示した．

さて，説明変数が 2 つ以上ある場合には，どのように描いたら良いのであろうか．図 7.6 左に示した JMP の「予測プロファイル」では，甲羅の幅を 27 cm，体重を 2 kg に固定した場合にそれぞれのポアソン回帰曲線と 95%信頼区間が示されている．更に体重を 3 kg に変化させた場合は図 7.6 中，4 kg に変化させた場合は図 7.6 右に図示されている．SAS を含む他の統計ソフトで，このような結果の解釈に有益な表示にこれまで出会ったことがない．

Excel を用いて，JMP による「予測プロファイル」を再現することにより，JMP 内部での計算方法を再現してみる．表 7.9 で推定された回帰パラメータ $\hat{\boldsymbol{\beta}}$ は，

$$\hat{\boldsymbol{\beta}} = [-1.2952 \quad 0.0461 \quad 0.4470]^T$$

である．切片を $x_0 = 1$，甲羅の幅を $x_1 = 27$ cm とした場合に，体重を x_2 =（1，2，3，4，5 kg）と変化させた場合の推定値 $\hat{y}$ は，

$$\hat{y}_{x_2=1} = \exp[1\times(-1.2952) + 27\times 0.0461 + 1\times 0.4470] = \exp(0.3958) = 1.4856$$
$$\hat{y}_{x_2=2} = \exp[1\times(-1.2952) + 27\times 0.0461 + 2\times 0.4470] = \exp(0.8428) = 2.3228$$
$$\vdots$$
$$\hat{y}_{x_2=5} = \exp[1\times(-1.2952) + 27\times 0.0461 + 5\times 0.4470] = \exp(2.1837) = 8.8791$$

として計算される．それぞれの推定値の分散 $Var(\ln \hat{y})$ は，パラメータの共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ としたときに表 7.9 から

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = \begin{array}{|ccc|}\hline \mathbf{0.8080} & -0.0412 & 0.1156 \\ -0.0412 & \mathbf{0.0022} & -0.0067 \\ 0.1156 & -0.0067 & \mathbf{0.0252} \\ \hline\end{array}$$

なので，体重が x_2=1 kg の場合に，$\boldsymbol{x}$=[1　27　1] として，次の 2 次形式で計算することができる．

$$Var(\ln \hat{y}_{x_2=1}) = \boldsymbol{x}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}^T = 0.0703$$

推定値 $\hat{y}_{x_2=1}$ =1.4856 の 95%信頼区間は，

$$(U95\%\ L95\%,) = \exp\left[\ln \hat{y}_{x_2=1} \pm 1.96\sqrt{Var(\ln \hat{y}_{x_2=1})}\right]$$
$$= \exp\left[0.3958 \pm 1.96\sqrt{0.0703}\right] = (0.8836, 2.4978)$$

として計算される．逆に体重を x_2=2 kg に固定して甲羅の幅を x_1 =（21，24，27，30，33 cm）と変化させた場合も同様の計算方法によって推定値および 95%信頼区間を計算することができる．

表 7.10 に予測プロファイルを計算するための Excel シートを示す．推定値 $\hat{\boldsymbol{\beta}}$ を列ベクトル，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を与えて任意の $\boldsymbol{x}$=[x_0　x_1　x_2] に対する計算シートとなっている．甲羅の幅を x_1 =27 cm と固定し x_2 =（1，2，3，4，5 kg）と変化させた場合，体重を x_2 =2 kg と固定し x_1 =（21，24，27，30，33 cm）と変化させた場合の計算結果が示されている．体重を x_2 =3 kg と変える場合には，Excel シー上で（3，3，3，3，3 kg）と上書きすれば，再計算が行なわれる．

表 7.10　予測プルファイルの計算のための Excel シート

$\boldsymbol{x}$		推定値 $\boldsymbol{\beta}$	共分散 $\boldsymbol{\Sigma}$	切片	甲羅の幅	体重	
x_0	切片	-1.2952	β_0	**0.8080**	-0.0412	0.1156	
x_1	甲羅の幅	0.0461	β_1	-0.0412	**0.0022**	-0.0067	
x_2	体重	0.4470	β_2	0.1156	-0.0067	**0.0252**	
x_0	x_1	x_2	ln y^	Var(ln y^)	y^	L95%	U95%
1	**28.3**	**3.05**	1.3720	0.0026	**3.9433**	3.5670	4.3592
1	**27**	1	0.3958	0.0703	1.4856	0.8836	2.4978
1	**27**	2	0.8428	0.0125	2.3228	1.8653	2.8927
1	**27**	3	1.2898	0.0051	3.6319	3.1574	4.1778
1	**27**	4	1.7367	0.0480	5.6788	3.6963	8.7245
1	**27**	5	2.1837	0.1412	8.8791	4.2511	18.5457
1	21	**2**	0.5663	0.0391	1.7618	1.1959	2.5954
1	24	**2**	0.7046	0.0061	2.0230	1.7352	2.3584
1	27	**2**	0.8428	0.0125	2.3228	1.8653	2.8927
1	30	**2**	0.9810	0.0583	2.6672	1.6618	4.2809
1	33	**2**	1.1193	0.1434	3.0626	1.4581	6.4324

最初の行の $\boldsymbol{x}$ = [1　28.3　3.05] は，表 1.45 で与えられたデータリストの最初のデータである．推定値と 95%信頼区間は，JMP による予測プロファイルの結果と一致する．

Excel シートでの計算式を次に示す．

$\ln \hat{y}_i$ = Mmult($\boldsymbol{x}_i$の範囲, $\hat{\boldsymbol{\beta}}$の範囲)

$Var(\ln \hat{y}_i)$ = Mmult(Mmult($\boldsymbol{x}_i$の範囲, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$の範囲), Transpose($\boldsymbol{x}_i$の範囲))

$$\hat{y}_i = \exp(\ln \hat{y}_i)$$
$$L95\% = \exp(\ln \hat{y}_i - 1.96\sqrt{Var(\ln \hat{y}_i)})$$
$$U95\% = \exp(\ln \hat{y}_i + 1.96\sqrt{Var(\ln \hat{y}_i)})$$

これらの予測プロファイルを Excel の散布図の機能を使って作図した結果を図 7.7 に示す．この図は，図 7.6 の左端の図に対応する．

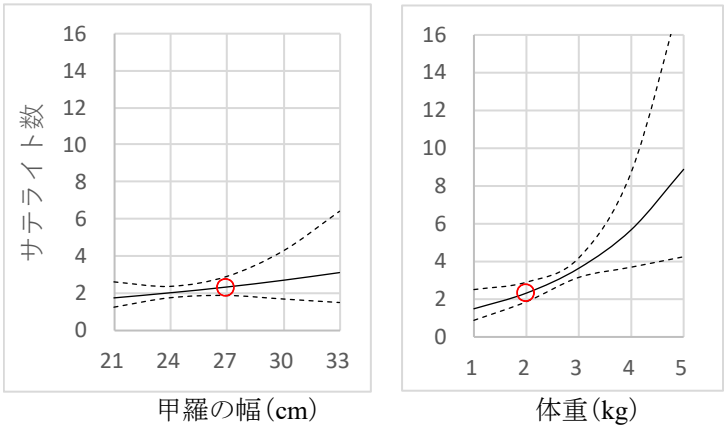


図 7.7　体重を 2 kg に固定した場合の甲羅の幅のプロファイル（右）と
甲羅の幅を 27 cm に固定した場合の体重のプロファイル（左）

交互作用（甲羅の色×体重）を含めたポアソン重回帰

さて，甲羅の色が暗くなるにつれて棘の破損が多くなり，サテライト数の平均が減ることを表 7.8 で示した．では，甲羅の色と体重を組み合わせた場合に，何らかの関連が見出されるのであろうか．名義あるいは順序尺度の甲羅の色と連続尺度の体重の 2 変数がサテライト数に及ぼす影響を観察すためには，図 7.8 に示すように JMP の「二変量の関係」を用いて，「甲羅の色」で「グループ」化し，「層別確率楕円」を描くことにより概観できる．

甲羅の色が「やや明るい」場合には，確率楕円に左右の振れがないので，サテライト数は体重に関連しないようであり，「中ぐらい，やや暗い，暗い」場合は，やや正の相関を持つように傾いており，体重が増えればサテライト数が増えるようである．

この様な関連を，ポアソン回帰で見出すためには，甲羅の色について何らかの数値を使いデザイン行列化し，体重との交互作用を含めたポアソン重回帰を行う必要がある．JMP の一般化線形モデルでは，名義尺度に対しては対比型のデザイン行列を自動生成するので，「モデル効果の構成」で表 7.11 に示すように（甲羅の色，体重，甲羅の色*体重）を設定すればよい．なお，順序尺度として JMP で設定した場合には，対比型とはならないので注意が必要である．

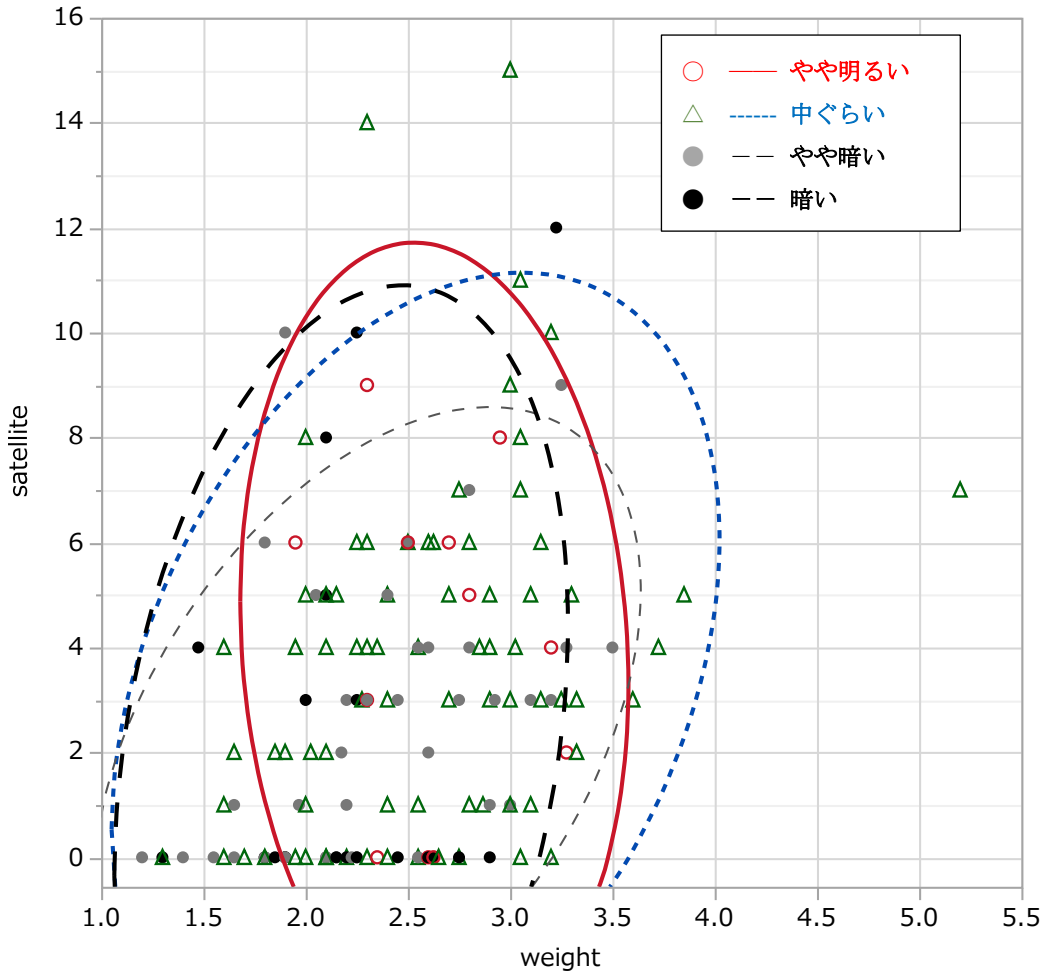


図 7.8　体重とサテライト数に対する甲羅の色による層別確率楕円

解析モデルに多水準の名義尺度が含まれ，さらに交互作用が含めると，解析モデルのデザイン変数が膨張する．パラメータの推定結果を表 7.12 に示すが，切片を含めて 1+3+1+3=8 変数となり，このままでは，結果の解釈は困難を極める．そこで，「予測プロファイル」の機能を用いて，図 7.9 に示すように甲羅の色ごとの体重の増加によるサテライト数との関連を概観する．

表 7.11　JMP の一般化線形モデルにおける交互作用の設定

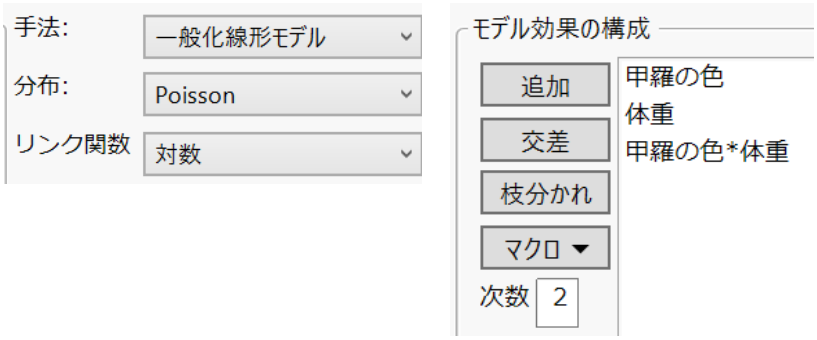


表 7.12　甲羅の色と体重の交互作用を含めた対数リンクでのポアソン重回帰

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-0.2778	0.3450	0.6530	0.4191
甲羅の色[1:やや明るい]	2.2221	0.7978	7.5086	0.0061*
甲羅の色[2:中ぐらい]	0.2010	0.3797	0.2812	0.5959
甲羅の色[3:やや暗い]	-1.1855	0.4865	6.0352	0.0140*
体重	0.5463	0.1344	16.0804	<.0001*
甲羅の色[1:やや明るい]*体重	-0.7518	0.3050	6.1530	0.0131*
甲羅の色[2:中ぐらい]*体重	-0.0646	0.1456	0.1967	0.6574
甲羅の色[3:やや暗い]*体重	0.3820	0.1870	4.2010	0.0404*

過分散の調整を行っていないので p 値は小さ目になっている

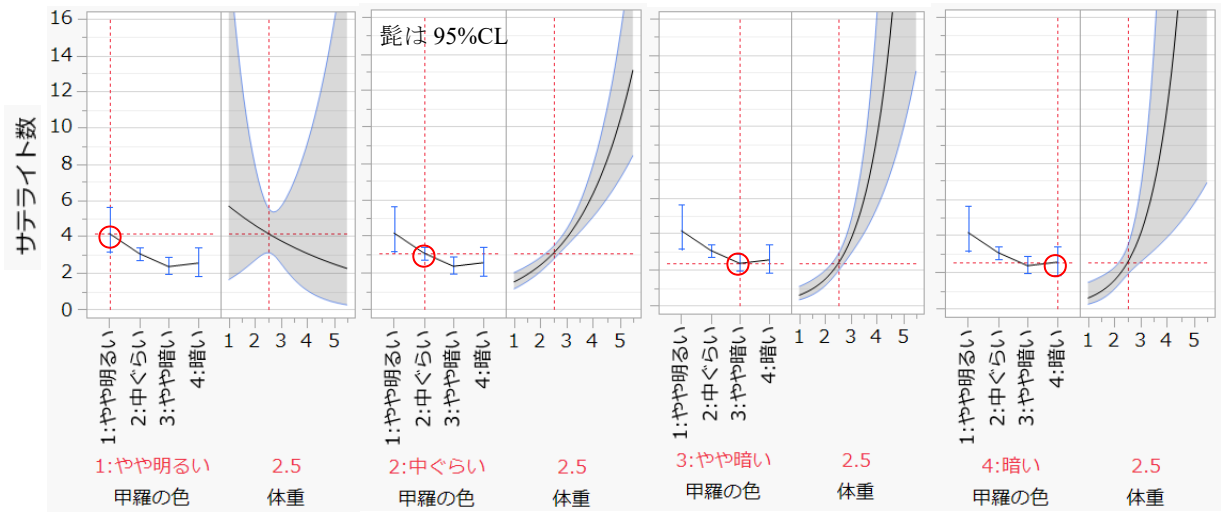


図 7.9　甲羅の色別の体重とサテライト数の関連

この結果は，図 7.8 の層別確率楕円で観察し見いだした結果を支持する結果となっている．甲羅の色が「やや明るい」場合は，体重とサテライト数の関連は，マイナスの傾きで，95%信頼区間からプラスの傾きも起こりえる結果となっており，一定の傾向は見いだせない．「中ぐらい」以上では，体重とサテライト数との関連はプラスの傾きで，95%信頼区間からも明らかにプラスの傾きが支持される．

Excel による質的変数を含む予測プロファイル

JMP の名義尺度のダミー変数（デザイン変数）は，表 7.13 に示すように（1，-1）対比型とする．体重と甲羅の色の交互作用は，体重とそれぞれのダミー変数の積となる．表 7.12 に対数リンクによるポアソン回帰の推定結果，図 7.9 に予測プロファイルを示した．変数の数は多くなるが，予測プロファイル作成のための計算方法は，2 変数のポアソン回帰の場合と考え方は同じである．

表 7.13　（1，−1）対比型のダミー変数（デザイン変数）

甲羅の色	x_1	x_2	x_3
1:やや明るい	1	0	0
2:中ぐらい	0	1	0
3:やや暗い	0	0	1
4:暗い	-1	-1	-1

Excel による予測プロファイルも変数が増えると煩雑になるが，基本は 2 変数の場合と同じである．表 7.12で得られた推定値，さらに JMP で出力した共分散行列の結果をコピーし，表 7.14 に示すようにパラメータの共分散行列の枠にペーストする．

甲羅の色が「1：やや明るい」場合には，表 7.13 から（$x_1=1$，$x_2=0$，$x_3=0$），甲羅の色が「2：中ぐらい」場合には，（$x_1=0$，$x_2=1$，$x_3=0$），となり，体重を $x_4=$（1，2，3，4，5 kg）と変化させ，交互作用（$x_5=x_1x_4$，$x_6=x_2x_4$，$x_7=x_3x_4$）を計算した推定値結果が示されている．さらに，体重を 2.5 kg に固定し，甲羅の色を（1，2，3，4 kg）と変化させた場合，甲羅の色が「4：暗い」場合は，（$x_1=-1$，$x_2=-1$，$x_3=-1$）であるが，それらの交互作用が計算されている．

表 7.14　予測プルファイルの計算のための Excel シート

		推定値	共分散		甲羅の色			体重	甲羅の色×体重		
	項	$\boldsymbol{\beta}$^	$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^)	切片	A_1	A_2	A_3	W	A_1×W	A_2×W	A_3×W
x_0	切片	-0.278	β^$_0$	**0.119**	0.140	-0.106	-0.060	-0.046	-0.052	0.041	0.024
x_1	A_1	2.222	β^$_1$	0.140	**0.637**	-0.152	-0.199	-0.052	-0.241	0.056	0.074
x_2	A_2	0.201	β^$_2$	-0.106	-0.152	**0.144**	0.048	0.041	0.056	-0.054	-0.020
x_3	A_3	-1.185	β^$_3$	-0.060	-0.199	0.048	**0.237**	0.024	0.074	-0.020	-0.089
x_4	W体重	0.546	β^$_4$	-0.046	-0.052	0.041	0.024	**0.018**	0.019	-0.017	-0.010
x_5	A_1×W	-0.752	β^$_5$	-0.052	-0.241	0.056	0.074	0.019	**0.093**	-0.021	-0.028
x_6	A_2×W	-0.065	β^$_6$	0.041	0.056	-0.054	-0.020	-0.017	-0.021	**0.021**	0.008
x_7	A_3×W	0.382	β^$_7$	0.024	0.074	-0.020	-0.089	-0.010	-0.028	0.008	**0.035**

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	ln y^	$Var(\ln y$^)	y^	L95%	U95%
1	**1**	0	0	1	1	0	0	1.739	0.405	5.690	1.635	19.803
1	**1**	0	0	2	2	0	0	1.533	0.075	4.633	2.713	7.913
1	**1**	0	0	3	3	0	0	1.328	0.044	3.773	2.498	5.699
1	**1**	0	0	4	4	0	0	1.122	0.314	3.072	1.024	9.212
1	**1**	0	0	5	5	0	0	0.917	0.883	2.501	0.396	15.786
1	0	**1**	0	1	0	1	0	0.405	0.022	1.499	1.119	2.008
1	0	**1**	0	2	0	2	0	0.887	0.007	2.427	2.068	2.848
1	0	**1**	0	3	0	3	0	1.368	0.004	3.928	3.492	4.419
1	0	**1**	0	4	0	4	0	1.850	0.013	6.359	5.082	7.957
1	0	**1**	0	5	0	5	0	2.332	0.035	10.293	7.130	14.861
1	1	0	0	**2.5**	2.5	0	0	1.431	0.022	4.181	3.128	5.589
1	0	1	0	**2.5**	0	2.5	0	1.127	0.004	3.087	2.746	3.471
1	0	0	1	**2.5**	0	0	2.5	0.857	0.010	2.357	1.930	2.879
1	**-1**	**-1**	**-1**	**2.5**	-2.5	-2.5	-2.5	0.936	0.024	2.550	1.887	3.447
	甲羅の色			体重	甲羅の色×体重			推定値	分散	推定値	95%信頼区間	

推定値，分散，95%信頼区間の計算は，表 7.10 の Excel シートで示したと同様に次に示す計算式が用いられている．

$$\ln \hat{y}_i = \mathrm{Mmult}(\boldsymbol{x}_i\text{の範囲}, \hat{\boldsymbol{\beta}}\text{の範囲})$$
$$(1\times 8) \qquad (8\times 1)$$

$$Var(\ln \hat{y}_i) = \mathrm{Mmult}(\mathrm{Mmult}(\boldsymbol{x}_i\text{の範囲}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{の範囲}), \mathrm{Transpose}(\boldsymbol{x}_i\text{の範囲}))$$
$$(1\times 8) \qquad (8\times 8) \qquad (8\times 1)$$

$$\hat{y}_i = \exp(\ln \hat{y}_i)$$
$$L95\% = \exp(\ln \hat{y}_i - 1.96\sqrt{Var(\ln \hat{y}_i)})$$
$$U95\% = \exp(\ln \hat{y}_i + 1.96\sqrt{Var(\ln \hat{y}_i)})$$

甲羅の色を「1 ：やや明るい」に固定し，体重を（1，2，3，4，5 kg）と変化させた場合について図 7.10(左)，甲羅の色を「2：中ぐらい」に固定し，体重を（1，2，3，4，5 kg）と変化させた場合について，図 7.10(中)，体重を 2.5 kg に固定した場合の甲羅の色のプロファイ（1，2，3，4）を変化させた場合について図 7.10(右）に示す．

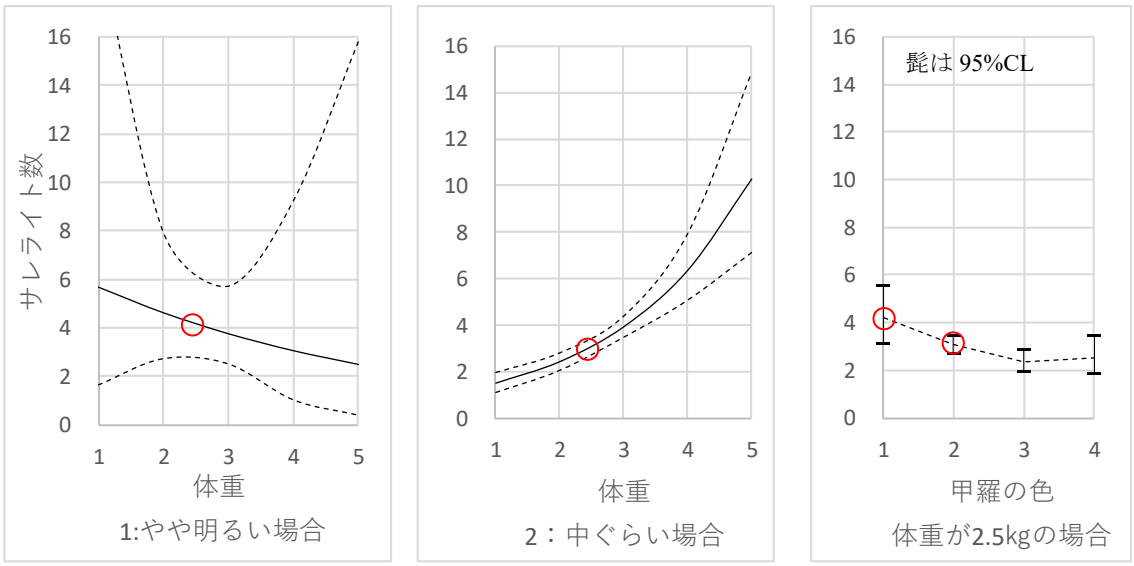


図 7.10　甲羅の色を「1：やや明るい」，「2：中ぐらい」に固定した場合のた場合の体重のプロファイル，体重を 2.5 kg に固定した場合の甲羅の色のプロファイル

交互作用（後体部の棘×体重）を含めたポアソン重回帰

後体部の棘の状態は，甲羅の色によって破損が進行することを表 7.8 で明らかにした．甲羅の色が「中くらい」の場合には，後部の棘が「正常」と「両方破損」に分かれているので，サテライト数との関連を甲羅の色が「中くらい」に限定して関連を調べた結果を図 7.11 に示す．

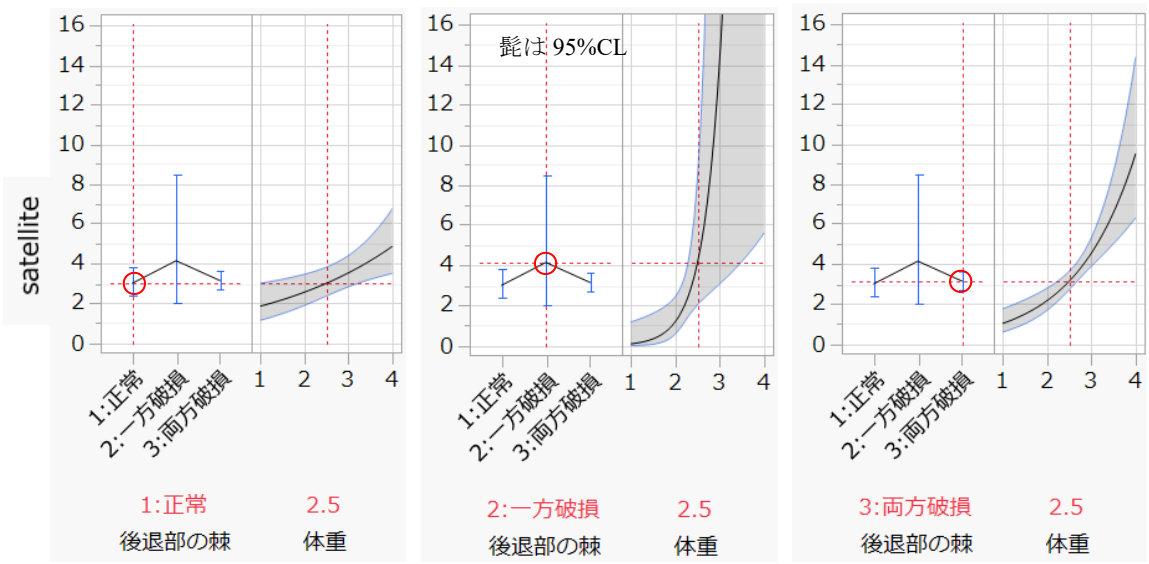


図 7.11　甲羅の色が「中ぐらい」の後部の棘の状態別の体重とサテライト数の関連

甲羅の色が「中ぐらい」で後体部の棘が「正常」の場合に体重が増えればサテライト数も微増する．「一方破損」および「両方破損」では，体重が増えた場合にサテライト数が急増する．表 7.8 から，甲羅の色が「やや明るい」場合には，後体部の棘は 12 匹中 9 匹が「正常」で，図 7.9 から体重が増えてもサテライト数は増えない．甲羅の色が「中ぐらい」に変化すると，体重が増加するとサテライト数も大幅に増える．更に色が「やや暗い，暗い」場合には，更に体重が増えるにつれて，サテライト数が増えるとも言えるが，体重が小さい場合には，サテライト数が減少することが読み取れる．

グラフ・ビルダーによる探索解析的

交互作用が疑われるような観察データに対し，探索的な解析を行うためには，各種のグラフ表示が欠かせない．これまでも JMP の多彩なグラフ表示を活用し，カブトカニの各種の変数とサテライト数の関連を浮き彫りにしてきたが，満足できるものではなかった．全体を俯瞰できるように結果を 1 枚のグラフで表わすことは，可能なのだろうか．JMP の新しい作図機能である「グラフ・ビルダー」を用いた結果を図 7.12 に示す．

この図から，これまでの探索的解析の結果がより鮮明に浮彫される．サテライト数は，甲羅の色が暗くなるにつて後方の棘の破損が進み，それに伴い，体重の軽い雌ほど連結する雄のサテライト数が減少することが読み取れる．甲羅の色が暗くなり，後部の棘の状態が悪くなる加齢現象により，体重の軽い雌ほど連結する雄のサテライト数が減少すると解される．そのため，ゼロ・カウントが多い過分散となったと推測される．

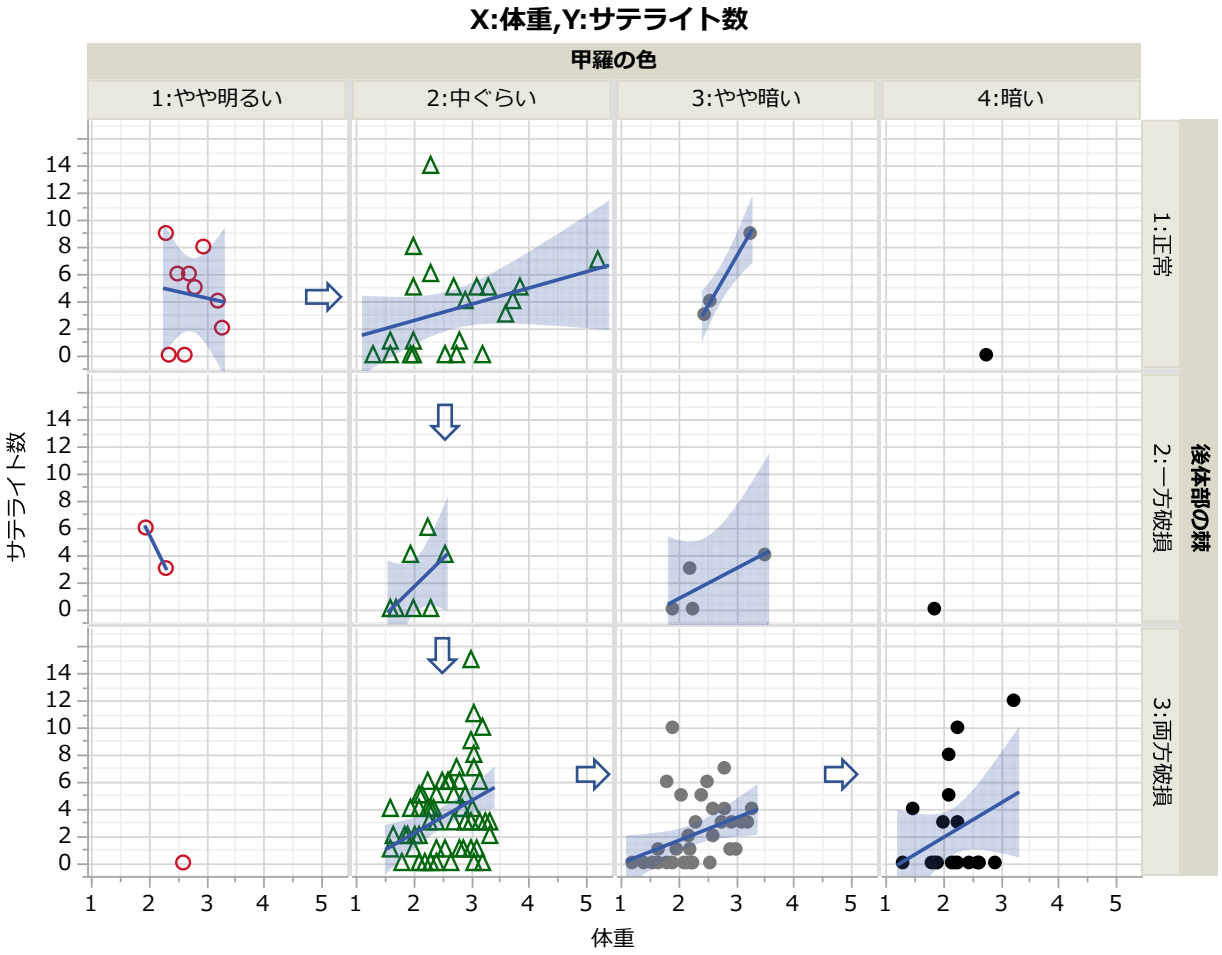


図 7.12　甲羅の色・棘の状態による層別散布図

各セルの中の回帰直線と 95%信頼区間の表示は，通常の回帰分析の結果で，ポアソン回帰の結果ではない．

全データに対するサテライト数の分布について，第 6.6 節で示したようにゼロ過剰ポアソン分布よりも，さらに第 6.7 節で示したようにゼロ過剰ガンマ・ポアソン分布のあてはめが良好であったが，それらの分布を用いた回帰分析には難点がある．これは，図 7.4 にも示したように，甲羅の幅が大きい場合，および，体重が重い場合にはサテライト数のゼロが存在しなくなるので，ゼロ過剰ガンマ・ポアソン分布を仮定して回帰分析を行うと，体重が重い場合にも過剰なゼロが存在を仮定することになり，現実のデータとの乖離を無視できなくなるためである．

対数リンクによるポアソン回帰は，元データには指数曲線のあてはめ，両辺に対数を取り線形化するモデルであり，ゼロ・データに対しては対数変換が行なわれないように調整する仕組みになっている．この仕組みは，一般化線形モデルで分布を正規とし，対数リンクとした場合でも適用され，ゼロを含むようなデータに対し指数曲線をあてはめることが可能とな

る．なお，ポアソン回帰を行っても過分散が解消されないような場合に，正規分布を仮定し，対数リンクによる指数曲線をあてはめる場合にも，ゼロ・データに対する調整が行なわれる．

過剰なゼロが，どのような状況で発生するかを念頭にし，「甲羅の色」，「後体部の棘」とサテライト数の関係から，甲羅の色が暗くなるにつれゼロ・カウントが増加するが，後体部の棘については，関連が見いだされなかった．さらに，甲羅の色と後体部の棘を組み合わせても過分散は解消しなかった．

甲羅の幅と体重の 2 変数間には 0.89 と高い相関があり，2 変数のポアソン回帰に引き続き，図 7.6 に示したように体重を段階的に変化させた場合の甲羅の幅の推定曲線と 95%信頼区間のプロファイルから，甲羅の幅をポアソン回帰の説明変数に加える必要がないことが，視覚的にも見いだされた．もちろん，2 変数のポアソン回帰の尤度比検定で，甲羅の幅の p 値は 0.3257 と有意ではないことからも推測されることではあるが，JMP の予測プロファイル機能は，視覚的に変数相互の関連を見い出し，より具体的な相互関係の理解するために有益である．

予測プロファイル機能により，図 7.9 に示したように 4 水準の甲羅の色と体重の 2 変数に交互作用を加えたポアソン回帰で，甲羅の色が「やや明るい」場合に，体重が増えてもサテライト数が増えないことが図示され，甲羅の色が「中ぐらい，やや暗い，暗い」場合とは，全く異なるプロファイルであることが明示された．他方，図 7.11 に示すように後体部の棘と体重の関連には，交互作用を示唆するような兆候は見いだせなかった．

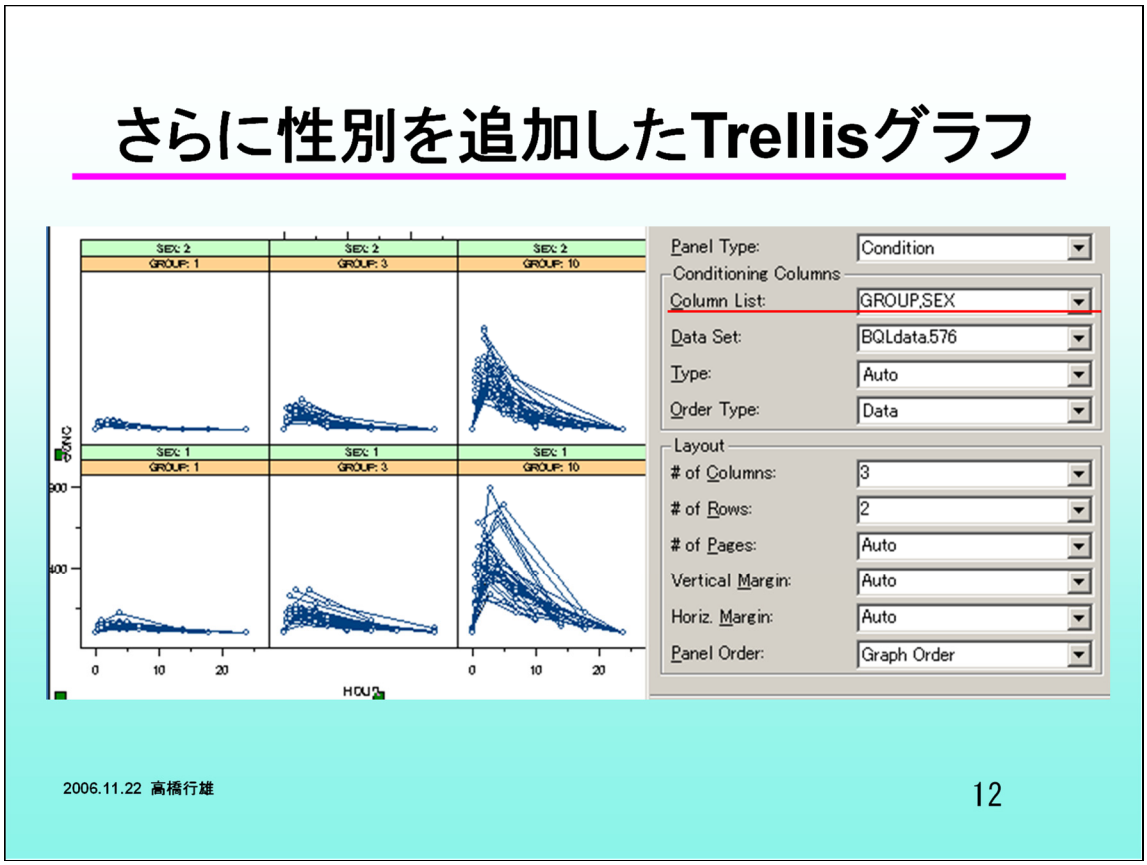
甲羅の色と後体部の棘に体重，さらにそれらの交互作用を含めたポアソン回帰は，観察データなので，データが不均一であり，解を得ることができなかった．これらの変数とサテライト数の関連を見出すためには，図 7.12 に示したように JMP のグラフ・ビルダーが役に立つ．最初に体重とサテライト数の散布図を描き，回帰直線と 95%信頼区間を上書きする．ここまでならば，JMP の伝統的な二変量の関係での対応と同じであるが，これに 4 水準の甲羅の色，3 水準の後体部を組み合わせた 4×3 の場合についてタイル状に体重とサテライト数の回帰直線と 95%信頼区間を並べて表示できた．

グラフ・ビルダーで対数リンクのポアソン回帰が実施できれば申し分ないのであるが，残念ながら現在のバージョン 14 では対応していない．伝統的な回帰分析であっても，名義尺度の水準ごとの散布図行列上に回帰直線の 95%信頼区間が表示されるだけでも，結果を総合的

に俯瞰するために有益である．これに類似する機能が S プラスにあり，以前は愛用していたのであるが，JMP グラフ・ビルダーは，S プラスの機能を大幅に凌駕する探索的な統計解析を支援するツールとして優れている．

S-PLUS の Trellis（格子）グラフ

JMP でグラフ・ビルダーが提供されたときに，S-PLUS の Trellis（格子）グラフを思い出した．Trellis（格子）グラフの有用性については，2006 年の S-PULS ユーザ・コンファレンスで「SAS ユーザのための S-Plus 活用術」を発表した．Web で検索すると（株）NTT データ数理システムのサイトに当時の資料が掲載されているのが見出された．ポアソン回帰の事例ではないが，得られたデータおよび結果のグラフ化の方法について参考にしてもらいたい．久保訳（2009），「R グラフィックス，第 4：lattice パーケージ」も同様と思われる．



http://www.msi.co.jp/splus/usersCase/medical/pdf/06takappt.pdf　2020/05/15 アクセス

株式会社中外臨床研究センター 様
pdf SASユーザのためのS-PLUS活用術で新薬のスピーディーな臨床開発に役立てる

高橋（2006），「SAS ユーザのための S-Plus 活用術」
http://www.msi.co.jp/splus/usersCase/medical/pdf/05chugai.pdf　2020/05/15 アクセス

7.3. 殺人被害者数に関する AICc を用いた分布の同定

Agresti (2013),「Categorical Data Analysis 3rd ed.」の「Section 14.4　Negative Binomial Regression」には,「殺人被害者数に関する調査データ」に対して過分散を考慮した解析結果が示されている．また，このデータを引用して，蓑谷（2013),「一般線形モデルと生存時間解析」の「第 6.5 節　負の 2 項回帰モデル」で，このデータを引用して論じている．どちらの著書でも，各種の過分散モデルをあてはめ，期待度数と回帰パラメータを主体にした記述がなされている．

JMP によるポアソン回帰

調査は，被検者 1,308 人に対して,「過去 12 か月以内に、殺人の被害者であることを個人的に何人知っていますか」と質問した結果である．表 7.15 は，被検者を（黒人と白人）に層別した結果である．分散と平均の比が，それぞれ（2.2027，1.6828）と 1 を大きく超えていることから調査データに特有の過分散が，全体でも層別した場合でも起きている．

表 7.15　何人の被害者を知っていますか

被害者数	白人	黒人	全体
y	*n*	*n*	*n*
0	1070	119	1189
1	60	16	76
2	14	12	26
3	4	7	11
4	0	3	3
5	0	2	2
6	1	0	1
計	1149	159	1308
平均	0.0923	0.5220	0.1445
分散	0.1552	1.1498	0.2951
分散/平均	1.6828	2.2027	2.0423
1以上の割合	6.9%	25.2%	9.1%

白人を $x_1 = 0$，黒人を $x_1 = 1$ とする（0，1）型デザイン行列とし，分布を Poisson，恒等リンクによるポアソン回帰の結果を表 7.16 に示す．切片の推定値 $\hat{\beta}_0 = 0.0923$ は，表 7.15 で示した白人の平均値であり，x に対する推定値 $\hat{\beta}_1 = 0.4298$ は，黒人の平均値 0.5220 人から白人の平均値 0.0923 との差である．もちろん，推定値 $\hat{y}$ は，白人の場合は，$\hat{\beta}_0 = 0.0923$，黒人の場合は,

$$\hat{y}_{\text{黒人}} = \hat{\beta}_0 + 1 \times \hat{\beta}_1 = 0.0923 + 1 \times 0.4298 = 0.5220 \text{ 人}$$

である．適合度の Pearson のカイ 2 乗は 2279.8732 と自由度 1306 に対して大きく $p < 0.0001$ となり，過分散であることが確認される．

表 7.16　白人 vs 黒人に関するポアソン回帰

手法: 一般化線形モデル
分布: Poisson
リンク関数 恒等
☐ 過分散に基づく検定と信頼区間

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)
Pearson	2279.8732	1306	<.0001*
デビアンス	844.7073	1306	1.0000

AICc
1121.9990

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	0.0923	0.0090	106.0000	<.0001*
x	0.4298	0.0580	118.0931	<.0001*

そこで，パラメータの推定値の標準誤差を過分散の調整により大きくする．JMP の「過分散に基づく検定と信頼区間」を考慮した解析を行った結果を表 7.17 に示す．過分散は，Pearson のカイ 2 乗値 2279.8732 を自由度 1306 で割った 1.7457 と推定されている．

表 7.17　白人 vs 黒人に関する過分散を考慮したポアソン回帰

手法: 一般化線形モデル
分布: Poisson
リンク関数 恒等
☑ 過分散に基づく検定と信頼区間

適合度統計量	カイ2乗	自由度	p値	過分散
Pearson	2279.8732	1306	<.0001*	1.7457
デビアンス	844.7073	1306	1.0000	

AICc
646.4465

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値(Prob>ChiSq)
切片	0.0923	0.0118	60.7209	<.0001*
x	0.4298	0.0766	67.6483	<.0001*

この過分散 1.7457 を使い表 7.16 のパラメータ推定値 $\hat{\beta}_1 = 0.4298$ の標準誤差 $SE(\hat{\beta}_1) = 0.0580$ を元の分散に戻し，標準誤差を計算し直した結果が表 7.17 に

$$SE(\hat{\beta}_1') = \sqrt{0.0580^2 \times 1.7457} = 0.0766$$

と計算され，尤度比カイ 2 乗=118.0931 は，過分散で除した

$$\text{尤度比カイ2乗}' = 118.0931 / 1.7457 = 67.6483$$

結果となっている．過分散を調整しても白人と黒人間に知っている殺人の被害者数には明らかな差である．

ポアソン回帰における過分散の調整は，簡便な方法で魅力的ではあるが，元の分布をポアソン分布と仮定したままであり，便宜的な方法である．そこで，ゼロ過剰ポアソン分布，ガンマ・ポアソン分布（負の 2 項分布），ゼロ過剰ガンマ・ポアソン分布（負の 2 項分布）を仮定する回帰分析を行い，ポアソン分布を仮定した回帰分析の場合と比較する．

Excel によるポアソン回帰

まず，表 7.17 に示した JMP でのポアソン回帰を Excel によって再現する．白人を $x_1=0$，黒人を $x_1=1$ とする（0，1）型デザイン行列とし，ポアソン回帰によるあてはめを行った結果を表 7.18 に示す．切片の推定値 $\hat{\beta}_0=0.0923$ は，表 7.15 で示した白人の平均値であり，x_1 に対する推定値 $\hat{\beta}_1=0.4298$ は，黒人の平均値 0.5220 から白人の平均値 0.0923 との差である．もちろん，推定値 $\hat{y}$ は，白人の場合は，$\hat{\beta}_0=0.0923$，黒人の場合は，

$$\hat{y}_{\text{黒人}}=\hat{\beta}_0+1\times\hat{\beta}_1=0.0923+1\times0.4298=0.5220$$

である．ポアソン確率は，$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1+x_{1,i}$ として Excel の関数を使い，

$$P_i^{Poisson}=\text{Poisson.Dist}(y_i,\hat{y}_i,false)$$

で計算されている．この確率を用いて，推定人数

$$\hat{n}_i=n_iP_i^{Poisson}$$

を計算し，（$n_i-\hat{n}_i$）により，推定人数の偏差によりあてはめの性能を可視化している．

表 7.18　ポアソン分布を仮定した回帰

				G_0	β_0^=	**0.0923**							
				G_1-G_0	β_1^=	**0.4298**						飽和モデル	
人種	i	x_0	x_1	y	n	y^	$P^{Poisson}$	$\ln L_i^{Pois.}$	n^	n - n^	χ^2	P	$\ln L_i$
白人	1	1	0	0	1070	0.0923	0.9119	-98.71	1047.74	22.26	98.71	1.0000	0.00
G_0	2	1	0	1	60	0.0923	0.0841	-148.53	96.66	-36.66	535.91	0.3679	-60.00
	3	1	0	2	14	0.0923	0.0039	-77.73	4.46	9.54	552.31	0.2707	-18.30
	4	1	0	3	4	0.0923	0.0001	-36.13	0.14	3.86	366.60	0.2240	-5.98
	5	1	0	4	0	0.0923	0.0000	0.00	0.00	0.00	0.00	0.1954	0.00
	6	1	0	5	0	0.0923	0.0000	0.00	0.00	0.00	0.00	0.1755	0.00
	7	1	0	6	1	0.0923	0.0000	-20.97	0.00	1.00	378.32	0.1606	-1.83
黒人	8	1	1	0	119	0.5220	0.5933	-62.12	94.34	24.66	62.12	1.0000	0.00
G_1	9	1	1	1	16	0.5220	0.3097	-18.75	49.25	-33.25	7.00	0.3679	-16.00
	10	1	1	2	12	0.5220	0.0808	-30.18	12.85	-0.85	50.22	0.2707	-15.68
	11	1	1	3	7	0.5220	0.0141	-29.85	2.24	4.76	82.34	0.2240	-10.47
	12	1	1	4	3	0.5220	0.0018	-18.90	0.29	2.71	69.52	0.1954	-4.90
	13	1	1	5	2	0.5220	0.0002	-17.12	0.03	1.97	76.83	0.1755	-3.48
	14	1	1	6	0	0.5220	0.0000	0.00	0.00	0.00	0.00	0.1606	0.00
				N=	1308	(-2) $\ln L^{Poisson}$ =		**1117.99**			*Pearson* χ^2=	2279.87	(-2) $\ln L$= **273.28**
				k=	2	AICc^{Poisson} =		1122.00			df=1306, p=	0.0000	AICc= 277.28
													デビアンス= **844.71**

統計的な評価としては，それぞれの対数尤度

$$\ln L_i^{Poisson}=n_i\ln P_i^{Poisson}$$

を求め，それらの合計 $\ln L$ の負の 2 倍 $(-2)\ln L$ は，

$$(-2)\ln L^{Poisson}=\sum_i \ln L_i^{Poisson}=1117.99$$

として計算され，AICc^{Poisson} は，$N=1{,}308$，$k=2$ として

$$\begin{aligned}\mathrm{AICc}^{Poisson} &= (-2)\ln L + 2k + 2k(k+1)/(N-k-1)\\ &= 1117.99 + 2\cdot 2 + 2\cdot 2\cdot(2-1)/(1308-2-1)\\ &= 1122.00\end{aligned}$$

であり，表 7.16 に示した JMP での結果に一致する．適合度統計量の Pearson のカイ 2 乗値

$$\text{Pearsonのカイ2乗}=\sum_i \chi_i^2 = \sum_i n_i \frac{(y_i-\hat{y})^2}{\hat{y}} = 2279.87$$

によって計算されている．これに対し，適合度検定のデビアンスが 844.7073 と全く異なり，過分散ではないとの判断になる．第 11.4 節で詳細に示すが，飽和モデルのマイナス 2 倍の対数尤度を計算すると 273.28 となり，完全モデルの 1117.99 との差が 844.71 と Excel での計算結果と一致する．SAS の GENMOD プロシジャを使う場合には，どちらの過分散を使うか注意が必要である．

ゼロ過剰ポアソン回帰

ゼロ過剰ポアソン回帰は, ゼロ人（$y_i=0$）場合の過剰な割合を ω とし, ゼロ人でない（$y_i\neq 0$）場合の割合 $(1-\omega)$ に対してポアソン分布を次のように

$$\begin{aligned} y_i=0 &: \quad P_i^{ZP} = \hat{\omega} + (1-\hat{\omega})\cdot \mathrm{Poisson.Dist}(y_i, \hat{y}_i, false)\\ y_i\neq 0 &: \quad P_i^{ZP} = (1-\hat{\omega})\cdot \mathrm{Poisson.Dist}(y_i, \hat{y}_i, false)\end{aligned}$$

仮定して計算する．推定したいパラメータ（$\hat{\omega}$，$\hat{\beta}_0$，$\hat{\beta}_1$）は，適当な初期値を設定し，Excel

表 7.19　Excel によるゼロ過剰ポアソン回帰

					ω^=	**0.7708**	*poisson zero*				
				G_0	β_0^=	**0.4167**	*Intercept*				
				G_1-G_0	β_1^=	**1.4101**	*x*				
人種	*i*	x_0	x_1	*y*	*n*	*y*^	P^{ZP}	$\ln L_i^{ZP}$	*n*^	*n* - *n*^	χ^2
白人	1	1	0	0	1070	0.4167	0.9219	-87.03	1059.25	10.75	445.86
G_0	2	1	0	1	60	0.4167	0.0630	-165.91	72.35	-12.35	48.99
	3	1	0	2	14	0.4167	0.0131	-60.67	15.07	-1.07	84.23
	4	1	0	3	4	0.4167	0.0018	-25.23	2.09	1.91	64.06
	5	1	0	4	0	0.4167	0.0002	0.00	0.22	-0.22	0.00
	6	1	0	5	0	0.4167	0.0000	0.00	0.02	-0.02	0.00
	7	1	0	6	1	0.4167	0.0000	-13.72	0.00	1.00	74.81
黒人	8	1	1	0	119	1.8268	0.8077	-25.42	128.42	-9.42	217.39
G_1	9	1	1	1	16	1.8268	0.0674	-43.16	10.71	5.29	5.99
	10	1	1	2	12	1.8268	0.0616	-33.45	9.79	2.21	0.20
	11	1	1	3	7	1.8268	0.0375	-22.99	5.96	1.04	5.27
	12	1	1	4	3	1.8268	0.0171	-12.20	2.72	0.28	7.76
	13	1	1	5	2	1.8268	0.0063	-10.15	0.99	1.01	11.02
	14	1	1	6	0	1.8268	0.0019	0.00	0.30	-0.30	0.00
				N=	1308		(-2) $\ln L^{ZP}$ =	**999.86**		*Pearson* χ^2=	965.58
				k=	3		AICc^{ZP} =	**1005.88**		*df*=1305, *p*=	1.0000

のソルバーにて，$(-2)\ln L$ を最小化するように（$\hat{\omega}$，$\hat{\beta}_0$，$\hat{\beta}_1$）を変化させて求める．表 7.19 に示すように，（$\hat{\omega}$=0.7708，$\hat{\beta}_0$=0.4167，$\hat{\beta}_1$=1.4101）が得られる．マイナス 2 倍の対数尤度 $(-2)\ln L$ は，

$$(-2)\ln L^{ZP} = \sum_i \ln(L_i) = 999.86$$

として計算され，AICc^{ZP} は，$N=1{,}308$，$k=3$ として

$$\begin{aligned}\mathrm{AICc}^{ZP} &= (-2)\ln L + 2k + 2k(k+1)/(N-k-1)\\ &= 1117.99 + 2\cdot 3 + 2\cdot 3\cdot(3-1)/(1308-3-1)\\ &= 1005.88\end{aligned}$$

とポアソン回帰の場合の $\mathrm{AICc}(Poisson)=1122.00$ に比べて大幅に減少している．

SAS/GENMOD によるゼロ過剰ポアソン回帰

JMP の一般化線形モデルでは，ゼロ過剰ポアソン回帰がサポートされていないので，SAS の GENMOD プロシジャにより結果の検証を行う．分布の設定は，dist=zip を使う．

```
Titel2 '<<< ゼロ過剰 Poisson >>>' ;
proc genmod data=d01 ;   /* zero Poisson */
     freq  n ;
     zeromodel  ;
     model y = x / dist=zip link=identity ;
     output out=out03 xbeta=xbeta pzero=poisson_zero ;  run ;
     proc print data=out03 ; run ;
```

表 7.20　SAS GENMOD によるゼロ過剰ポアソン分布を仮定した回帰

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept	1	0.4167	0.0730	0.2737	0.5597	32.61	<.0001
x	1	1.4101	0.2303	0.9587	1.8615	37.48	<.0001
尺度	0	1.0000	0.0000	1.0000	1.0000		
AICC（小さいほどよい）			1005.8798				

Obs	y	G	x	n	xbeta	poisson_zero
1	0	G0	0	1070	0.41669	0.77078
:						
8	0	G1	1	119	1.82678	0.77078
:						

推定値は，$\hat{\beta}_0$=0.4167，$\hat{\beta}_1$=1.4101 と Excel の結果と一致し，ゼロ過剰割合は pzero オプションの出力で *poisson_zero*=0.77078 であり，Excel の $\hat{\omega}$=0.7708 に一致する．AICc も 1005.88 と一致することが確認された．

ガンマ・ポアソン回帰（負の 2 項回帰）

負の 2 項分布は，第 6.1 節の式（6.2）で出現確率π，および，成功数kとしたときに，失敗数yの分布とし，次のようにガンマ関数を用い

$$NegBinom(y;k,\pi)=\frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\pi^k(1-\pi)^y \qquad (7.1)$$

と定義されていた．この式のパラメータを負の 2 項分布の期待値μ（位置パラメータ）および分散に関連する過分散σ（形状パラメータ）となるように変換する．成功の確率πを期待値μとkで，

$$\pi=\frac{k}{\mu+k}$$

で置き換え，

$$GammaPoisson(y;k,\mu)=\frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y \qquad (7.2)$$

さらに，kを$1/\sigma$で置き換え，整理すると

$$GammaPoisson(y;\mu,\sigma)=\frac{\Gamma(y+1/\sigma)}{\Gamma(y+1)\Gamma(1/\sigma)}\cdot\frac{(\mu\sigma)^y}{(1+\mu\sigma)^{y+1/\sigma}} \qquad (7.3)$$

が得られる．ここで，パラメータμをガンマ・ポアソン回帰の場合に，

$$\mu=\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{1,i}$$

とする．

表 7.18 に示したポアソン回帰をガンマ・ポアソン回帰となるように分布の確率計算を変更した結果を表 7.21 に示す．白人を$x_1=0$，黒人を$x_1=1$とする（0，1）型デザイン行列とし，切片の推定値$\hat{\beta}_0=0.0923$は，表 7.15 で示した白人の平均値であり，x_1に対する推定値$\hat{\beta}_1=0.4298$は，黒人の平均値 0.5220，白人の平均値 0.0923 との差である．もちろん，推定値$\hat{y}$は，白人の場合は，$\hat{\beta}_0=0.0923$，黒人の場合は，

$$\hat{y}_{黒人}=\hat{\beta}_0+1\times\hat{\beta}_1=0.0923+1\times0.4298=0.5220$$

である．ここまでは，ポアソン回帰の結果と同じである．

ガンマ・ポアソン回帰の場合のパラメータは，ポアソン回帰の場合の（β_0，β_1）に加えてσが加わる．表 7.21 には，マイナス 2 倍の対数尤度$(-2)\ln L^{GP}$が最小になるように求めた

$\hat{\sigma}=4.9429$ が結果として示されている．白人の $y_1=0$ の場合については，$\hat{y}_1=0.0923$ なので，ガンマ・ポアソン分布の確率 P_1^{GP} は，

$$GammaPoisson(y_1=0;\hat{y}_1=0.0923,\sigma=4.9429)=\frac{\Gamma(y_1+1/\sigma)}{\Gamma(y_1+1)\Gamma(1/\sigma)}\cdot\frac{(\hat{y}_1\sigma)^{y_1}}{(1+\hat{y}_1\sigma)^{y_1+1/\sigma}}$$

$$=\frac{\Gamma(0+1/4.9429)}{\Gamma(0+1)\Gamma(1/4.9429)}\cdot\frac{(0.0923\times4.9429)^0}{(1+0.0923\times4.9429)^{0+1/4.9429}}$$

$$=\frac{4.5354}{1\times4.5354}\cdot\frac{1}{1.0790}=0.9268$$

として計算されている．これら確率を用いて，推定人数

$$\hat{n}_i=n_iP_i^{GP}$$

を計算し，$n_i-\hat{n}_i$ により，推定人数の偏差によりあてはめの性能を可視化している．

表 7.21　ガンマ・ポアソン分布を仮定した回帰

					β_0^=	**0.0923**	*Intercept*				
				G_0	β_1^=	**0.4298**	*x*				
				G_1-G_0	σ^=	**4.9429**	*Dispersion*				
人種	*i*	x_0	x_1	*y*	*n*	*y*^	P^{GP}	$\ln L_i{}^{GP}$	*n*^	*n* - *n*^	χ^2
白人	1	1	0	0	1070	0.0923	0.9268	-81.33	1064.90	5.10	67.80
G_0	2	1	0	1	60	0.0923	0.0587	-170.09	67.47	-7.47	368.07
	3	1	0	2	14	0.0923	0.0111	-63.07	12.70	1.30	379.34
	4	1	0	3	4	0.0923	0.0025	-23.90	2.92	1.08	251.78
	5	1	0	4	0	0.0923	0.0006	0.00	0.73	-0.73	0.00
	6	1	0	5	0	0.0923	0.0002	0.00	0.19	-0.19	0.00
	7	1	0	6	1	0.0923	0.0000	-10.00	0.05	0.95	259.84
黒人	8	1	1	0	119	0.5220	0.7726	-30.71	122.84	-3.84	17.35
G_1	9	1	1	1	16	0.5220	0.1126	-34.94	17.91	-1.91	1.96
	10	1	1	2	12	0.5220	0.0488	-36.24	7.76	4.24	14.03
	11	1	1	3	7	0.5220	0.0258	-25.60	4.11	2.89	23.00
	12	1	1	4	3	0.5220	0.0149	-12.62	2.37	0.63	19.42
	13	1	1	5	2	0.5220	0.0090	-9.42	1.43	0.57	21.46
	14	1	1	6	0	0.5220	0.0056	0.00	0.90	-0.90	0.00
				N=	1308		(-2) $\ln L^{GP}$ =	**995.80**		*Pearson* χ^2=	**1424.03**
				k=	3		$AICc^{GP}$=	**1001.82**		*df*=1305, *p*=	0.0114

統計的な評価としては，それぞれの対数尤度

$$\ln L_i=n_i\ln(P_i^{GP})$$

を求め，それらの合計 $\ln L$ の負の 2 倍 $(-2)\ln L$ は，

$$(-2)\ln L^{GP}=\sum_i\ln(L_i)=995.80$$

として計算され，AICc^{GP} は，$N=1{,}308$，$k=3$ として

$$\begin{aligned}\mathrm{AICc}^{GP} &= (-2)\ln L + 2k + 2k(k+1)/(N-k-1)\\ &= 995.80 + 2\cdot 3 + 2\cdot 3\cdot(3-1)/(1308-3-1)\\ &= 1001.82\end{aligned}$$

となる．ポアソン回帰の場合の $\mathrm{AICc}^{Poisson}=1122.00$ に比べて大幅な減少となっている．

SAS の GENMOD プロシジャにより結果の検証を行う．分布の設定は，負の 2 項分布 negbin を使う．

```
Title2 '<<< 負の 2 項分布 ガンマ・ポアソン >>>' ;
proc genmod data=d01 ;   /* negbin */
     freq  n ;
     model y = x / dist=negbin  link=identity ;
     run ;
```

表 7.22　SAS GENMOD による負の 2 項分布（ガンマ・ポアソン分布）を仮定した回帰

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept	1	0.0923	0.0108	0.0711	0.1134	72.80	<.0001
x	1	0.4298	0.1090	0.2162	0.6433	15.56	<.0001
Dispersion	1	4.9429	1.0005	3.3242	7.3497		
AICC（小さいほどよい）			1001.8163				

推定値は，$\hat{\beta}_0$=0.0923，$\hat{\beta}_1=0.4298$ と Excel の結果と一致し，Dispersion=4.9429 は，Excel の $\hat{\sigma}=4.9429$ に一致する．AICc^{GP} も 1001.82 と一致することが確認された．

ゼロ過剰ガンマ・ポアソン回帰

ゼロ過剰ガンマ・ポアソン回帰は，ゼロ人（$y_i=0$）場合の過剰な割合を ω とし，ゼロ人でない（$y_i\neq 0$）場合の割合 $(1-\omega)$ に対してガンマ・ポアソン分布を次のように

$$y_i=0\ :\quad P_i^{ZGP}=\hat{\omega}+(1-\hat{\omega})\cdot GammaPoisson(y_i;\hat{y}_i,\hat{\sigma})$$
$$y_i\neq 0\ :\quad P_i^{ZGP}=(1-\hat{\omega})\cdot GammaPoisson(y_i;\hat{y}_i,\hat{\sigma})$$

過程して計算する．推定したいパラメータ（$\hat{\omega}$，$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\sigma}$）は，適当な初期値を設定し，Excel のソルバーにて，$(-2)\ln L^{ZGP}$ を最小化するようにパラメータを変化させて求める．表 7.23

に示すように，（$\hat{\omega}$=0.6152， $\hat{\beta}_0$=0.2424， $\hat{\beta}_1=1.0172$， $\hat{\sigma}=1.0190$）が得られ， $(-2)\ln L$ は，

$$(-2)\ln L^{ZGP}=\sum_i \ln(L_i{}^{ZGP})=994.74$$

として計算され，AICc^{ZGP} は， $N=1{,}308$， $k=4$ として

$$\begin{aligned}\text{AICc}^{ZGP}&=(-2)\ln L+2k+2k(k+1)/(N-k-1)\\&=994.74+2\cdot4+2\cdot4\cdot(4-1)/(1308-4-1)\\&=1002.77\end{aligned}$$

とガンマ・ポアソン回帰の場合の AICc^{GP} =1001.82 に比べてわずかに大きくなっている．

表 7.23　ゼロ過剰ガンマ・ポアソン分布を仮定した回帰

					ω^=	**0.6152**	*pzero*				
				G_0	β_0^=	**0.2424**	*Intercept*				
				G_1-G_0	β_1^=	**1.0172**	*x*				
					σ^=	**1.0190**	*Dispersion*				
人種	*i*	x_0	x_1	*y*	*n*	*y*^	P^{ZGP}	$\ln L_i{}^{ZGP}$	*n*^	*n* - *n*^	χ^2
白人	1	1	0	0	1070	0.2424	0.9251	-83.34	1062.91	7.09	207.96
G_0	2	1	0	1	60	0.2424	0.0602	-168.59	69.19	-9.19	113.96
	3	1	0	2	14	0.2424	0.0118	-62.14	13.58	0.42	143.11
	4	1	0	3	4	0.2424	0.0023	-24.26	2.67	1.33	100.65
	5	1	0	4	0	0.2424	0.0005	0.00	0.53	-0.53	0.00
	6	1	0	5	0	0.2424	0.0001	0.00	0.10	-0.10	0.00
	7	1	0	6	1	0.2424	0.0000	-10.93	0.02	0.98	109.69
黒人	8	1	1	0	119	1.2596	0.7864	-28.60	125.03	-6.03	65.64
G_1	9	1	1	1	16	1.2596	0.0944	-37.77	15.01	0.99	0.37
	10	1	1	2	12	1.2596	0.0526	-35.35	8.36	3.64	2.29
	11	1	1	3	7	1.2596	0.0294	-24.70	4.67	2.33	7.37
	12	1	1	4	3	1.2596	0.0164	-12.33	2.61	0.39	7.83
	13	1	1	5	2	1.2596	0.0092	-9.38	1.46	0.54	9.73
	14	1	1	6	0	1.2596	0.0052	0.00	0.82	-0.82	0.00
				N=	1308		(-2) $\ln L^{ZGP}$ =	**994.74**		*Pearson* χ^2=	**768.61**
				k=	4		AICc^{ZGP} =	**1002.77**		*df*=1304, *p*=	1.0000

SAS の GENMOD プロシジャにより結果の検証を行う．分布の設定は，ゼロ過剰負の 2 項分布 zinb オプションを使う．

```
Titel2 '<<< ゼロ過剰 負の 2 項分布 ゼロ過剰 ガンマ・ポアソン >>>' ;
proc genmod data=d01 ;   /* zero negbin */
     freq  n ;
     zeromodel  ;
     model y=x / dist=zinb link=identity ;
     output out=out04 pred=pred pzero=pzero ; run ;
     proc print data=out04 ; run ;
```

表 7.24　SAS GENMOD によるゼロ過剰負の 2 項分布を仮定した回帰

最大尤度パラメータ推定値の分析							
パラメータ	自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept	1	0.2424	0.1239	−0.0004	0.4851	3.83	0.0504
x	1	1.0172	0.4483	0.1386	1.8959	5.15	0.0233
Dispersion	1	1.0190	1.1771	0.1059	9.8044		
AICC (小さいほどよい)			1002.7735				

Obs	y	G	x	n	pred	pzero
1	0	G0	0	1070	0.24236	0.61524
:						
8	0	G1	1	119	1.25960	0.61524
:						

推定値は，$\hat{\beta}_0$=0.2424，$\hat{\beta}_1$=1.0172 と Excel の結果と一致し，dispersio=1.0190 は，表 7.23 に示した Excel の $\hat{\sigma}$=1.0190 に一致する．AICc^{ZGP} も 1002.77 と一致することが確認された．

仮定した分布間の比較

これまでに取り上げたポアソン分布（*P*），ゼロ過剰ポアソン分布（*ZP*），ガンマ・ポアソン分布（*GP*），ゼロ過剰ガンマ・ポアソン分布（*ZGP*）を仮定した回帰分析のあてはめの良さについて検討する．表 7.25 に示すように，ポアソン分布を仮定した場合の $(n_i - \hat{n}_i)$ は，$y_i = 0$ の場合にプラス 22.26 人であり，ゼロ過剰ポアソン分布を仮定した場合には，プラス 10.75 人と精度が向上し，AICc での比較でも 1122.00 から 1105.88 と大幅に減少し，あてはめ精度の向上が図られた．

ガンマ・ポアソン分布（*GP*）を仮定した場合は，$y_i = 0$ の場合の $(n_i - \hat{n}_i)$ は，プラス 5.10 人とさらに小さくなり，AICc も 1105.88 から 1001.821 と 4.06 の減少となっている．ガンマ・ポアソン分布（*GP*）を仮定した場合に比べ，ゼロ過剰ガンマ・ポアソン分布（*ZGP*）を仮定した場合には，マイナス 2 倍の対数尤度 $(-2)\ln L$ は，わずかに増えるが，パラメータ数が 4 となり，AICc は増加している．これらの結果から，ガンマ・ポアソン分布（*GP*）を仮定した回帰が尤もあてはまりがよいとの結果となる．

表 7.25　仮定した 4 分布の性能比較

			Poisson		**ゼロ過剰*Poisson***		**ガンマ*Poisson***		**ゼロ過剰*GP***	
人種	*y*	*n*	*n*^	*n* - *n*^	*n*^	*n* - *n*^	*n*^	*n* - *n*^	*n*^	*n* - *n*^
白人	0	1070	1047.74	22.26	1059.25	10.75	1064.90	5.10	1062.91	7.09
G0	1	60	96.66	-36.66	72.35	-12.35	67.47	-7.47	69.19	-9.19
	2	14	4.46	9.54	15.07	-1.07	12.70	1.30	13.58	0.42
	3	4	0.14	3.86	2.09	1.91	2.92	1.08	2.67	1.33
	4	0	0.00	0.00	0.22	-0.22	0.73	-0.73	0.53	-0.53
	5	0	0.00	0.00	0.02	-0.02	0.19	-0.19	0.10	-0.10
	6	1	0.00	1.00	0.00	1.00	0.05	0.95	0.02	0.98
黒人	0	119	94.34	24.66	128.42	-9.42	122.84	-3.84	125.03	-6.03
G1	1	16	49.25	-33.25	10.71	5.29	17.91	-1.91	15.01	0.99
	2	12	12.85	-0.85	9.79	2.21	7.76	4.24	8.36	3.64
	3	7	2.24	4.76	5.96	1.04	4.11	2.89	4.67	2.33
	4	3	0.29	2.71	2.72	0.28	2.37	0.63	2.61	0.39
	5	2	0.03	1.97	0.99	1.01	1.43	0.57	1.46	0.54
	6	0	0.00	0.00	0.30	-0.30	0.90	-0.90	0.82	-0.82
			(-2) ln *L*=	1117.99		999.86		995.798		994.743
			AICc=	**1122.00**		**1005.88**		**1001.82**		**1002.77**
	AICc順位			**4**		**3**		**1**		**2**
	パラメータ数			2		3		3		4

どのような分布を仮定してポアソン回帰をしたらよいのだろうか. なやましい問題である. 目の前にあるデータだけで決めたとすると，同様な調査を再度行った場合に，そのたびごとに分布の同定を行なうのであろうか. データには誤差的変動が付きまとっていることも考慮すると，目の前のデータだけで分布の同定は，不確性に振り回されることになる.

これまでも種々のカウント・データの例示から，調査データから得られるカウント・データは，過分散になりがちであり，単純にポアソン回帰をあてはめると *p* 値を低めに推定するバイアスが入り込むことを注意してきた. この例のようにデータ数が多ければ, *p* 値を出すまでもなく明らかな差であるような場合に，どのような分布があてはまるのかの議論は非生産的でもある.

調査データは，常に探索的解析の要素があり，この例であれは，性別・年齢階層などにより，知っている被害者の平均値がどの様に変化するのか，その変化は統計的に意味のあるものなのか，あるいは，誤差変動の範囲内なのか，その判断に際し過分散の大きさを考慮するのが現実的と思われる.

8. 2 本の回帰直線の比較

第 8 章は，2 本の回帰直線をあてはめた後の各種の推定法を扱う．第 1.8 節で細菌を用いた用量反応試験データについてポアソン回帰による効力比の解析法を示した．通常の回帰分析での効力比の解析法は，偏差平方和をベースにしており，ポアソン回帰の場合に適用できないので，パラメータの共分散行列を使うことを示してきた．通常の回帰分析での効力比の解析にも同様にパラメータの共分散行列を使って解析することができる．そこで，偏差平方和ベースの解析法に代え，パラメータの共分散行列を用いた定式化を行い，ポアソン回帰と共通な汎用的な解析法であることを示す．

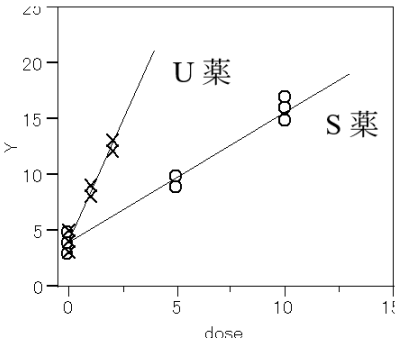
8.1. 共通の切片を持つ回帰直線の傾きの比較

複数の薬物の直線的な用量反応を比較する際に，投与量 x が 0 mg/kg での反応 Y が複数の薬物で共通と見なせる場合，切片を共通にする複数の回帰直線の同時あてはめが行なわれる．複数の直線をあてはめた後に，基準となる薬物 S との回帰直線の傾きの比（勾配比）を求め，未知の物質 U との薬効の関係を計量化する方法を取り上げる．

佐久間（1977），「薬効評価 - 計画と解析 - I 」の第 8 章の［例 6］に，細菌の増殖の事例がある．この事例は，表 8.1 に示すように 5 点法に準じて blank の $\mathcal{B}$，nicotinic acid $\mathcal{S}$ の 5，10 μ g/mL，粗物質 $\mathcal{U}$ の 1，2 mg/mL について菌の増殖を濁度で観測した．いずれも 4 回のくり返し観測を行なったデータである．なお，この事例は，佐久間著，五所・酒井・佐藤・竹内編（2017），「新版 薬効評価」の第 7 章の［例 7.5］にも再掲されている．また，高橋（2004），「各種の効力比の統計を支える非線形最小 2 乗法入門」にも取り上げられている．

表 8.1　標準薬 S 薬と未知物質 U 薬の細菌増殖［佐久間(1977)，例 6］

		繰返			
substanse	*dose*	1	2	3	4
blank	0	3	4	4	5
nicotinic acid	5	10	9	9	10
S薬	10	17	16	15	15
unknown	1	8	8	9	8
U薬	2	12	13	12	13



この実験の目的は，S 薬に対する U 薬の効力比をそれぞれの薬剤の回帰直線の傾きの比から推定し，その 95%信頼区間を求めることである．これは，「勾配比検定法」として生物検定法の代表的な方法であるが，伝統的にシグマ表記による計算で定式化されていて，ポアソン回帰の場合に応用できない．正規分布を仮定し，デザイン行列を用いた最小 2 乗法による解析方法は，最尤法によるポアソン回帰と多くの共通点を持つので，各種の推定の問題について応用できる．

デザイン行列を用いたパラメータの推定

第 3.5 節の「切片を共通とする場合」の考え方により，第 4 章で示した Excel の行列計算を用いた回帰分析の方法を拡張する．第 4 章では，説明変数が 1 変数の単回帰分析に対する方法であったが，説明変数が 2 変数となる重回帰分析に対しても計算方法は，全く同じである．このような途切れのない応用のためには，シグマによる計算に代えてデザイン行列による計算方法が必須の基礎知識である．

表 8.2 示した S 薬と U 薬の細菌増殖データのデザイン行列は，第 3.5 節の表 3.26 の形式のデザイン行列と同じ形式である．表 3.26 では，T 薬と S 薬それぞれに濃度 0 が別々に示されているが，表 8.1 では「blank」として S 薬と U 薬の共通の 0 μg/mL となっている．どちらの薬剤でもデザイン行列の設定は「切片」に 1 を設定し，S 薬の場合 x_1 は S 薬の濃度，x_2 は 0 とする．U 薬の場合は x_1 は 0，x_2 は U 薬の濃度とする．

デザイン行列を用いた「回帰パラメータの推定」および「分散分析表」については第 4.5 節に準じている．

1）表 8.1 のデータを表 8.2 に示すように行方向に展開し，$\boldsymbol{Y}$ の平均値 $\bar{y}=10.00$ を計算し，列方向に張り付ける．データ $\boldsymbol{Y}$ の自由度はデータ数の 20 である．平均の自由度は，用いているパラメータ数が 1 つなので 1 となる．

2）偏差 $(y_i-\bar{y})$ を計算し，さらに平方を SumSq()関数で足し合わせて平方和 $S_T=322.00$ を計算する．自由度は，データの自由度 20 から，平均の自由度 1 を引いて 19 となる．

3）デザイン行列 $\boldsymbol{X}$ および反応 $\boldsymbol{Y}$ から，$\boldsymbol{X}^T\boldsymbol{X}$，$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$，$\boldsymbol{X}^T\boldsymbol{Y}$，$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ を順次計算する．Excel による行列計算については第 4.1 節に詳しく説明しているが，$\boldsymbol{X}^T\boldsymbol{X}$ の場合であれば，行列の転置を行う Transpose()関数，行列の積を計算する Mmult()関数を使う．

$\boldsymbol{X}^T\boldsymbol{X}$ = Mmult(Transpose($\boldsymbol{X}$ の範囲)， $\boldsymbol{X}$ の範囲)

表 8.2　標準薬 S 薬と未知物質 U 細菌増殖の回帰分析

			デザイン行列 $\boldsymbol{X}$			$\boldsymbol{Y}$		S_T		S_R	S_e
i	*substanse*	*dose*	x_0	x_1:x_S	x_2:x_U	y	$\bar{y}$	$y-\bar{y}$	$\hat{y}$	$\hat{y}-\bar{y}$	$y-\hat{y}$
1	blank	0	1	0	0	3	10.00	-7.00	3.89	-6.11	-0.89
2			1	0	0	4	10.00	-6.00	3.89	-6.11	0.11
3			1	0	0	4	10.00	-6.00	3.89	-6.11	0.11
4			1	0	0	5	10.00	-5.00	3.89	-6.11	1.11
5	nicotinic acid	5	1	5	0	10	10.00	0.00	9.76	-0.24	0.24
6	S薬		1	5	0	9	10.00	-1.00	9.76	-0.24	-0.76
7			1	5	0	9	10.00	-1.00	9.76	-0.24	-0.76
8			1	5	0	10	10.00	0.00	9.76	-0.24	0.24
9		10	1	10	0	17	10.00	7.00	15.62	5.62	1.38
10			1	10	0	16	10.00	6.00	15.62	5.62	0.38
11			1	10	0	15	10.00	5.00	15.62	5.62	-0.62
12			1	10	0	15	10.00	5.00	15.62	5.62	-0.62
13	unknown	1	1	0	1	8	10.00	-2.00	8.21	-1.79	-0.21
14	U薬		1	0	1	8	10.00	-2.00	8.21	-1.79	-0.21
15			1	0	1	9	10.00	-1.00	8.21	-1.79	0.79
16			1	0	1	8	10.00	-2.00	8.21	-1.79	-0.21
17		2	1	0	2	12	10.00	2.00	12.52	2.52	-0.52
18			1	0	2	13	10.00	3.00	12.52	2.52	0.48
19			1	0	2	12	10.00	2.00	12.52	2.52	-0.52
20			1	0	2	13	10.00	3.00	12.52	2.52	0.48
						10.00		**322.00**		**314.1143**	**7.8857**
						平均		S_T		S_R	S_e
				自由度 *df*		20	1	20-1	3	3-1	20-3

		$\boldsymbol{X^TX}$	
x_0	**20**	60	12
x_1	60	**500**	0
x_2	12	0	**20**

		$(\boldsymbol{X^TX})^{-1}$	
x_0	**0.1786**	-0.0214	-0.1071
x_1	-0.0214	**0.0046**	0.0129
x_2	-0.1071	0.0129	**0.1143**

	$\boldsymbol{X^TY}$	$\boldsymbol{\beta}=(\boldsymbol{X^TX})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$	
x_0	200	β_0^ =	3.8929
x_1	820	β_1^ =	1.1729
x_2	133	β_2^ =	4.3143

パラメータの共分散行列

$\boldsymbol{\Sigma}(\boldsymbol{\beta}^\wedge)=(\boldsymbol{X^TX})^{-1}\sigma^{\wedge 2}$

β_0^	**0.0828**	-0.0099	-0.0497
β_1^	-0.0099	**0.0021**	0.0060
β_2^	-0.0497	0.0060	**0.0530**

Excel の行列計算式

$\boldsymbol{X^TX}$=Mmult(Transpose($\boldsymbol{X}$), $\boldsymbol{X}$)

$(\boldsymbol{X^TX})^{-1}$=Minvese($\boldsymbol{X^TX}$)

引数はセル範囲で設定する

分散分析表

要因	平方和	自由度	平均平方	
回帰 R	314.1143	2	157.0571	
誤差 e	7.8857	17	**0.4639**	: $\sigma^{\wedge 2}$
全体 T	322.0000	19		

パラメータの推定値

		係数	分散	*SE*	*t*
	β_0^	3.8929	0.0828	0.2878	13.53
S	β_1^	1.1729	0.0021	0.0460	25.47
U	β_2^	4.3143	0.0530	0.2302	18.74

注）Excel の「分析ツール」の「回帰分析」を使えば，この表に示した「分散分析表」および「パラメータの推定値」が得られる．ただし，応用の元となる「パラメータの共分散行列」を得るためには，デザイン行列を用いた計算が必要となる．

4）反応 $\boldsymbol{Y}$ の推定値を $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ により計算し，平均値からの偏差 $(\hat{y}_i - \bar{y})$ の平方を足し合わせて平方和 $S_R = 314.1143$ を計算する．推定値 $\hat{y}_i$ の自由度は，パラメターとしてデザイン行列の 3 変数を用いているので 3 となり，偏差 $(\hat{y}_i - \bar{y})$ の回帰の平方和 S_R の自由度は，平均の自由度 1 を引いて 2 となる．

5）観測値 y_i と推定値 $\hat{y}_i$ の偏差 $(y_i - \hat{y}_i)$ の平方和（誤差平方和） $S_e = 7.8857$ を計算する．自由度は， y_i の自由度 20 から $\hat{y}_i$ の自由度 3 を引いて 17 となる．

6）これらの平方和（S_T， S_R， S_e）と自由度の計算結果から，「分散分析表」を作成し，誤差分散 $\hat{\sigma}^2 = 0.4639$ を得る．

7）自由度については，平方和の計算に準じてそれぞれの自由度の差から次のように求める．

S_R は， $\hat{y}_i$ の推定に用いたデザイン行列の列数 3 と平均の 1 の差で 2,

S_e は，データ数 20 とデザイン行列の列数 3 の差で 17,

S_T は，データ数 20 と平均の 1 との差で 19 となる．

8）パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は， $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ に誤差分散 $\hat{\sigma}^2 = 0.4639$ を掛け， $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$ で求められる．

9）パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の対角要素が，（$\hat{\beta}_0$， $\hat{\beta}_1$， $\hat{\beta}_2$）の分散であるので，それらの平方根をとりパラメータの *SE* とし, *t* 値を計算し「パラメータの推定値」を完成させる．

注）手順 8）を除いて他の手順は，Excel の「回帰分析」で代替できる．

傾きの差の 95%信頼区間

この実験で求めたいのは 2 本の回帰直線の傾きの比であるが，まず傾きの差の 95%信頼区間の推定方法について示す．推定されたパラメータから S 薬とU薬の回帰直線は，

$$\text{S 薬：}\quad \hat{y}_{S,i} = \hat{\beta}_0 + \hat{\beta}_1 x_{S,i} = 3.8929 + 1.1729 x_{S,i}$$

$$\text{U 薬：}\quad \hat{y}_{U,i} = \hat{\beta}_0 + \hat{\beta}_2 x_{U,i} = 3.8929 + 4.3143 x_{U,i}$$

であり，その差は,

$$(\hat{\beta}_2 - \hat{\beta}_1) = 4.3143 - 1.1729 = 3.1414$$

となる．その 95%信頼区間は，分散 $Var(\hat{\beta}_2 - \hat{\beta}_1)$ をパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いて

$$\begin{aligned} Var(\hat{\beta}_2 - \hat{\beta}_1) &= Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_2, \hat{\beta}_1) + Var(\hat{\beta}_1) \\ &= 0.0530 - 2 \times 0.0060 + 0.0021 \\ &= 0.0432 \end{aligned}$$

として求め,

$$
\begin{aligned}
(L95\%, U95\%) &= (\hat{\beta}_2 - \hat{\beta}_1) \pm t_{0.05}(17)\sqrt{Var(\hat{\beta}_2 - \hat{\beta}_1)} \\
&= 3.1414 \pm 2.1098 \times \sqrt{0.0432} \\
&= (2.7029, 3.5800)
\end{aligned}
$$

と推定される．傾きの差で結果を評価することも可能であるが，得られた推定値に対して「用量が 1 mg/ml における濁度の差」というような回りくどい説明となってしまう．

傾きの比

効力比 ρ は，2 種類の化学物質の効果を比較する際に，基準とする S 薬のある濃度 x_S での反応と同じ反応を得るために必要となる U 薬の濃度 x_U とするために係数（効力比）として，次のように

$$
x_U = \rho x_S
$$

定義される．S 薬の濁度が $y^{(10)} = 10$ となる濃度は，

$$
y^{(10)} = \hat{\beta}_0 + \hat{\beta}_1 x_S
$$

$$
\hat{x}_S = \frac{y^{(10)} - \hat{\beta}_0}{\hat{\beta}_1} = \frac{10 - 3.8929}{1.1729} = 5.2071
$$

であり，U 薬の場合は，

$$
\hat{x}_U = \frac{y^{(10)} - \hat{\beta}_0}{\hat{\beta}_2} = \frac{10 - 3.8929}{4.3143} = 1.4156
$$

従って，効力比 $\hat{\rho}$ は，

$$
\hat{\rho} = \frac{\hat{x}_U}{\hat{x}_S} = \frac{(10 - \hat{\beta}_0)/\hat{\beta}_1}{(10 - \hat{\beta}_0)/\hat{\beta}_2} = \frac{\hat{\beta}_2}{\hat{\beta}_1} = \frac{4.3143}{1.1729} = 3.6784
$$

となり，傾きの比に帰着する．切片が共通で，傾きだけが異なる場合，効力比 $\hat{\rho}$ は，濃度 x に依存せず，傾きの比となることから，「勾配比検定」として知られている．ここで定式化したのは，誤差分布に正規分布を仮定した最小 2 乗法による回帰分析による場合であるが，誤差にポアソン分布を仮定した最尤法によるポアソン回帰でも手順は，全く同じである．

効力比の近似の 95%信頼区間

効力比の近似の 95%信頼区間は，第 1.8 節で示したと同様のデルタ法の手順で求めることができる．まず，効力比 ρ

$$
\hat{\rho} = \frac{\hat{\beta}_2}{\hat{\beta}_1} \ : \ \frac{\text{U薬の傾き}}{\text{S薬の傾き}}
$$

を $\hat{\beta}_0$， $\hat{\beta}_1$，および $\hat{\beta}_2$ で偏微分すると

$$
d_0 = \frac{\partial \hat{\rho}}{\partial \hat{\beta}_0} = 0
$$

$$d_1 = \frac{\partial \hat{\rho}}{\partial \hat{\beta}_1} = \frac{-\hat{\beta}_2}{\hat{\beta}_1^2} = \frac{-4.3143}{1.1729^2} = -3.1363$$

$$d_2 = \frac{\partial \hat{\rho}}{\partial \hat{\beta}_2} = \frac{1}{\hat{\beta}_1} = 0.8526$$

が得られる．これらを行ベクトル $\boldsymbol{d}$

$$\boldsymbol{d} = [d_0 \quad d_1 \quad d_2] = [0 \quad -3.1363 \quad 0.8526]$$

としてまとめる．パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を挟み込むデルタ法にて効力比 $\hat{\rho}$ の分散

$$Var(\hat{\rho}) = \boldsymbol{d}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{d}^T$$

	$\boldsymbol{d}$								
	d_0	d_1	d_2	共分散 $\boldsymbol{\Sigma}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{2\wedge}$			$\boldsymbol{d}^T$		$Var(\rho\wedge)$
=	0.0000	-3.1363	0.8526	**0.0828**	-0.0099	-0.0497	0.0000	=	0.0275
				-0.0099	**0.0021**	0.0060	-3.1363		
				-0.0497	0.0060	**0.0530**	0.8526		
				$\beta_0\wedge$	$\beta_1\wedge$	$\beta_2\wedge$			

を求める．得られた分散 $Var(\hat{\rho})$ を用い 95%信頼区間を

$$\begin{aligned}(L95\%, U95\%) &= \hat{\rho} \pm t(0.05,17)\sqrt{Var(\hat{\rho})} \\ &= 3.6784 \pm 2.1098 \times \sqrt{0.0275} \\ &= (3.3286,\ 4.0283)\end{aligned}$$

として求める．信頼区間が 1 を含んでいないので，統計的には有意であると言える．

2 本の回帰直線の傾きの比の 95%信頼区間を求めるために，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いたデルタ法で求めた．この方法は，第 8.3 節でも示すように最尤法によるポアソン回帰による比の 95%信頼区間の推定にも適用できる汎用的な方法である．

ソルバーを用いた正確な 95%信頼区間

効力比の正確な 95%信頼区間の算出は，伝統的には定式化されてはいるが複雑で難解であり，避けて通りたい．これは，元になる計算原理から計算公式の導出が技巧的であり，説明を簡略化するために，「この式で与えられる」との紋切り型の記述となっていることも起因する．そこで，パラメータ共分散行列を用いた 2 次方程式の解を Excel のソルバーで解いて，正確な勾配比の 95%信頼区間を算出する方法を示すことにより，見通しを良くしたい．

推定された傾き $\hat{\beta}_1$ および $\hat{\beta}_2$ の分散を $Var(\hat{\beta}_1)$ と $Var(\hat{\beta}_2)$，共分散を $Cov(\hat{\beta}_1, \hat{\beta}_2)$ とする．勾配比の定義式から，次の関数 $\psi = \rho\beta_1 - \beta_2$ を考える． $\hat{\beta}_1$ と $\hat{\beta}_2$ が不偏推定量なので， ψ の期待値は，

$$E(\psi) = \rho\hat{\beta}_1 - \hat{\beta}_2 = 0$$

となる．ψ の分散は，パラメータの共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ とし，$E(\psi)$ について $\hat{\boldsymbol{\beta}}$ の係数は．$\boldsymbol{g}=\begin{bmatrix}0 & \rho & -1\end{bmatrix}$ となるので，$Var(\hat{\psi})$ は

$$\begin{aligned} Var(\hat{\psi}) &= \boldsymbol{g}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{g}^T \\ &= \rho^2 Var(\hat{\beta}_1) - 2\rho Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2) \end{aligned}$$

のように ρ に関して 2 次式となる．そのとき，$\hat{\beta}_1$ と $\hat{\beta}_2$ は，正規分布に従うと仮定されるので，ψ は，同様に正規分布に従い

$$z_\alpha = \frac{\rho\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{Var(\hat{\psi})}}$$

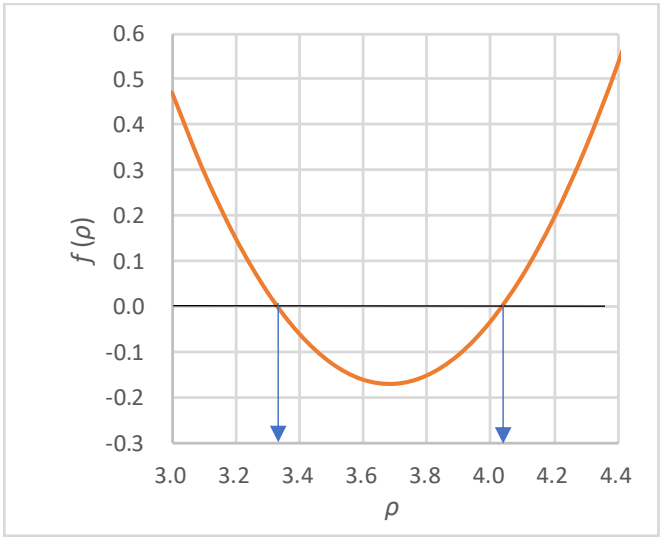
は，標準正規分布に従う．従って，z_α を自由度 df の t 分布の両側 α 点であるとしたときに，ρ の $100\cdot(1-\alpha)\%$ 信頼区間は，不等式

$$\left|\rho\hat{\beta}_1 - \hat{\beta}_2\right| \le z_\alpha\sqrt{Var(\hat{\psi})}$$

で与えられる．ここで ρ を変化させ等式が成り立つ場合の ρ が，推定された効力比 $\hat{\rho}$ の正確な 95%信頼区間となる．そこで，両辺を 2 乗し右辺を移項して等式とすると，

$$\begin{aligned} f(\rho) &= (\rho\hat{\beta}_1 - \hat{\beta}_2)^2 - z_\alpha^2 Var(\hat{\psi}) \\ &= (\rho\hat{\beta}_1 - \hat{\beta}_2)^2 - z_\alpha^2[\rho^2 Var(\hat{\beta}_1) - 2\rho Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2)] \end{aligned}$$

となる．この関数 $f(\rho)$ は，2 つの 2 次式を複合したもので，全体としても ρ に関して 2 次式である．図 8.1 に ρ を 3.0～5.0 と変化させ，$z_\alpha = t(0.05,17)$ とし，$f(\rho)$ の計算結果を図示する．図に示すように $\hat{\rho}$ の解は 2 つあり効力比の正確な 95%信頼区間の下限と上限となる．



ρ	$(\rho\hat{\beta}_1-\hat{\beta}_2)^2$	$z_\alpha^2 Var(\psi)$	$f(\rho)$
3.0	0.6332	0.1616	0.4715
3.1	0.4603	0.1621	0.2982
3.2	0.3149	0.1627	0.1522
3.3	0.1970	0.1636	0.0335
3.3333	0.1639	0.1639	**0.0000**
3.4	0.1066	0.1646	-0.0579
3.6	0.0085	0.1672	-0.1587
3.8	0.0203	0.1705	-0.1502
4.0	0.1422	0.1746	-0.0324
4.0356	0.1754	0.1754	**0.0000**
4.10	0.2445	0.1770	0.0675
4.2	0.3742	0.1795	0.1947
4.4	0.7162	0.1851	0.5311
4.6	1.1683	0.1915	0.9768

図 8.1 勾配比 ρ を変化させたときの $f(\rho)$ の 2 次曲線から求めた正確な 95%信頼区間

Excel のソルバーを使って効力比の推定値 $\hat{\rho}=3.6784$ よりも小さい方へ $f(\rho)$ がマイナスからプラスに変化する $f(\rho:3.3)=0.0335$ に対し，ソルバーで $f(\rho)=0$ となるように $\rho=3.3$ を変化させて 95%信頼区間の下限 $\hat{\rho}_{L95\%}=3.3333$ を得ることができる．推定値 $\hat{\rho}=3.6784$ よりも大

きい方へく $f(\rho)$ がマイナスからプラスに変化する $\rho=4.1$ に対し $f(\rho)=0.06750$ が 0 となるように $\rho=4.1$ をソルバーで変化させれば，95%信頼区間の上限 $\rho_{U95\%}=4.0356$ も得ることができる．ただし，z_α^2 は，自由度 17 の t 分布の 5%点 $t_\alpha^2=2.1098$ としている．この結果から勾配比の正確な 95%信頼区間は，

$$(L95\%, U95\%)=(3.3333,\ 4.0356)$$

として得られる．

2 次式の解を用いた正確な 95%信頼区間

伝統的には，$f(\rho)$ に含まれる 2 つの 2 次式を ρ について整理し，分散，共分散を偏差平方和に落とし込み，2 次式の解の公式で解いた複雑な計算式が天下り的に示されており，理解の妨げになってきた．どのようなものかは，杉本 ()，「統計学入門，13.3 節　勾配比検定法」に詳しく述べられている．同じデータが用いられているので，計算結果は一致している．

$f(\rho)$ の分散および共分散をそのまま用いて，2 次式の解の公式による効力比の正確な 95%信頼区間を求める方法については，第 8.3 節に示してあるので，ここでは Excel での計算結果のみを示す．

2次式の解の公式を用いた正確な95%信頼区間の計算								
		共分散 $\boldsymbol{\Sigma}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$			$t_\alpha=$	2.1098	$df=$	17
	β_0^ 3.8929	**0.0828**	-0.0099	-0.0497	$a=$	18.3771	$\rho=$	**3.6784**
S	β_1^ 1.1729	-0.0099	**0.0021**	0.0060	$b=$	-10.0670	L95%=	**3.3333**
U	β_2^ 4.3143	-0.0497	0.0060	**0.0530**	$c=$	1.3662	U95%=	**4.0356**
		β_0^	β_1^	β_2^				

非線形回帰による効力比の 95%信頼区間の推定

ポアソン回帰への応用のために Excel の行列計算による効力比の 95%信頼区間の推定方法を示したのであるが，JMP の「非線形回帰のあてはめ」を用いれば，効力比の正確な 95%信頼区間の推定が表 8.3 に示すように直接求めることができる［高橋（2004）］．

表 8.3　JMP の「非線形回帰のあてはめ」を用いた効力比の 95%信頼区間の直接推定

解

SSE	DFE	MSE	RMSE
7.8857	17	0.4639	0.6811

パラメータ	推定値	近似標準誤差	下側信頼限界	上側信頼限界
β0	3.8929	0.2878	3.2856	4.5001
β1	1.1729	0.0460	1.0757	1.2700
ρ	3.6784	0.1658	3.3333	4.0356

解法: 解析 Gauss-Newton

8.2. 切片は異なるが共通の傾きを持つ 2 本の回帰直線

薬物の薬理作用は，対数用量に対してシグモイド曲線状になることが経験的に知られている．シグモイド曲線の中ほどは，直線的であることを利用して，複数の薬物の効力を比較する方法が平行線検定法として定式化されている．効力比は，同じ反応が得られる用量の比として定義されているので，基準となる薬物 S と未知の物質 U の対数用量について平行な直線をあてはめ，ある反応 y_0 となる S 薬の常用対数用量 $\hat{x}_S$ と U 薬の常用対数用量 $\hat{x}_U$ を推定し，

$$(10\wedge \hat{x}_{S,y_0}) = \rho(10\wedge \hat{x}_{U,y_0})$$

となる ρ が効力比である．これは，第 8.1 節の勾配比の場合と同じ考え方である．

平行な直線をあてはめる方法は，共分散分析として定式化されている．この方法は，2 群の反応の比較をする実験で，測定したい反応に明らかに影響することが分かっているが，実験をする際に実験者が制御できないような変数を共変量とする方法である．ただし，効力比 ρ を求めるため，S 薬，U 薬それぞれに対する用量は，実験者が能動的に設定するので「共変量」ではなく「制御因子」である．とはいえ，統計的には共分散分析と全く同じである．異なるのは，統計量として効力比 ρ とその 95%信頼区間の推定を主たる目的していることである．

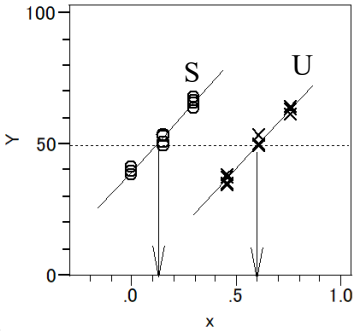
平行線検定法

佐久間（1977）の第 8 章の［例 1］に，モルモット回腸の収縮試験の例がある．Superfusion（還流）法で，S を histamine の 1，$\sqrt{2}$，$2\,\mu$g/L，U を histamine 様物質の $2\sqrt{2}$，4，$4\sqrt{2}$ μg/L として，乱塊法の割りつけに従い，モルモット回腸についての収縮を観測した．いずれも抗 histamine 剤の mepyramine で桔抗される．用量 x はメタメーターで $\log_{10}$ 濃度である．なお，この事例は，佐久間ら（2017）の 7 章の［例 7.1］にも再掲されている．表 8.4 にデータおよび平行線のあてはめを示す．

表 8.4　ヒスタミンとヒスタミン様物質に対するモルモット回腸の収縮量（単位 mm）

		$\log_{10}$ *dose*	繰返し			
substanse	*dose*	x	1	2	3	4
histamine	1.00	0.0000	42	40	39	40
S薬	1.41	0.1492	51	53	50	54
	2.00	0.3010	67	68	66	64
U histamine	2.83	0.4518	37	38	35	34
U薬	4.00	0.6021	49	50	49	53
	5.66	0.7528	63	61	64	63

［佐久間（1977），例 1］



この実験は，平行線検定法として知られている生物検定法の代表的な方法であるが，シグマ表記による偏差平方和に基づく計算法で定式化されていて，ポアソン回帰の場合に応用できない．デザイン行列を用いた共分散分析は，効力比 ρ の 95%信頼区間が容易に計算でき，ポアソン回帰の場合にも応用できる．効力比の推定は，先人たちによって定式化された回帰パラメータの分散・共分散を用いた応用問題である．ここでは，ポアソン回帰による効力比の前振りなので，詳しくは，Web 上の公開資料，橘田・福島（2013），「効力比の推定」を参照のこと．平行線検定法の応用例については，原田（2017），「平行線検定を利用した薬物の効力比較」が詳しい．なお，高橋（2004）では，非線形回帰 SAS の NLIN プロシジャを用いて，平行線検定法による効力比 ρ とその 95%信頼区間を直接推定する方法が示されている．

デザイン行列は，第 3.5 節の表 3.28 と同様の（1，1） 型であり，表 8.5 に示すように，共通の傾きを持つが，S 薬と U 薬に対して別々の切片が推定できるようなデザイン行列となっている．

$$y_i = \beta_{0S} x_{0S,i} + \beta_{0U} x_{0U,i} + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \text{正規分布}$$

デザイン行列は勾配比の推定のための表 8.2 とは異なるが，計算手順は同様である．

1）表 8.4 のデータを表 8.5 に示すように行方向に展開する．Excel の分析ツールの回帰分析により分散分析表と回帰パラメータの推定を行う．誤差平方和 $S_e = 55.1482$，誤差分散 $\hat{\sigma}^2 = 2.6261$，回帰パラメータ $(\hat{\beta}_{0S},\ \hat{\beta}_{0U},\ \hat{\beta}_1) = (39.6823,\ -3.1030,\ 87.6251)$ が得られる．

2）デザイン行列を $\boldsymbol{X}$ とし，$\boldsymbol{X}^T\boldsymbol{X}$，$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$ を計算する．

3）パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の対角要素の平方根を SE として計算し，Excel の標準誤差と一致することを確認する．

Y 軸方向の差

得られた回帰式は，

$$\text{S 薬：}\quad \hat{y}_{S,i} = 39.6823 + 87.6251 x_i$$
$$\text{U 薬：}\quad \hat{y}_{U,i} = -3.1030 + 87.6251 x_i$$

である．S 薬薬に対する U 薬の効果を推定するために，Y 軸方向の 2 本の回帰直線の差は，切片の差

$$\begin{aligned}(\hat{y}_U - \hat{y}_S) &= (\hat{\beta}_{0U} - \hat{\beta}_{0S}) \\ &= -3.1030 - 39.6823 \\ &= -42.7853\end{aligned}$$

となる．差の分散は，

表 8.5　モルモット回腸の収縮量を用いた平行線のあてはめ

	薬		デザイン行列 $\boldsymbol{X}$			$\boldsymbol{Y}$
i	剤	$dose$	x_{0S}	x_{0U}	x	y
1	S薬	1.00	1	0	0.0000	42
2			1	0	0.0000	40
3			1	0	0.0000	39
4			1	0	0.0000	40
5		1.41	1	0	0.1492	51
6			1	0	0.1492	53
7			1	0	0.1492	50
8			1	0	0.1492	54
9		2.00	1	0	0.3010	67
10			1	0	0.3010	68
11			1	0	0.3010	66
12			1	0	0.3010	64
13	U薬	2.83	0	1	0.4518	37
14			0	1	0.4518	38
15			0	1	0.4518	35
16			0	1	0.4518	34
17		4.00	0	1	0.6021	49
18			0	1	0.6021	50
19			0	1	0.6021	49
20			0	1	0.6021	53
21		5.66	0	1	0.7528	63
22			0	1	0.7528	61
23			0	1	0.7528	64
24			0	1	0.7528	63
					$x=\log_{10}(dose)$	

分散分析表（分析ツール：回帰分析，定数0をオン）

	自由度	変動	分散	分散比
回帰	3	65880.9	21960.3	8362.3
残差	21	55.1482	**2.6261**	
合計	24	65936.0		

	係数	標準誤差	t	P-値
切片	0	#N/A	#N/A	#N/A
x0S	**39.6823**	0.6181	64.2019	0.0000
x0U	**-3.1030**	1.6871	-1.8392	0.0801
x	**87.6251**	2.6916	32.5548	0.0000

$\boldsymbol{X}^T\boldsymbol{X}$=Mmult(Transpose($\boldsymbol{X}$), $\boldsymbol{X}$)

x_{0S}	**12.0000**	0.0000	1.8010	$z_\alpha : t_\alpha$= **2.0796**
x_{0U}	0.0000	**12.0000**	7.2267	
x	1.8010	7.2267	**4.9848**	

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$=Minvese($\boldsymbol{X}^T\boldsymbol{X}$)

x_{0S}	**0.1455**	0.2493	-0.4140	
x_{0U}	0.2493	**1.0839**	-1.6614	
x	-0.4140	-1.6614	**2.7588**	

パラメータの共分散 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$

					SE
x_{0S}	**0.3820**	0.6548	-1.0873	$\hat{\beta}_{0S}$	**0.6181**
x_{0U}	0.6548	**2.8463**	-4.3630	$\hat{\beta}_{0U}$	**1.6871**
x	-1.0873	-4.3630	**7.2448**	$\hat{\beta}_1$	**2.6916**
	$\hat{\beta}_{0S}$	$\hat{\beta}_{0U}$	$\hat{\beta}_1$		

$$
\begin{aligned}
Var(\hat{y}_U - \hat{y}_S) &= Var(\hat{\beta}_{0U} - \hat{\beta}_{0S}) \\
&= Var(\hat{\beta}_{0U}) - 2Cov(\hat{\beta}_{0U}, \hat{\beta}_{0S}) + Var(\hat{\beta}_{0S}) \\
&= 0.3820 - 2 \times 0.6548 + 2.8463 \\
&= 1.9187
\end{aligned}
$$

となり，差の 95%信頼区間は，

$$
\begin{aligned}
(L95\%, U95\%) &= (\hat{\beta}_{0U} - \hat{\beta}_{0S}) \pm t_{0.05}(21)\sqrt{Var(\hat{\beta}_{0U} - \hat{\beta}_{0S})} \\
&= -42.7853 \pm 2.0796 \times \sqrt{1.9187} \\
&= (-45.6659, -39.9047)
\end{aligned}
$$

と計算できる．ただし，S 薬に対する U 薬の効果の比較をする際に，「同じ用量を投与したときに，作用が U 薬は，S 薬に対して –42.7853 mm である」などの表現になり，冗長で歯切れが悪い．

効力比

効力比は，2 種類の化学物質の効果を比較する際に，基準とする化合物 S のある用量 $(10\hat{}\,x_S)$ での反応と同じ反応を得るために必要となる化合物 U の用量 $(10\hat{}\,x_U)$ としたときの両者の効

力比をρとしたときに

$$(10\wedge x_S)=\rho\cdot(10\wedge x_U)$$

で定義される．標準薬 S の濁度が $y_0=50$ となる対数用量は,

$$50=39.6823+87.6251x_S$$

$$\hat{x}_S=\frac{50-39.6823}{87.6251}=0.1177,\quad 10\wedge 0.1177=1.3114\,\mu g/L$$

であり，化合物 U の場合は,

$$\hat{x}_U=\frac{50-(-3.1030)}{87.6251}=0.6060,\quad 10\wedge 0.6060=4.0367\,\mu g/L$$

従って，効力比 $\hat{\rho}$は,

$$\hat{\rho}=\frac{10\wedge\hat{x}_S}{10\wedge\hat{x}_U}=\frac{1.3114}{4.0367}=0.3249$$

となる. 収縮量が $y_0=50$ として効力比を求めたが, y_0のままで, 式を整理すると y_0が消えて, 次式が得られる.

$$\begin{aligned}\hat{\rho}&=\frac{10\wedge[(y_0-\hat{\beta}_{0S})/\hat{\beta}_1]}{10\wedge[(y_0-\hat{\beta}_{0U})/\hat{\beta}_1]}\\&=10\wedge\left(\frac{-\hat{\beta}_{0S}+\hat{\beta}_{0U}}{\hat{\beta}_1}\right)\\&=10\wedge\left(\frac{-39.6823-3.1030}{87.6251}\right)\\&=10\wedge(-0.4883)=0.3249\end{aligned}$$

効力比が $\hat{\rho}=0.3249$ と 1 よりも小さいので，未知のヒスタミン U 薬は標準のヒスタミン S 薬に比べて効力が弱いことになる.

効力比の近似の 95%信頼区間

効力比 ρの 95%信頼区間は，対数用量での効力比 ρ' の 95%信頼区間として求め 10^（べき乗）で元の用量に戻す．対数用量での効力比 ρ' は,

$$\log_{10}\hat{\rho}=\hat{\rho}'=\frac{-\beta_{0S}+\beta_{0U}}{\beta_1}$$

なので対数効力比 ρ' を $\hat{\beta}_{0S}$， $\hat{\beta}_{0U}$ および $\hat{\beta}_1$ で偏微分すると

$$d_{0S}=\frac{\partial\hat{\rho}'}{\partial\hat{\beta}_{0S}}=\frac{-1}{\hat{\beta}_1}=-0.0114$$

$$d_{0U}=\frac{\partial\hat{\rho}'}{\partial\hat{\beta}_{0U}}=\frac{1}{\hat{\beta}_1}=0.0114$$

$$d_1 = \frac{\partial \hat{\rho}'}{\partial \hat{\beta}_1} = \frac{\hat{\beta}_{0S} - \hat{\beta}_{0U}}{\hat{\beta}_1^{\,2}} = 0.0056$$

が得られる．それらを行ベクトル $\boldsymbol{d}$

$$\boldsymbol{d} = [d_{0S} \quad d_{0U} \quad d_1] = [-0.0114 \quad 0.0114 \quad 0.0056]$$

として，2 次形式のデルタ法にて効力比 $\hat{\rho}$ の分散

$$Var(\hat{\rho}) = \boldsymbol{d}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{d}^T$$

		$\boldsymbol{d}$							
	d_{0S}	d_{0U}	d_1	$\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{2\hat{}}$			$\boldsymbol{d}^T$		$Var(\rho')$
=	-0.0114	0.0114	0.0056	**0.3820**	0.6548	-1.0873	-0.0114	=	5.8236E-05
				0.6548	**2.8463**	-4.3630	0.0114		
				-1.0873	-4.3630	**7.2448**	0.0056		
				$\beta_{0S}\hat{}$	$\beta_{0U}\hat{}$	$\beta_1\hat{}$			

を求める．対数 95%信頼区間は，

$$\begin{aligned}\log_{10}(L95\%, U95\%) &= \hat{\rho}' \pm t(0.05, 20)\sqrt{Var(\hat{\rho}')} \\ &= -0.4883 \pm 2.0796 \times \sqrt{5.8236 \times 10^{-5}} \\ &= (-0.5041, -0.4724)\end{aligned}$$

として求める．元の用量に戻すと

$$\hat{\rho} = 10\,\hat{}\,(-0.4883) = 0.3249$$
$$L95\% = 10\,\hat{}\,(-0.5041) = 0.3132$$
$$U95\% = 10\,\hat{}\,(-0.4724) = 0.3370$$

が得られる．

ソルバーを用いた正確な 95%信頼区間

平行線検定法の効力比の正確な信頼区間の算出は，定式化されてはいるが複雑で難解であり，避けて通りたくなる．そこで，勾配比検定の場合と同様に 2 次方程式の解を Excel のソルバーで解いて，正確な勾配比の 95%信頼区間を算出する方法を示す．

平行線検定法での効力比の正確な 95%信頼区間を求めたい．対数効力比は，次式で与えられているので，

$$\rho' = \frac{-\beta_{0S} + \beta_{0U}}{\beta_1} = \frac{-39.6823 - 3.1030}{87.6251} = -0.4883$$

式を変形し，次の関数 $\psi = \beta_{0S} - \beta_{0U} + \rho'\beta_1$ を考える．推定されたパラメータ $\hat{\beta}_{0U}, \hat{\beta}_{0S}, \hat{\beta}_1$ が $\beta_{0U}, \beta_{0S}, \beta_1$ の不偏推定量となるので，ψ の期待値は，

$$E(\psi) = \hat{\beta}_{0S} - \hat{\beta}_{0U} + \rho'\hat{\beta}_1 = 0$$

となる．また，ψ の分散は，$\hat{\beta}_{0S}, \hat{\beta}_{0U}, \hat{\beta}_1$ の係数を $\boldsymbol{g} = [1 \quad -1 \quad \rho']$，$\hat{\beta}_{0U}, \hat{\beta}_{0S}, \hat{\beta}_1$ の共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ としたときに $E(\psi)$ は，$E(\psi) = \boldsymbol{g}\hat{\boldsymbol{\beta}}$ と表すことができ，分散 $Var(\hat{\psi})$ は

$$Var(\hat{\psi}) = \boldsymbol{g}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{g}^T$$
$$= Var(\hat{\beta}_{0S}) + Var(\hat{\beta}_{0U}) + (\rho')^2 Var(\hat{\beta}_1)$$
$$-2Cov(\hat{\beta}_{0S}, \hat{\beta}_{0U}) + 2\rho' Cov(\hat{\beta}_{0S,}\, \hat{\beta}_1) - 2\rho' Cov(\hat{\beta}_{0U}, \hat{\beta}_1)$$

					$\boldsymbol{\Sigma}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{2\wedge}$			
$\boldsymbol{\Sigma}(\boldsymbol{\beta}\wedge)=$	$\boldsymbol{Var}(\beta_{0S}\wedge)$	$Cov(\beta_{0S}\wedge, \beta_{0U}\wedge)$	$Cov(\beta_{0S}\wedge, \beta_1\wedge)$	=	0.3820	0.6548	-1.0873	$\beta_{0S}\wedge$
	$Cov(\beta_{0U}\wedge, \beta_{0S}\wedge)$	$\boldsymbol{Var}(\beta_{0U}\wedge)$	$Cov(\beta_{0U}\wedge, \beta_1\wedge)$		0.6548	2.8463	-4.3630	$\beta_{0U}\wedge$
	$Cov(\beta_1\wedge, \beta_{0S}\wedge)$	$Cov(\beta_1\wedge, \beta_{0U}\wedge)$	$\boldsymbol{Var}(\beta_1\wedge)$		-1.0873	-4.3630	7.2448	$\beta_1\wedge$
					$\beta_{0S}\wedge$	$\beta_{0U}\wedge$	$\beta_1\wedge$	

で与えられる．このとき，$\hat{\beta}_{0U}, \hat{\beta}_{0S}, \hat{\beta}_1$ は，正規分布に従うと仮定されるので，ψ は，同様に正規分布に従い，

$$z = \frac{\hat{\psi}}{\sqrt{Var(\hat{\psi})}} = \frac{\hat{\beta}_{0S} - \hat{\beta}_{0U} + \rho'\hat{\beta}_1}{\sqrt{Var(\hat{\psi})}}$$

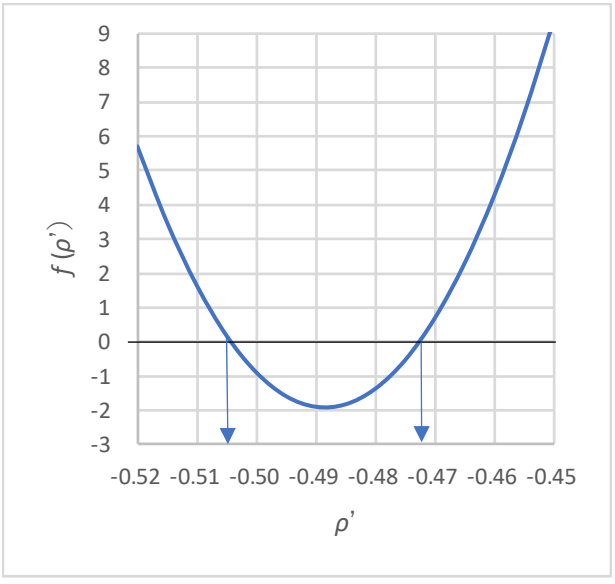
は，標準正規分布に従う．従って，z_α を標準正規分布の両側 α 点としたときに，ρ' の $100\cdot(1-\alpha)\%$ 信頼区間は，

$$\left|\hat{\beta}_{0S} - \hat{\beta}_{0U} + \rho'\hat{\beta}_1\right| \le z_\alpha Var(\hat{\psi})$$

で与えられる．ここで ρ' を変化させ等式が成り立つ場合の ρ' が効力比 $\hat{\rho}'$ の正確な 95%信頼区間となる．そこで，両辺を 2 乗して右辺を移項して等式

$$f(\rho') = (\hat{\beta}_{0S} - \hat{\beta}_{0U} + \rho'\hat{\beta}_1)^2 - z_\alpha^2 Var(\hat{\psi})$$

とする．この関数 $f(\rho')$ は，ρ' に関する 2 つの 2 次式の複合式であり，$\hat{\rho}'$ の解は 2 つあり効力比の 95%信頼区間の下限と上限ととなる．



g_1	g_2	g_3: ρ'	$(\boldsymbol{g}\boldsymbol{\beta}\wedge)^2$	$z_\alpha{}^2 Var(\psi\wedge)$	$f(\rho')$
1	-1	-0.5200	7.7270	2.0372	5.6899
1	-1	-0.5150	5.4832	2.0167	3.4665
1	-1	-0.5100	3.6233	1.9978	1.6255
1	-1	-0.5050	2.1473	1.9804	0.1669
1	-1	**-0.5043**	1.9782	1.9782	**0.0000**
1	-1	-0.5000	1.0552	1.9647	-0.9094
1	-1	-0.4950	0.3471	1.9505	-1.6034
1	-1	-0.4900	0.0228	1.9378	-1.9150
1	-1	-0.4850	0.0824	1.9267	-1.8443
1	-1	-0.4800	0.5260	1.9172	-1.3912
1	-1	-0.4750	1.3534	1.9093	-0.5558
1	-1	**-0.4725**	1.9059	1.9059	**0.0000**
1	-1	-0.4700	2.5648	1.9029	0.6619
1	-1	-0.4650	4.1601	1.8981	2.2620
1	-1	-0.4600	6.1392	1.8948	4.2444
1	-1	-0.4550	8.5023	1.8931	6.6092
1	-1	-0.4500	11.2493	1.8930	9.3563

図 8.2 効力比 ρ' を変化させた場合の $f(\rho')$ の 2 次曲線と正確な 95%信頼区間の推定

Excel のソルバーを使って推定値 $f(\rho')$ が 0 をまたぐ近傍の $\rho' = -0.5050$ に着目し，セルの $f(\rho') = 0.1669$ が指定値 0 になるように $\rho' = -0.5050$ を変化させれば 95%信頼区間の下限 $\hat{\rho}'_{L95\%} = -0.5043$ を得ることができる．次に推定値 $f(\rho')$ が 0 をまたぐ近傍の $\rho' = -0.4750$ に着目し，セルの $f(\rho') = -0.5558\,0$ が 0 となるように $\rho' = -0.4750$ を変化させれば，95%信頼区間の上限 $\rho'_{U95\%} = -0.4725$ を得ることができるただし， z_α^2 は，自由度 21 の t 分布の 5%点 $t_\alpha^2 = 2.0796$ としている．これらから対数効力比 $\rho' = -0.4883$ の正確な 95%信頼区間は

$$\log_{10}(L95\%, U95\%) = (-0.5043, -0.4725)$$

となり，10 の冪乗を取り効力比 $\rho = 0.3249$ の正確な 95%信頼区間

$$(L95\%, U95\%) = (0.3131, 0.3369)$$

が得られる．

伝統的には， $f(\rho')$ に含まれる 2 つの 2 次式を ρ' について整理し，分散，共分散を偏差平方和に落とし込み，2 次式の解の公式で解いた複雑な計算式が天下り的に示されており，理解の妨げになってきた．どのようなものかは，杉本（），「統計学入門，13.2 節 平行線検定法」に詳しく述べられている．同じデータが用いられているので，計算結果は一致している．

非線形回帰による効力比の 95%信頼区間の推定

ポアソン回帰への応用のために Excel の行列計算による効力比の 95%信頼区間の推定方法を示したのであるが，JMP の非線形回帰を用いれば，対数での効力比の正確な 95%信頼区間の推定が行なえる［高橋（2004）］.

表 8.6　JMP の非線形回帰のあてはめを用いた対数効力比の 95%信頼区間の直接推定

解

SSE	DFE	MSE	RMSE	
55.1482	21	2.6261	1.6205	
パラメータ	**推定値**	**近似標準誤差**	**下側信頼限界**	**上側信頼限界**
β0S	39.6823	0.6181	38.3969	40.9677
β0U	-3.1030	1.6871	-6.6115	0.4055
ρ'	-0.4883	0.0076	-0.5043	-0.4725

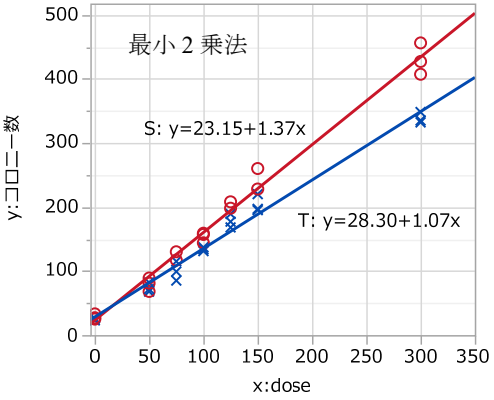
解法: 解析 Gauss-Newton

8.3. ポアソン回帰による勾配の比による効力比の推定

第 1.8 節では，細菌を用いた用量反応試験について概説し，SAS の GENMOD プロシジャを用いたポアソン回帰を行い，共分散行列を用いたデルタ法による効力比の近似計算の結果を示した［富山・杉本（2004)］．表 8.7 は，第 1.8 節で示した Ames 試験の結果であり，散布図上に示した S 薬と T 薬それぞれに各用量の分散の違いを無視した最小 2 乗法による回帰直線を引いた結果である．各用量の分散と平均の比が，S 薬と T 薬を込みにした 14 用量群の平均で 0.77 と 1 よりも小さいが，比の 95%信頼区間が 0.40～1.14 と 1 を包含していることからポアソン分布にしたがっていると判断し，ポアソン回帰を行う妥当性について示した．

表 8.7　Ames 試験での変異コロニー数の比較（表 1.27 再掲）

濃度	陽性対照 S			代替物質 T		
mg/plate	変異コロニー数			変異コロニー数		
0	27	33	25	23	26	26
50	68	89	81	68	82	72
75	131	130	117	99	85	115
100	144	157	159	137	131	134
125	199	208	198	189	177	168
150	260	229	228	197	195	220
300	427	407	456	335	332	348



誤差分布の同定の難しさ

得られた実験データからだけで，分布を同定することの困難さは，第 1.7 節で示したように，4 群の各 50 個のデータの場合でも，適合度の検定の p 値からは，「ポアソン分布」のあてはめは否定できないので「ポアソン分布らしい」とのあいまいな結論であった．「正規分布があてはめられるのか」の検定でも，第 4 群のみが「正規分布があてはまっているとは言えない」との茫洋とした結論であり，分布の同定には，常にあいまいさが残る．他方，群間で分散が異なるかの検定では，分散が 4 群間で等しいとは言えないとの結果であったが，ポアソン分布のあてはめを肯定的に示しているわけではない．

用量が増えるに従いデータの変動が大きくなる場合に，標準偏差を平均で割ったパーセント表示の変動係数 CV が一定であるような場合には，対数変換することにより標準偏差が同一な正規分布になることが知られている．S 群の場合は，(14.7%，13.4%，6.2%，5.3%，2.7%，7.6%，5.7%）用量が増えるに従い CV は減少傾向で，CV 一定とは言い難い．

では，各用量の分散が異なるかのバートレットの検定を行うと，S 群では $p=2.518$，T 群では $p=0.2140$ となり，分散が異なるとは言えない．これは，各用量のデータが 3 個しかないために検出できにくくなっていることによる．このように得られたデータだけから分布の同定は，困難を極めるので，過去の実験データあるいは観察データから総合的な判断が求められる．

実用的には，何らかの回帰直線（曲線）の個別データの 95%信頼区間を描き散布図上の点が適切に包含されているかで判断することを第 1.9 節の表 1.35，第 1.13 節の図 1.15 などで示してきた．表 8.7 の変異コロニー数について，最小 2 乗法で求めた回帰直線の残差プロットを図 8.3 に示す．

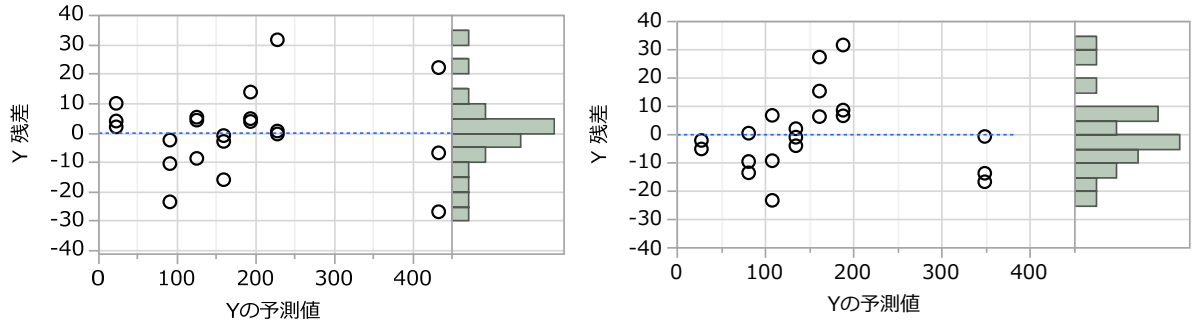


図 8.3　最小 2 乗法による残差プロット

散布図上の点は，個別データの 95%信頼区間内に収まってはいるが，残差プロットは扇型となっており，用量が少ない場合と多い場合に同程度の幅を与えていることに疑問が生ずるが，ポアソン回帰の適用に不慣れであれば，通常の回帰分析の適用を問題視するほどではない．

群ごとのポアソン回帰

群ごとに Excel シート上で反復重み付き回帰によるポアソン回帰を行い，得られた回帰パラメータおよび共分散行列を用いて個別データの 95%信頼区間を計算し，散布図上に回帰直線と個別の 95%信頼区間を上書きする．表 8.8 は，反復重み付き回帰の Excel の計算シートで，適当な初期値を $(m-1)$ にセットすると (m) に新たなパラメータが，

$$\hat{\boldsymbol{\beta}}^{(m)}=[(\boldsymbol{X}*\boldsymbol{w})^T\boldsymbol{X}]^{-1}(\boldsymbol{X}*\boldsymbol{w})^T\boldsymbol{Z}$$

によって計算される．パラメータ間の差がなくなるまで，(m) のパラメータの値のみを $(m-1)$ にペーストを繰り返した結果である．

表 8.8　S 薬に対する反復重み付き回帰によるポアソン回帰

			$X^T WX=(X*w)^T X$			$\Sigma(\beta\hat{}):(X^T WX)^{-1}$			$X^T WZ=(X*w)^T Z$	
			0.2276	11.2591		**6.8168**	-0.0490			21.0000
			11.2591	1566.42		-0.0490	**0.0010**			2400.00
				$(m-1)\ \beta_0\hat{}=$	**25.5584**		$(m)\ \beta_0\hat{}=$	**25.5584**	平方和	
				$\beta_S\hat{}=$	**1.3484**		$\beta_S\hat{}=$	**1.3484**	0.0000	
	薬	X			推定値	重み	リンク	反復推定値		
i	剤	x_0	x_1	Y	$Y\hat{}$	$w=1/Y\hat{}$	$Z=Y$	$Z\hat{}$	$Y\hat{}-Z\hat{}$	
1	S	1	0	27	25.56	0.0391	27	25.56	0.000	
2	S	1	0	33	25.56	0.0391	33	25.56	0.000	
3	S	1	0	25	25.56	0.0391	25	25.56	0.000	
4	S	1	50	68	92.98	0.0108	68	92.98	0.000	
5	S	1	50	89	92.98	0.0108	89	92.98	0.000	
6	S	1	50	81	92.98	0.0108	81	92.98	0.000	
7	S	1	75	131	126.69	0.0079	131	126.69	0.000	
8	S	1	75	130	126.69	0.0079	130	126.69	0.000	
9	S	1	75	117	126.69	0.0079	117	126.69	0.000	
10	S	1	100	144	160.40	0.0062	144	160.40	0.000	
11	S	1	100	157	160.40	0.0062	157	160.40	0.000	
12	S	1	100	159	160.40	0.0062	159	160.40	0.000	
13	S	1	125	199	194.11	0.0052	199	194.11	0.000	
14	S	1	125	208	194.11	0.0052	208	194.11	0.000	
15	S	1	125	198	194.11	0.0052	198	194.11	0.000	
16	S	1	150	260	227.83	0.0044	260	227.83	0.000	
17	S	1	150	229	227.83	0.0044	229	227.83	0.000	
18	S	1	150	228	227.83	0.0044	228	227.83	0.000	
19	S	1	300	427	430.09	0.0023	427	430.09	0.000	
20	S	1	300	407	430.09	0.0023	407	430.09	0.000	
21	S	1	300	456	430.09	0.0023	456	430.09	0.000	

表 8.9 に濃度 x を 0～350 mg/plate まで変化させた場合の S 薬のポアソン回帰直線に対する 95%信頼区間の推定値を示す．共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$，行ベクトルを $\boldsymbol{x}=[1 \quad x]$ とした場合の $\hat{y}=\boldsymbol{x}\hat{\boldsymbol{\beta}}$ に対する分散は，$Var(\hat{y})=\boldsymbol{x}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}^T$ であり，$\boldsymbol{x}=[1 \quad 50]$ の場合は，

$$
\begin{aligned}
Var(\hat{y}_{x=50}) &= \boldsymbol{x}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}^T \\
&= Var(\hat{\beta}_0)+2\times 50\times Cov(\hat{\beta}_0,\hat{\beta}_1)+50^2\times Var(\hat{\beta}_1) \\
&= 6.8168+2\times 50\times(-0.0490)+50^2\times 0.0010 \\
&= 4.39
\end{aligned}
$$

となり，回帰の 95%信頼区間は，$\hat{y}\pm 1.96\sqrt{Var(\hat{y})}$ となり，個別データの 95%信頼区間は，$\hat{y}$ の分散はポアソン回帰の場合は $\hat{y}$ なので，

$$
\begin{aligned}
(\text{個別}L95\%,\ U95\%) &= \hat{y}_{x=50}\pm 1.96\sqrt{\hat{y}+Var(\hat{y}_{x=50})} \\
&= 92.98\pm 1.96\sqrt{92.98+4.39} \\
&= (73.64,\ 112.32)
\end{aligned}
$$

として推定される．

表 8.9　S 薬に対するポアソン回帰の 95%信頼区間

				回帰		個別	
x_0	x_1	y^	$Var(y$^$)$	L95%	U95%	L95%	U95%
1	0	25.56	6.82	20.44	30.68	14.41	36.71
1	25	59.27	4.99	54.89	63.65	43.56	74.98
1	**50**	92.98	**4.39**	88.87	97.09	**73.64**	**112.32**
1	75	126.69	5.04	122.29	131.09	104.20	149.19
	:						
1	350	497.52	93.87	478.53	516.50	449.85	545.18

表 8.9 に示した S 薬の個別データの 95%信頼区間を図 8.4 左に示す．濃度が低い場合には幅が狭まり，濃度が高い場合には，逆に広がっていることが読み取れる．図 8.4 右は T 薬についての結果であり，S 薬と同様の個別データの 95%信頼区間となっている．

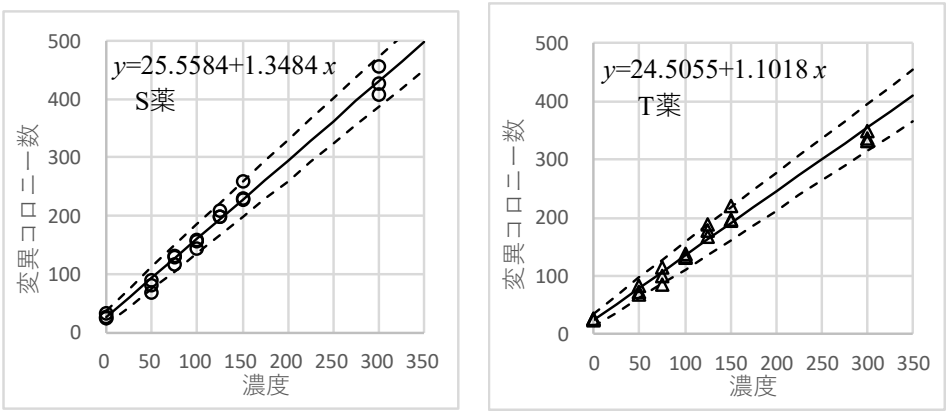


図 8.4　S 薬および T 薬に対するポアソン回帰直線の個別データの 95%信頼区間

散布図上のポイントは，個別の 95%信頼区間内のギリギリの範囲となっていて，ポアソン回帰の適用の妥当性が示されている．このように個別データの 95%信頼区間を描くことにより，通常の回帰分析に対してポアソン回帰がより妥当であるとの確証が得られる．

切片を共通とするポアソン回帰

S 薬と T 薬の傾きは 1.3484 と 1.1018 のように異なるが，切片の違いは 25.5584 と 24.5055 とわずかであり，変異コロニー数のプロット点は，ほぼ個別の 95%信頼区間内に含まれるので，切片を 2 群で共通とし，傾きだけが異なる回帰直線のあてはめを行い，勾配比検定を行う．ポアソン回帰の推定方法として，表 8.10 に示すように反復重み付き回帰による方法は，変数が増えた場合にデザイン行列の変数を増やし，推定値の計算式を変更するだけで済むので，対数尤度を用いる場合に比べ容易である．

表 8.10 は，適当な初期値 $\hat{\boldsymbol{\beta}}=[25\quad 1\quad 1]^T$ をセットして重み付き回帰の結果として $\hat{\boldsymbol{\beta}}=[24.7993\quad 1.3546\quad 1.0997]^T$ を得た結果である．あてはまりの良さは，反復 $(m-1)$ の回帰係数から計算された予測値 $\hat{\boldsymbol{Y}}$ と，反復 (m) の回帰係数から計算された予測値 $\hat{\boldsymbol{Z}}$ の差の平方和であり，5.91E+04 とかなり大きい．

表 8.10　反復重み付き回帰による切片を共通にするポアソン回帰（初期値）

			$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}=(\boldsymbol{X}^*\boldsymbol{w})^T\boldsymbol{X}$				$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$				$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z}=(\boldsymbol{X}^*\boldsymbol{w})^T\boldsymbol{Z}$	
			0.5207	14.4907	14.4907		**3.1781**	-0.0226	-0.0226		48.5	
			14.4907	2037.73	0.0000		-0.0226	**0.0007**	0.0002		3119.7	
			14.4907	0.0000	2037.73		-0.0226	0.0002	**0.0007**		2600.3	
						$(m\text{-}1)\ \hat{\beta}_0=$	**25.0000**		$(m)\ \hat{\beta}_0=$	**24.7993**		
						$\hat{\beta}_1=$	**1.0000**		$\hat{\beta}_1=$	**1.3546**	平方和	
						$\hat{\beta}_2=$	**1.0000**		$\hat{\beta}_2=$	**1.0997**	**5.91E+04**	
	薬		—— $\boldsymbol{X}$ ——					重み	リンク			
i	剤	$dose$	x_0	x_1	x_2	$\boldsymbol{Y}$	$\boldsymbol{Y}\hat{}$	$\boldsymbol{w}=1/\boldsymbol{Y}\hat{}$	$\boldsymbol{Z}=\boldsymbol{Y}$	$\boldsymbol{Z}\hat{}$	$\boldsymbol{Y}\hat{}-\boldsymbol{Z}\hat{}$	
1	S	0	1	0	0	27	25.00	0.040	27	24.80	0.201	
2	S	0	1	0	0	33	25.00	0.040	33	24.80	0.201	
3	S	0	1	0	0	25	25.00	0.040	25	24.80	0.201	
4	S	50	1	50	0	68	75.00	0.013	68	92.53	-17.529	
5	S	50	1	50	0	89	75.00	0.013	89	92.53	-17.529	
6	S	50	1	50	0	81	75.00	0.013	81	92.53	-17.529	
:												
40	T	300	1	0	300	335	325.00	0.003	335	354.72	-29.723	
41	T	300	1	0	300	332	325.00	0.003	332	354.72	-29.723	
42	T	300	1	0	300	348	325.00	0.003	348	354.72	-29.723	

得られたパラメータ $\hat{\boldsymbol{\beta}}$ を $(m-1)$ のセルに値だけをペーストして，平方和が 0 に近づくまでペーストを繰り返した過程を表 8.11 に示す．反復の 3 回目の結果は（25.0403, 1.3521, 1.0980）で，平方和は 2.66E-03 と 0 に近ずき，反復の 4 回目の結果のパラメータは小数点以下 4 桁まで一致しているが，平方和はさらに小さく 4.15E-06 となっている．反復をさらに繰り返すことにより，平方和は小さくなるが，パラメータが小数点以下 4 桁まで一致したときに繰返しを終了するとすれば，反復 4 で収束したとみなせる．さらに反復を繰り返すと表 8.12 に示すように平方和は 1.40E–10 と更に小さくなる．

表 8.11　反復重み付き回帰の収束過程

反復 m	$\beta_0\hat{}$	$\beta_1\hat{}$	$\beta_2\hat{}$	平方和
0	25.0000	1.0000	1.0000	
1	24.7993	1.3546	1.0997	5.91E+04
2	25.0503	1.3520	1.0979	1.72E+00
3	**25.0403**	**1.3521**	**1.0980**	**2.66E-03**
4	25.0407	1.3521	1.0980	4.15E-06
5	25.0407	1.3521	1.0980	6.47E-09

表 8.12　反復重み付き回帰による切片を共通にするポアソン回帰（収束値）

$X^T WX=(X^*w)^T X$			$\Sigma(\beta\hat{})=(X^T WX)^{-1}$			$X^T WZ=(X^*w)^T Z$
0.4805	11.2676	13.4178	**3.2931**	-0.0237	-0.0235	42.0
11.2676	1566.33	0.0000	-0.0237	**0.0008**	0.0002	2400.0
13.4178	0.0000	1879.77	-0.0235	0.0002	**0.0007**	2400.0

$(m\text{-}1)\,\beta_0\hat{}=$	**25.0407**		$(m)\,\beta_0\hat{}=$	**25.0407**	
$\beta_1\hat{}=$	**1.3521**		$\beta_1\hat{}=$	**1.3521**	平方和
$\beta_2\hat{}=$	**1.0980**		$\beta_2\hat{}=$	**1.0980**	**1.40E-10**

効力比および近似の 95%信頼区間

効力比の求め方，近似の 95%信頼区間，正確な 95%信頼区間の求め方は，第 8.1 節と同じ考え方で求められる．S 薬の傾きは $\hat{\beta}_1=1.3521$，T 薬の傾きは $\hat{\beta}_2=1.0980$ であり，効力比 $\hat{\rho}$ は，

$$\hat{\rho}=\frac{\hat{\beta}_2}{\hat{\beta}_1}=\frac{1.0980}{1.3521}=0.8121$$

となる．デルタ法による近似の 95%信頼区間を計算するために，効力比 $\hat{\rho}$ をパラメータ（$\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$）で偏微分した結果は，

$$d_0=\frac{\partial\hat{\rho}}{\partial\hat{\beta}_0}=0$$

$$d_1=\frac{\partial\hat{\rho}}{\partial\hat{\beta}_1}=\frac{-\hat{\beta}_2}{\hat{\beta}_1^{\,2}}=\frac{-1.0980}{1.3521^2}=-0.6006$$

$$d_2=\frac{\partial\hat{\rho}}{\partial\hat{\beta}_2}=\frac{1}{\hat{\beta}_1}=\frac{1}{1.3521}=0.7396$$

となり，

$$\boldsymbol{d}=[0.0000\quad -0.6006\quad 0.7396]^T$$

として，$\hat{\rho}$ の分散は，

	$\boldsymbol{d}$								
	d_0	d_1	d_2	共分散 $\Sigma=(X^T WX)^{-1}$			$\boldsymbol{d}^T$		$Var(\rho)$
$Var(\rho\hat{})=$	0.0000	-0.6006	0.7396	**3.2931**	-0.0237	-0.0235	0.0000	=	0.0005
				-0.0237	**0.0008**	0.0002	-0.6006		
				-0.0235	0.0002	**0.0007**	0.7396		
				$\beta_0\hat{}$	$\beta_1\hat{}$	$\beta_2\hat{}$			

と推定され，95%信頼区間は，

$$\begin{aligned}(L95\%, U95\%) &= \hat{\rho}\pm 1.96\sqrt{Var(\hat{\rho})}\\ &= 0.8121\pm 1.96\times 0.0229\\ &= (0.7672, 0.8569)\end{aligned}$$

となり，1 を含まないので，代替物質 T 薬は，標準品 S 薬に比べて変異原性が有意に減弱していると判断される．このように，第 8.1 節で示した最小 2 乗による勾配比の近似 95%信頼区間の計算式が，反復重み付きによるポアソン回帰の場合でも応用できるには，どちらも計算過程で得られるパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いているためである．

効力比の正確な 95%信頼区間

効力比の正確な 95%信頼区間の求め方は，第 8.1 節と同じ考え方で求められる．推定された傾き $\hat{\beta}_1$ および $\hat{\beta}_2$ の分散を $Var(\hat{\beta}_1)$ と $Var(\hat{\beta}_2)$，共分散を $Cov(\hat{\beta}_1, \hat{\beta}_2)$ とする．勾配比の定義式から，次の関数 $\psi = \rho\beta_1 - \beta_2$ を考える．$\hat{\beta}_1$ と $\hat{\beta}_2$ が β_1 と β_2 の不偏推定量となるので，ψ の期待値は，

$$E(\psi) = \rho\hat{\beta}_1 - \hat{\beta}_2 = 0$$

となり，ψ の分散は，共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ とし $E(\psi)$ の係数を $\boldsymbol{g} = \begin{bmatrix} 0 & \rho & -1 \end{bmatrix}$ とすれば，

$$\begin{aligned} Var(\hat{\psi}) &= \boldsymbol{g}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{g}^T \\ &= \rho^2 Var(\hat{\beta}_1) - 2\rho Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2) \end{aligned}$$

と ρ に関して 2 次式となる．そのとき，$\hat{\beta}_1$ と $\hat{\beta}_2$ は，正規分布に従うと仮定されるので，ψ は，同様に正規分布に従い

$$z = \frac{\rho\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{Var(\hat{\psi})}}$$

は，標準正規分布に従う．従って，z_α を正規分布の両側 α 点であるとしたときに，ρ の $100\cdot(1-\alpha)$% 信頼区間は，不等式

$$\left|\rho\hat{\beta}_1 - \hat{\beta}_2\right| \le z_\alpha\sqrt{Var(\hat{\psi})}$$

で与えられる．ここで ρ を変化させ等式が成り立つ場合の ρ が推定された効力比 $\hat{\rho}$ の正確な 95%信頼区間となる．そこで．両辺を 2 乗し，右辺を移項して等式とすると，2 つの 2 次式の複合式

$$f(\rho) = (\rho\hat{\beta}_1 - \hat{\beta}_2)^2 - z_\alpha^2 Var(\hat{\psi}) = 0$$

となる．

第 8.1 節では，$f(\rho)$ が 0 になるように ρ を Excel のソルバーで変化させ正確な（*L*95%, *U*95%）を求めたのであるが，ここでは，ρ についての 2 次式に展開し，2 次式の解の公式で直接求めることにする．$Var(\psi)$ を $f(\rho)$ に代入すると

$$f(\rho) = \rho^2\hat{\beta}_1^2 - 2\rho\hat{\beta}_1\hat{\beta}_2 + \hat{\beta}_2^2 - z_\alpha^2\left[\rho^2 Var(\hat{\beta}_1) - 2\rho Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2)\right] = 0$$

となり，式を整理すると，次のように ρ に関する 2 次式が得られる．

$$\left[\hat{\beta}_1^2 - Var(\hat{\beta}_1)z_\alpha^2\right]\rho^2 + \left[-2\hat{\beta}_1\hat{\beta}_2 + 2Cov(\hat{\beta}_1,\hat{\beta}_2)z_\alpha^2\right]\rho + \left[\hat{\beta}_2^2 - Var(\hat{\beta}_2)z_\alpha^2\right] = 0$$

この 2 次方程式の 2 つの根は，ρ のための 95%信頼区間となる．

2 次式 $a + b\rho + c\rho^2 = 0$ の 係数 a，b，c は，それぞれ，

$$a = \hat{\beta}_2^2 - Var(\hat{\beta}_2)z_\alpha^2,$$
$$b = -2\hat{\beta}_1\hat{\beta}_2 + 2Cov(\hat{\beta}_1,\hat{\beta}_2)z_\alpha^2,$$
$$c = \hat{\beta}_1^2 - Var(\hat{\beta}_1)z_\alpha^2$$

となり，式は複雑であるが，$\hat{x}_{L95}$ に関して 2 次式

$$a + b\hat{\rho} + c\hat{\rho}^2 = 0$$

となる．従って，2 次式の解の公式

$$(L95\%, U95\%) = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c}$$

により ρ の 95%信頼区間を求めることができる．解は 2 つあるが，小さい方が $\hat{\rho}_{L95}$ となり，大きい方が $\hat{\rho}_{U95}$ となる．

正確な $(L95\%, U95\%) = (0.7682, 0.8580)$

表 8.13　2 次式の解の公式による効力比の正確な 95%信頼区間

		パラメータの共分散行列 $\boldsymbol{\Sigma}=(-\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$		
$\beta\hat{}_0=$	25.041	3.1732	-0.0225	-0.0226
$\beta\hat{}_1=$	**1.3521**	-0.0225	**0.0008**	0.0002
$\beta\hat{}_2=$	**1.0980**	-0.0226	**0.0002**	**0.0007**
		$\beta\hat{}_0$	$\beta\hat{}_1$	$\beta\hat{}_2$

$z_\alpha=$	**1.96**			
$a=$	1.2030	$=\hat{\beta}_2^2 - Var(\hat{\beta}_2)z_\alpha^2$	$\rho=$	**0.8121**
$b=$	-2.9680	$=-2\hat{\beta}_1\hat{\beta}_2 + 2Cov(\hat{\beta}_1,\hat{\beta}_2)z_\alpha^2$	$L95\%=$	0.7682
$c=$	1.8251	$=\hat{\beta}_1^2 - Var(\hat{\beta}_1)z_\alpha^2$	$U95\%=$	0.8580
2次式の解:		$\hat{\rho}_{95\%} = (-b \pm \sqrt{b^2 - 4ac}) / 2c =$	(0.7682,	0.8580)

ソルバーを用いた正確な 95%信頼区間の推定

第 8.1 節の図 8.1 で示したと同様にポアソン回帰で得られたパラメータの推定値 $\hat{\boldsymbol{\beta}}$ および共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用い，図 8.5 に示すように勾配比 ρ を 0.70～0.95 変化させたときの $f(\rho)$ を計算する．

$$f(\rho)=(\rho\hat{\beta}_1-\hat{\beta}_2)^2-z_\alpha^2 Var(\hat{\psi})$$

ソルバーを用い，2 次式 $f(\rho)$ がマイナスからプラスに代わる $\rho=0.76$ および $\rho=0.86$ の $f(\rho)$ がゼロになるように ρ を変化させた結果が，（0.7682， 0 0.8580）となり，2 次式の解の公式で求めた正確な 95%信頼区間が推定されている．

ρ	$(\rho\hat{\beta}_1-\hat{\beta}_2)^2$	$z_\alpha^2 Var(\psi)$	$f(\rho)$
0.70	0.0230	0.0033	**0.0197**
0.72	0.0155	0.0034	0.0121
0.74	0.0095	0.0034	0.0061
0.7682	0.0035	0.0035	**0.0000**
0.78	0.0019	0.0036	-0.0017
0.80	0.0003	0.0036	-0.0034
0.82	0.0001	0.0037	-0.0036
0.84	0.0014	0.0038	-0.0024
0.8580	0.0039	0.0039	**0.0000**
0.88	0.0084	0.0039	0.0045
0.90	0.0141	0.0040	0.0101
0.92	0.0213	0.0041	0.0172
0.94	0.0299	0.0042	0.0257
0.95	0.0348	0.0043	0.0305

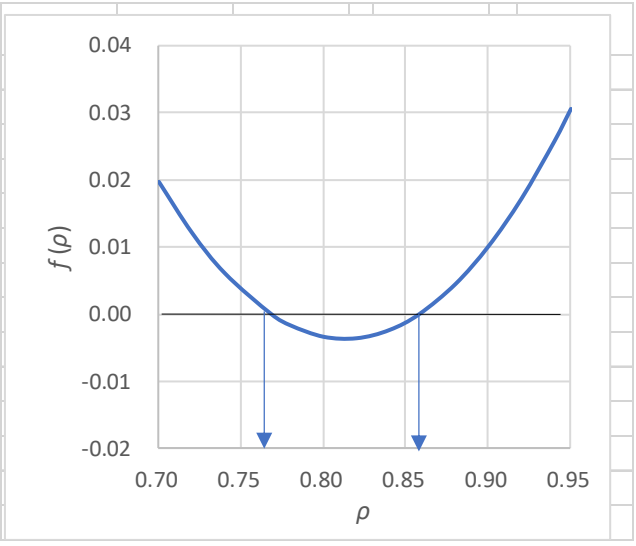


図 8.5　勾配比 ρ を変化させたときの $f(\rho)$ の 2 次曲線から求めた正確な 95%信頼区間

勾配比 $\rho=0.7$ の場合の $f(\rho)$ の計算は，次に示すように Excel で計算され，フィルハンドルで計算式をコピーしている．

$$\begin{aligned}
f(\rho)&=(\rho\hat{\beta}_1-\hat{\beta}_2)^2-z_\alpha^2 Var(\hat{\psi})\\
f(0.7)&=(0.7\times 1.3521-1.0980)^2-1.96^2 Var(\hat{\psi})\\
&=0.0230-1.96^2[\rho^2 Var(\hat{\beta}_1)-2\rho Cov(\hat{\beta}_1,\hat{\beta}_2)+Var(\hat{\beta}_2)]\\
&=0.0230-1.96^2[0.7^2\times 0.0008-2\times 0.7\times 0.0002+0.0007]\\
&=0.0230-0.0033\\
&=0.0197
\end{aligned}$$

9. 花数を共変量とした種子数の探索的ポアソン回帰

ポアソン回帰に関連する論文を Web で検索し，データリストを含む探索的な解析が行なわれている下野（2010）の「R を用いた一般化線形モデル（回帰係数編）：カウントデータを例に」を見出した．仮想データとして，3 地域（Region：A，B，C）において 2 つの生育環境（Habitat：Dry，Wet）に分布する植物種各 10 個体を選び，全体で 60 個のデータで，個体あたりの生産花数（Flower No），生産種子数（Seed No）を調査した結果が示されている．生産種子数が地域や生育環境といった要因によって違いがあるか Excel，JMP，SAS を用いて別の切り口から探索的なポアソン回帰を試みる．また，R の負の二項回帰の関数 glm.nb()関数の出力結果に示されている各種のデビアンスについても Excel による計算結果との対比を行っている．

9.1. データの概観

下野（2010）は，種子数が花数によって影響を受けるので，花数をオフセットとした過分散を調整したポアソン回帰，さらに負の 2 項分布を使った解析結果を示し，カウント・データの解析について総合的に論じている．

第 1.5 節の「冠動脈心疾患の死亡者数」で示したように，オフセットは，観測された死亡者数の部分母集団の人数の対数として扱われている．花数は，種子数に対して影響を与えると推測されるので，オフセットとして扱うことに違和感がある．花数をオフセットとするモデルとすることは可能であるが，観測された反応変数としての側面もある．そこで，花数を共変量とした探索的な解析を行い，さらに，花数をオフセットとした Excel，JMP，SAS による解析結果を示すことにした．なお，R 言語のユーザで，無償の OnDemad SAS に興味を持たれた場合には，臨床評価研究会（ACE）基礎解析分科会（2017），「新版 実用 SAS 生物統計ハンドブック」に，SAS と R による多くの解析例が並列的に示されプログラムもダウンロードできるので参考にしてもらいたい．

表 9.1 に下野（2010）で示されているデータを示す．生育環境（Habitat：Dry，Wet）を表側に，地域（Region：A，B，C）を表頭にし，植物種各 10 個体あたりの生産花数（Flower No），および，生産種子数（Seed No）を示す．全体で 60 個の仮想データである．

表 9.1　生育環境 Habitat と地域別 Region の花数および種子数のデータ［下野（2010）］

	Region A		Region B		Region C	
Habitat	FlowerNo	SeedNo	FlowerNo	SeedNo	FlowerNo	SeedNo
Dry	3	57	1	22	3	67
	1	19	1	21	3	59
	1	11	2	36	1	10
	1	12	1	24	2	45
	3	46	3	45	1	8
	3	51	2	35	3	60
	2	21	3	53	3	47
	2	29	2	25	1	9
	2	28	2	23	3	79
	2	46	3	56	1	15
平均		32.00		34.00		39.90
Wet	2	26	2	42	4	97
	3	61	1	27	2	45
	2	35	3	101	2	38
	4	81	3	149	3	68
	3	85	2	56	1	28
	1	34	1	29	1	33
	1	25	1	35	2	55
	1	31	2	49	1	14
	2	48	3	86	4	129
	2	70	3	73	3	101
平均		49.60		64.70		60.80

下野は，オフセットが有効なデータとして

1）ある面積の中に生えていた雑草個体数
2）ある土壌体積中に含まれていた埋土種子数
3）あるサイズの個体が生産した種子数
4）ある花数の個体が生産した種子数

などを挙げている．1）と 2）は，（ある面積中，ある土壌の体積中）は，その場所に存在する各個体に対する部分母集団と見なし，オフセットとして扱うことが有効と思われる．しかし，3）と 4）は，何らかの処置により（あるサイズの個体，ある花数）自体も影響を受けると思われるので，部分母集団と見なすのではなく共変量としてモデルの中に含めることが望ましいと思われる．

下野は，図 9.1 に示すような生育環境別 地域別の種子数の平均の線グラフ，花数別の種子数のボックス・プロットを示している．なお，線グラフは，生育環境と地域について JMP の「モデルのあてはめ」で通常の線形モデルによる 2 元配置分散分析を行い「交互作用プロファイル」による結果であり，ボックス・プロットは，花数について「二変量の関係」で作成した結果である．

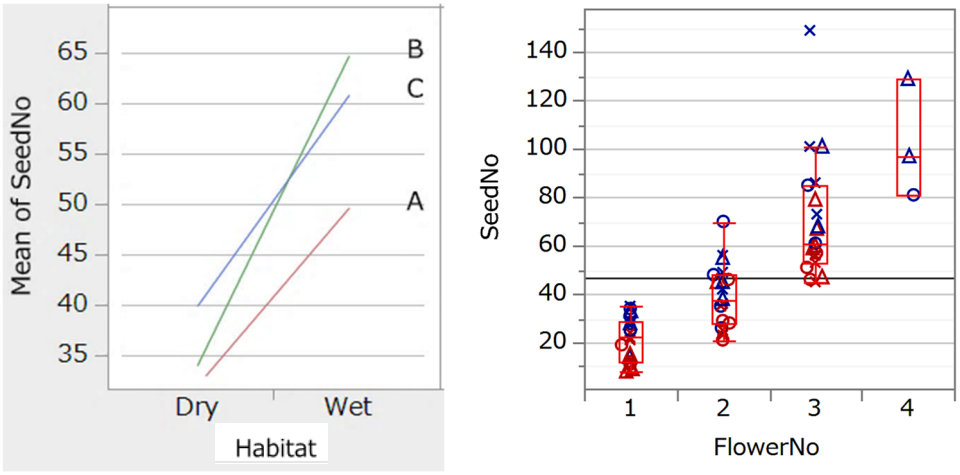


図 9.1　生育環境と種子数の関係（左）および花数と種子数の関係（右）

これらの図から，生育環境が Wet の場合は，Dry に比べて種子数は明らかに多く，地域 B と地域 C は，地域 A に比べて種子数が大きめであることが読み取れる．また，地域 B は，Wet 環境で他の地域に比べ種子数が多く，量的な交互作用が示唆されている．ボックス・プロットからは，花数が増えた場合に種子数の平均の増大に伴い，分散も増大する傾向が観察される．

なお，量的な交互作用とは，傾きが同じ方向を向いており，誤差的な変動が大きめに出ているとも解されるような傾きの違いである．量的な交互作用に対し質的な交互作用は，傾きがプラス方向とマイナス方向のような極端な場合も含め，偶然の変動とは思われないような傾きの差がある場合である．

JMP のグラフ・ビルダーを用い，図 9.1 に示した別々の図を一まとめにしたのが図 9.2 である．Y 軸方向に 2 水準の生育環境，X 軸方向に 3 水準の地域に分割し，分割された区画の中で，さらに X 軸方向に花数を，Y 軸方向に種子数とした散布図を描き，花数ごとにボックス・プロットを上書きし，さらに平滑線をあてはめた結果である．この結果から，花数の増加による種子数の増加は，やや指数関数的な増加であることが読み取れる．

第 7.2 節で例示したように，S-PLUS にも同様な Trellis（格子）グラフ作成機能があり重宝していた．R 言語でも Trellis（格子）グラフのパッケージが提供されているようだが使用経験はない．詳細は，Murrell 著・久保訳（2009），「R グラフィックス　第 4 章　Trellis 作図：lattice パッケージ」を参照のこと．

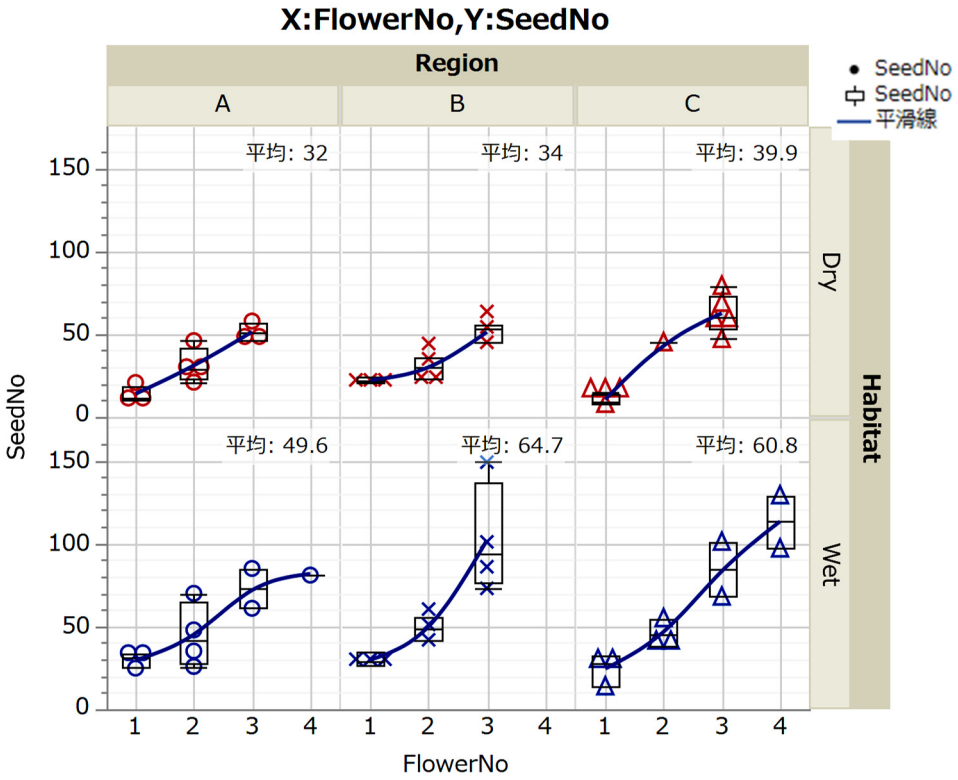


図 9.2　生育環境別 地域別 花数別の種子数に対する平滑線のあてはめ

JMP のグラフ・ビルダーによる作図は，基本の散布図（X 軸：花数，Y 軸：種子数）を作成し，その上に「平滑線」を上書きさせ，「ボックス・プロット」も上書きさせる．さらに，X 軸のグループとして地域を，Y 軸のグループとして育成環境を設定する．なお，これらは，すべて画面上の操作（GUI）で行うことができる．

図 9.2 に対応して地域別・生育環境別・花数別の分散/平均の比を計算し過分散の程度を表 9.2 に示す．分散/平均の比は，1 以下の場合もあるが，それぞれの地域の分散/平均の比の平均は（3.06，2.43，3.26）と過分散となっている．

表 9.2　生育環境別 地域別 花数別の種子数の分散/平均の比

	Flower	Region A				Region B				Region C			
Habitat	No	*n*	平均	分散	分散/平均	*n*	平均	分散	分散/平均	*n*	平均	分散	分散/平均
Dry	1	3	14.00	19.00	1.36	3	22.33	2.33	0.10	4	10.50	9.67	0.92
	2	4	31.00	112.67	3.63	4	29.75	44.92	1.51	1	45.00	.	.
	3	3	51.33	30.33	0.59	3	51.33	32.33	0.63	5	62.40	137.80	2.21
Wet	1	3	30.00	21.00	0.70	3	30.33	17.33	0.57	3	25.00	97.00	3.88
	2	4	44.75	364.92	8.15	3	49.00	49.00	1.00	3	46.00	73.00	1.59
	3	3	73.00	288.00	3.95	4	102.25	1102.3	10.78	2	84.50	544.50	6.44
	4	1	81.00	.	.	.	.	.	.	2	113.00	512.00	4.53
(分散/平均)比の平均					3.06				2.43				3.26

9.2. JMP のポアソン回帰による探索的解析

実験計画法の観点からこのデータは，地域（A，B，C）を標示因子，生育環境（Dry，Wet）を制御因子，花数を共変量とする 3×2 の要因配置モデルと見なすことができる．共変量としての花数がなければ，ごく普通の繰り返しが 10 の 2 元配置型モデルとなる．地域と生育環境に交互作用が統計的に検出されたとしても地域は標示因子（変量効果，ランダムに選ばれた場所）であり，地域によって育成環境による種子数に差があった（交互作用あり）場合であっても，主たる解析では，交互作用を含まない解析を優先すべきである．だだし，副次的な解析としては，交互作用の確認を行うことは必要である．

交互作用モデル

JMP のポアソン回帰で過分散を考慮し，

モデル： 種子数 ＝ 生育環境 ＋ 地域 ＋ 生育環境×地域 ＋ 花数

分布：Poisson， リンク関数：対数， 過分散：あり

による解析で交互作用（生育環境×地域）を確認する．表 9.3 に示すように，交互作用の p 値は 0.1565 であり，図 9.1 では，交互作用があるかと示唆したが，統計的には有意な差ではなかった．図 9.2 のボックス・プロットから，（育成環境：Wet，地域：B，花数：3）において外れ値的な種子数 149 のデータがあり，これが平均値が大きくなっている原因である．

表 9.3　交互作用（生育環境×地域）を含めたポアソン回帰

モデル全体の検定

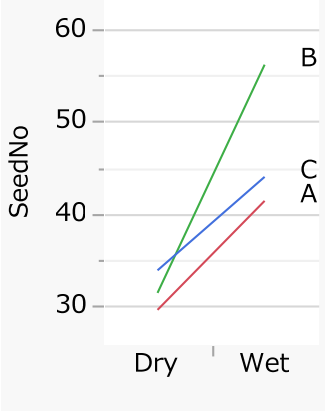
モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	120.5426	241.0852	6	<.0001*
完全	74.3134			
縮小	194.8559			

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)	過分散
Pearson	181.6694	53	<.0001*	3.4277
デビアンス	179.8630	53	<.0001*	

AICc
167.4502

効果の検定

要因	自由度	尤度比カイ2乗	p値(Prob>ChiSq)
Habitat	1	29.5774	<.0001*
Region	2	4.0718	0.1306
Habitat*Region	2	3.7089	0.1565
FlowerNo	1	181.8276	<.0001*



この図は，花数を共変量としてモデルに含めているので，図 9.1 とは形状が異なる

注）過分散を反映．交互作用プロファイルは，花数を 2 と固定した場合である．

JMP のポアソン回帰では,「予測プロファイル」の出力に引き続き表 9.3 右に示すように「交互作用プロファイル」を出力し，どのような交互作用なのか手軽に可視化することができる．

表 9.4 に交互作用を含むパラメータの推定値および Excel で整形した共分散行列を示す．推定値に有意な差が有るか否かは分かるが，対数リンクでの結果なので，元の種子数に戻さないと理解しづらい．さらに，何らかのグラフ表示をしなければ結果の解釈は困難であり，どのような計算を JMP の内部で行っているのか, それをどのようにグラフ化しているのか Excel を用いた検討を行う．

表 9.4　交互作用含むモデルの推定値およびパラメータの共分散行列

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	2.6060	0.1063	423.6291	<.0001*
Habitat[Dry]	-0.1970	0.0367	29.5774	<.0001*
Region[A]	-0.0938	0.0528	3.2199	0.0727
Region[B]	0.0893	0.0513	2.9845	0.0841
Habitat[Dry]*Region[A]	0.0284	0.0527	0.2895	0.5905
Habitat[Dry]*Region[B]	-0.0941	0.0510	3.4300	0.0640
FlowerNo	0.5211	0.0398	181.8276	<.0001*

パラメータの共分散行列 $\Sigma(\hat{\beta})$	切片	Habitat [Dry]	Region [A]	Region [B]	Habitat [Dry] *Region[A]	Habitat [Dry] *Region[B]	Flower No
切片	**0.0113**	-0.0002	-0.0001	-0.0006	0.0000	0.0004	-0.0040
Habitat[Dry]	-0.0002	**0.0013**	0.0000	0.0001	0.0002	0.0000	0.0002
Region[A]	-0.0001	0.0000	**0.0028**	-0.0014	0.0006	-0.0004	0.0001
Region[B]	-0.0006	0.0001	-0.0014	**0.0026**	-0.0004	0.0007	0.0002
Habitat[Dry]*Region[A]	0.0000	0.0002	0.0006	-0.0004	**0.0028**	-0.0014	0.0000
Habitat[Dry]*Region[B]	0.0004	0.0000	-0.0004	0.0007	-0.0014	**0.0026**	-0.0001
FlowerNo	-0.0040	0.0002	0.0001	0.0002	0.0000	-0.0001	**0.0016**

パラメータの共分散行列 $\Sigma(\hat{\beta})$ は，JMPの出力結果をExcelに取り込み整形した表である

JMP で質的変数を解析モデルに含めた場合には，（1，–1）対比型のデザイン行列の変数が生成される．表 9.5 に示すように 2 水準の生育環境の場合は，内部のコード順が（Dry，Wet）なので，Dry の場合に 1，Wet の場合に −1 となる Habitat[Dry]変数が生成される．地域のように 3 水準の場合は，内部コード順が（A，B，C）なので，A の場合（1，0），B の場合（0，1），C の場合（−1，−1）となるような 2 組の Region[A]，Region[B] 変数が生成される．交互作用は，取り上げた 2 つの名義尺度変数から生成される変数の総当たりによる Habitat[Dry]*Region[A]，Habitat[Dry]*Region[B] 変数が生成される．なお，統計ソフトによってデフォルトで生成されるデザイン行列のダミー変数の型は異なるので注意が必要である．

表 9.5　JMP の内部で生成される（1，–1）対比型デザイン変数

Habitat	Habitat [Dry]
Dry	1
Wet	-1

Region	Region [A]	Region [B]
A	1	0
B	0	1
C	-1	-1

Habitat	Region	Habitat [Dry] *Region [A]	Habitat [Dry] *Region [B]
Dry	A	1	0
	B	0	1
	C	-1	-1
Wet	A	-1	0
	B	0	-1
	C	1	1

表 9.6 に内部で生成されたデザイン行列の変数に対応する推定値 $\hat{\boldsymbol{\beta}}$，花数を 2 に固定した生育環境と地域の組み合わせに対するデザイン行列の変数 $\boldsymbol{x}$ を示す．生育環境が Dry，地域が A，花数が 2 つの場合は，

$$\boldsymbol{x}_{Dry,A,2} = [1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 2]$$

であり，推定値 $\hat{\boldsymbol{\beta}}$ を表 9.4 の列ベクトルを転置して行ベクトル

$$\hat{\boldsymbol{\beta}}^T = [2.6060 \quad -0.1970 \quad -0.0938 \quad 0.0893 \quad 0.0284 \quad -0.0941 \quad 0.5211]$$

として示す．推定値 $\hat{y}_{\mathrm{Dry,A,2}}$ は，Excel の行列関数を用いて

$$\begin{aligned}\hat{y}_{\mathrm{Dry,A,2}} &= \exp(\boldsymbol{x}_{\mathrm{Dry,A,2}}\hat{\boldsymbol{\beta}}) \\ &= \exp(\mathrm{Mmult}(\boldsymbol{x}_{\mathrm{Dry,A,2}}\text{の範囲},\ \mathrm{Transpose}(\hat{\boldsymbol{\beta}}^T\text{の範囲}))) \\ &= \exp(3.3857) = 29.5390\end{aligned}$$

表 9.6　交互作用を含むモデルの推定値

			Habitat	Region		交互作用					
生育		切片	Dry	A	B	Dry*A	Dry*B	花数	対数	指数	分散
環境	地域	2.6060	-0.1970	-0.0938	0.0893	0.0284	-0.0941	0.5211	推定値	推定値	*Var*
Dry	A	**1**	**1**	**1**	**0**	**1**	**0**	**2**	**3.3857**	**29.5390**	**0.0109**
Wet	"	1	-1	1	0	-1	0	**2**	3.7230	41.3864	0.0075
Dry	B	1	1	0	1	0	1	**2**	3.4463	31.3851	0.0102
Wet	"	1	-1	0	1	0	-1	**2**	4.0285	56.1779	0.0056
Dry	C	1	1	-1	-1	-1	-1	**2**	3.5213	33.8288	0.0090
Wet	"	1	-1	-1	-1	1	1	**2**	3.7841	43.9945	0.0070

と計算している．図 9.3 に示すように JMP の予測プロファイルを用いて種子数 $\hat{y}_{\mathrm{Dry,A,2}} = 29.5390$ が推定されている．推定値の 95%信頼区間（24.0829, 36.2312）は，表 9.4 に示されているパラメータの共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ としたときに $\ln \hat{y}_{\mathrm{Dry,A,2}}$ の分散を，

$$\begin{aligned}Var(\ln \hat{y}_{\mathrm{Dry,A,2}}) &= \boldsymbol{x}_{\mathrm{Dry,A,2}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}^T_{\mathrm{Dry,A,2}} \\ &= \mathrm{Mmult}(\mathrm{Mmult}(\boldsymbol{x}_{\mathrm{Dry,A,2}}\text{の範囲}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{の範囲}), \mathrm{Transpose}(\boldsymbol{x}_{\mathrm{Dry,A,2}}\text{の範囲})) \\ &= 0.0109\end{aligned}$$

として，95%信頼区間は，

$$95\%\,\mathrm{CL} = \exp(\ln \hat{y}_{\mathrm{Dry,A,2}} \pm 1.96\sqrt{Var(\ln \hat{y}_{\mathrm{Dry,A,2}})}$$
$$= \exp(3.8358 \pm 1.96\sqrt{0.0109}$$
$$= (24.0829,\quad 36.2312)$$

と推定されている．

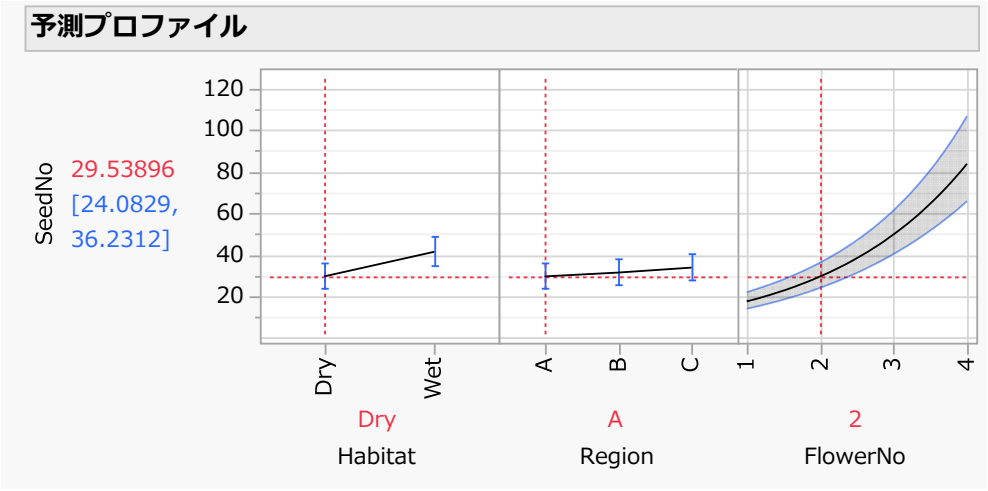


図 9.3　生育環境 Dry　地域 A　花数 2 に固定した場合の予測プロファイル

表 9.6 には，$\hat{y}_{\mathrm{Wet,A,2}}$，$\hat{y}_{\mathrm{Dry,B,2}}$，$\hat{y}_{\mathrm{Wet,B,2}}$，$\hat{y}_{\mathrm{Dry,C,2}}$，および，$\hat{y}_{\mathrm{Wet,C,2}}$ についても推定されているので，これらの推定値を用いて Excel の「折れ線グラフ」により表 9.3 の中の図に示した交互作用プロファイルを図 9.4（中）に示すように再現することができる．さらに，花数を変化させることにより図 9.4（左）および（右）の交互作用プロファイルも元データにグラフが連動させているので自動的に得られる．

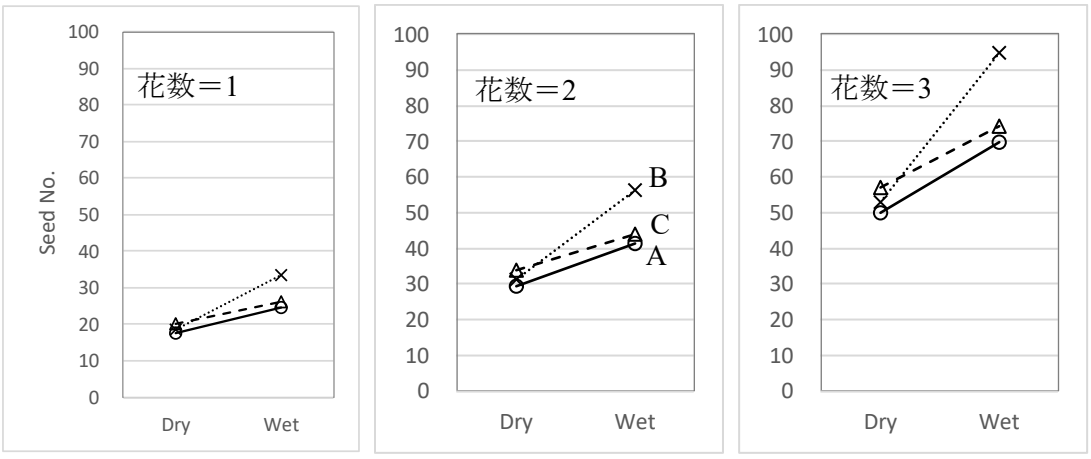


図 9.4　Excel の折れ線グラフによる生育環境による地域別の種子数の交互作用プロファイル

図 9.2 のグラフ・ビルダーで作成したグラフは，花数を X 軸にした種子数のデータのプロットであるが，因子間の関連については見えにくい．図 9.4 では，交互作用モデルに即した花数ごとの生育環境と地域における種子数の関係が明瞭に把握できる．

主効果モデル

交互作用が統計的には検出されなかったので，解析モデルから交互作用項を除いた主効果モデル

モデル： 種子数 ＝ 生育環境 ＋ 地域 ＋ 花数

分布：Poisson，リンク関数：対数，過分散：あり

による解析した結果を表 9.7 に示す．

表 9.7　過分散を反映した主効果モデルによる検討

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値)
差分	113.2155	226.4310	4	<.0001*
完全	72.6557			
縮小	185.8712			

適合度統計量	カイ2乗	自由度	p値	過分散
Pearson	197.6378	55	<.0001*	3.5934
デビアンス	192.5761	55	<.0001*	

AICc
158.8964

効果の検定

要因	自由度	尤度比カイ2乗	p値
Habitat	1	28.9356	<.0001*
Region	2	5.7869	0.0554
FlowerNo	1	171.8171	<.0001*

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	2.6197	0.1080	418.5343	<.0001*
Habitat[Dry]	-0.1984	0.0373	28.9356	<.0001*
Region[A]	-0.1016	0.0526	3.8146	0.0508
Region[B]	0.1126	0.0505	4.8943	0.0269*
FlowerNo	0.5160	0.0404	171.8171	<.0001*

地域 Region の p 値は，0.0554 と微妙な大きさであるのに対し，育成環境 Habitat には，明らかな統計的な差が検出されている．これらの要因を組み合わせた予測プロファイルおよび 2 つの要因を組み合わせた“交互作用”プロファイルを用いて結果の吟味を行う．

図 9.5 に示すように花数を 2 に固定した場合に，種子数が最も多くなるのは，育成環境が Wet，地域が B であり，$\hat{y}_{\mathrm{Wet,B,2}} = 52.5984$ と推定されており，その 95%信頼区間は（45.8902, 60.2873）となる．この予測プロファイルの中で，ポインターで育成環境が Dry の位置で選択すると他の予測プロファイルが，下方にシフトし，推定値も $\hat{y}_{\mathrm{Dry,B,2}} = 35.3673$ と更新される．

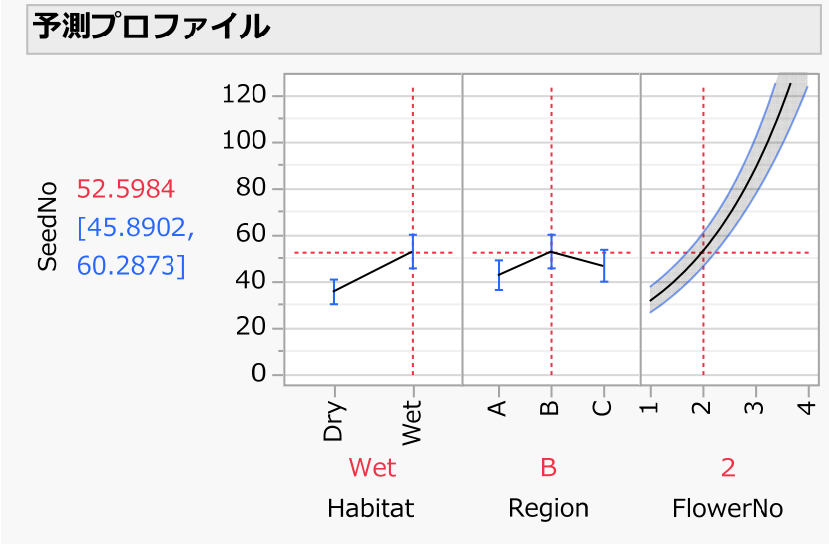


図 9.5　JMP による生育環境 Wet 地域 B 花数 2 の場合の種子数の推定

JMP の予測プロファイルの動的な機能は，探索的解析のための優れものであるので，Excel によって再現してみよう．表 9.8 に示すように，主効果モデルのパラメータの推定値，

$$\hat{\boldsymbol{\beta}}^T = [2.6197 \quad -0.1984 \quad -0.1016 \quad 0.1126 \quad 0.5160]$$

および，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を Excel に取り込んで，交互作用が有る場合と同様に対数推定値，対数分散，指数推定値，$L95\%$，$U95\%$を計算した結果を示す．

表 9.8　主効果モデルによる予測プロファイルの計算シート

					Habitat	Region		花数				指数	
		生育		切片	Dry	A	B	2	対数	対数	指数	95%信頼区間	
作図用		環境	地域	2.6197	-0.1984	-0.1016	0.1126	0.5160	推定値	分散	推定値	$L95\%$	$U95\%$
1	1	Dry	B	1	**1**	0	1	2	3.5658	0.0058	35.3673	30.48	41.04
2	2	Wet	"	1	**-1**	0	1	2	**3.9627**	**0.0048**	**52.5984**	**45.89**	**60.29**
1	1	Wet	A	1	-1	**1**	**0**	2	3.7484	0.0058	42.4540	36.59	49.26
2	2	"	B	1	-1	**0**	**1**	2	**3.9627**	**0.0048**	**52.5984**	**45.89**	**60.29**
3	3	"	C	1	-1	**-1**	**-1**	2	3.8391	0.0056	46.4820	40.16	53.79
		Wet	B	1	-1	0	1	**1**	3.4467	0.0081	31.3974	26.33	37.44
		"	"	1	-1	0	1	**2**	**3.9627**	**0.0048**	**52.5984**	**45.89**	**60.29**
		"	"	1	-1	0	1	**3**	4.4786	0.0049	88.1154	76.83	101.06
		"	"	1	-1	0	1	**4**	4.9946	0.0082	147.615	123.62	176.27

	共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$	切片	Habitat Dry	RegionA	RegionB	Flower No
JMPの結果を	切片	**0.0117**	-0.0002	-0.0001	-0.0008	-0.0041
Excelに取り込み	HabitatDry	-0.0002	**0.0014**	0.0000	0.0000	0.0002
整形した表である	RegionA	-0.0001	0.0000	**0.0028**	-0.0014	0.0001
	RegionB	-0.0008	0.0000	-0.0014	**0.0025**	0.0003
	FlowerNo	-0.0041	0.0002	0.0001	0.0003	**0.0016**

この予測プロファイルの計算シートは，Excel の行列関数を 1 行目で

$$\ln \hat{y} = \text{Mmult}(\text{デザイン変数}\boldsymbol{x}\text{の範囲，Transpose(固定した}\hat{\boldsymbol{\beta}}^T\text{の範囲)})$$

$$= \text{Mmult (F8:J8, Transpose(\$F\$7:\$J\$7)) = 3.5658}$$

のように変数の範囲を選択した計算式を，2 行目以後にフィルハンドルで数式をコピーしている．回帰パラメータは，範囲を選択した後に「F4」キーでセルアドレスに $ を付けて位置を固定している．このような計算に不慣れな場合には，第 4.5 節の「デザイン行列を用いた回帰分析の実際」を参照してもらいたい．対数分散は，1 行目で次のように計算式をセットし

$$Var(\ln \hat{y}) = \text{Mmult(Mmult(}\boldsymbol{x}\text{の範囲, 固定した}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{ の範囲), Transpose(}\boldsymbol{x}\text{の範囲))}$$

$$=\text{Mmult(Mmult(F8:J8 , \$K\$18:\$O\$22), Transpose(F8:J8)) = 0.0058}$$

2 行目以後をフィルハンドルで数式をコピーしている．なお，これらの計算方法は，通常の回帰分析でも同じである．

図 9.5 に示した JMP の予測プロファイルと同様に，育成環境 Habitat を変化させ，推定値と 95%信頼区間を $\hat{y}_{\text{Dry,B,2}} = 35.3673$（30.48，41.04），$\hat{y}_{\text{Wet,B,2}} = 52.5984$（45.89，60.29）を計算する．次に，地域 Region を変化させ，さらに花数を変化させた推定値と 95%信頼区間を計算する．図 9.6 に示した予測プロファイルは，Excel の「散布図」を使い，推定値を最初に描き，順次 95%信頼区間のための数値を「データの選択」機能を使い「追加」し，「データ系列の書式設定」で形を整えた結果である．

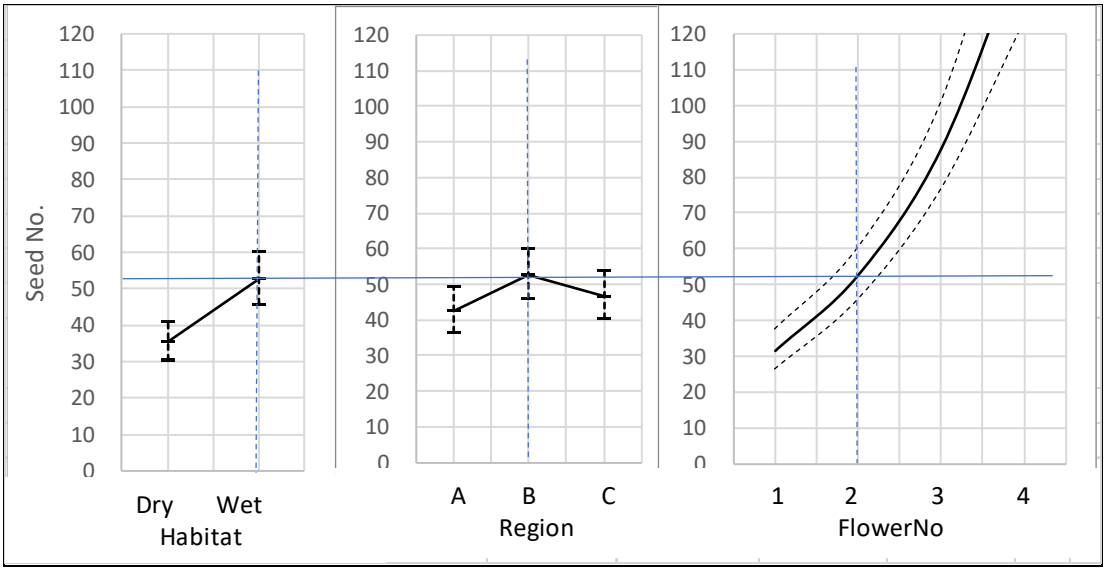


図 9.6　Excel による予測プロファイル（生育環境 Wet 地域 B 花数 2）に固定

2 つの要因を組み合わせた JMP の交互作用プロファイルを図 9.7 に示す．交互作用が，解析モデルに含まれていなくても 2 つの要因の組み合わせた場合の予測値を推定している．推定結果は予測プロファイルで設定した要因の水準の値を反映している．

環境要因 Habitat と地域 Region の組み合わせは，図 9.6 で花数を 2 としたので，花数が 2 の場合の予測値となっている．花数を変化させると，その花数を反映した推定値が示される．環境要因と花数の組み合わせは，地域 B に固定した推定値であるので，他の地域を選択すると上下に変化する．地域と花数の組み合わせは，育成環境 Wet に固定した場合の推定値で Dry に変化させると下方に移動する．

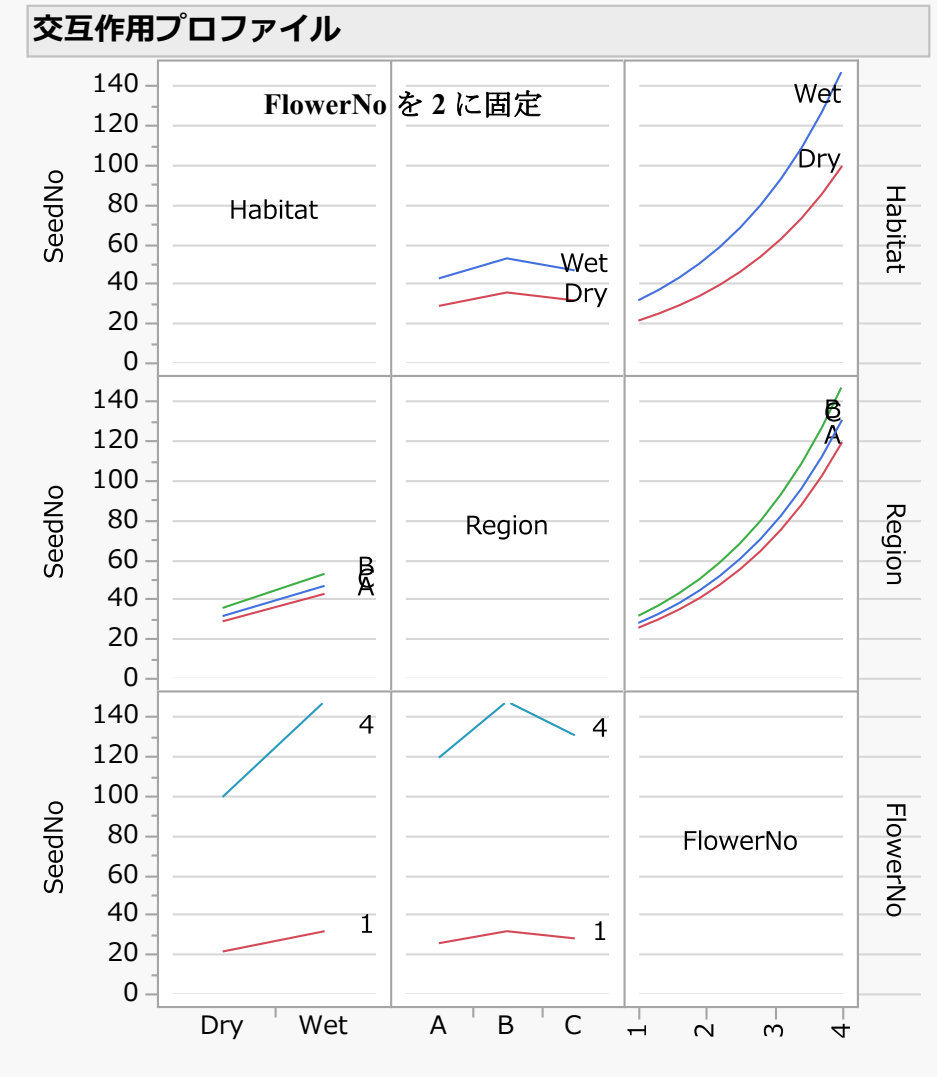


図 9.7　JMP の主効果モデルにおける交互作用プロファイル
連続変数の花数と質的変数間の交互作用の表示は，（最小値 1 と最大値 4）となっている．変更はできるが，表示できるのは 2 本に限定されている．

9.3. 無償版 SAS の GENMOD プロシジャによる主効果モデル

無償版 OnDemand SAS の GENMOD プロシジャを用い，JMP でのポアソン回帰の結果を検証しつつ，使用法の解説を行う．SAS は，DATA ステップで SAS データセットを作成し，PROC ステップで GENMOD などのプロシジャを起動する．表 9.1 で示したデータリストをできるだけコンパクトにするために，（育成環境別 地域別）を先頭に，

（花数$_1$　種子数$_1$），（花数$_2$　種子数$_2$），...，（花数$_{10}$　種子数$_{10}$）

のように横並びにすると datalines 以下に示した 6 行で 60 組の全データを表すことができる．

```
Title "Poisson_S9_1_F_Seed.sas   2019/12/05 Y.Takahashi" ;
data d01 ;
  input Habitat$ Region$ @@ ;
  do No=1 to 10 ;
    input  FlowerNo SeedNo @@ ;
    ln_FloweNo = log(FlowerNo);  output ;
    end;
datalines ;
Dry  A   3 57   1 19   1  11   1  12   3 46   3 51   2 21   2 29   2  28   2  46
Dry  B   1 22   1 21   2  36   1  24   3 45   2 35   3 53   2 25   2  23   3  56
Dry  C   3 67   3 59   1  10   2  45   1  8   3 60   3 47   1  9   3  79   1  15
Wet  A   2 26   3 61   2  35   4  81   3 85   1 34   1 25   1 31   2  48   2  70
Wet  B   2 42   1 27   3 101   3 149   2 56   1 29   1 35   2 49   3  86   3  73
Wet  C   4 97   2 45   2  38   3  68   1 28   1 33   2 55   1 14   4 129   3 101
;
proc print data=d01 ;  run;
```

1）横並びのデータリストを縦方向に並べ直すために，まず，input　Habitat$　Region$ @@ ; により，文字型の変数として Habitat と Region を読み込み，@@ ; により改行せずにその位置で読み込みポインターを待機させる．Do No=1 to 10 ; により，end;　まで 10 回繰返しが行なわれる．なお，Habitat$ の $ は，文字型のデータであるとの定義である．

2）Do No=1 to 10 ; の次の Input　FlowerNo　SeedNo　@@ ; により データリストから FlowerNo SeedNo を読込み　@@ ; により「読み込みポインター」改行せずにその位置で待機させる．

3）読み込んだ　FlowerNo　を用いて　オフセット　ln_FloweNo = log(FlowerNo); を計算する．

4）Output ;　により，既に読み込んだ Habitat, Region に加え，制御変数としての No ，読み込んだ　FlowerNo, SeedNo 計算された ln_FloweNo を SAS データセット d01 に出力する．

5）Do No=1 to 10 ; で No=2　として繰り返し，　No=10 を実行すると 1 行目のデータリストが尽きる．

6) 次の行に読み込みポインターが移動し，次のデータリストに対し，1) から 5) を繰り返す．

7) これをデータリストが尽きるまで繰り返す．

8) Proc print data=d01 ； run; により SAS データセット d01 の 60 行分が表 9.9 に示すように出力される．

表 9.9　作成された SAS データセットを proc print で出力した結果

Obs	Habitat	Region	No	FlowerNo	SeedNo	ln_FloweNo
1	Dry	A	1	3	57	1.09861
2	Dry	A	2	1	19	0.00000
3	Dry	A	3	1	11	0.00000
4	Dry	A	4	1	12	0.00000
:						
60	Wet	C	10	3	101	1.09861

SAS の GENMOD プロシジャは，ありとあらゆる一般化線形モデルによる解析が行なえるようになっている．JMP の一般化線形モデルは，GENMOD プロシジャに限らず 10 分の 1 程度の解析機能に絞られているが，GUI 操作で扱えるグラフ表示に優れている．

```
Title2 '<<< poisson 過分散 >>>' ;
proc genmod data=d01  ;
     class  Habitat  Region / param=effect ;
     model  SeedNo = Habitat  Region  FlowerNo
            / dist=poisson  link=log  scale=Pearson  type3;
     run ;
```

9) Proc genmod data=d01 ； は，SAS データセット d01 に対しての解析の指示である．

10) Class Habitat Region は，名義尺度変数として扱うことを宣言している．

11) Param=effect ； は，(1, -1) 対比型の変数の内部生成を指示しているオプションである．

12) なお，Param オプションにより名義尺度の最初の水準を基準として，他の水準との「差」が得られるような (0, 1) 型ダミー変数を生成することも可能である．ディフォルトは最後の水準を基準とする (1, 0) 型のダミー変数となる．

13) Model SeedNo = Habitat Region FlowerNo; により回帰式を設定し，後のオプションで各種の解析モデルを規定する．

14) dist=poisson によりポアソン回帰を設定している．JMP ではサポートされていない（負の 2 項分布，ゼロ過剰ポアソン，ゼロ過剰負の 2 項分布）なども扱える．

15) Link=log は対数リンクの設定である．

16) Scale=Pearson は，過分散の調整にピアソン残差を使う指示で，SAS では，デビエンス残差を指定することもできる．

17) Type3 オプションは，JMP と同様の「効果の検定」を出力するためのオプションである．

表 9.10 に示すのは，JMP と同様の「効果の検定」であり，表 9.7 に示した JMP の結果に相当する．JMP の出力に加えて，誤差の自由度を考慮した F 検定が追加されている．

表 9.10　SAS による JMP と同様の「効果の検定」

Type 3 分析の LR 統計量						
要因	分子の自由度	分母の自由度	F 値	Pr > F	カイ 2 乗	Pr > ChiSq
Habitat	1	55	28.94	<.0001	28.94	<.0001
Region	2	55	2.89	0.0639	5.79	0.0554
FlowerNo	1	55	171.82	<.0001	171.82	<.0001

表 9.11 に示す「パラメータの推定値」は，表 9.7 に示した JMP の結果に相当する．大きく異なるのは，SAS では推定値の標準誤差を用いた Wald 検定となっていて，JMP の尤度比検定の結果とは微妙に異なることに注意が必要である．過分散の調整ための推定値は，「尺度（形状と同義語）」の行の推定値の欄の 1.8956 である．JMP では，表 9.7 の「適合度の統計量」の過分散 3.5934 の平方根 1.8956 に一致する．

表 9.11　SAS によるポアソン回帰の対比型デザイン変数による推定値

最大尤度パラメータ推定値の分析								
パラメータ		自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept		1	2.6197	0.1080	2.4080	2.8314	588.43	<.0001
Habitat	Dry	1	−0.1984	0.0373	−0.2716	−0.1253	28.24	<.0001
Region	A	1	−0.1016	0.0526	−0.2047	0.0014	3.74	0.0532
Region	B	1	0.1126	0.0505	0.0137	0.2116	4.98	0.0257
FlowerNo		1	0.5160	0.0404	0.4368	0.5951	163.25	<.0001
尺度		0	1.8956	0.0000	1.8956	1.8956		

Note:尺度パラメータは Pearson カイ 2 乗/DOF の平方根により推定されています。

SAS の GENMOD プロシジャでは，これまで示してきた予測プロファイル，交互作用プロファイルの作図が行なえないのは残念なことであるが，表 9.12 に示すように covb オプションでパラメータの共分散行列を得ることができるので，Excel に取り込んで予測プロファイルおよび交互作用プロファイルを作成する必要がある．

表 9.12　GENMOD プロシジャによるパラメータの共分散行列（covb オプション）

推定値の共分散行列					
	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.01166	−0.000243	−0.000069	−0.000764	−0.004096
Prm2	−0.000243	0.001394	0.0000101	9.4424E−6	0.0002299
Prm3	−0.000069	0.0000101	0.002764	−0.001375	0.0000977
Prm4	−0.000764	9.4424E−6	−0.001375	0.002549	0.0002708
Prm5	−0.004096	0.0002299	0.0000977	0.0002708	0.001631

このパラメータの共分散行列は，JMP の出力を Excel に取り込んだ表 9.8 の結果と一致する．SAS の GENMOD プロシジャの Lsmeans ステートメントなどの各種の推定機能を発揮させるためには，デザイン変数の設定を（1，–1）対比型からデフォルトの最後の水準を基準とする（1，0）型でないと利用できない．また，表 9.8 および図 9.6 で示した予測プロファイルは，共変量としての FlowerNo も含めている．SAS の GENMOD プロシジャの推定は，質的因子の各水準を対象としているために，共変量についての推定は対象外である．

ただし，表 9.13 に示すように推定したい（生育環境，地域，花数）のカテゴリを 61 行目以後に追加し，output ステートメントで，表 9.14 に示すように推定値および 95%信頼区間を出力することができる．この出力を用いて予測プロファイルを作成することはできるが，パラメータの共分散行列を使って Excel で計算する方が生産的である．ただし，表 9.8 の中段に示した Excel での計算を検証するためには有益である．

表 9.13　推定したい予測プロファイルデータの追加

Obs	Habitat	Region	No	FlowerNo	SeedNo	ln_FloweNo
:						
60	Wet	C	10	3	101	1.09861
61	Wet	B	1	1	.	0
62	Wet	B	1	2	.	0.69315
63	Wet	B	1	3	.	1.09861
64	Wet	B	1	4	.	1.38629

表 9.14　追加データに対する予測値と 95%信頼区間の計算結果（表 9.8 の中ほどに対応）

Obs	Habitat	Region	No	FlowerNo	SeedNo	ln_FloweNo	y_est	ln_y_est	L95	U95
61	Wet	B	1	1	.	0.00000	31.397	3.44673	26.330	37.440
62	Wet	B	1	2	.	0.69315	52.598	3.96269	45.890	60.287
63	Wet	B	1	3	.	1.09861	88.115	4.47865	76.831	101.057
64	Wet	B	1	4	.	1.38629	147.615	4.99461	123.619	176.269

9.4. 花数をオフセットとしたポアソン回帰

ポアソン回帰でのオフセットは，第 1.5 節の冠動脈心疾患の死亡者数で示したように，年齢階層別の部分母集団の大きさが人口統計学的に既知である場合に，稀に発生する死亡者数を統計モデルによって調整するためである．共変量は，反応変数に明らかに影響を及ぼすことが分かっているが，事前にコントロールできない変数であり，種子数に対して花数は，共変量と位置づけられる．花数をオフセットとしても扱えるが，単位花数に換算した場合には，2×3 の要因配置型のポアソン回帰となり，花数と種子数の関連が見えにくくなる．

下野（2010）と同様に，過分散を無視し，花数の対数をオフセットとした JMP によるポアソン回帰を行い結果の吟味を行う．表 9.15 に示すように地域は（A，B，C）の 3 水準なので，（0，1）型ダミー変数とし，*RegionB* 変数は「B」の場合に 1，それ以外は 0，*RegionC* 変数は「C」の場合に 1，それ以外は 0，生育環境は（Dry，Wet）の 2 水準なので，*HabitatWet* 変数は「Wet」の場合に 1，それ以外は 0 とする．

表 9.15　最初の水準を基準とする（0，1）型ダミー変数（デザイン変数）

Region	*RegionB*	*RegionC*		Habitat	*HabitatWet*
A	0	0		Dry	0
B	1	0		Wet	1
C	0	1			

モデル式は

$$y_i = Flower_i \cdot \exp[\beta_0 + \beta_1 \cdot (RegionB_i) + \beta_2 \cdot (RegionC_i) + \beta_3 \cdot (HabitatWet_i)] + \varepsilon_i$$

ただし，　$\varepsilon_i \sim Poisson$分布

であり，両辺に対数を取り推定値の形式として

$$\ln \hat{y}_i = \ln(Flower_i) + \hat{\beta}_0 + \hat{\beta}_1 \cdot (RegionB_i) + \hat{\beta}_2 \cdot (RegionC_i) + \hat{\beta}_3 \cdot (HabitatWet_i)$$

であり，表 9.16 に過分散を無視した結果を示す．推定値に対する標準誤差は，過分散が全く考慮されておらず，*RegionC* の推定値は，*RegionA*（切片）との対数での差で $\hat{\beta}_2 = 0.1352$　に対

表 9.16　過分散を無視したポアソン回帰の結果

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値	下側信頼限界	上側信頼限界
切片	2.7422	0.0425	2463.6554	<.0001*	2.6581	2.8247
RegionB	0.1903	0.0473	16.2424	<.0001*	0.0976	0.2831
RegionC	0.1352	0.0471	8.2693	0.0040*	0.0430	0.2277
HabitatWet	0.4390	0.0389	130.8425	<.0001*	0.3629	0.5156

して $p=0.0040$ と有意な差となっている．ところが，図 9.6 で Wet における *RegionA* と *RegionC* を比べても，平均的な差はデータの変動に比べても有意な差とは，とても見えない．これは，ポアソン回帰の場合に分散を推定値と同じと見なしているために，過分散を全く考慮しないためである．

表 9.17 に示すように，適合度統計量の Pearson カイ 2 乗値は 183.5650 となり，自由度 56 のカイ 2 乗分布から $p<0.0001$ となり，過分散であることが分かる．Pearson カイ 2 乗を自由度で割った過分散の係数 $\varphi=3.2779$ を考慮した結果が示されている．結果として RegionC の RegionA に対する差の標準誤差は，

$$SE=\sqrt{3.2779\times 0.0471}=0.0853$$

と 2 倍弱大きくなり $p=0.1122$ と有意な差ではなくなっている．

表 9.17　過分散を考慮したポアソン回帰の結果

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	22.5930	45.1861	3	<.0001*
完全	77.8913			
縮小	100.4843			

適合度統計量	カイ2乗	自由度	p値	過分散
Pearson	183.5650	56	<.0001*	3.2779
デビアンス	181.0588	56	<.0001*	

AICc
166.8937

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値	下側信頼限界	上側信頼限界
切片	2.7422	0.0770	751.5849	<.0001*	2.5887	2.8905
RegionB	0.1903	0.0857	4.9550	0.0260*	0.0227	0.3587
RegionC	0.1352	0.0853	2.5227	0.1122	-0.0316	0.3029
HabitatWet	0.4390	0.0705	39.9160	<.0001*	0.3016	0.5780

推定値は，対数変換された結果であり，このままでは見通しが悪いので，Excel の行列関数を用いて生育環境別 地域別 花数別 の対数での推定値を計算し，指数を取って元のスケールに戻す必要がある．表 9.18 に示すように，対数推定値の Excel での計算は，$\ln\hat{y}_{\mathrm{dry,A,1}}$ の場合，$RegionB=0$，$RegionC=0$，$HabitatWet=0$ なので，

$$\begin{aligned}\ln\hat{y}_{\mathrm{dry,A,1}}&=\ln(1)+2.7422+0.1903\times 0+0.1352\times 0+0.4390\times 0\\&=2.7422\end{aligned}$$

であり，指数を取って

$$\hat{y}_{dry,A,1} = \exp(2.7422) = 15.52$$

が計算されている．表の中の花数は，「1」としているが，Excel シート上では，これを 2 および 3 変えれば他の花数についての計算が自動的に得られるようにしてあり，表 9.20 はこれを用いた表である．

表 9.18　各種の推定値の計算

		offset		Region		Habitat					
生育		**花数**	切片	B	C	Wet	対数	対数	指数		
環境	地域	**1**	2.7422	0.1903	0.1352	0.4390	推定値	分散	推定値	*L*95%	*U*95%
Dry	A	**1**	**1**	**0**	**0**	**0**	**2.7422**	**0.0059**	**15.52**	13.35	18.05
	B	**1**	1	1	0	0	2.9325	0.0052	18.77	16.29	21.63
	C	**1**	1	0	1	0	2.8774	0.0052	17.77	15.42	20.47
Wet	A	**1**	1	0	0	1	3.1812	0.0047	24.08	21.04	27.55
	B	**1**	1	1	0	1	3.3715	0.0040	29.12	25.71	32.99
	C	**1**	1	0	1	1	3.3164	0.0039	27.56	24.37	31.17

表 9.19　パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ （JMP の出力結果を整形）

共分散	切片	RegionB	RegionC	HabitatWet
切片	**0.005924**	-0.004017	-0.003987	-0.003078
RegionB	-0.004017	**0.007338**	0.004017	0.000000
RegionC	-0.003987	0.004017	**0.007273**	-0.000049
HabitatWet	-0.003078	0.000000	-0.000049	**0.004968**

育成環境が Dry で地域が A の対数推定値は，2.7422 であり，その分散は，デザイン変数のベクトルを $\boldsymbol{x}_{Dry,A,1} = [1 \quad 0 \quad 0 \quad 0]$ とし，パラメータの共分散行列を $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ としたときに，次の 2 次形式で

$$Var(\ln \hat{y}_{\mathrm{dry,A,1}}) = \boldsymbol{x}_{\mathrm{Dry,A,1}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}^{T}_{\mathrm{Dry,A,1}}$$

$$= \mathrm{Mmult}(\mathrm{Mmult}(\boldsymbol{x}_{\mathrm{Dry,A,1}}\text{の範囲}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{の範囲}), \mathrm{Transpose}(\boldsymbol{x}_{\mathrm{Dry,A,1}}\text{の範囲}))$$

=	1	0	0	0	**0.005924**	-0.004017	-0.003987	-0.003078	1	=	0.0059
					-0.004017	**0.007338**	0.004017	0.000000	0		
					-0.003987	0.004017	**0.007273**	-0.000049	0		
					-0.003078	0.000000	-0.000049	**0.004968**	0		

と計算されている．指数推定値

$$\hat{y}_{\mathrm{dry,A,1}} = \exp(\ln \hat{y}_{\mathrm{dry,A,1}})$$

$$= \exp(2.7422) = 15.52$$

に対して，95%信頼区間は，

$$95\%\,\mathrm{CI} = \exp\left[\ln \hat{y}_{\mathrm{dry,A,1}} \pm 1.96\sqrt{Var(\ln \hat{y}_{\mathrm{dry,A,1}})}\right]$$
$$= \exp\left[2.7422 \pm 1.96\sqrt{0.0059}\right] = (13.35, 18.05)$$

として計算されている．

花数=1 の場合について，Excel の「折れ線グラフ」を用いて地域 A に対する生育環境の折れ線を描き，「データの選択」により地域 B および地域 C を上書きした結果を図 9.8（左）に示す．表 9.18 の花数の欄を「花数=2」および「花数=3」に変更して得られた折れ線グラフを図 9.8（中），図 9.8（右）に示す．このように何らかの統計モデルで推定された結果についてグラフ化することは，結果の解釈するために役に立つ．

表 9.20　花数を変え場合の推定値

		offset		offset		offset	
生育		**花数**	指数	**花数**	指数	**花数**	指数
環境	地域	**1**	推定値	**2**	推定値	**3**	推定値
Dry	A	**1**	15.52	**2**	31.04	**3**	46.56
	B	**1**	18.77	**2**	37.55	**3**	56.32
	C	**1**	17.77	**2**	35.54	**3**	53.30
Wet	A	**1**	24.08	**2**	48.15	**3**	72.23
	B	**1**	29.12	**2**	58.24	**3**	87.37
	C	**1**	27.56	**2**	55.12	**3**	82.68

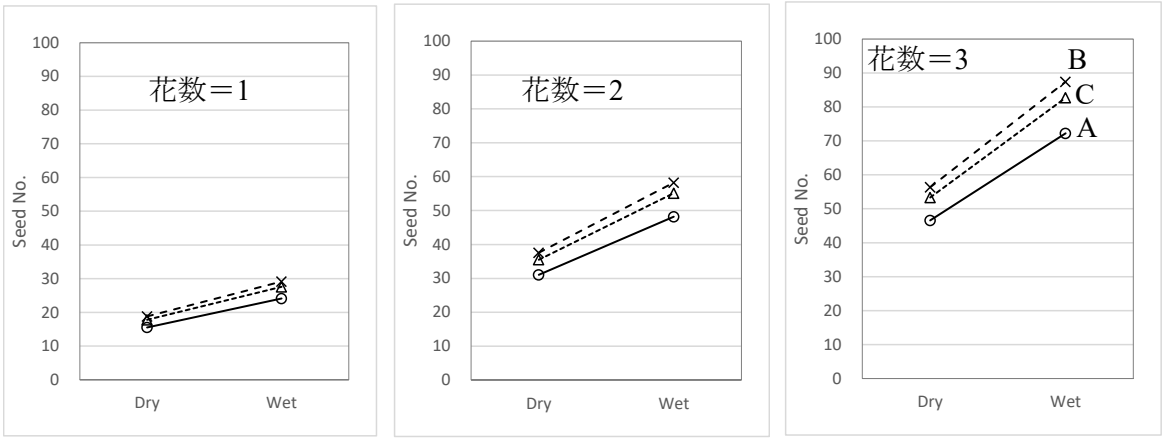


図 9.8　花数オフセットを考慮した組み合わせプロファイル

9.5. 花数をオフセットとした負の 2 項回帰への適用

下野（2010）は，R の glm.nb()関数による負の 2 項分布をあてはめた一般化線形モデルの解析結果も示しているので，Excel での負の 2 項回帰および SAS の GENMOD プロシジャを使い，結果の吟味を行う．

負の 2 項分布のパラメータ変換

負の 2 項分布は，第 6.1 節で示したように，成功の確率を π としたときに一定の成功数 k が得られるまでの試行を行った場合の失敗の数 Y の分布として定義されている．ポアソン分布のパラメータは，平均 μ であり過分散を考慮した場合に，負の 2 項回帰が適していると言っても理解に苦しむのではないだろうか．

第 6.1 節では，負の 2 項分布の整数で定義されている成功数 k を実数化し，組合せ数の計算で用いられている階乗をガンマ関数に置き換え，成功の確率 π を期待値 μ と k で

$$\pi = \frac{k}{\mu + k}, \quad k = \frac{1}{\sigma}$$

置き換え，ポアソン分布のパラメータの期待値 μ と整合性が取れるようにする．さらに，k を $k = 1/\sigma$ で置き換えることにより過分散の（形状）パラメータ σ が定義される．これにより，ポアソン回帰に後付けで過分散の係数でパラメータの標準誤差を割り増しする便宜的な方法に代え，回帰パラメータとして推定が可能となる．

位置パラメータを μ，形状（尺度）パラメータ σ を用いて負の 2 項分布（ガンマ・ポアソン分布）を表すことができることを式（6.6）で示した．ガンマ・ポアソン分布の計算で，σ が小さくなると $\Gamma(1/\sigma)$ が増大するために計算不能となるので，式（6.11）に示した Excel の対数ガンマ関数 Gammaln() 関数を使って，計算する必要がある．

$$GammaPoisson(y;\mu,\sigma) = \frac{\Gamma(y+1/\sigma)}{\Gamma(y+1)\Gamma(1/\sigma)} \cdot \frac{(\mu\sigma)^y}{(1+\mu\sigma)^{y+1/\sigma}}$$

$$= \exp\left[\ln\Gamma(y+1/\sigma) - \ln\Gamma(y+1) - \ln\Gamma(1/\sigma) + y\ln(\mu\sigma) - (y+1/\sigma)\ln(1+\mu\sigma)\right]$$

従って，一般化線形モデルで分布を負の 2 項分布に設定した場合には，過分散を考慮したガンマ・ポアソン回帰を適用したことに対応する．

負の 2 項回帰で，分布を負の 2 項分布に設定したといっても，元々のパラメータとしての成功確率 π，成功数 k としたときの失敗数 Y の分布との理解することに違和感を持つ．そもそも成功数 k は，整数として定義されていて Excel の負の 2 項分布 NegBinom.dist() 関数で

$k=1/\sigma$ などの実数での計算は不能である．位置パラメータを μ，形状（尺度）パラメータを σ としたときの負の 2 項分布に代わる分布としてガンマ・ポアソン分布が知られている．したがって，負の 2 項回帰よりも，ガンマ・ポアソン回帰とすべきと思うが，これらが混在して使われている．

負の 2 項回帰の更なる応用に関しては，南・Lennert-Cody（2013），「ゼロの多いデータの解析：負の 2 項回帰モデルによる傾向の過大推定」を参照のこと．

Excel によるポアソン回帰

Excel により，一足飛びにガンマ・ポアソン回帰を行う前に，最初にポアソン回帰を行うことにする．その Excel シートに追加変更をし，負の 2 項分布を仮定したガンマ・ポアソン回帰を行う．第 7.3 節の事例で示したように，対数尤度の推定結果により AIC あるいは AICc を求め，モデルのあてはまりについて比較検討も容易にできる．

表 9.1 に示したデータリストを表 9.21 に示すように行方向に展開し，表 9.15 に示した最初の水準を基準とするデザイン変数とする．Region は，［A：（0，0），B：（1，0），C：（0，1）］とする（0，1）型ダミー（デザイン）変数（*RegioB*，*RegioC*）とし，Habitat は，（Dry : 0，Wet : 1）とするダミー変数 *HabitatWet* とし，切片 1 を含めてデザイン行列を生成する．

対数推定値を

$$\begin{aligned}\ln \hat{y}_i &= \ln(FlowerNo_i) + \boldsymbol{x}_i\hat{\boldsymbol{\beta}} \\ &= \ln(FlowerNo_i) + \hat{\beta}_0 + \hat{\beta}_1(RegionB_i) + \hat{\beta}_2(RegionC_i) + \hat{\beta}_3(HabitatWet_i)\end{aligned}$$

で計算する．指数推定値を $\hat{y}_i = \exp(\ln \hat{y}_i)$ で計算し，$y_i = SeedNo_i$ としてポアソン確率を Excel の関数

$$\begin{aligned}P_i &= \frac{\hat{y}_i^{\,y_i} e^{-\hat{y}_i}}{y_i!} \\ &= \text{Poisson.Dist}(y_i, \hat{y}_i, false)\end{aligned}$$

を使い計算する．適当に設定した切片モデル $\hat{\boldsymbol{\beta}}^T = [3 \quad 0 \quad 0 \quad 0]^T$ を初期値に与え，対数尤度 $\ln L$ を計算すると

$$\ln L = \sum_{i=1}^{60} \ln(P_i) = -344.2485$$

となるので，$\ln L$ を最大化するように $\hat{\boldsymbol{\beta}}^T = [3 \quad 0 \quad 0 \quad 0]^T$ を変化させると，

$$\hat{\boldsymbol{\beta}}^T = [2.7422 \quad 0.4390 \quad 0.1903 \quad 0.1352]^T$$

画得られ，表 9.17 に示した JMP で推定されてパラメータに一致する．

過分散パラメータを推定するためにピアソンのカイ 2 乗の計算を

$$\text{Pearson}カイ2乗 = \sum_{i=1}^{60} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} = 183.5651$$

計算し，自由度（60－4）で割り，過分散パラメータ 3.2779 を計算する．ここまでが，ガンマ・ポアソン回帰を Excel で行うための準備であり，表 9.17 の JMP での結果に一致する．

表 9.21　ポアソン回帰による過分散パラメータの推定

						$\hat{\beta}_0$		切片	**2.7422**	変化させるセル			過分散
						$\hat{\beta}_1$	*RegionB*		**0.1903**			対数尤度	**3.2779**
						$\hat{\beta}_2$	*RegionC*		**0.1352**			$\ln L$	P.カイ2乗
						$\hat{\beta}_3$	*HabitatWet*		**0.4390**		最大化	**-255.3235**	183.5650
No	Habitat	Region	切片	*RegionB*	*RegionC*	*HabitatWet*	*FlowerNo*	*SeedNo*	対数推定値	指数推定値	*Poisson* 確率	対数尤度 i	Pearson カイ2乗
1	Dry	A	1	0	0	0	3	57	3.8408	46.56	0.0177	-4.0331	2.3398
2	Dry	A	1	0	0	0	1	19	2.7422	15.52	0.0633	-2.7593	0.7799
3	Dry	A	1	0	0	0	1	11	2.7422	15.52	0.0573	-2.8591	1.3168
4	Dry	A	1	0	0	0	1	12	2.7422	15.52	0.0741	-2.6018	0.7987
5	Dry	A	1	0	0	0	3	46	3.8408	46.56	0.0585	-2.8385	0.0068
：													
58	Wet	C	1	0	1	1	1	14	3.3164	27.56	0.0018	-6.3222	6.6719
59	Wet	C	1	0	1	1	4	129	4.7027	110.24	0.0077	-4.8621	3.1922
60	Wet	C	1	0	1	1	3	101	4.4150	82.68	0.0060	-5.1216	4.0590

Excel によるガンマ・ポアソン回帰（負の 2 項回帰）

ポアソン回帰のための Excel の計算シートがあれば，若干の追加でガンマ・ポアソン回帰の計算シートを作成することができる．過分散パラメータ σ を追加し，ポアソン確率の計算をガンマ・ポアソン確率に変更する．なお，ガンマ・ポアソン確率は，

$$GammaPoisson(y_i; \hat{y}_i, \sigma)$$
$$= \exp\left[\ln\Gamma(y_i + 1/\sigma) - \ln\Gamma(y_i + 1) - \ln\Gamma(1/\sigma) + y_i \ln\left(\hat{y}_i\sigma\right) - (y_i + 1/\sigma)\ln\left(1 + \hat{y}_i\sigma\right)\right]$$

で与えられる．

ピアソンのカイ 2 乗値は，分散がポアソン分布の場合は $Var(\hat{y}_i) = \hat{y}_i$ であるが，ガンマ・ポアソン分布の場合は，式（6.10）に示したように

$$Var(\hat{y}_i) = \hat{y}_i(1 + \hat{y}_i\hat{\sigma})$$

となるので，分母を

$$\text{Pearson}カイ2乗 = \sum_{i=1}^{60} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i(1 + \hat{y}_i\hat{\sigma})}$$

のように変更する．

初期値を適当に設定した切片モデル $[\hat{\boldsymbol{\beta}} \quad \hat{\sigma}]^T = [3.00 \quad 0.00 \quad 0.00 \quad 0.00 \quad 0.10]^T$ とすると，対数尤度 $\ln L$ は，

$$\ln L^{GP} = \sum_{i=1}^{60} \ln(P_i^{GP}) = -246.1822$$

となるので，$\ln L$ 最大すするように $[\hat{\boldsymbol{\beta}} \quad \hat{\sigma}]^T$ を変化させると

$$[\hat{\boldsymbol{\beta}} \quad \hat{\sigma}]^T = [2.7477 \quad 0.4530 \quad 0.1629 \quad 0.0927 \quad 0.0429]^T$$

と，表 9.22 に示すような結果を得る．対数尤度は，$\ln L = -226.2417$ とポアソン回帰の場合に比べ 29.0818 大きくなっている．ピアソンのカイ 2 乗も 60.6644 となり，自由度とほぼ同じ大きさとなり，あてはまりは極めてよい．

表 9.22　ガンマ・ポアソン（負の 2 項）回帰

					$\hat{\beta}_0$		切片=	**2.7477**	変化させるセル				
					$\hat{\beta}_1$		*RegionB*=	**0.1629**					
					$\hat{\beta}_2$		*RegionC*=	**0.0927**				対数尤度	
					$\hat{\beta}_3$		*HabitatWet*=	**0.4530**				$\ln L$	カイ2乗
							σ^=	**0.0429**			最大化	**-226.2417**	60.6644
No	*Habitat*	*Regio*	切片	*RegionB*	*RegionC*	*HabitatWet*	*FlowerNo*	*SeedNo*	対数推定値	指数推定値	ガンマ *Poisson*	対数尤度 *i*	Pearson カイ2乗
1	Dry	A	1	0	0	0	3	57	3.8463	46.82	0.0204	-3.8917	0.7363
2	Dry	A	1	0	0	0	1	19	2.7477	15.61	0.0552	-2.8961	0.4422
3	Dry	A	1	0	0	0	1	11	2.7477	15.61	0.0611	-2.7946	0.8146
4	Dry	A	1	0	0	0	1	12	2.7477	15.61	0.0701	-2.6577	0.4993
5	Dry	A	1	0	0	0	3	46	3.8463	46.82	0.0339	-3.3844	0.0048
:													
58	Wet	C	1	0	1	1	1	14	3.2935	26.94	0.0120	-4.4258	2.8833
59	Wet	C	1	0	1	1	4	129	4.6797	107.74	0.0098	-4.6245	0.7465
60	Wet	C	1	0	1	1	3	101	4.3921	80.81	0.0105	-4.5603	1.1304

SAS の GENMOD プロシジャによる負の 2 項分布を用いた場合

SAS の GENMOD プロシジャで，名義尺度の最初の水準を基準とするデザイン変数を作成のために，class ステートメントのオプションを param=ref ref=first と設定する．また，負の 2 項分布の設定は，dist=negbin によって設定できる．

```
Title2 '<<< 負の 2 項分布 offset >>>' ;
proc genmod data=d01  ;
     class Habitat Region / param=ref ref=first ;
     model SeedNo = Region Habitat
          / dist=negbin link=log  type3  offset=ln_FloweNo  covb ;
     run ;
```

SAS による実行結果を次に示す．

<<< 負の 2 項分布 offset >>>
GENMOD プロシジャ

モデルの情報	
データセット	WORK.D01
分布	Negative Binomial
リンク関数	Log
従属変数	SeedNo
オフセット変数	ln_FloweNo
読み込んだオブザベーション数	60
使用されたオブザベーション数	60

分類変数に対してデザイン変数が，指示通り最初の水準を基準とした変数になっていることが確認できる．

分類変数の水準の情報			
分類	値	デザイン変数	
Habitat	Dry	0	
	Wet	1	
Region	A	0	0
	B	1	0
	C	0	1

各種の適合度の基準が出力されている．Pearson のカイ 2 乗は，60.6645 と Excel の結果 60.6647 と計算誤差範囲である．完全対数尤度は，–226.2417 と一致している．

適合度評価の基準			
基準	自由度	値	値/自由度
デビアンス	56	61.2206	1.0932
Scaled デビアンス	56	61.2206	1.0932
Pearson カイ 2 乗	56	60.6645	1.0833
Scaled Pearson カイ 2 乗	56	60.6645	1.0833
対数尤度		8440.6031	
完全対数尤度		−226.2417	
AIC (小さいほどよい)		462.4833	
AICC (小さいほどよい)		463.5945	
BIC (小さいほどよい)		472.9551	

パラメータの推定値，標準誤差などが出力されている．表 9.22 に示した Excel の推定値と一致している．Dispersion の推定値 0.0429 は，Excel の過分散パラメータ σ に一致している．なお，R でのパラメータは，σ ではなく，$k = 1/\sigma = 1/0.0429 = 23.3289$ に対応する．

最大尤度パラメータ推定値の分析								
パラメータ		自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept		1	2.7477	0.0699	2.6106	2.8847	1544.19	<.0001
Region	B	1	0.1629	0.0831	−0.0000	0.3258	3.84	0.0500
Region	C	1	0.0927	0.0833	−0.0705	0.2560	1.24	0.2655
Habitat	Wet	1	0.4530	0.0678	0.3202	0.5859	44.66	<.0001
Dispersion		1	0.0429	0.0124	0.0243	0.0757		

Note:負の 2 項の分散性パラメータは最尤法により推定されました。

Type3 の対数尤度は，Region の場合は，全体の尤度から Region を除いた（切片，Habitai）での尤度との差から計算され，全体の尤度から Habitat を除いた（切片，Region）での尤度との差から計算されている．

Type 3 分析の LR 統計量			
要因	自由度	カイ 2 乗	Pr > ChiSq
Region	2	3.71	0.1565
Habitat	1	33.87	<.0001

パラメータの共分散行列を covb オプションで出力している．この行列を Excel に取り込むことにより各種の推定が行なえ，種々のグラフの作成を可能にする．パラメータ推定値の Habitat の標準誤差 0.0678 の 2 乗が，Pram4 の対角要素 0.004596 になる．

推定値の共分散行列					
	Prm1	Prm2	Prm3	Prm4	Dispersion
Prm1	0.004889	−0.003502	−0.003589	−0.002467	−4.521E−6
Prm2	−0.003502	0.006910	0.003568	−0.000116	−0.000013
Prm3	−0.003589	0.003568	0.006937	0.0000375	−0.000065
Prm4	−0.002467	−0.000116	0.0000375	0.004596	0.0000318
Dispersion	−4.521E−6	−0.000013	−0.000065	0.0000318	0.0001544

下野の R の glm.nb による結果

下野（2010）で示されている R の glm.nb による結果を次に引用し，Excel および SAS での結果と対比する．

Excel と SAS での結果と照合を試みる．Deviance Residuals は，-2.1987～2.5954 の範囲なので良くあてはまっていることが分かる．SAS では，

```
output out=out1    stdresdev=stdresdev ;
```

をステートメントで挿入し，out1 ファイルを用いて再計算することで結果が得られる．結果の表示は，R の方がコンパクトである．

```
Call:
glm.nb(formula = SeedNo ~ Region + Habitat + offset(log(FlowerNo)),
            data = seed, init.theta = 23.32892933, link = log)

Deviance Residuals:
                  Min         1Q     Median         3Q        Max
               -2.1987    -0.7527    -0.0828     0.6518     2.5954

Coefficients:
              Estimate  Std.Error    z value   Pr(>|z|)
(Intercept)    2.74768    0.07013     39.178    < 2e-16 ***
RegionB        0.16289    0.08303      1.962     0.0498 *
RegionC        0.09274    0.08335      1.113     0.2658
HabitatWet     0.45303    0.06787      6.675   2.47e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(23.3289) family taken to be 1)

    Null deviance: 110.540  on 59  degrees of freedom
Residual deviance:  61.221  on 56  degrees of freedom
AIC: 462.48

Number of Fisher Scoring iterations: 1

            Theta:          23.33
            Std. Err.:       6.74

             2 x log-likelihood:  -452.483
```

下野（2010）：第 6 図 負の 2 項分布を当てはめた一般化線形モデルの解析結果を引用.

パラメータの推定値は，一致しているが，標準偏差は，小数点 4 桁目で計算誤差が見いだされる．この原因は，収束の判定基準の設定が，統計ソフト間で異なっていることの反映であろう．

R の Dispersion parameter for Negative Binomial (23.3289) は，SAS の Dispersion=0.0429 の逆数に一致する．これは，SAS では，負の二項分布のパラメータ k を $k=1/\sigma$ と置き換えているのに対し，R では，パラメータとして実数とした k を用いているためである．Null devience: 110.540 は SAS で見いだせないが，Residual devience:61.221 は，SAS の「適合度評価の基準」の中の Scaled デビアンスに一致する．AIC:462.48 は，SAS の AIC（小さいほどよい):462.4833 に一致する．また，2×log-liklifoof：-452.483 は，「完全対数尤度」 -226.2417 の 2 倍に等しい.

R でも SAS などの統計ソフトで出力された結果に基づいて何らかの考察をしようと思ったときに，統計ソフトの中でどのようなデザイン変数（ダミー変数）が生成されているのかを把握し，出力結果を Excel に取り込み，対数リンクの場合であれば指数を計算するなどの計算を行い，グラフ化することが望まれる．

デビアンス（Deviance）

R の出力に，3 箇所の Deviance が登場する．最初は，「Deviance Residuals」で，モデルのあてはまりの検討のために，四分位統計量などが出力されている．最後に「Residual deviance」が現われる．その前には，「Null deviance」があり，これらが何を意味しているのか，説明だけ聞いてもなかなか理解しづらい．第 11 章で「デビアンス・逸脱度・テコ比・4 種の残差」についてあらためて取り上げる．第 1.9 節の「デビアンス」をベースに若干の解説を加えると，「Null deviance」は，切片のみのモデル「縮小モデル」と「完全モデル」の対数尤度の「差分」に対応している．通常の回帰分析の分散分析表では，「回帰」の F 値に対応する．

通常の回帰分析では，回帰直線のあてはまりの検討に「残差プロット」が使われる．Excel の分析ツールの回帰分析に「残差のグラフ作成」のほかに「標準化された残差」などもあるが，Deviance Residuals（デビアンス残差）とは，いったい何なのだろうか．通常の残差ならば，観測データと推定値の差であることは自明であるが，「標準化された残差」も何となく理解できる範囲ではあるが，「デビアンス残差」は，残差らしきものらしいが，「デビアンス」とは，何なのか．「逸脱度」と言われても，ますます理解に苦しむことになる．なぜ「通常の残差」ではいけないのだろうか．「デビアンス残差」の計算方法を通じて理解しなければ，身につかない．

ピアソン残差は，観測データ $y_1 = 57$ に対し，推定値 $\hat{y}_1 = 46.82$ の差に対し，$Var(\hat{y}_1)$ の平方根で除した

$$\text{Pearson残差}_1 = \frac{y_1 - \hat{y}_1}{\sqrt{\hat{y}_1(1+\hat{y}_1\hat{\sigma})}} = \frac{57 - 46.82}{\sqrt{46.82 \times (1 + 46.82 \times 0.0429)}} = 0.8581$$

として定義されており，この 2 乗が Pearson カイ 2 乗 0.7363 であるが，ポアソン分布は右に裾を引いているので，推定値より小さい場合に残差が過小評価されやすくなる欠点が内在する．逆に，推定値より大きい場合には，過大評価となる．デビアンス残差は，観測データ $y_1 = 57$ に対し，推定値も $\hat{y}_1 = 57$ とした飽和モデルの場合の対数尤度と，完全モデルでの対数尤度の差に着目した方法である．表 9.23 に各種のモデルの場合の対数尤度の計算結果を示す．

表 9.23　Devianse Residuals vs. Residual deviance

							$\hat{\beta}_0$		切片	**2.7477**		2.7477			**46.8333**	
							$\hat{\beta}_1$	*RegionB*		**0.1629**		0.1629			0.1629	*Null*
							$\hat{\beta}_2$	*RegionC*		**0.0927**		0.0927		*R. dev.*	0.0927	切片
							$\hat{\beta}_3$	*HabitatWet*		**0.4530**		0.0927	$y_i\hat{}=y_i$	残差 デ	0.0927	縮小
								σ		**0.0429**	完全	0.0429	飽和	ビアンス	**0.3460**	モデル
									$\ln L^{完全}$	**-226.24**	$\ln L^{飽和}$	**-195.63**		**61.2208**	$\ln L^{縮小}$	**-278.22**
No	*H.*	*R.*	切片	*R. B*	*R. C*	*H. W*	*F. No*	*S. No*	指数推定値	ガンマ *Poisson*	完全	ガンマ *Poisson*	飽和	デビアンス 残差	ガンマ Poisson	縮小
								y_i	$y_i\hat{}$	$P^{完全}$	$\ln L_i^{完全}$	$P^{飽和}$	$\ln L_i^{飽和}$	*D. Res.*		$\ln L_i^{縮小}$
1	Dry	A	1	0	0	0	3	**57**	46.82	0.02	**-3.89**	0.03	**-3.56**	**0.8112**	0.0107	**-4.54**
2	Dry	A	1	0	0	0	1	19	15.61	0.06	-2.90	0.07	-2.70	0.6341	0.0142	-4.25
3	Dry	A	1	0	0	0	1	11	15.61	0.06	-2.79	0.10	-2.32	-0.9745	0.0090	-4.71
4	Dry	A	1	0	0	0	1	12	15.61	0.07	-2.66	0.09	-2.38	-0.7492	0.0098	-4.63
5	Dry	A	1	0	0	0	3	46	46.82	0.03	-3.38	0.03	-3.38	-0.0694	0.0139	-4.28
58	Wet	C	1	0	1	1	1	14	26.94	0.01	-4.43	0.08	-2.48	-1.9723	0.0001	-9.13
59	Wet	C	1	0	1	1	4	129	107.74	0.01	-4.62	0.01	-4.29	0.8171	0.0000	-16.91
60	Wet	C	1	0	1	1	3	101	80.81	0.01	-4.56	0.02	-4.07	0.9935	0.0000	-10.56

表 9.22 の完全モデルの結果に飽和モデルの結果を加えている．

「飽和モデル」の 1 行目は，観測データ $y_1 = 57$ に対し，推定値も $\hat{y}_1 = 57$ とした場合であり，対数尤度 $\ln L_1^{飽和} = -3.56$ が計算されている．デビアンス残差 *Devianse Residuals* は，これらの差から，第 11 章で示す計算式を用い，

式（11.1）　$d_1 = 2[\ln(L_1^{飽和}) - \ln(L_1^{完全})] = 2\times[-3.56-(-3.89)] = 0.66$

式（11.2）　$\varepsilon_1^{(D)} = \mathrm{Sign}(y_1 - \hat{y}_1)\sqrt{d_1} = \mathrm{Sign}(57-46.82)\sqrt{0.66} = 0.8112$

として計算されている．「飽和モデル」が常に「完全モデル」の場合よりも確率が大きい（等しい場合もある）ので，$\mathrm{Sign}(y_i - \hat{y}_i)$ によって観測値と予測値の符号でプラスマイナスを $\sqrt{d_1}$ に付けている．このデビアンス残差の中央値は –0.0828 となり R での結果に一致する．

Residual deviance : 残差デビアンス（逸脱度）は，飽和モデルの対数尤度 $\ln L^{飽和} = -195.63$ と完全モデルの対数尤度 $\ln L^{完全} = -226.24$ の差の–2 倍で定義されている．実際に計算すると

$$\begin{aligned} \textit{Residual deviance} &= -2(\ln L^{飽和} - \ln L^{完全}) \\ &= -2\times[-195.6313-(-226.2417)] = 61.2208 \end{aligned}$$

となる．もちろんデビアンス残差 *Devianse Residuals* の平方和

$$\begin{aligned} \textit{Residual deviance} &= \sum_{i=1}^{20}(\varepsilon_i^{(D.R.)})^2 \\ &= 0.8112^2 + 0.6341^2 + \cdots + 0.9934^2 = 61.2208 \end{aligned}$$

に等しい. いずれにしても, 対数尤度を自ら計算することが, 理解の向上となる. *Null Deviance* は，縮小モデルと完全モデルの対数尤度の差の–2 倍で定義されている．実際に計算すると

$$\begin{aligned} \textit{Null deviance} &= 2(\ln L^{Null\cdot 切片\cdot 縮小} - \ln L^{完全}) \\ &= 2\times[-278.2248-(-226.2417)] = 103.9663 \end{aligned}$$

となるのだが，R の出力の 110.540 とは一致しない．原因の究明は，今後の課題である．

第 9.2 節で，過分散を考慮した場合の各種の推定値の計算を表 9.6 に示した．負の 2 項回帰の場合も同様に推定可能である．回帰パラメータは，Excel，SAS，および，R で同じ推定値

$$\hat{\boldsymbol{\beta}} = [2.7477 \quad 0.1629 \quad 0.0927 \quad 0.4530]^T$$

が得られている．パラメータの共分散行列は，SAS での結果を Excel に取り込み，花数が 1 とした場合について，表 9.24 に示すように推定値と 95%信頼区間が求められる．

表 9.24　負の 2 項回帰での各種の推定結果

			offset		Region		Habitat					
	生育		**花数**	切片	B	C	Wet	対数	対数	指数		
x	環境	地域	**1**	2.7477	0.1629	0.0927	0.4530	推定値	分散	推定値	*L*95%	*U*95%
1	Dry	A	**1**	1	0	0	0	2.7477	0.0049	15.61	13.61	17.90
2	"	B	**1**	1	1	0	0	2.9106	0.0048	18.37	16.04	21.04
3	"	C	**1**	1	0	1	0	2.8404	0.0046	17.12	14.98	19.57
4	Wet	A	**1**	1	0	0	1	3.2007	0.0046	24.55	21.51	28.02
5	"	B	**1**	1	1	0	1	3.3636	0.0042	28.89	25.44	32.82
6	"	C	**1**	1	0	1	1	3.2935	0.0044	26.94	23.66	30.67

共分散$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$	切片	RegionB	RegionC	HabitatWet
切片	**0.0049**	-0.0035	-0.0036	-0.0025
RegionB	-0.0035	**0.0069**	0.0036	-0.0001
RegionC	-0.0036	0.0036	**0.0069**	0.0000
HabitatWet	-0.0025	-0.0001	0.0000	**0.0046**

（生育環境：Dry，地域：A）の場合の対数推定値のデザイン変数 $\boldsymbol{x}_1$ は，

$$\boldsymbol{x}_1 = [1 \quad 0 \quad 0 \quad 0]$$

なので，花数を 1 とした場合の対数推定値は，

$$\ln \hat{y}_{Dry,A} = \ln(1) + \boldsymbol{x}_1 \hat{\boldsymbol{\beta}} = 2.7477$$

と計算され，対数分散は，

$$Var(\ln \hat{y}_{Dry,A}) = \boldsymbol{x}_1 \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_1^T = 0.0049$$

であり，指数を取った推定値および 95%信頼区間は，

$$\hat{y}_{Dry,A} = \exp(\ln \hat{y}_{Dry,A}) = 15.61$$

$$(L95\%, \quad U95\%) = \exp\left[\ln \hat{y}_{Dry,A} \pm 1.96 \sqrt{Var(\ln \hat{y}_{Dry,A})} \right]$$

$$= (13.61, \quad 17.90)$$

として計算されている．

なお，過分散を考慮した簡便的な表 9.18 の推定結果と負の 2 項回帰での表 9.24 の推定結果を比較すると，ほぼ同程度の推定値と分散が得られていることが確認される．

10. オフセットを含む探索的ポアソン回帰

ポアソン回帰の特徴的な事例は，観察対象の部分母集団のサイズが既知で，ある一定期間に発現する事象がカウントされるような場合である．これまでに取り上げた事例は，心疾患による死亡，癌の発生など人数で，対象となる部分母集団の人数が人口統計学的に得られる事例であった．本章で取り上げるのは，McCullagh and Nelder（1989）, Generalized Linear Models 2nd ed. の第 6.3.2 節で取り上げられている貨物船の損傷数データである．この事例は，欠測セルがある 5×4×2 要因配置型デザインであり，交互作用も含めて探索的な解析を試みる．

10.1. 貨物船の損傷数（5×4×2 要因配置，対数リンク，オフセット）

McCullagh ら（1989）の「貨物船の前方部の損傷数」データを表 10.1 に示す．このデータは，ロイドの J.Crilley および L.N.Heminway によって提供されたもので，貨物船の船種，建造年，運行年の 3 因子の要因配置型で，それらの組み合わせセル毎に運航期間中に起きた貨物船の前方部への損傷数について集計された結果である．

それぞれの因子の水準は，

船種，　5 水準：A，B，C，D，E
建造年，4 水準：1960-64，1965-69，1970-74，1975-79
運行年，2 水準：1960-74，1975-79

であり，5×4×2 の 40 セルごとに，総運行月数に対する損傷数がカウントされている．なお，同一の船での複数回の損傷も含まれている．ただし，建造年が 1975-79 の場合に 1960-74 での運行年セルは，必然的に欠測値となる．船種が E タイプの建造年 1960-74 の場合の 1975-79 運行年のセルは，データが得られなかったための欠測値となっている．損傷数を運行月数で割り 1,000 月あたりの損傷数（損傷千月比）を算出し加えてある．

表 10.1 に示した損傷数のカウント・データは，運行年 1960-74 と 1975-79 の分類は，総運行月数が同程度になるように分類されたのであろうか．建造年の区切りは，5 年単位となっており，元々のデータは，船種，建造の年月，損傷があった年月，それまでの運行月数，廃船の

年月，集計時までの運行月数などと思われるが，解析上の便宜を図るための集計結果が示されていると解釈される．

表 10.1　船舶の前方部への損傷数

		運行年 1960-74				運行年 1975-79			
船種	建造年	No	運行月数 n_i	損傷数 y_i	損傷千月比	No	運行月数 n_i	損傷数 y_i	損傷千月比
A	1960-64	1	127	0	0.00	2	63	0	0.00
	1965-69	3	1,095	3	2.74	4	1,095	4	3.65
	1970-74	5	1,512	6	3.97	6	3,353	18	5.37
	1975-79	7	-	-*	-	8	2,244	11	4.90
B	1960-64	9	44,882	39	0.87	10	17,176	29	1.69
	1965-69	11	28,609	58	2.03	12	20,370	53	2.60
	1970-74	13	7,064	12	1.70	14	13,099	44	3.36
	1975-79	15	-	-*	-	16	7,117	18	2.53
C	1960-64	17	1,179	1	0.85	18	552	1	1.81
	1965-69	19	781	0	0.00	20	676	1	1.48
	1970-74	21	783	6	7.66	22	1,948	2	1.03
	1975-79	23	-	-*	-	24	274	1	3.65
D	1960-64	25	251	0	0.00	26	105	0	0.00
	1965-69	27	288	0	0.00	28	192	0	0.00
	1970-74	29	349	2	5.73	30	1,208	11	9.11
	1975-79	31	-	-*	-	32	2,051	4	1.95
E	1960-64	33	45	0	0.00	34	-	-**	-
	1965-69	35	789	7	8.87	36	437	7	16.02
	1970-74	37	1,157	5	4.32	38	2,161	12	5.55
	1975-79	39	-	-*	-	40	542	1	1.85
		計	88,911	139		計	74,663	217	
千月比損傷数=1,000×(y_i/n_i)			*必然的に空のセル，**誤って空のセル						

運行月数は，最も少ないもので 45 月，最大のもので 44,882 月であり，損傷数を運行月数で除して基準化し，通常の線形モデルを適用したくなる．図 10.1 に示すように 1,000 月あたりの損傷数を四捨五入し整数化して分布の形状を確認すると分布は，指数分布的である．また 1,000 月あたりの 34 セルの組み合わせの損傷数の平均は 3.1765，分散は 11.7861 であり，その比は，

$$\text{分散} \ / \ \text{平均} = 11.7861 / 3.1765 = 3.71$$

3.71 と 1.0 よりもかなり大きく，過分散となっている．

また，2 値反応としてロジスティック回帰を適用することも可能であるが，2 項分布を仮定することも躊躇される．損傷率は，ほとんどが 1 パーセント以下で，延べ運行月数 163,574 月に対して延べ損傷数 356 件と，平均損傷率は 0.22 パーセントである．過分散を念頭におき，運行月数の対数をオフセットとしたポアソン回帰の適用が望ましいと思われる．

図 10.1　整数化した千月比の損傷数に対するポアソン分布のあてはめ

損傷数を y_i とし，デザイン行列の変数を（$x_0, x_1, \ldots, x_m$）としたときに，通常の重回帰のモデル式は，

$$y_i = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon_i \quad \varepsilon_i \sim \text{正規分布} \tag{10.1}$$

である．運行月数を n_i としたときに，対数リンクのポアソン回帰のモデル式は，

$$y_i = n_i \exp\left(\beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right) + \varepsilon_i \quad \varepsilon_i \sim \text{ポアソン分布} \tag{10.2}$$

となる．線形モデルにするために，両辺を n_i で除し，対数を取る．この際に誤差 ε_i が和の形式で残り，扱いがややこしくなるので，推定値として，次の線形式

$$\ln\left(\frac{\hat{y}_i}{n_i}\right) = \hat{\beta}_0 x_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \tag{10.3}$$

とする．損傷数 $\hat{y}_i$ を運行数 n_i で除し，月あたりの損傷数に対して，対数を取っている．左辺を差 $(\ln \hat{y}_i - \ln n_i)$ として $-\ln n_i$ を右辺に移項し，

$$\ln \hat{y}_i = \ln n_i + \hat{\beta}_0 x_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \tag{10.4}$$

を得る．ここでの $\ln n_i$ が，いわゆるオフセット項である． $\ln n_i$ は，実数でカウント・データではないので，式（10.4）を $\hat{y}_i$ について解いた式

$$\hat{y}_i = n_i \cdot \exp(\hat{\beta}_0 x_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m) \tag{10.5}$$

にし， y_i に対する母数を $\hat{y}_i$ としたポアソン分布の確率 P_i を求め，対数尤度を $\ln L_i$

$$\ln L_i = \ln P_i = \ln\left[Poisson\left(y_i; \hat{y}_i\right)\right] \tag{10.6}$$

を計算する．それらの和を取った対数尤度 $\ln L$

$$\ln L = \sum_{i=1}^{n} \ln L_i \tag{10.7}$$

を最大にするような（$\hat{\beta}_0$，$\hat{\beta}_1,\ldots,\hat{\beta}_m$）を求める．対数での式（10.4）で（$x_0=1, x_1=0,\ldots,x_m=0$）とすると $\ln\hat{\beta}_0$ は，いわゆる切片であり，式

$$\ln\hat{y}_i = \ln n_i + (\hat{\beta}_0\times 1+\hat{\beta}_1\times 0+\cdots+\hat{\beta}_m\times 0)$$

$$\ln\hat{y}_i = \ln n_i + \hat{\beta}_0 \qquad (10.8)$$

における $\ln n_i$ は，$\ln\hat{y}_i$ を推定する際に，$\hat{\beta}_0$ に $\ln n_i$ を加えた，あるいは，基準となる切片を $\hat{\beta}_0$ とし，$\ln n_i$ ずらした（オフセットした）切片を用いることになる．

第 2.6 節では，オーストラリアのある地方の冠動脈心疾患の死亡者数に対して，対数目盛上の基準となるポアソン回帰直線の切片をオフセットしたグラフを図 2.6 に例示したので，参考にしてもらいたい．

推定値（$\hat{\beta}_0$，$\hat{\beta}_1,\ldots,\hat{\beta}_m$）を求めるための数値計算の方法は，2 通りが定式化されている．第 1 の方法は，第 2.6 節で示した方法，すなわち式（2.7）の対数尤度 $\ln L$ をパラメータで 2 階の偏微分行列（ヘッセ行列，マイナスを付けて情報行列）を用いるニュートン・ラフソン法である．Excel のソルバーを用いることにより，2 階の偏微分行列を用いずに対数尤度 $\ln L$ を最大化してパラメータの推定が行なえ，第 10.5 節で例示する．

第 2 の方法は，第 5.3 節で示した反復重み付き回帰を用いる方法である．この方法は，式（5.3）を用いる．誤差 ε_i がポアソ分布に従うとの仮定は，反復計算のための回帰式と重みの計算式の中に組み込まれている．この方法は，第 1 の方法に比べ技巧的であるが，計算量が少ない利点がある．

なお，第 2 の方法による反復重み付き回帰による方法が，一般化線形モデルの由来である．ただし，モデル式を線形化できない打切りデータを含む寿命データのワイブル回帰などは，線形化できないので，第 1 の方法による方法で定式化されている．また，第 7.3 節で扱ったゼロ過剰ポアソン回帰なども一般化線形モデルとはならないので，第 1 の方法によっている．

10.2. 主効果モデルの適用

(0,1) 型デザイン変数（最初の水準を基準）

McCullagh ら(1989) では，運行月数の対数をオフセット，誤差分布をポアソン分布，リンク関数を対数としている．質的変数を表 10.2 に示すように最初の水準を基準とする（0，1）型のデザイン変数（ダミー変数）にし，過分散を考慮したポアソン回帰の結果が示されている．

表 10.2　質的変数の最初の水準を基準としたデザイン行列のための変数

船種	x_B	x_C	x_D	x_E
A	**0**	**0**	**0**	**0**
B	**1**	0	0	0
C	0	**1**	0	0
D	0	0	**1**	0
E	0	0	0	**1**

建造年	x_{C65}	x_{C70}	x_{C75}
60-64	**0**	**0**	**0**
65-69	**1**	0	0
70-74	0	**1**	0
75-79	0	0	**1**

運行年	x_{Op75}
60-74	**0**
75-79	**1**

デザイン行列を用いたポアソン回帰のためのデータリストを表 10.3 に示す．運行月数の対数をオフセットとし，主効果モデル（船種，建造年，運行年）に対するデザイン変数（$x_0, x_B, x_C, x_D, x_E, x_{C65}, x_{C70}, x_{C75}, x_{Op75}$）が示されている．

質的変数を含むポアソン回帰を統計ソフトで行う場合に，統計ソフトの内部で生成される量的変数（ダミー変数）について注意を払う必要がある．これは，統計ソフトが出力する結果の解釈に際して必要不可欠なためである．McCullagh らで使用されている統計ソフトは，彼らが推奨している（S，GLIM，Minitab）と思われる．これらの統計ソフトは，質的変数に対し，最初の水準を基準とした量的変数（ダミー変数）がデフォルトで使用されているのであろう．S の後の R も同様である．最初の水準を基準（対照群）とした場合のデザイン変数は，推定されたパラメータが，「最初の水準（対照群）」との差となり，t 検定あるいはカイ 2 乗検定の結果が，対照群とそれぞれの群の差についての検定統計量として使える利便性がある．

量的変数しか受け付けないポアソン回帰に対し，質的変数の各水準に対し，何らかの量的変数のセットを与える必要がある．この量的変数が，通常ダミー変数と称されている．本書では，ポアソン回帰を行うために必要なデザイン行列（計画行列）に必要不可欠な変数なので，ダミー変数ではなくデザイン変数として統一的に使用してきた．これは，連続変数と質的変数との交互作用は，(0，1）ではなく，連続変数となりダミー変数と言い難いことも理由の一つである．また，デザイン行列のためのダミー変数と言うのも違和感があり，切片も含めて全てデザイン変数で統一してきた．

表 10.3　船舶の前方部への損傷数データに対するデザイン行列

No	船種	建造年度	運行年度	運行月数n_i	損傷数y_i		デザイン行列 $\boldsymbol{X}$								
							x_0	x_B	x_C	x_D	x_E	x_{C65}	x_{C70}	x_{C75}	x_{Op75}
1	**A**	**60-64**	**60-74**	127	0		**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
2			75-79	63	0		1	0	0	0	0	0	0	0	1
3		65-69	60-74	1,095	3		1	0	0	0	0	1	0	0	0
4			75-79	1,095	4		1	0	0	0	0	1	0	0	1
5		70-74	60-74	1,512	6		1	0	0	0	0	0	1	0	0
6			75-79	3,353	18		1	0	0	0	0	0	1	0	1
7		75-79	60-74	-	-	*	1	0	0	0	0	0	0	1	0
8			75-79	2,244	11		1	0	0	0	0	0	0	1	1
9	B	60-64	60-74	44,882	39		1	1	0	0	0	0	0	0	0
10			75-79	17,176	29		1	1	0	0	0	0	0	0	1
11		65-69	60-74	28,609	58		1	1	0	0	0	1	0	0	0
12			75-79	20,370	53		1	1	0	0	0	1	0	0	1
13		70-74	60-74	7,064	12		1	1	0	0	0	0	1	0	0
14			75-79	13,099	44		1	1	0	0	0	0	1	0	1
15		75-79	60-74	-	-	*	1	1	0	0	0	0	0	1	0
16			75-79	7,117	18		1	1	0	0	0	0	0	1	1
17	C	60-64	60-74	1,179	1		1	0	1	0	0	0	0	0	0
18			75-79	552	1		1	0	1	0	0	0	0	0	1
19		65-69	60-74	781	0		1	0	1	0	0	1	0	0	0
20			75-79	676	1		1	0	1	0	0	1	0	0	1
21		70-74	60-74	783	6		1	0	1	0	0	0	1	0	0
22			75-79	1,948	2		1	0	1	0	0	0	1	0	1
23		75-79	60-74	-	-	*	1	0	1	0	0	0	0	1	0
24			75-79	274	1		1	0	1	0	0	0	0	1	1
25	D	60-64	60-74	251	0		1	0	0	1	0	0	0	0	0
26			75-79	105	0		1	0	0	1	0	0	0	0	1
27		65-69	60-74	288	0		1	0	0	1	0	1	0	0	0
28			75-79	192	0		1	0	0	1	0	1	0	0	1
29		70-74	60-74	349	2		1	0	0	1	0	0	1	0	0
30			75-79	1,208	11		1	0	0	1	0	0	1	0	1
31		75-79	60-74	-	-	*	1	0	0	1	0	0	0	1	0
32			75-79	2,051	4		1	0	0	1	0	0	0	1	1
33	E	60-64	60-74	45	0		1	0	0	0	1	0	0	0	0
34			75-79	-	-	**	1	0	0	0	1	0	0	0	1
35		65-69	60-74	789	7		1	0	0	0	1	1	0	0	0
36			75-79	437	7		1	0	0	0	1	1	0	0	1
37		70-74	60-74	1,157	5		1	0	0	0	1	0	1	0	0
38			75-79	2,161	12		1	0	0	0	1	0	1	0	1
39		75-79	60-74	-	-	*	1	0	0	0	1	0	0	1	0
40			75-79	542	1		1	0	0	0	1	0	0	1	1
			計	163,574	356										

統計ソフト SAS では，質的変数に対してデフォルトの設定では，最後の水準を基準とした（1，0）型のデザイン変数が生成され，JMP では，（1，-1）対比型のデザイン変数が生成される．ただし，SAS では，内部で生成されるデザイン変数を意識しなくても，元の水準に対する推定値を最小 2 乗平均（Lsmeans）として推定する機能が充実している．

デザイン変数を自らデータとして生成することは，表計算ソフト Excel が普及する以前には煩雑であったが，表 10.3 に示すように Excel のコピー&ペーストで比較的容易に（0，1）から成るデザイン行列が作成できるようになった．このように自ら能動的に作成したデザイン行列を用い，統計ソフトによる解析をすることは，ブラックボックス的となりがちな統計ソフトの解析結果に対し理解を深めることが期待される．さらに，このデザイン行列を用い，Excel のみで統計ソフトと同様の解析を試みることは，さらなる理解を深めることになる．

表 10.3 のデザイン行列を用いた JMP によるポアソン回帰の結果を表 10.4 示す．適合度統計量の Pearson のカイ 2 乗統計量は 42.2753 で自由度が 25 なので，その比は 1.6910 と過分散となっている．パラメータの推定値は，運行月数をオフセットにしたので，一月あたりの損傷数の対数に対する推定値となっている．

表 10.4　損傷数に対するポアソン回帰（最初を基準，過分散あり）

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	31.8251	63.6503	8	<.0001*
完全	40.3787			
縮小	72.2038			

適合度統計量	カイ2乗	自由度	p値	過分散
Pearson	42.2753	25	0.0168*	1.6910
デビアンス	38.6951	25	0.0395*	

AICc
110.3226

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-6.4059	0.2828	958.7224	<.0001*
x_B	-0.5433	0.2309	5.0023	0.0253*
x_C	-0.6874	0.4279	2.8890	0.0892
x_D	-0.0760	0.3779	0.0408	0.8399
x_E	0.3256	0.3067	1.1062	0.2929
x_C65	0.6971	0.1946	13.4958	0.0002*
x_C70	0.8184	0.2208	13.9492	0.0002*
x_C75	0.4534	0.3032	2.1554	0.1421
x_Op75	0.3845	0.1538	6.3040	0.0120*

表 10.4 のパラメータの推定値は，一月あたりの対数の損傷数なので解釈が困難である．表 10.5 に示すように基準となる最初の水準の（船種：A，建造年：60-64，運行年：60-74）の推定値は，切片の推定値 $\hat{\beta}_0 = -6.4059$ に等しくなる．他の水準は，最初の水準との差の推定値であり，最初の水準の推定値を加えることにより，推定値が得られる．

船種 B：　船種 A＋（船種 B－船種 A）＝–6.4059 – 0.5433 = –6.9492

船種 C：　船種 A＋（船種 C－船種 A）＝–6.4059 – 0.6874 = –7.0933

：

建造年 65-79：建造年 65-79＋（建造年 60-64－建造年 65-79）＝–6.4059 + 0.6971 = –5.7088

：

運行年 75-79：運行年 60-74＋（運行年 75-79－運行年 60-74）＝–6.4059 + 0.3845 = –6.0214

元の水準での一月当たりの対数の損傷数に，指数を取り一月当たりの損傷数を計算する．少数点以下 2 桁がゼロなので 1,000 を掛けて損傷千月比の推定値を計算する．

表 10.5　パラメータの推定値に対する解釈

					JMP の結果(表10.4)				Excelでの換算			
No	因子	変数・パラメータ			推定値 β	標準誤差	尤度比カイ2乗	p 値	月あたりln損傷数	月あたり損傷数	損傷千月比	
0		切片	x_0	$\hat{\beta}_0$	**-6.4059**	0.2828	958.72	<.0001	-6.4059	0.0017	**1.6518**	
	船種	(A)	-		**0**				-6.4059	0.0017	**1.6518**	基準
1		B	x_B	$\hat{\beta}_1$	-0.5433	0.2309	5.0023	0.0253	-6.9492	0.0010	**0.9594**	
2		C	x_C	$\hat{\beta}_2$	-0.6874	0.4279	2.8890	0.0892	-7.0933	0.0008	0.8306	
3		D	x_D	$\hat{\beta}_3$	-0.0760	0.3779	0.0408	0.8399	-6.4819	0.0015	1.5310	
4		E	x_E	$\hat{\beta}_4$	0.3256	0.3067	1.1062	0.2929	-6.0803	0.0023	2.2874	
	建造	(60-64)	-		**0**				-6.4059	0.0017	1.6518	基準
5	年	65-69	x_{C65}	$\hat{\beta}_5$	0.6971	0.1946	13.4958	0.0002	-5.7088	0.0033	**3.3168**	
6		70-74	x_{C70}	$\hat{\beta}_6$	0.8184	0.2208	13.9492	0.0002	-5.5875	0.0037	3.7445	
7		75-79	x_{C75}	$\hat{\beta}_7$	0.4534	0.3032	2.1554	0.1421	-5.9525	0.0026	2.5994	
	運行	(60-74)	-		**0**				-6.4059	0.0017	1.6518	基準
8	年	75-79	x_{Op75}	$\hat{\beta}_8$	0.3845	0.1538	6.3040	0.0120	-6.0214	0.0024	**2.4262**	
オフセット: ln(運行月数),					Pearson適度度 カイ2乗=42.2753, df=25,				過分散 42.2753/25 = 1.6910			

船種 A の 1,000 月あたりの損傷数は，切片の対数の推定値が –6.4059 で，指数を取った 0.0017 件が月あたりの損傷数となり，

$$\hat{y}^{(1000)}_{\mathrm{A,C60,Op60}} = \exp(\hat{\beta}_0)\times 1000$$
$$= \exp(-6.4059)\times 1000 = 1.6518$$

1,000 月あたりに換算すると 1.6518 件と推定される．

船種 B は，切片と x_B 以外が全て 0 なので，

$$\hat{y}^{(1000)}_{\mathrm{B,C60,Op60}} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_B)\times 1000$$
$$= \exp(-6.4059 - 0.5433\times 1)\times 1000 = 0.9594$$

と推定される．

建造年 65-69 の推定値は，デザイン変数の基準を最初の水準としているので，建造年 60-64 との差であり，船種 A，運行年 60-74 と固定した場合に，

$$\hat{y}^{(1000)}_{\mathrm{A,C65,Op60}} = \exp(-6.4059 + 0.6971) \times 1000 = 3.3168$$

と推定される．

運行年 75-79 の推定値は，運行年 60-74 との差で，船種 A，運行年 60-74 とした場合に，

$$\hat{y}^{(1000)}_{\mathrm{A,C60,Op75}} = \exp(-6.4059 + 0.3845) \times 1000 = 2.4262$$

と推定される．

これらの結果から, 船種別の 1,000 月あたりの損傷数は, 船種 C が 0.8306 件と最も少なく，船種 E が 2.2874 件と最も多いと推定される．建造年の場合は，60-64 年が 1.6518 件と最も少なく，70-74 年が 3.7445 件と最も多いことがわかる．運行年では，75-79 年の 2.4262 件が，60-74 年の 1.6518 件に比べて多くなっている．

多くの統計ソフトでは，質的変数を説明変数とした場合に水準数に応じたダミー変数を内部で自動生成し，それらを用いて何らかのデザイン行列を用いた計算が行なわれる．どのようなダミー変数を内部で生成するかは，開発者の考え方によって異なる．

SAS の GENMOD プロシジャでのデフォルトは，質的変数に対し最後の水準を基準とする（1，0）型のダミー変数であり，最初の水準に変更したい場合は，classs ステートメントのオプションで ref=first とすれば（0，1）型となる．ただし，デフォルト以外のダミー変数を生成すると，最小 2 乗平均（Lsmeans）などの計算がサポートされなくなるデメリットがある．

JMP の一般化線形モデルでは，（1，-1）対比型のダミー変数がデフォルトであり，設定の変更はできない．そのため，最初の水準を基準にしたダミー変数にしたければ，これまでに示したように，自ら（0，1）型のダミー変数を生成しなければならない．その結果として，JMP が標準的に提供している予測プロファイルなどのグラフ表示ができなくなる．

（1, −1）対比型のデザイン行列

表 10.4 で示した結果は，質的変数に対し（0，1）型のデザイン変数を自ら生成して JMP のポアソン回帰を適用した結果であり，（0，1）型のデザイン変数のままでは，JMP の予測プロファイルが機能しない．そこで，質的変数に対しては，JMP の内部でデザイン変数を自動生成させることにより，予測プロファイルおよび交互作用プロファイルが取りまとめられて表示される．表 10.6 に（1, –1）対比型のデザイン行列のための変数を示す．なお，対比型とは，1 の反対の性質を持つ数値として–1 が使われているためであり．したがって，対比型の変数は，足して 0 となる性質を持っている．

表 10.6　質的変数の（1，–1）対比型のデザイン変数（ダミー変数）

船種	x'_A	x'_B	x'_C	x'_D
A	**1**	0	0	0
B	0	**1**	0	0
C	0	0	**1**	0
D	0	0	0	**1**
E	**–1**	**–1**	**–1**	**–1**

建造年	x'_{C60}	x'_{C65}	x'_{C70}
60-64	**1**	0	0
65-69	0	**1**	0
70-74	0	0	**1**
75-79	**–1**	**–1**	**–1**

運行年	x'_{Op60}
60-74	**1**
75-79	**–1**

JMP による主効果モデルは，表 10.7 に示すように因子ごとの尤度比カイ 2 乗検定に引き続き，（1，–1）対比型のデザイン変数に対応する推定値が出力される．表 10.4 と比較すると「モデル全体の検定」は，差分の尤度比カイ 2 乗値が 63.6503 と同じ結果となっている．次の「効果の検定」は，質的変数ごとの尤度比カイ 2 乗検定で，要因配置型における一般的な分散分析表と同様な形式となっている．

表 10.7　損傷数に対するポアソン回帰（対比型，過分散を考慮）

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	31.8251	63.6503	8	<.0001*
完全	40.3787			
縮小	72.2038			

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)	過分散
Pearson	42.2753	25	0.0168*	1.6910
デビアンス	38.6951	25	0.0395*	

AICc
110.3226

効果の検定

要因	自由度	尤度比カイ2乗	p値
船種	4	13.9977	0.0073*
建造年	3	18.5735	0.0003*
運行年	1	6.3040	0.0120*

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-5.9176	0.1280	57532.562	<.0001*
船種[A]	0.1962	0.1954	0.9809	0.3220
船種[B]	-0.3471	0.1459	5.2948	0.0214*
船種[C]	-0.4912	0.3148	2.8802	0.0897
船種[D]	0.1203	0.2782	0.1808	0.6707
建造年[60-64]	-0.4922	0.1505	11.2088	0.0008*
建造年[65-69]	0.2049	0.1181	2.9823	0.0842
建造年[70-74]	0.3262	0.1228	6.8496	0.0089*
運行年[60-74]	-0.1922	0.0769	6.3040	0.0120*

「パラメータ推定値」は，全く異なる．表 10.4 での切片 –6.4059 は，切片以外のデザイン変数が全て 0 の場合であり，船種がA，建造年が 1960–64，運行年が 1960–74 の場合の推定値となっている．表 10.7 の切片は，$\hat{\beta}'_0 = -5.9176$ であり（船種：A，建造年：60–64，運行年：60–74）の (1, –1) 対比型デザイン変数は（$x'_A = 1, x'_B = 0, x'_C = 0, x'_D = 0, x'_{C60} = 1, x'_{C65} = 0, x'_{C75} = 0,$ $x'_{Op60} = 1$）なので，表 10.4 の場合と同様の推定は，

$$\begin{aligned}\ln \hat{y}_{\mathrm{A,C60,Op60}} &= \hat{\beta}'_0 + \hat{\beta}'_1 x'_A + \hat{\beta}'_5 x'_{C60} + \hat{\beta}'_8 x'_{Op60} \\ &= -5.9176 + 0.1962 - 0.4922 - 0.1922 \\ &= -6.4059\end{aligned}$$

とすることにより同じ結果が得られる．このように，推定値は名義尺度の各水準に与えるデザイン変数によって全く異なるので，パラメータの推定値を用いた結果の解釈には細心の注意が必要である．

表 10.8 に対比型のデザイン変数を使った場合のパラメータの推定値を用い, 船種, 建造年, および運行年の各水準の推定値（月あたりの対数損傷数）を計算した結果を示す. 切片–5.9176

表 10.8　対比型デザイン変数による主効果モデル

変数番号		変数・パラメータ			JMP推定値	月あたりln損傷数		月あたり損傷数	損傷千月比	
0		切片	x'_0	$\hat{\beta}'_0$	**-5.9176**	**-5.9176**		0.0027	***2.6915***	
1	船種	A	x'_A	$\hat{\beta}'_1$	0.1962	-5.7214		0.0033	3.2751	幾何
2		B	x'_B	$\hat{\beta}'_2$	-0.3471	-6.2648	平均	0.0019	1.9022	平均
3		C	x'_C	$\hat{\beta}'_3$	-0.4912	-6.4088	**-5.9176**	0.0016	1.6470	***2.6915***
4		D	x'_D	$\hat{\beta}'_4$	0.1203	-5.7974		0.0030	3.0355	
		(E)		負の和	**0.5218**	-5.3958		0.0045	4.5354	
5	建造年	60-64	x'_{C60}	$\hat{\beta}'_5$	-0.4922	-6.4099		0.0016	1.6452	幾何
6		65-69	x'_{C65}	$\hat{\beta}'_6$	0.2049	-5.7128	平均	0.0033	3.3036	平均
7		70-74	x'_{C70}	$\hat{\beta}'_7$	0.3262	-5.5915	**-5.9176**	0.0037	3.7296	***2.6915***
		(75-79)		負の和	**-0.0388**	-5.9565		0.0026	2.5890	
8	運行年	60-74	x'_{Op60}	$\hat{\beta}'_8$	-0.1922	-6.1099	平均	0.0022	2.2208	幾何平均
		(75-79)		負の和	**0.1922**	-5.7254	**-5.9176**	0.0033	3.2620	***2.6915***

は，それぞれの質的変数の損傷数の対数の推定値の平均であり，1,000 月あたりの損傷数は，

切片： $\exp(-5.9176) \times 1000 = 2.6915$ 件

となる．船種 E の推定値は，表 10.6 左から切片に船種 A～D の負の和を加えることが示されていて，

船種 E＝切片－船種 A－船種 B－船種 C－船種 D－船種 E

$$\begin{aligned}\ln \hat{y}_{\mathrm{E},\cdot,\cdot} &= \hat{\beta}'_0 - \hat{\beta}'_1 x'_A - \hat{\beta}'_2 x'_B - \hat{\beta}'_3 x'_C - \hat{\beta}'_4 x'_D \\ &= -5.9176 - 0.1962 - (-0.3471) - (-0.4912) - 0.1203 \\ &= -5.9176 + 0.5218 \\ &= -5.3958\end{aligned}$$

と推定される．損傷千月比は指数を取り 1,000 倍した結果である．建造年 および 運行年の場合も同様にして推定する．損傷数は，表 10.4 と全く異なるが，それぞれの水準の大小関係は保たれている．

月当たりの損傷数の対数について，対比型のデザイン変数なので船種 A～E の平均 –5.9176 では，切片と等しくなる．また，当然のことながら損傷千月比の幾何平均は，切片の損傷千月比と等しくなる．

このように，統計ソフトから出力されたポアソン回帰の結果は，Excel に取り込み何らかの計算を加えて結果の解釈をする必要が常にある．

予測プロファイル

JMP のデフォルトの（1，-1）対比型のダミー変数でポアソン回帰を行った場合には，主効果に関する予測プロファイルおよび交互作用プロファイルが出力される．まず，予測プロファイルの結果を図 10.2 に示す．この図は，（船種を A，建造年 60-64，運行年 60-74）を基準

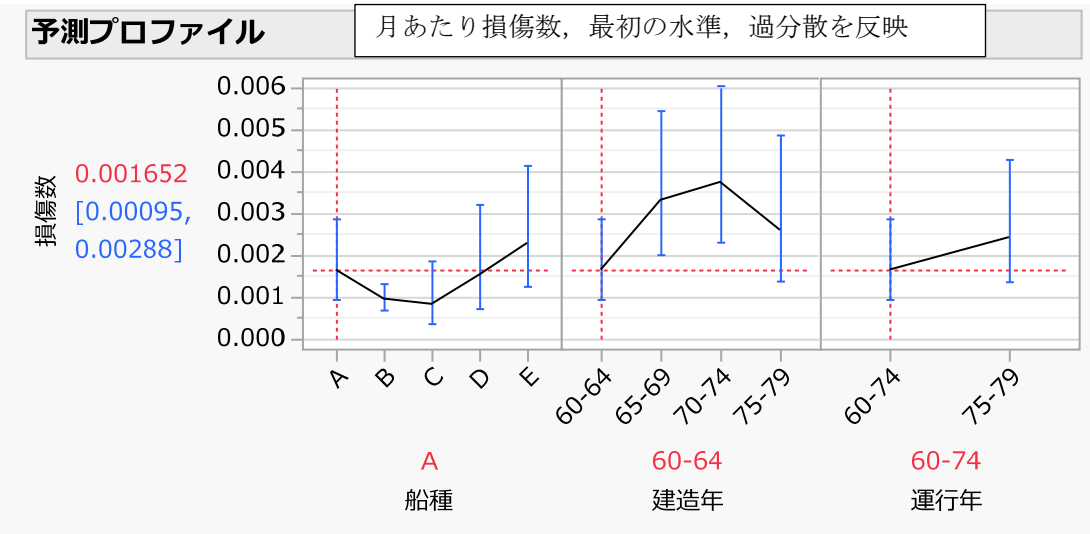


図 10.2　JMP の予測プロファイル（船種 A，建造年 60-64，運行年 60-74）を基準

とした場合が示されている．損傷数の推定値は 0.001652，95%信頼区間が（0.00095，0.00288）と推定されている．これは，推定された月あたりの損傷数の対数に対し，指数を取ったもので，exp(–6.4059) = 0.001652 であり，1,000 月あたりでは 1.6552 件となる．なお，図 10.2 の予測プロファイルでの船種（B，C，D，E）の推定値は，表 10.5 に示したように（船種 A，建造

年 60-64，運行年 60-74）を基準とした推定値（0.0010，0.0008，0.0015，0.0023）が表示されている．同様に建造年（65-69，70-74，75-79）の推定値は（0.0033，0.0037，0.0026）が，運行年（75-79）の推定値は（0.0024）である．

JMP で出力されている予測プロファイル上で，マウス操作により表示内容を変更することができる．それぞれの変数の最大となっている水準（船種 E，建造年 70-74，運行年 75-79）にマウスを移動し選択すると図 10.3 が得られる．推定値として 0.007617, 1,000 月あたり 7.617 件が得られている．また，最小となる水準の組み合わせ（船種 C, 建造年 60–64, 運行年 60–74）となる推定値を得ることも容易にできる．

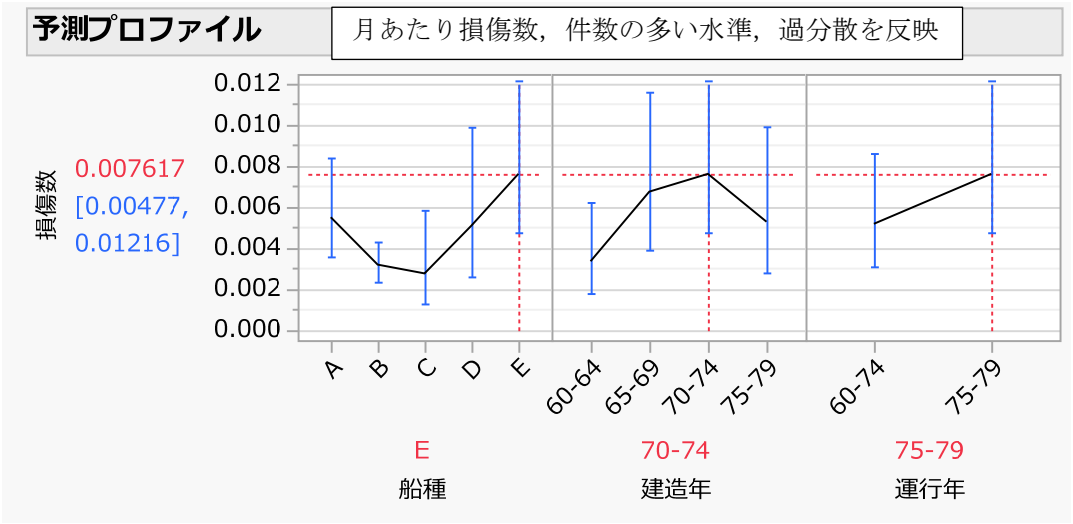


図 10.3　JMP の予測プロファイル（船種 E，建造年 70-74，運行年 75-79）を基準

図 10.3 の予測プロファイルでの船種（A，B，C，D）の推定値は，（船種 E，建造年 70-74，運行年 75-79）に固定した場合のそれぞれの船種の推定値 0.007617 となっていて，図 10.2 とは異なる．このように主効果の推定値は，固定されるものではなく，他の条件により変化する相対的なものであることに注意が必要である．

交互作用プロファイル

予測プロファイルに引き続き，交互作用プロファイルを作成する．交互作用を含まない主効果モデルなので，“交互作用” はないが，図 10.4 に示すように 3 つの質的変数を 2 つごとに組み合わせた推定値のプロファイルが得られる．全体を俯瞰するためには有益であり，主効果に対する予測プロファイルと合わせて使うと効果的である．

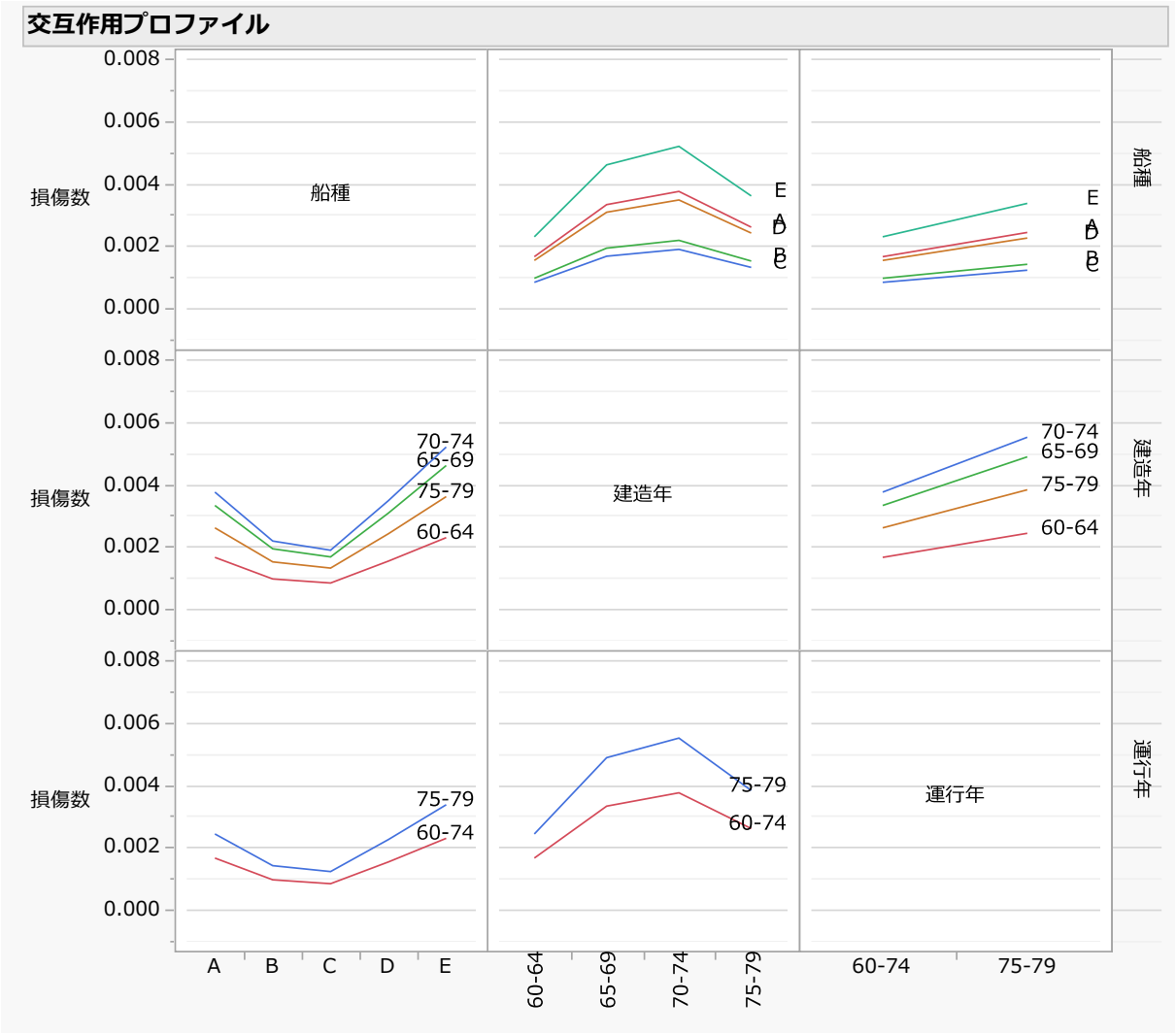


図 10.4　交互作用プロファイル（船種を A，建造年 60-64，運行年 60-74）を基準

10.3.　Excel による（0，1）型ダミー変数での予測プロファイル

JMP の予測プロファイルは，結果を解釈するために有益であるが，JMP 以外では見かけたことがないので，Excel による作図法を示す．95%信頼区間も含めるためには，パラメータの共分散行列を用いた計算が必要となるのであるが，まず，推定値のみの折れ線グラフを示す．

表 10.5 から因子と水準，損傷千月比を抜き出して作図用の Excel データとする．3 因子の水準と推定値を選択し，「折れ線」グラフを選択する．2 つの因子をまたいで繋がっている線を因子の切れ目のポイントのみを選択し「線なし」とする．船種は質的変数なので点線に変更した結果を図 10.5 に示す．Excel のグラフなので推定値を変えれば，自動的に更新される．

因子	水準	千月比 損傷数
船種	A	**1.65**
	B	0.96
	C	0.83
	D	1.53
	E	2.29
建造	60-64	**1.65**
年	65-69	3.32
	70-74	3.74
	75-79	2.60
運行	60-74	**1.65**
年	75-79	2.43

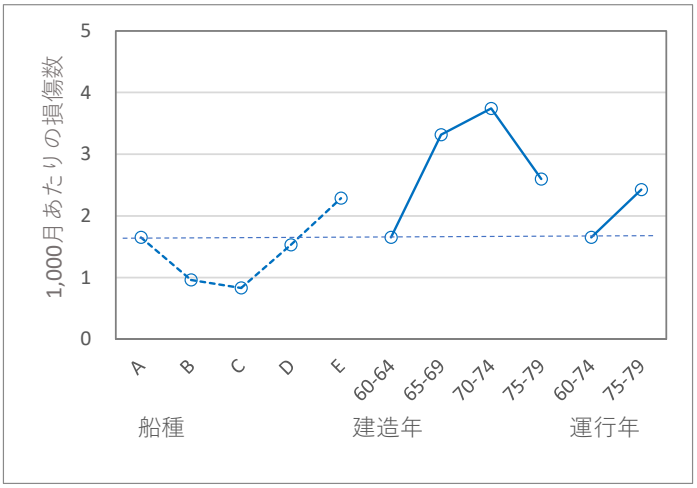


図 10.5　Excel による予測プロファイル（船種を A，建造年 60–64,，運行年 60–74）を基準

図 10.5 の予測プロファイルに 95%信頼区間の重ね書きを試みよう．JMP では，自動的に 95%信頼区間を表示しているが，どのような計算を行っているのであろうか．表 10.3 に示したデザイン行列 $\boldsymbol{X}$ の各行のベクトル $\boldsymbol{x}_i$ とパラメラータの推定値 $\hat{\boldsymbol{\beta}}$ の積から推定値は，次のように

$$\ln(\hat{y}_i / n_i) = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}$$

で求められる．この $\ln(\hat{y}_i / n_i)$ の分散 $Var[\ln(\hat{y}_i / n_i)]$ は，パラメラータの推定値 $\hat{\boldsymbol{\beta}}$ のパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を使って

$$Var[\ln(\hat{y}_i / n_i)] = \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T$$

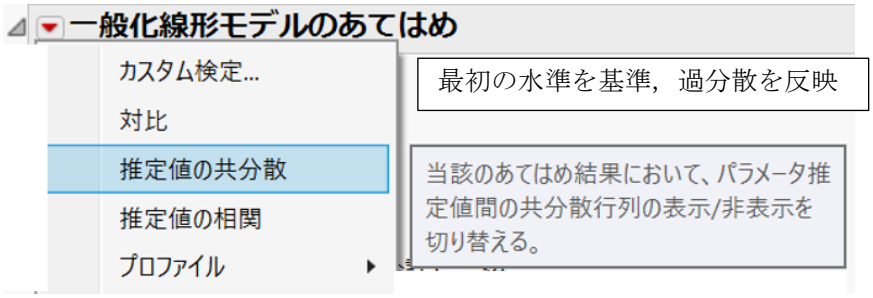
によって推定することができる．パラメータの共分散行列は，JMP のプルダウンメニューで「推定値の共分散」を選択することによって得られる．第 10.7 節で示す SAS の GENMOD プロシジャの場合であれば，Model ステートメントで covb オプションを追加することで得られ

る．使っている統計ソフトでパラメータの共分散行列が得られない場合は，第 12 章でデザイン行列および重みベクトルを用いた Excel による計算方法を示すので，参考にしてもらいたい．

ザイン行列の各行の行ベクトル $\boldsymbol{x}_i$ と $\hat{\boldsymbol{\beta}}$ の積から推定値 $\ln(\hat{y}_i / n_i) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ を求め，その分散をパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を使って $\boldsymbol{x}_i\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i^T$ によって推定する方法は，第 4.5 節「デザイン行列を用いた回帰分析の実際」で丁寧に示した．この方法は，複数の変数に対する回帰分析にも，ポアソン回帰にも適用できる基本的な方法である．なお，第 1.4 節「人工データ（恒等リンク，3 水準)」の「共分散行列を用いた 95%信頼区間の計算」は，恒等リンクのポアソン回帰の場合で活用事例である．

推定値 $\ln(\hat{y}_i / n_i)$ の分散が求まれば，95%信頼区間の推定は，容易である．表 10.9 に示すように「推定値の共分散」を JMP から Excel に取り込む．デフォルトの結果の表示では，小数点以下 4 桁程度であるが，8 桁程度に変更し，Excel に取り込んだ後に桁数の表示を縮めることが，計算精度を保つために必要である．

表 10.9　推定値（パラメータ）の共分散行列の獲得



推定値の共分散　パラメータの共分散行列

共分散

	切片	x_B	x_C	x_D	x_E	x_C65	x_C70	x_C75	x_Op75
切片	0.0800	-0.0530	-0.0458	-0.0396	-0.0408	-0.0266	-0.0343	-0.0344	-0.0094
x_B	-0.0530	0.0533	0.0428	0.0390	0.0404	0.0038	0.0138	0.0160	0.0009
x_C	-0.0458	0.0428	0.1831	0.0384	0.0412	0.0030	0.0043	0.0126	-0.0002
x_D	-0.0396	0.0390	0.0384	0.1428	0.0387	0.0020	0.0024	-0.0111	-0.0003
x_E	-0.0408	0.0404	0.0412	0.0387	0.0941	-0.0002	-0.0025	0.0049	0.0013
x_C65	-0.0266	0.0038	0.0030	0.0020	-0.0002	0.0379	0.0272	0.0281	-0.0036
x_C70	-0.0343	0.0138	0.0043	0.0024	-0.0025	0.0272	0.0487	0.0367	-0.0090
x_C75	-0.0344	0.0160	0.0126	-0.0111	0.0049	0.0281	0.0367	0.0919	-0.0147
x_Op75	-0.0094	0.0009	-0.0002	-0.0003	0.0013	-0.0036	-0.0090	-0.0147	0.0237

ここでのデザイン行列のベクトル $\boldsymbol{x}_i$ は 1 行 9 列，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は 9 行 9 列，転置した $\boldsymbol{x}_i^T$ は 9 行 1 列であり， $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の要素を $C_{j,k}$ としたときに， $Var[\ln(\hat{y}_i / n_i)]$ は，

$$
\begin{aligned}
Var[\ln(\hat{y}_i / n_i)] &= \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T \\
&= \begin{bmatrix} x_{i,0} & x_{i,1} & \cdots & x_{i,8} \end{bmatrix} \begin{bmatrix} C_{0,0} & C_{0,1} & \cdots & C_{0,8} \\ C_{1,0} & C_{1,1} & \cdots & C_{1,8} \\ \vdots & \vdots & \ddots & \vdots \\ C_{8,0} & C_{8,1} & \cdots & C_{8,8} \end{bmatrix} \begin{bmatrix} x_{i,0} \\ x_{i,1} \\ \vdots \\ x_{i,8} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=0}^{8} x_{i,j} C_{j,0} & \sum_{j=0}^{8} x_{i,j} C_{j,1} & \cdots & \sum_{j=0}^{8} x_{i,j} C_{j,8} \end{bmatrix} \begin{bmatrix} x_{i,0} \\ x_{i,1} \\ \vdots \\ x_{i,8} \end{bmatrix} \\
&= \sum_{j=0}^{8} x_{i,j} C_{j,0} x_{i,0} + \sum_{j=0}^{8} x_{i,j} C_{j,1} x_{i,1} + \cdots + \sum_{j=0}^{8} x_{i,j} C_{j,8} x_{i,8} \\
&= \sum_{j=0}^{8} \sum_{k=0}^{8} x_{i,j} C_{j,k} x_{i,k}
\end{aligned}
$$

で与えられる．このように， $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の全ての要素 $C_{j,k}$ に対して，その添え字に対応する $x_{i,j}$ と $x_{i,k}$ を掛けて加えたものになる．この式は， $x_{i,j}$ についての 2 次の項の和の形式となっているので，「2 次形式」と言われている．Excel での行列計算が，実際にはどのようなものなのか，シグマを用いた計算式と対比することによって理解を深めてもらいたい．

表 10.10 の下側に（0，1）型のデザイン行列から得られたポアソン回帰パラメータの推定値 $\hat{\boldsymbol{\beta}}^T$ およびパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を示す．表 10.10 の上側には，船種 A，建造年 60-64，運行年 60-74 を基準とした場合の各因子の水準の 1,000 月当たりの損傷数の推定値と 95%信頼区間を計算した結果である．

大変な計算のように思われるかもしれないが，第 4.5 節では 2×2 のパラメータの共分散行列に対して示した Excel の行列計算の計算式と，全く同じ形式であり，推定値 $\ln(\hat{y}_i / n_i)$ の分散が

$$
\begin{aligned}
Var[(\ln(\hat{y}_i / n_i)] &= \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T \\
&= \mathrm{Mmult}(\mathrm{Mmult}(\boldsymbol{x}_i\text{の範囲}，\ \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{の範囲}), \mathrm{Transpose}(\boldsymbol{x}_i\text{の範囲}))
\end{aligned}
$$

にて求まれば，95%信頼区間の推定値の計算は，選択の範囲を 2×2 から 9×9 に代えるだけで 2×2 の場合と同様に計算ができる．

95%信頼区間を求めるためには，推定値の分散を求めることが必須であるが，一般的には，1 次回帰式のあてはめが主体で，3×3 のパラメータの共分散行列を必要とする 2 次回帰式の 95%信頼区間の分散の計算事例を見いだすことができない．2 次式は，切片を含めて 3 変数で

表 10.10　予測プロファイルの 95%信頼区間の計算シート（0，1）型

			—— 船種 ——				—建造年—			運行年	月あたり	対数	—— 千月比 ——		
		切片	B	C	D	E	65-69	70-74	75-79	75-79	対数	分散	損傷数	信頼区間	
因子	水準	x_0	x_B	x_C	x_D	x_E	x_{C65}	x_{C70}	x_{C75}	x_{Op75}	損傷数	Var	$\hat{y}^{(1000)}$	L95%	U95%
船種	A	**1**	0	0	0	0	0	0	0	0	-6.4059	0.0800	1.6518	0.9490	2.8750
	B	1	**1**	0	0	0	0	0	0	0	-6.9493	0.0273	0.9593	0.6939	1.3263
	C	1	0	**1**	0	0	0	0	0	0	-7.0933	0.1715	0.8306	0.3689	1.8702
	D	1	0	0	**1**	0	0	0	0	0	-6.4819	0.1435	1.5309	0.7285	3.2170
	E	1	0	0	0	**1**	0	0	0	0	-6.0803	0.0925	2.2874	1.2601	4.1522
建造	60-64	1	0	0	0	0	0	0	0	0	-6.4059	0.0800	1.6518	0.9490	2.8750
年	65-69	1	0	0	0	0	**1**	0	0	0	-5.7088	0.0645	3.3168	2.0160	5.4569
	70-74	1	0	0	0	0	0	**1**	0	0	-5.5875	0.0600	3.7444	2.3167	6.0521
	75-79	1	0	0	0	0	0	0	**1**	0	-5.9525	0.1030	2.5994	1.3856	4.8765
運行	60-74	1	0	0	0	0	0	0	0	0	-6.4059	0.0800	1.6518	0.9490	2.8750
年	75-79	1	0	0	0	0	0	0	0	**1**	-6.0214	0.0848	2.4262	1.3710	4.2936

		—— 船種 ——				—— 建造年 ——			運行年
	切片	B	C	D	E	65-69	70-74	75-79	75-79
	x_0	x_B	x_C	x_D	x_E	x_{C65}	x_{C70}	x_{C75}	x_{Op75}
$\hat{\boldsymbol{\beta}}^T$	**-6.4059**	**-0.5434**	**-0.6874**	**-0.0760**	**0.3256**	**0.6971**	**0.8184**	**0.4534**	**0.3845**
共分散$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
x_0	**0.0800**	-0.0530	-0.0458	-0.0396	-0.0408	-0.0266	-0.0343	-0.0344	-0.0094
x_B	-0.0530	**0.0533**	0.0428	0.0390	0.0404	0.0038	0.0138	0.0160	0.0009
x_C	-0.0458	0.0428	**0.1831**	0.0384	0.0412	0.0030	0.0043	0.0126	-0.0002
x_D	-0.0396	0.0390	0.0384	**0.1428**	0.0387	0.0020	0.0024	-0.0111	-0.0003
x_E	-0.0408	0.0404	0.0412	0.0387	**0.0941**	-0.0002	-0.0025	0.0049	0.0013
x_{C65}	-0.0266	0.0038	0.0030	0.0020	-0.0002	**0.0379**	0.0272	0.0281	-0.0036
x_{C70}	-0.0343	0.0138	0.0043	0.0024	-0.0025	0.0272	**0.0487**	0.0367	-0.0090
x_{C75}	-0.0344	0.0160	0.0126	-0.0111	0.0049	0.0281	0.0367	**0.0919**	-0.0147
x_{Op75}	-0.0094	0.0009	-0.0002	-0.0003	0.0013	-0.0036	-0.0090	-0.0147	**0.0237**

月当たり対数推定値：$\ln(\hat{y}_i / n_i) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ =Mmult($\boldsymbol{x}_i$ の範囲，Transpose($\hat{\boldsymbol{\beta}}^T$ の範囲)

分散：$Var(\ln \hat{y}_i / n_i) = \boldsymbol{x}_i\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i^T$ =Mmult(Mmult($\boldsymbol{x}_i$ の範囲，$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の範囲)，Transpose($\boldsymbol{x}_i$ の範囲))

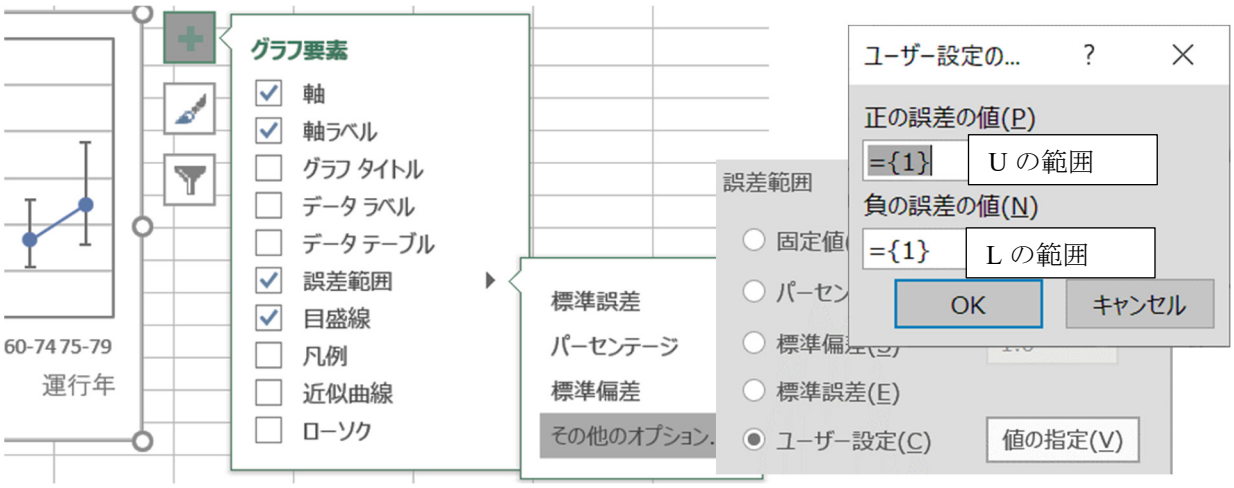
月当たり推定値；$\hat{y}_i^{(1000)} = \exp[(\ln(\hat{y}_i / n_i)] \times 1{,}000$

1,000 月あたりの 95%信頼区間：$95\%\text{CL} = \exp\left\{\ln(\hat{y}_i / n_i) \pm 1.96\sqrt{Var[(\ln(\hat{y}_i / n_i)]}\right\} \times 1{,}000$

あるのに対し，ここでの事例は 9 変数となっていて，Excel 行列関数なしには，各種の条件における 95%信頼区間の計算は絶望的である．なお，2 次回帰式の 95%信頼区間については，第 12.4 節を参照のこと．

95%信頼区間の計算は，それらの結果のグラフ表示とセットでなければ，計算した苦労が浮かばれない．また，グラフ化に手間がかかると避けたくもなるが，Excel の折れ線グラフで手軽に作図できる．なお，Excel の折れ線グラフで誤差範囲の設定の「標準誤差」は，表示されているデータ全体から計算された標準誤差を一律に表示するのでまったく使いものにならない．このようなまがい物の機能に惑わされることなく，「その他のオプション」で「ユーザ設定」を使って正しい 95%信頼区間の作図を心がけてもらいたい．

Excel の折れ線グラフに 95%信頼区間を上書きするためには，「グラフの要素➔誤差範囲➔その他のオプション➔ユーザ設定➔値の推定➔［正の誤差の値(P)，負の誤差の値(N)］」で，（上限までの距離の範囲，下限までの距離の範囲）で設定を行う．



Excel の誤差範囲の設定は，範囲の中心点からの距離で与える仕様になっているので，図 10.6（左）に示すように

正の誤差範囲： $U=U95\%-$推定値

負の誤差範囲： $L=$推定値$-L95\%$

を改めて計算する必要がある．この結果を用いて図 10.6（右）に 95%信頼区間付き予測プロファイルを示す．この図は，図 10.2 の JMP での予測プロファイルと同等の結果を表している．ただし，図 10.2 の Y 軸の目盛りは，月あたりの損傷数であるのに対し，図 10.6 では，1,000 月あたりの損傷数となっている．慣れるまでは煩わしいができあがりは上々である．このように，解析目的にそったグラフの作成は，統計ソフトの出力を Excel に取り込み，必要な計算を追加し，Excel による作図の連係プレーに挑戦してもらいたい．

		推定値	$y^{(1000)}$からの距離	
因子	水準	$y^{(1000)}$	L	U
船種	A	1.65	0.70	1.22
	B	0.96	0.27	0.37
	C	0.83	0.46	1.04
	D	1.53	0.80	1.69
	E	2.29	1.03	1.86
建造	60-64	1.65	0.70	1.22
年	65-69	3.32	1.30	2.14
	70-74	3.74	1.43	2.31
	75-79	2.60	1.21	2.28
運行	60-74	1.65	0.70	1.22
年	75-79	2.43	1.06	1.87

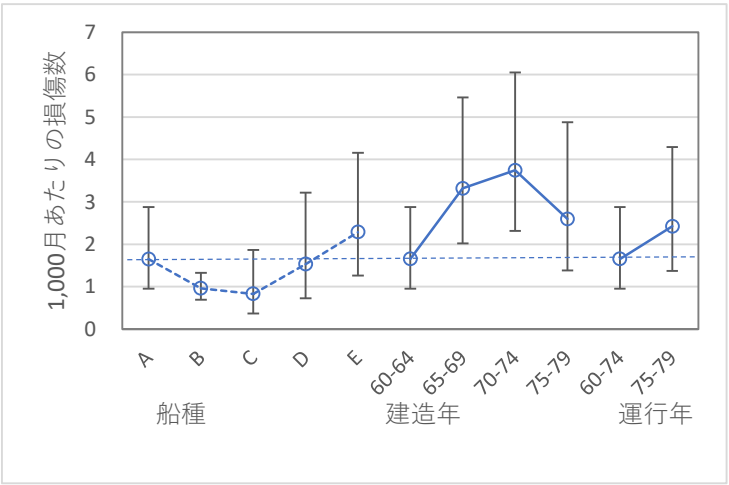


図 10.6　Excel の折れ線グラフによる 95%信頼区間付き予測プロファイル

10.4.　交互作用の検討

観察データに対して，主効果モデルで検討した結果，過分散であるということは，ある船種の，ある建造年の，ある運行年にかぎって損傷を受けやすい建造方法が取られている可能性が McCullagh ら（1989）で指摘されている．このような疑問に対して，交互作用を含めたポアソン回帰を適用して，結果を吟味する必要性がある．交互作用として（船種×建造年）と（船種×運行年）を含めたモデルを適用する．なお，（建造年×運行年）は，組み合わせがない場合（欠測セル，ミッシング・セル）があり，交互作用としては含めないことにする．なお，交互作用として（建造年×運行年）をモデルに含めた場合には，推定値に「バイアスあり」との警告が出される．

解析モデルは，次の式

ln 損傷数＝ln 運行月数＋ln(切片＋船種＋建造年＋運行年
＋船種×建造年＋船種×運行年)

とする．交互作用の自由度は，船種が 4，建造年が 3 なので，船種×建造年は 4×3＝12 となり，運行年の自由度は 1 なので，船種×運行年は 4×1＝4 となる．データ数が 34 なので，切片と主効果の自由度は 9 で，交互作用の自由度 16 を加えると 25 となり，誤差の自由度は 34－25＝9 となる．JMP での結果を表 10.4 に示す．　Pearson のカイ 2 乗が 9.8680 と自由度 9 とほぼ等しくなり過分散が解消している．

表 10.11　交互作用モデル 1（過分散なし）

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	68.9038	137.8075	24	<.0001*
完全	53.1936			
縮小	122.0974			

適合度統計量	カイ2乗	自由度	p値
Pearson	9.8680	9	0.3613
デビアンス	8.5208	9	0.4826

AICc
318.8873

効果の検定

要因	自由度	尤度比カイ2乗	p値
船種	4	1.7671	0.7785
建造年	3	13.2437	0.0041*
運行年	1	0.1822	0.6695
船種*建造年	12	25.2354	0.0137*
船種*運行年	4	6.0661	0.1943

交互作用が，統計的に有意となったのは，（船種×建造年）であり，有意ではない（船種×運行年）をモデルから除く．ただし，（運行年）と（船種）の主効果も有意ではないからとの理由で除いてはならない．これは，交互作用（船種×運行年）をモデルに入れた結果として，主効果が有意ではなくなったとも考えられるからである．同様に，交互作用（船種×建造年）が有意であるが，主効果（船種）は有意ではないからといってモデルから除いてしまうと，交互作用が適切に評価されなくなる恐れがある．なお，2 因子交互作用は，2 つの因子の主効果の存在下で定義されるものである．

表 10.12 に交互作用（船種×建造年）と主効果（船種，建造年，運行年）に対するポアソン回帰の結果を示す．Pearson のカイ 2 乗値は自由度 13 に対して 17.3680 とやや大きくなるが，統計的には有意でなはない．

表 10.12　交互作用モデル 2（過分散なし）

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	65.8707	131.7415	20	<.0001*
完全	56.2267			
縮小	122.0974			

適合度統計量	カイ2乗	自由度	p値
Pearson	17.3680	13	0.1830
デビアンス	14.5869	13	0.3338

AICc
231.4534

効果の検定

要因	自由度	尤度比カイ2乗	p値
船種	4	1.9918	0.7373
建造年	3	12.8690	0.0049*
運行年	1	10.6215	0.0011*
船種*建造年	12	24.1082	0.0197*

図 10.7 に（船種 E，建造年 70-74,，運行年 75-79）とした予測プロファイルを示す．予測プロファイルに引き続き交互作用プロファイルを描くこともできるが，水準の数が多いこともあり，煩雑で読み取りにくいので，図 10.7 の予測プロファイル上で，船種を A から E まで順次選択した時の建造年の予測プロファイルの変化を詳細に検討する．

図 10.8 に示すように，船種 A，B，C では，建造年の予測プロファイルは，月あたりの損傷数に若干の上下はあるもののほぼ同様の推移であり，交互作用はないものと推測される．船種 D では，建造年 70-74 に損傷数が急上昇し，建造年 75-79 には，船種 A，B，C と同程度の

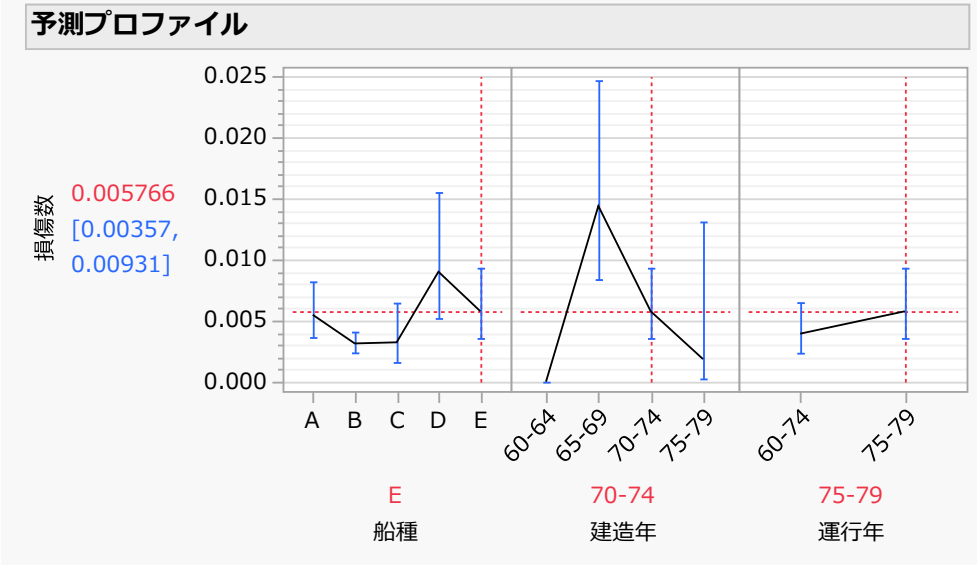


図 10.7　予測プロファイル（船種を E，建造年 70-74,，運行年 75-79）を基準

レベルに落ち着いている．船種 E は，建造年 65-69 に驚異的な損傷数の急上昇があり，建造年 70-74 では，半減平均的な損傷数となっている．建造年 75-79 では，さらに低いレベルまで損傷数が減っている．

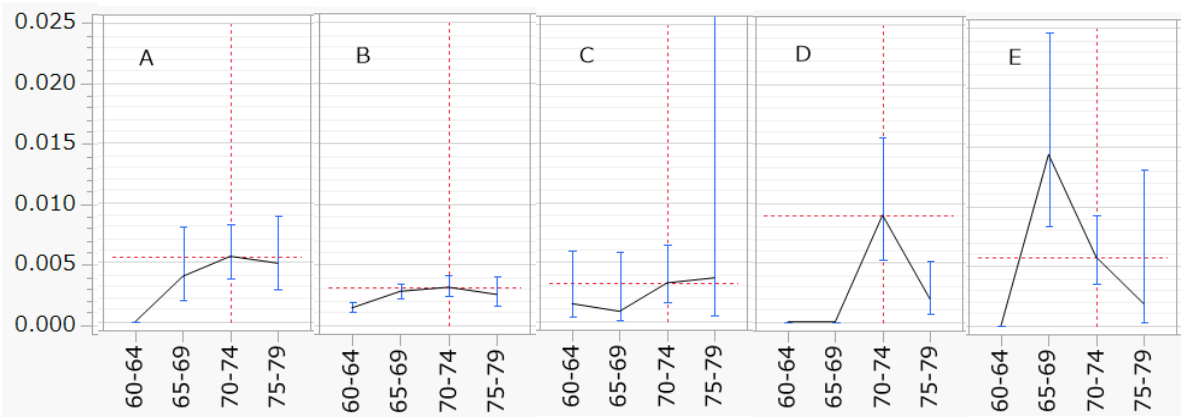


図 10.8　船種を変化させた場合の建造年の予測プロファイル

交互作用を含むモデルの予測プロファイルの推定値を報告書に表として示したい．主効果モデルの場合は，第 10.3 節で「Excel による予測プロファイル」で，共分散行列を JMP から取得して推定値および 95%信頼区間の計算を行い，折れ線グラフにより作図した結果を示してきたが，交互作用を含む場合には，デザイン変数が 21 なので，21×21 の行列となり，Excel による対応は非現実的である．

図 10.8 で示した予測プロファイルは，運行年 75-79 に固定し，船種を A～E と変化させつつ建造年別の推定値と 95%信頼区間を描いている．これらの推定値を得るためには，JMP の

ポアソン回帰の追加機能である「列の保存➔（予測式，平均の信頼区間)」を使うことにより得ることができる．

表 10.13 に示すように，解析用のデータの末尾に推定したい条件についてのデータを追加することにより，自動的に推定値，および 95%信頼区間を得ることができる．JMP の予測プロファイルは，月あたりの故障数となっているので，運行月数として 1 とする．もちろん対数は 0.0 になる．表 10.13 には，追加レコードも含めて JMP ファイルを Excel に取り込み，折れ線グラフ用の長さ（L，U）を計算した結果が示されている．これらを用いて図 10.9 に示すように，JMP と同様な交互作用プロファイルをが得られる．

表 10.13　追加レコードに対する JMP による予測値と 95%信頼区間の計算

No	船種	建造年	運行年	運行月数	ln運航月数	損傷数	目盛り	交互作用予測値	L95%	U95%	長さL	長さU
1	A	60-64	60-74	127	4.84	0		0.00	0.00			
2	A	60-64	75-79	63	4.14	0		0.00	0.00			
3	A	65-69	60-74	1095	7.00	3		2.83	1.33	6.02		
:												
38	E	70-74	75-79	2161	7.68	12		12.46	7.72	20.13		
39	E	75-79	60-74	.								
40	E	75-79	75-79	542	6.30	1		1.00	0.14	7.10		
	A	60-64	75-79	1	0.00		A60	**0.0000**	0.0000		**0.0000**	
追	A	65-69	75-79	1	0.00		A65	**0.0038**	0.0018	0.0080	**0.0020**	**0.0042**
加	A	70-74	75-79	1	0.00		A70	**0.0055**	0.0037	0.0082	**0.0018**	**0.0027**
レ	A	75-79	75-79	1	0.00		A75	**0.0049**	0.0027	0.0089	**0.0022**	**0.0040**
コ	B	60-64	75-79	1	0.00		B60	**0.0014**	0.0011	0.0019	**0.0003**	**0.0005**
-	B	65-69	75-79	1	0.00		B65	**0.0028**	0.0022	0.0035	**0.0006**	**0.0007**
ド	:						:					
	E	70-74	75-79	1	0.00		E70	**0.0058**	0.0036	0.0093	**0.0022**	**0.0035**
	E	75-79	75-79	1	0.00		E75	**0.0018**	0.0003	0.0131	**0.0015**	**0.0113**

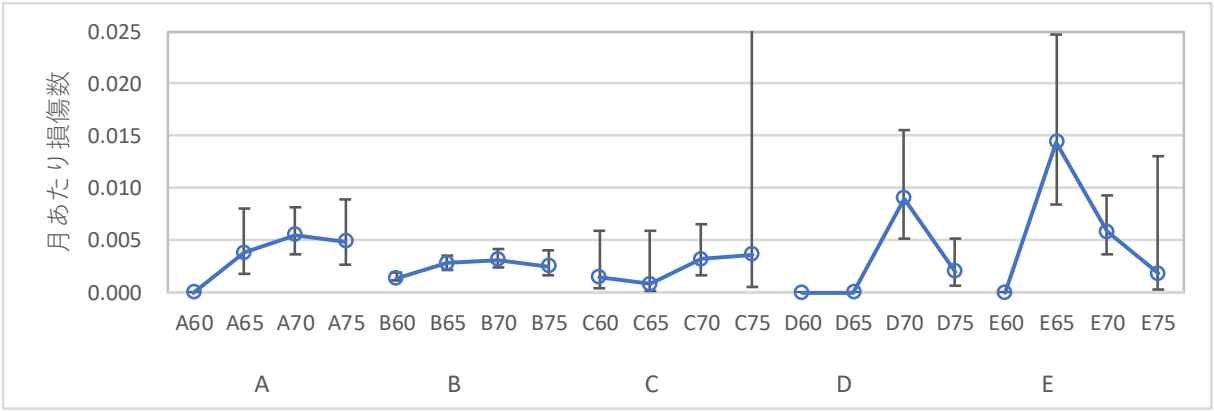


図 10.9　Excel による船種を変化させた場合の建造年の交互作用予測プロファイル

10.5.　主効果モデルを活用した新たな交互作用の可視化の試み

観察データに対する交互作用を含めたポアソン回帰は，因子間の特異的な関連を見極めることが必要である．しかし，船種に対する損傷数の解析の本来の目的は，保険会社の視点からは，直近の損傷数のデータから，近未来の損傷数を予測し，保険料を適切に算定するのが目的と思われる．

ある年に建造されたある船舶の種類によって，実際の運航年に起きた損傷数が，平均的な損傷数に比較して多くなった場合には，次年度の保険料を高く設定する必要があり，損傷数が平均的な損傷数に比べて減少傾向にあるならば．保険料を安く設定することも必要と思われる．

船舶の種類と建造年との間に交互作用があるということは，ある船種の損傷数ついて特異的な変動があることが示唆されたことになる．どのような特異的な変動かは，図 10.8 の船種別の建造年による損傷数の推移によって，船種 D と船種 E に特異的な変動があることが認識されるが，この図は予測プロファイルを手作業で変化させて，切り貼り作業で完成させたものであり，全体を概観するには難点がある．

伝統的に交互作用の検出には，分散分析表の分散比の F 検定で行ない，有意ならばグラフに示し目視による解釈ををするのが常である．ポアソン回帰の場合であれば，表 10.11 および表 10.12 に示したように，尤度比カイ 2 乗検定の結果を分散比の F 検定と同様に示してきたのであるが，グラフ表示を試みるも隔靴掻痒のごとくである．

どのような交互作用なのかを手軽にすばやく認識するために，主効果モデルでの予測値 $\ln \hat{y}_i$ を，損傷千月比に換算した予測値 $\hat{y}_i^{(1,000)}$（以下，主効果予測値とする）を計算する．元データの損傷数から計算した損傷千月比 $y_i^{(1,000)}$（以下，損傷千月比とする）と対比することにより，交互作用が浮き彫りできるのではないかと思われた．

主効果モデルの予測値 $\hat{y}_i^{(1,000)}$ は，表 10.3 のデザイン行列に対し，表 10.4 に示した JMP で得られたのポアソン回帰のパラメータの推定値 $\hat{\boldsymbol{\beta}}$，デザイン行列の i 行目のベクトル $\boldsymbol{x}_i$ を用い，表 10.14 に示すように

$$\hat{y}_i^{(1,000)} = \frac{\exp\left[\ln(n_i) + (\boldsymbol{x}_i\hat{\boldsymbol{\beta}})\right]}{n_i} \times 1000$$

によって計算した結果である．

表 10.14　主効果モデルにおける予測値の計算

	船	建造	運行	運行	損傷	損傷千月比	デザイン行列									予測値		主効果予測値
No	種	年度	年度	月数n_i	数y_i	$y_i^{(1000)}$	x_0	x_B	x_C	x_D	x_E	x_{C65}	x_{C70}	x_{C75}	x_{Op75}	$\hat{\boldsymbol{\beta}}$		$\hat{y}_i^{(1000)}$
1	A	60-64	60-74	**127**	0	0.0000	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**-6.4059**	=	**1.6518**
2			75-79	63	0	0.0000	1	0	0	0	0	0	0	0	1	**-0.5433**		2.4262
3		65-69	60-74	1,095	3	2.7397	1	0	0	0	0	1	0	0	0	**-0.6874**		3.3168
4			75-79	1,095	4	3.6530	1	0	0	0	0	1	0	0	1	**-0.0760**		4.8718
5		70-74	60-74	1,512	6	3.9683	1	0	0	0	0	0	1	0	0	**0.3256**		3.7445
6			75-79	3,353	18	5.3683	1	0	0	0	0	0	1	0	1	**0.6971**		5.5000
7		75-79	60-74	-	-	-	1	0	0	0	0	0	0	1	0	**0.8184**		-
8			75-79	2,244	11	4.9020	1	0	0	0	0	0	0	1	1	**0.4534**		3.8181
9	B	60-64	60-74	44,882	39	0.8689	1	1	0	0	0	0	0	0	0	**0.3845**		0.9594
10			75-79	17,176	29	1.6884	1	1	0	0	0	0	0	0	1			1.4091
:																		
33	E	60-64	60-74	45	0	0.0000	1	0	0	0	1	0	0	0	0			2.2874
34			75-79	-	-	-	1	0	0	0	1	0	0	0	1			-
35		65-69	60-74	789	7	8.8720	1	0	0	0	1	1	0	0	0			4.5932
36			75-79	437	7	-	1	0	0	0	1	1	0	0	1			6.7466
37		70-74	60-74	1,157	5	4.3215	1	0	0	0	1	0	1	0	0			5.1855
38			75-79	2,161	12	5.5530	1	0	0	0	1	0	1	0	1			7.6166
39		75-79	60-74	-	-	-	1	0	0	0	1	0	0	1	0			-
40			75-79	542	1	1.8450	1	0	0	0	1	0	0	1	1			5.2874
				$\hat{y}_i^{(1000)}$ =(exp(ln(n_i) +Mmult($\boldsymbol{x}_i$の範囲, $\hat{\boldsymbol{\beta}}$の範囲)) / n_i *1000														

JMP のデータセットにもこれと同じ結果が出力されおり，図 10.10 に示すようにグラフ・ビルダーで「船種別・運行年別・建造年別の主効果予測値 $\hat{y}_i^{(1,000)}$ の推移図を描き，さらに損傷千月比 $y_i^{(1,000)}$ を上書きした結果である．平均的な損傷数に対応するのが主効果予測値 $\hat{y}_i^{(1,000)}$ の推移であり，実際の損傷千月比 $y_i^{(1,000)}$ により，特異的な変動が浮き彫りされている．

1）多くの場合，1,000 月あたり 5 件以下の損傷数である．
2）船種 C タイプの建造年 1970-74 の 1960-74 の運行年の千月比の損傷数は，7.66 件となるが，1975-79 の運行年では，1.03 件と減少している．
3）船種 D タイプの建造年 1970-74 は，運行年に関わらず千月あたりの損傷数が多い．
4）船種Ｅタイプは，建造年 1965-70 は，運行年に関わらず千月あたりの損傷数が 8.87, 16.02 と多く，船種Ｅタイプの建造年 1970-74 は，建造年 1965-70 に比べて減少しているが，4.32, 5.55 と少なくはない．しかし，船種Ｅタイプの建造年 1975-79 の場合には，1.85 と減少している．

保険会社の観点から考察してみよう．

1）船種 A は，建造年 75-79 の運行年 75-79 の $y_i^{(1,000)}$ が「平均」より僅かに多くなっている．次年度にさらに多くなるか注意が必要であるが，保険料は据え置きとする．
2）船種 B は，常に「平均」的な推移であり，通常の保険料で据え置きとする．

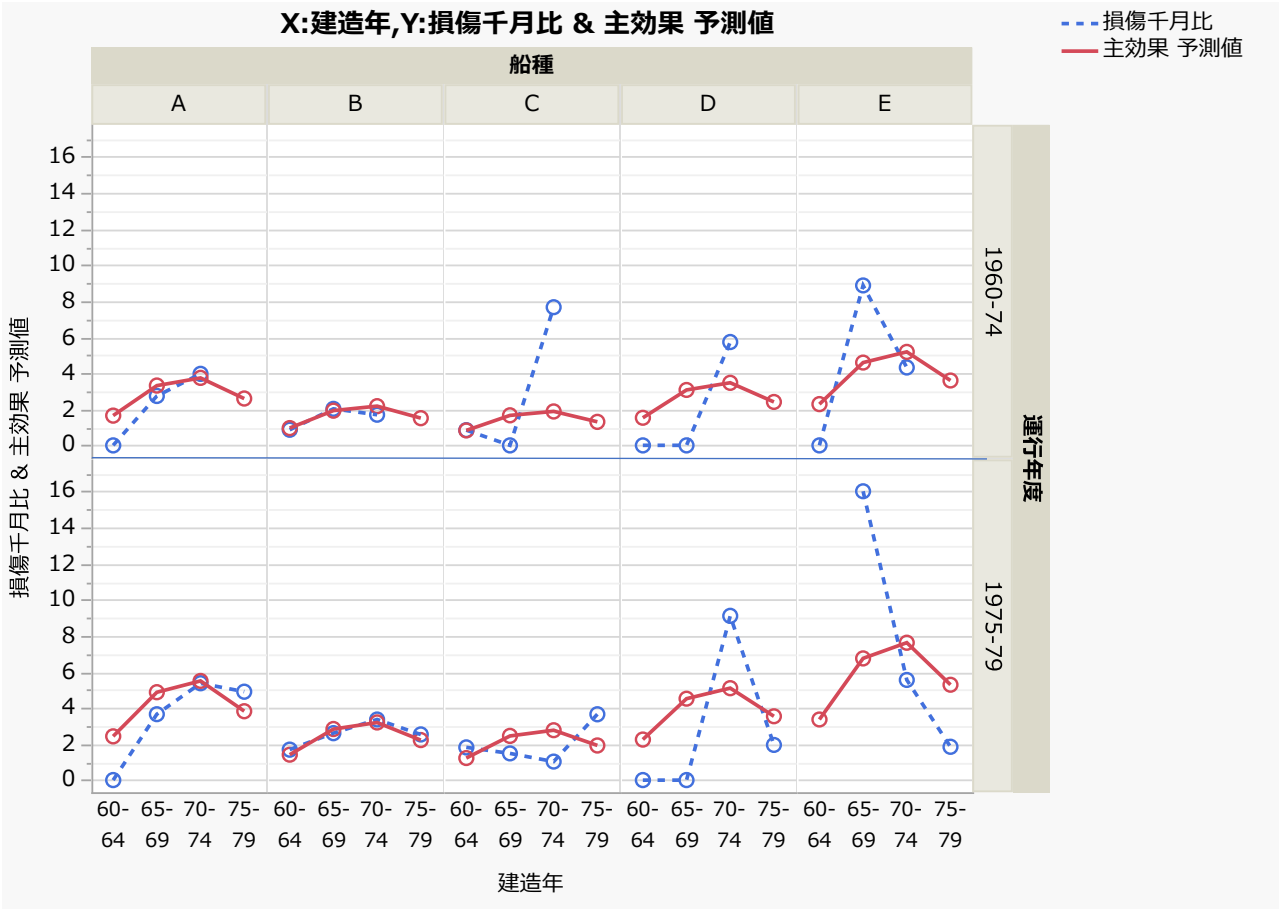


図 10.10　船種別，建造年別の運行年による 1,000 月あたりの損傷数の変化

3）船種 C は，建造年 70-74 の場合に運行年 60-74 で「平均」よりも明らかに $y_i^{(1,000)}$ が多くなったが，運行年 70-79 で $y_i^{(1,000)}$ が「平均」以下となっているので，何が原因なのかを検討の上，慎重に保険料の算定が必要であろう．

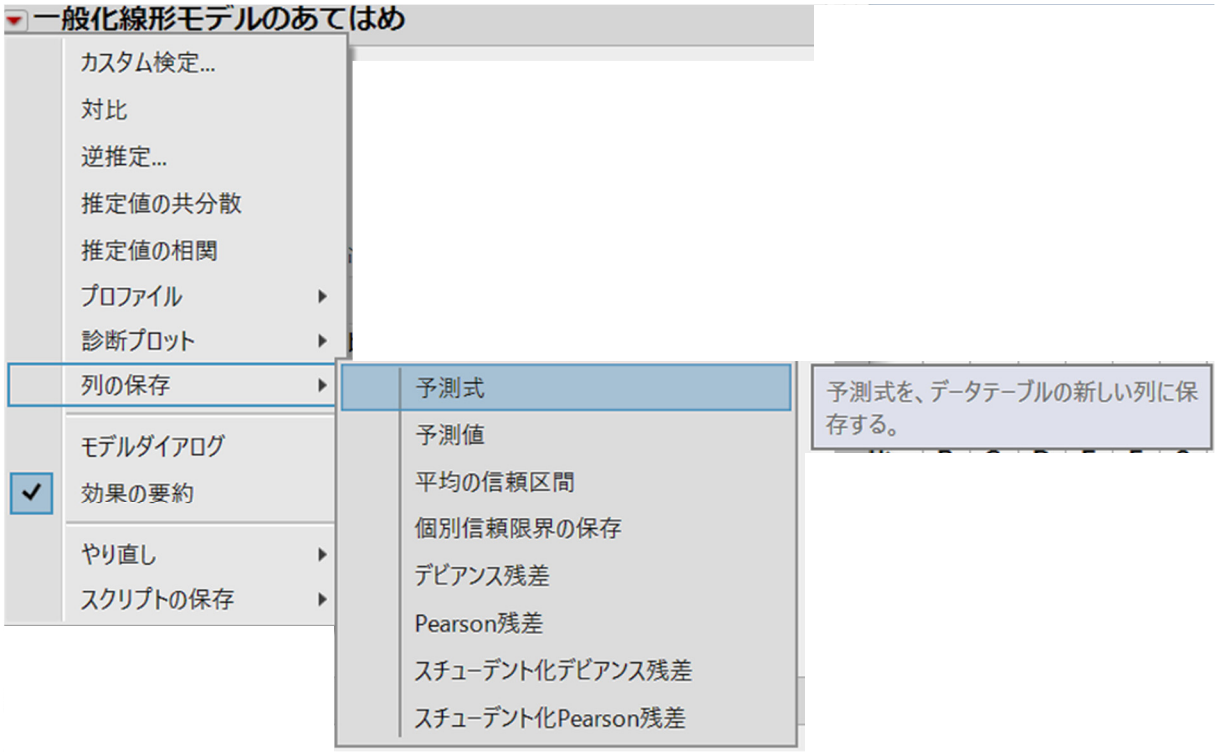
4）船種 D は，建造年 70-74 の場合の運行年 60-74 で $y_i^{(1,000)}$ が「平均」より明らかに多くなり，運行年 70-79 でも同様の傾向であり，保険料を高く設定してあるか，確認を要する．

5）船種 E は，建造年 65-69 の $y_i^{(1,000)}$ が運行年 60-74 でも「平均」より明らかに多く，運行年 70-79 では，さらに増加しており，次年度以降さらなる保険料の増額が必要である．建造年 70-74，建造年 75-79 については，「平均」的な保険料での継続としてよいだろう．

JMP による主効果モデルのポアソン回帰の予測値から $\hat{y}_i^{(1,000)}$ を求め，グラフ・ビルダーを用いて実際の損傷数 $y_i^{(1,000)}$ を重ね合わせた推移図が，結果を考察する際に有益であると思われる．統計ソフト S-PLUS では，JMP のグラフ・ビルダーに相当する Trellis（格子）グラフ機能が備わっているが，R 言語でも Trellis（格子）グラフのパッケージが提供されているよ

うだが使用経験はない［久保訳（2009)］．なお，SAS には，グラフ・ビルダーに代わるものがない．

Excel の折れ線グラフを用いて，図 10.10 に匹敵するものが容易できることを示そう．まず，主効果モデルで推定されたパラメータを用いた「予測値」を外部ファイルに出力する．JMP では，「列の保存➔予測式」で推定値 $\hat{y}_i$ が得られる．



元のデータと予測値を Excel に取り込み，損傷数 y_i を $y_i^{(1,000)}$ に，予測値 $\hat{y}_i$ を $\hat{y}_i^{(1,000)}$ に，次式で，

$$y_i^{(1,000)} = \frac{y_i}{n_i} \times 1{,}000\,, \qquad \hat{y}_i^{(1,000)} = \frac{\hat{y}_i}{n_i} \times 1{,}000$$

計算する．更に運用年ごとに上下に配置したいので，運用年別・船種別・建造年別にソートした結果を表 10.15 に示す．運行年毎に $y_i^{(1,000)}$ と $\hat{y}_i^{(1,000)}$ を選択し，折れ線グラフを作成し，線の色とマーカの書式を設定し，船種間の切れ目の線を消して，図 10.11 が完成する．

細々した交互作用の検討をせずに，主効果モデルによる予測値 $\hat{y}_i^{(1,000)}$ が図示されているので，損傷千月比 $y_i^{(1,000)}$ が予測値 $\hat{y}_i^{(1,000)}$ 同等レベルなのか，あるいは，大きく外れているが一目で把握でき，交互作用の解析結果と同様な判断が行なえる．私にとっても，このような予測プロファイルの作成は，これが初めてであり，新たな解析法として有益であることを認識した．

表 10.15　Excel の折れ線グラフのための Excel シート

運行年	船種	運行年	損傷数 $y_i^{(1000)}$	推定値 $\hat{y}_i^{(1000)}$
Op60-74	A	A60	0.00	1.65
		A65	2.74	3.32
		A70	3.97	3.74
		A75		2.60
	B	B60	0.87	0.96
		B65	2.03	1.93
		B70	1.70	2.17
		B75		1.51
	C	C60	0.85	0.83
		C65	0.00	1.67
		C70	7.66	1.88
		C75		1.31
	D	D60	0.00	1.53
		D65	0.00	3.07
		D70	5.73	3.47
		D75		2.41
	E	E60	0.00	2.29
		E65	8.87	4.59
		E70	4.32	5.19
		E75		3.60

運行年	船種	運行年	損傷数 $y_i^{(1000)}$	推定値 $\hat{y}_i^{(1000)}$
Op75-79	A	A60	0.00	2.43
		A65	3.65	4.87
		A70	5.37	5.50
		A75	4.90	3.82
	B	B60	1.69	1.41
		B65	2.60	2.83
		B70	3.36	3.19
		B75	2.53	2.22
	C	C60	1.81	1.22
		C65	1.48	2.45
		C70	1.03	2.77
		C75	3.65	1.92
	D	D60	0.00	2.25
		D65	0.00	4.52
		D70	9.11	5.10
		D75	1.95	3.54
	E	E60		3.36
		E65	16.02	6.75
		E70	5.55	7.62
		E75	1.85	5.29

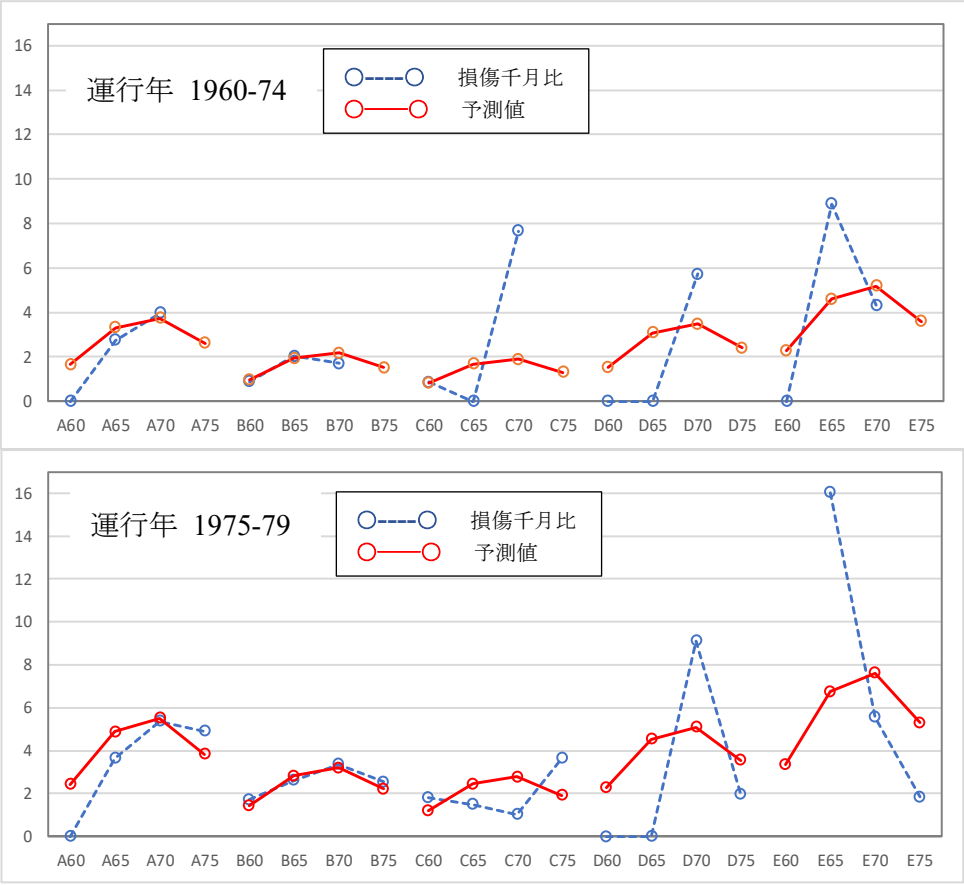


図 10.11　Excel による船種別，建造年別の運行年による損傷千月比と予測値との対比

10.6. Excel のソルバーによるオフセットを含むポアソン回帰

Excel のソルバーを用いてオフセットを含む場合のポアソン回帰については，これまでも取り上げてきたが，変数の数は切片を含め 3 変数までであり，更に変数が多い場合の適用可能性について検討する．

第 2 章では，対数尤度関数について 2 階の偏微分行列を用いたニュートン・ラフソン法を用いた反復計算を示した．ただし，変数が多くなると 2 階の偏微分行列の計算が変数の 2 乗に比例して増えるので，3 変数までは 3×3 の内 6 変数の計算であるが，4 変数となると 4×4 の 16 変数となり，Excel での対応は限界を超えている．

第 5 章では，反復重み付き回帰による方法を示したが，変数の数が増えてもデザイン行列 $\boldsymbol{X}$ に対する種々の行列計算なので，行列のサイズを変更するだけで対応可能であり，変数が増えたとしても対応は可能ではあるが，5 変数となると 5×5 の行列計算が必要となり，限界に近い．

第 2 章でも第 5 章でもパラメータの推定だけでなく，パラメータの共分散行列の推定も Excel で行っており，このために計算手順が込み入っている．これらは，ポアソン回帰を通じて一般化線形モデルの理論をきちっと学習するために適しているが，変数が多くなった場合には，勧められない．

第 10.5 節で示したように，95%信頼区間の算出せずに，モデルによる予測値と観測値の対応を主体にするような場合には，Excel のソルバーを用いることにより容易に各種のポアソン回帰のパラメータの推定を行うことができる．この方法ならば，これまでも示してきたガンマ・ポアソン回帰（負の 2 項分布），ゼロ過剰ポアソン回帰，ゼロ過剰ガンマ・ポアソン回帰のパラメータの推定でも，Excel のソルバーによる主効果モデルに対応可能である．

第 10.2 節の「主効果モデル」で質的変数に対するデザイン行列を使い，表 10.14 に示すように JMP で推定したパラメータを用いたオフセットがある場合のポアソン回帰の予測値の計算過程を示した．統計ソフトによるポアソン回帰を行う際に，どのようなデザイン行列が内部で生成されているのかを正確に知ることが，推定された結果の解釈に不可欠であることを繰返し示してきた．

生成されたデザイン行列を使い，Excel ソルバーを使ってポアソン回帰に限らず各種の回帰モデルに対してパラメータの推定を行うことに煩雑さはなく，生成されたデザイン行列に対応したパラメータの推定値が，どのような性質を持つのを理解し，結果の解釈に役立つ各種のグラフを試行錯誤的に作成することは，現実の問題に対して統計ソフトから出力される各種の結果を，解釈しやすいグラフを自在に作成するために役に立つ．

第 10.2 節の表 10.3 をベースに，Excel ソルバーを用い最尤法によるポアソン回帰を行うために拡張する．表 10.16 の「変化セル」から右側の列が表 10.13 に付け加えたものである．ポアソン確率 P_i は，「$\hat{y}_i$」を「平均」とした場合の「y_i」のポアソン分布の確率で Excel の Poisso.dist() 関数で計算した結果である．その確率 P_i の対数を取った「$\ln L_i$」がそれぞれの対数尤度となっている．「$\ln L_i$」を全て加えた結果が対数尤度「$\ln L = -125.245$」であり，これを最大にするようにパラメータ $\boldsymbol{\beta}$ を Excel のソルバーで変化させることにより最尤解が得られる．

初期値としてパラメータ（–6，0，0，0，0，0，0，0，0）を与えている．これは，$\beta_0 = -7$ とすると $\ln L = -224.95$，$\beta_0 = -5$ とすると $\ln L = -465.94$ となり，$\beta_0 = -6$ としたときが $\ln L = -125.245$ と大きくなったためである．さらに，その前後（-5.5，-6.5）などと動かして $\ln L$ が大きくなるか試し，適当な初期値を設定する．Excel のソルバーで $\ln L = -125.245$ のセルを「目的のセルの設定」とし，「変数セルの変更」欄で，パラメータ（–6，0，0，0，0，0，0，0，0）のセルを選択し，「解決」により，対数尤度 $\ln L$ を最大にするようなパラメータを瞬時に求めてくれる．

表 10.16　初期値に対する対数尤度

		建	運	運行	損傷		---船種---				--建造--			Op	変化		$\ln n_i +$	exp(ln	ポアソ	$\ln L$	最
	船	造	行	月数	数	x	B	C	D	E	65	70	75	65	セル		$\boldsymbol{x\beta}\hat{}=$	$y_i\hat{}$)	ン確率	-125.245	大
No	種	年	年	n_i	y_i	0	1	2	3	4	5	6	7	8	$\boldsymbol{\beta}\hat{}$		$\ln y_i\hat{}$	$y_i\hat{}$	P_i	$\ln L_i$	化
1	A	60-	60-	**127**	0	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**-6.0000**	=	**-1.16**	**0.3**	**0.730**	**-0.315**	セ
2			75-	63	0	1	0	0	0	0	0	0	0	1	**0.0000**		-1.86	0.2	0.855	-0.156	ル
3		65-	60-	1,095	3	1	0	0	0	0	1	0	0	0	**0.0000**		1.00	2.7	0.221	-1.510	
4			75-	1,095	4	1	0	0	0	0	1	0	0	1	**0.0000**		1.00	2.7	0.150	-1.898	
5		70-	60-	1,512	6	1	0	0	0	0	0	1	0	0	**0.0000**		1.32	3.7	0.091	-2.400	
6			75-	3,353	18	1	0	0	0	0	0	1	0	1	**0.0000**		2.12	8.3	0.001	-6.590	
8		75-	75-	2,244	11	1	0	0	0	0	0	0	1	1	**0.0000**		1.72	5.6	0.015	-4.188	
9	B	60-	60-	44,882	39	1	1	0	0	0	0	0	0	0	**0.0000**		4.71	111.3	0.000	-34.123	
10			75-	17,176	29	1	1	0	0	0	0	0	0	1	**0.0000**		3.75	42.6	0.006	-5.045	
:																					
37		70-	60-	1,157	5	1	0	0	0	1	0	1	0	0			1.05	2.9	0.092	-2.387	
38			75-	2,161	12	1	0	0	0	1	0	1	0	1			1.68	5.4	0.005	-5.204	
40		75-	75-	542	1	1	0	0	0	1	0	0	1	1			0.30	1.3	0.351	-1.048	

表 10.18 に対数尤度を最大化するようなパラメータ（–6.4059，–0.5433，–0.6874，–0.0759，0.3255，0.6972，0.8185，0.4534，0.3845）が推定されている．この結果が正しいことは，表 10.4 の JMP でのポアソン回帰の結果と一致することにより検証される．得られた $\ln L = -68.2808$ は，JMP の結果の対数のサマリー表の「完全」に一致している．

表 10.17　JMP による主効果モデルの対数尤度（過分散なし）

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	53.8166	107.6333	8	<.0001*
完全	68.2808			
縮小	122.0974			

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)
Pearson	42.2753	25	0.0168*
デビアンス	38.6951	25	0.0395*

AICc
162.0615

表 10.18　ソルバーによる対数尤度の最大化

		建	運	運行	損傷		---船種---				--建造--			Op	変化		$\ln n_i +$	exp(ln	ポアソ	$\ln L$	最
	船	造	行	月数	数	x	B	C	D	E	65	70	75	65	セル		$x\hat{\beta}=$	$y_i\hat{}$)	ン確率	-68.2808	大
No	種	年	年	n_i	y_i	0	1	2	3	4	5	6	7	8	$\hat{\beta}$		$\ln y_i\hat{}$	$y_i\hat{}$	P_i	$\ln L_i$	化
1	A	60-	60-	**127**	0	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**-6.4060**	=	**-1.56**	**0.2**	**0.811**	**-0.210**	セ
2			75-	63	0	1	0	0	0	0	0	0	0	1	**-0.5433**		-1.88	0.2	0.858	-0.153	ル
3		65-	60-	1,095	3	1	0	0	0	0	1	0	0	0	**-0.6873**		1.29	3.6	0.211	-1.554	
4			75-	1,095	4	1	0	0	0	0	1	0	0	1	**-0.0759**		1.67	5.3	0.163	-1.816	
5		70-	60-	1,512	6	1	0	0	0	0	0	1	0	0	**0.3256**		1.73	5.7	0.159	-1.839	
6			75-	3,353	18	1	0	0	0	0	0	1	0	1	**0.6972**		2.91	18.4	0.093	-2.374	
8		75-	75-	2,244	11	1	0	0	0	0	0	0	1	1	**0.8185**		2.15	8.6	0.087	-2.442	
9	B	60-	60-	44,882	39	1	1	0	0	0	0	0	0	0	**0.4535**		3.76	43.1	0.052	-2.950	
10			75-	17,176	29	1	1	0	0	0	0	0	0	1	**0.3845**		3.19	24.2	0.047	-3.052	
11		65-	60-	28,609	58	1	1	0	0	0	1	0	0	0			4.01	55.1	0.049	-3.025	
:																					
37		70-	60-	1,157	5	1	0	0	0	1	0	1	0	0			1.79	6.0	0.161	-1.829	
38			75-	2,161	12	1	0	0	0	1	0	1	0	1			2.80	16.5	0.059	-2.836	
40		75-	75-	542	1	1	0	0	0	1	0	0	1	1			1.05	2.9	0.163	-1.813	

Excel のソルバーを使うことにより，推定値は得られるが信頼区間を計算するために必須のパラメータの共分散行列が得られないのが残念である．ただし，主効果モデルでの予測値に対する観測値を同一の折れ線グラフで表わすことの有用性を第 10.5 節で示したが，Excel のソルバーの結果を用いても容易に実現できる．

10.7. SAS の GENMOD プロシジャを使った解析

SAS の GENMOD プロシジャが本格的に提供されたのは，2000 年ごろのバージョン 8 からである. その当時は, 実務上ロジスティック回帰に対する適用を主体にしており, 高橋 (2002), 「GENMOD プロシジャによる計数データの解析」では，LOGISTIC プロシジャとの使い分けについての検討を主体にし，ポアソン回帰について取り上げていなかった．無償版の OnDemand SAS が提供されるようになったので，これまでの章でも GENMOD プロシジャによるポアソン回帰の使い方を示してきた．ここでは，第 10.5 節で示した主効果モデルによる探索的な解析について GENMOD プロシジャを用いて，負の 2 項回帰（ガンマ・ポアソン回帰）への拡張も示す．無償版の SAS の利用に関しては，高波・舟尾（2016),「SAS Stadio によるやさしい統計データ分析」が詳しい．SAS と R と用いた各種の統計解析については，臨床評価研究会（ACE）基礎解析分科会（2017）に詳しく示されているが，残念ながらポアソン回帰についての記述はない.

SAS データセットの作成

SAS は，DATA ステップおよび PROC ステップの繰返しを基本としている．DATA ステップとは,「DATA」で始まるステップで，SAS データセットを作成するためのプログラミング言語である．PROC ステップとは，GENMOD プロシジャなど数百の解析用のプロシジャを起動するステップである．DATA ステップで解析用の SAS データセットを作成し，PROC ステップで解析した結果を SAS データセットへ出力し, DATA ステップで加工し, また別の PROC ステップでの解析を行う．このような SAS データセットを介した繰返が特徴である.

SAS の DATA ステップでは，表 10.3 に示すような Excel ファイルを読み込むこともできるが，必要最小限のデータリストから，プログラミング機能を用いて解析用の SAS データセットを作成するのが通常である．必要最小限なデータは，運行月数 n_i および損傷数 y_i であり，他の因子の水準は，プログラミング的に内部生成する．実際のデータリストは，datalines ステートメント以下の 5 行で，最初の行は船種 A で，その後に（n_1，y_1，n_2，y_2,...,n_8，y_8）が続いている．船種（A, B, C, D, E）も内部で生成できるが，行の識別子として付けてある.

SAS の DAT ステップの基本は，1 行分のデータを読み込み，何らかの処理をした後に最初のステートメントに戻り, 読み込むデータが尽きるまで続ける. 1 行の最初の船種の文字データを読み込み，建造年 Cons とし 4 年度分，運行年度 Oper とし 2 年度分，合計 8 セットの運行月数 Month，損傷数 Inci を順次読み込んで output ステートメントで，それ以前に設定され

た変数名とそのデータを SAS データセット d01 に順次出力する．1 行分の入力データリストから 8 行分の SAS データセットができあがり，5 行分のデータリストから，全体で 40 行分の SAS データセットが生成される．

DATA ステップで生成された SAS データセット d01 を確認するためには，proc print で SAS データセット名を data=d01 と指定して出力する．なお，Data ステップの詳細は，第 9.3 節を参照のこと．

```
Title "Poisson_S10_Ship_Incidents   2019/12/011 Y.Takahashi" ;
data d01 ;
  input Type$  @@ ;
  do Cons = "60-64", "65-69", "70-74", "75-79" ;
    do Oper = "60-74", "75-79" ;
      input  Month Inci   @@ ;
      ln_Month = log(Month) ; output ;
    end;
  end ;
datalines ;
A    127  0    63  0  1095  3  1095  4 1512  6  3353 18   .  .  2244 11
B  44882 39 17176 29 28609 58 20370 53 7064 12 13099 44   .  .  7117 18
C   1179  1   552  1   781  0   676  1  783  6  1948  2   .  .   274  1
D    251  0   105  0   288  0   192  0  349  2  1208 11   .  .  2051  4
E     45  0     .  .   789  7   437  7 1157  5  2161 12   .  .   542  1
;
proc print data=d01 ;  run;
```

Proc print による 40 行分の出力

Obs	Type	Oper	Cons	Month	Inci	ln_Month
1	A	60-64	60-74	127	0	4.8442
2	A	60-64	75-79	63	0	4.1431
3	A	65-69	60-74	1095	3	6.9985
4	A	65-69	75-79	1095	4	6.9985
:						
39	E	75-79	60-74	.	.	.
40	E	75-79	75-79	542	1	6.2953

過分散を考慮したポアソン回帰

GENMOD プロシジャで，分布を dist=poisson，リンク関数を link=log，過分散の指定を scale=pearson，オフセットを offset=ln_Month とし，推定値 $\ln \hat{y}_i$ を xbaata=xbeta，推定値 $\hat{y}_i$ を predicted=pred，スチューデント化デビアンス残差を stdresdev=s_dev を SAS データセット output

out=out01 で出力している．再度 proc print でこれらの統計量を出力し，proc univariate で s_dev についての基本統計量を計算している．

```
Title2 '<<< poisson offset ref=first 過分散 >>>' ;
proc genmod data=d01  ;
     class Type Cons Oper / ref=first ;
     model Inci = Type Cons Oper
          / dist=poisson  link=log type3
            scale=Pearson offset=ln_Month ;
      output  out=out01 xbeta=xbeta predicted=pred
                        stdresdev=s_dev ;
      run ;
proc print  data=out01 ;    run ;
proc univariate data=out01 ;
    var s_dev ;             run;
```

<<< poisson offset ref=first 過分散 >>>
GENMOD プロシジャ

適合度評価の基準			
基準	自由度	値	値/自由度
デビアンス	25	38.6951	1.5478
Scaled デビアンス	25	22.8828	0.9153
Pearson カイ 2 乗	25	42.2753	1.6910
Scaled Pearson カイ 2 乗	25	25.0000	1.0000
対数尤度		454.1742	
完全対数尤度		-68.2808	
AIC (小さいほどよい)		154.5615	
AICC (小さいほどよい)		162.0615	
BIC (小さいほどよい)		168.2988	

アルゴリズムは収束しました。

注） デビアンス=38.6951 は，過分散なしの場合．表 10.17 に一致．
Scaled デビアンスは，デビアンスを尺度 1.3004 の 2 乗=1.6910 で除した結果
Pearson カイ 2 乗=42.2753 は，過分散なしの場合．表 10.17 に一致．
Scaled Pearson カイ 2 乗=25.0000 は，42.2753/1.6910
対数尤度=454.1742 は，調査中
完全対数尤度＝–68.2808 は，過分散なしの場合．表 10.17 に一致．
AICC (小さいほどよい)=162.0615 は，過分散なしの場合．表 10.17 に一致．

最大尤度パラメータ推定値の分析								
パラメータ		自由度	推定値	標準誤差	Wald 95% 信頼限界		Wald カイ 2 乗	Pr > ChiSq
Intercept		1	−6.4059	0.2828	−6.9601	−5.8517	513.24	<.0001
Type	B	1	−0.5433	0.2309	−0.9960	−0.0907	5.54	0.0186
Type	C	1	−0.6874	0.4279	−1.5260	0.1512	2.58	0.1082
Type	D	1	−0.0760	0.3779	−0.8166	0.6646	0.04	0.8407
Type	E	1	0.3256	0.3067	−0.2756	0.9268	1.13	0.2885
Type	A	0	0.0000	0.0000	0.0000	0.0000	.	.
Cons	65−69	1	0.6971	0.1946	0.3157	1.0785	12.83	0.0003
Cons	70−74	1	0.8184	0.2208	0.3857	1.2511	13.74	0.0002
Cons	75−79	1	0.4534	0.3032	−0.1409	1.0477	2.24	0.1348
Cons	60−64	0	0.0000	0.0000	0.0000	0.0000	.	.
Oper	75−79	1	0.3845	0.1538	0.0830	0.6859	6.25	0.0124
Oper	60−74	0	0.0000	0.0000	0.0000	0.0000	.	.
尺度		0	1.3004	0.0000	1.3004	1.3004		

Note:尺度パラメータは Pearson カイ 2 乗/DOF の平方根により推定されています。

注） 推定値は，過分散の（なし，あり）に関係なく一致，表 10.4 および表 10.17 に一致.

標準偏差は過分散ありで調整．表 10.4 に一致．尺度（形状）1.3004 で除せば過分散なしの場合となる．尺度（形状）は，過分散パラメータ 1.6910 平方根.

Type 3 分析の LR 統計量						
要因	分子の自由度	分母の自由度	F 値	Pr > F	カイ 2 乗	Pr > ChiSq
Type	4	25	3.50	0.0212	14.00	0.0073
Cons	3	25	6.19	0.0027	18.57	0.0003
Oper	1	25	6.30	0.0189	6.30	0.0120

注） カイ 2 乗値は，表 10.7 の JMP の過分散がある場合の結果と一致.
P 値も表 10.7 に一致.

PROC GENMOD から出力された予測値など

Obs	Type	Cons	Oper	Month	Inci	ln_Month	pred	xbeta	s_dev
1	A	60−64	60−74	127	0	4.8442	0.2098	−1.56171	−0.50059
2	A	60−64	75−79	63	0	4.1431	0.1528	−1.87830	−0.42682
3	A	65−69	60−74	1095	3	6.9985	3.6319	1.28975	−0.28332
4	A	65−69	75−79	1095	4	6.9985	5.3346	1.67422	−0.51881
:									
39	E	75−79	60−74	.	.	.	.	.	.
40	E	75−79	75−79	542	1	6.2953	2.8658	1.05284	−1.07973

Proc GENMOD の Output ステートメントで出力された xbeta は推定損傷数の対数，pred は xbeta の指数で推定損傷数，s_dev は，スチューデント化デビアンス残差である．このデ

ータを Excel に取り込み，損傷数千月比を計算することにより，第 10.5 節の表 10.15 および図 10.11 で示した交互作用の新しい検討のためのグラフの作成が可能となる．

スチューデント化デビアンス残差（s_dev）についての統計量を，Proc UNIVARIATE で計算した結果の一部を次に示す．これにより，過分散を設定した主効果モデルに対する残差の評価が行なえる．最大値が 2.31407，最小値が-1.76687 となり，外れ値はないものと確認される．

s_dev についての PROC UNIVARIATE から出力されたリスト

極値			
最小値		最大値	
値	Obs	値	Obs
-1.76687	22	1.01410	10
-1.45675	32	1.32054	35
-1.33261	38	1.65314	36
-1.32681	19	1.76877	30
-1.07973	40	2.31407	21

負の 2 項回帰

GENMOD プロシジャで負の 2 項回帰を行うためには，分布の設定で dist=negbin とするだけで実行できる．

```
Title2 '<<< 負の 2 項回帰 offset >>>' ;
proc genmod data=d01  ;
     class Type Cons Oper / ref=first ;
     model Inci = Type Cons Oper
          / dist=negbin link=log type3 offset=ln_Month ;
      run ;
```

残念ながら，以下のメッセージがあり，結果を示すことができない変数を絞ると収束するので，データ数に対してパラメータ数が多いためと考えられる．

WARNING: 相対的な Hessian 収束基準 5.8797664802 は限界値
0.0001 を超えています。%w 収束は疑わしいです。

11. デビアンス・逸脱度・テコ比・4 種の残差

一般化線形モデルによるポアソン回帰分析では，通常の回帰分析とは結果の表記法がかなり異なり，なかなか馴染めないのではないだろうか．デビアンス・逸脱度は，その代表的な例であろう．逸脱度を理解するためには対数尤度と最尤法の理解も必要となり最小 2 乗法に慣れ親しんだきた人たちは，茫然自失の状態になるかもしれない．さらに，スチューデント化デビアンス残差に関連して「テコ比」も登場する．「テコ比」は，通常の回帰分析にも登場することもあるが，マイナーな存在である．そこで，通常の回帰分析と対比しつつポアソン回帰で使われている統計用語とその意味づけについてこれまで取り上げてきた事例を用いて関連付けを行う．

11.1. デビアンス

第 1.9 節では，久保（2012）で取り上げられている種子数のデータについて表 1.34 に Excel で計算した 3 種の対数尤度（縮小モデル，完全モデル，飽和モデル）を示し，デビアンスを飽和モデルの対数尤度 $\ln L^{飽和}$ から完全モデルの対数尤度 $\ln L^{完全}$ の差の 2 倍，84.9930 として示した．このデビアンス 84.9930 が，飽和モデルの自由度 100 に対し，完全モデルの自由度 2 との差の自由度 98 のカイ 2 乗分布に従うことから，上側確率が 0.8226 となり，100 個のパラメータで推定した飽和モデルに対し，2 個のパラメータで推定した完全モデルが統計的に遜色ないことを示した．

単にデビアンスと言う場合は，データに対して何らかの仮定をした 2 つのモデルから計算される対数尤度の差の 2 倍した統計量を意味する．ややこしいのは，統計ソフトを使うと「デビアンス残差」あるいは「スチューデント化デビアンス残差」なども出てくる．第 1.9 節で JMP によるポアソン回帰で出力される「スチューデント化デビアンス残差プロット」を図 1.9 に示したが，その計算方法を示さずに「ほとんどが（−2〜＋2）の範囲に入っていることからあてはめの妥当性が示されている」との解説をしている．なぜ，そのような判断ができるのか，本章で Excel による計算を通して理解を深めたい．

第 9.5 節では，R による負の 2 項回帰（ガンマ・ポアソン回帰）による結果を文献から引用した際に「Deviance Residuals］と「Residual deviance」について Excel での計算例を示し，「デ

ビアンス・逸脱度」について断片的な解説をした．しかし，他では意図的に使ってこなかった．それに代えてマイナス 2 倍の対数尤度，対数尤度の差の 2 倍，尤度比カイ 2 乗値を用いてきた．第 1.9 節では，Pearson のカイ 2 乗値の計算方法を示したが，デビアンスについては深入りせず，Pearson のカイ 2 乗値を多用してきた．これは，デビアンス残差の計算で用いている「飽和モデルの対数尤度」の概念が通常の最小 2 乗法の世界で，対応する概念がないために，あえて言及しなかった．

本章では，通常の回帰分析で標準的に使われている「分散分析表」の概念と比較しながらデビアンスついて解説する．さらに，通常の回帰分析で行われている残差に加え，スチューデント化残差（標準化残差）とスチューデント化デビアンス残差の対比，さらに（Peason 残差，スチューデント化 Peason 残差，デビアンス残差）について，統計ソフトの出力結果と対比しつつ Excel を用いた計算法を示し，それらの使い分けについて概説する．

Excel での各種の残差を計算する際に，厄介な問題に直面する．それは，テコ比 h_{ii} がハット行列 $\boldsymbol{H}$

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \qquad (11.1)$$

の対角要素と定義されているため，Excel の行列関数で対角要素を取り出すことが容易ではない．そのために，デザイン行列 $\boldsymbol{X}$ の i 行目のベクトル $\boldsymbol{x}_i$ を使った

$$h_{ii} = \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i^T \qquad (11.2)$$

計算により代替する．この式を使って $h_{1,1}$， $h_{2,2}$, . . . を列ベクトルとして得ることができる．このような，Excel による逐次的な計算手順を経験することにより，テコ比 h_{ii} の意味付けと活用方法について理解を深めてもらいたい．

テコ比 h_{ii} を使うことにより，通常の回帰分析でのスチューデント化残差（標準化残差）を計算することができるようになり，ポアソン回帰では，スチューデント化デビアンス残差を自ら計算できるようになる．

ハット行列を $\boldsymbol{H}$ としたのであるが，対数尤度の 2 階の偏微分行列をヘッセ行列 $\boldsymbol{H}$ と同じ記号を用いているが，伝統的な表記法であり文脈で区別ができるのであえて区別しない．

11.2.　通常の回帰分析におけるスチューデント化残差

回帰パラメータの推定

第 4 章で示した Excel の行列関数を用いた回帰分析について要点をまとめる．用いるデータは，第 1.4 節のポアソン回帰のための人工データである［ドブソン（2008)］．表 11.1 に示すように，まず，説明変数 X の平均からの偏差を計算し，それらの平方から平方和 S_{XX} を計算する．次に，反応変数 Y の偏差と X の偏差の積和 S_{XY} を計算し，それらから，回帰パラメータ傾き $\hat{\beta}_1$ を

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{24.000}{4.8889} = 4.9091$$

により計算する．説明変数 X の平均 $\bar{X}$ と反応変数 Y の平均 $\bar{Y}$ と傾き $\hat{\beta}_1$ を用いて回帰パラメータの切片 $\hat{\beta}_0$ を，

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 8.0 - 4.9091 \times 0.1111 = 7.4545$$

として計算する．なお，表中に $S_T = S_{YY}$ があるが，回帰パラメータの推定には使われない．

表 11.1　回帰パラメータの Excel シート上での推定（表 4.1 再掲）

i	X	Y	X偏差	X偏差^2	Y偏差	Y偏差^2	XY偏差		
1	-1	2	-1.1111	1.2346	-6.0000	36.0000	6.6667	β_1^=	4.9091
2	-1	3	-1.1111	1.2346	-5.0000	25.0000	5.5556	β_0^=	7.4545
3	0	6	-0.1111	0.0123	-2.0000	4.0000	0.2222		
4	0	7	-0.1111	0.0123	-1.0000	1.0000	0.1111		
5	0	8	-0.1111	0.0123	0.0000	0.0000	0.0000		
6	0	9	-0.1111	0.0123	1.0000	1.0000	-0.1111		
7	1	10	0.8889	0.7901	2.0000	4.0000	1.7778		
8	1	12	0.8889	0.7901	4.0000	16.0000	3.5556		
9	1	15	0.8889	0.7901	7.0000	49.0000	6.2222		
	0.1111	8.0000	0.0000	4.8889	0.0000	136.0000	24.0000		
	平均	平均	合計	平方和	合計	平方和	平方和		
	$\bar{X}$	$\bar{Y}$		S_{XX}		S_T=S_{YY}	S_{XY}		

Excel での一般的な計算方法は，セルに対し計算式を与え，そのセルをプルダウンして参照セルを変化させつつ計算式をコピーする．これは，Excel の画期的な計算機能であるが，操作ミスがあっても発見しづらい．セルごとの計算に代え，範囲を使った計算を使うのが確実である．たとえば，8 行 1 列の「X 偏差」は，=(***X*** の範囲－X の平均）のように行列計算の要領で一括計算ができる．「X 偏差^2」は，=(***X*** 偏差の範囲)^2 によって 8 行 1 列分の計算を一括して行っている．「XY 偏差」は，＝(***X*** 偏差 * ***Y*** 偏差) として一括計算している．範囲を指定した場合の四則演算は，対応するセル同士の計算となり，片方がスカラーのような場合は，相手のサイズに合わせてくれる．

分散分析表

表 11.2 に示すように分散分析表に必要な平方和の計算を逐次的に行った結果を示す. なお, 各種の平方和の計算は, SumSq() 関数を使用するのが効率的である.

総平方和　$S_T = \sum_{i=1}^{9}(Y_i - \bar{Y})^2 = \mathrm{SumSq}(\boldsymbol{Y}\text{の範囲} - \bar{Y}) = 136.0000$　$df = 9-1 = 8$

推定値　$\hat{\boldsymbol{Y}} = \boldsymbol{X}\,\hat{\boldsymbol{\beta}} = \mathrm{Mmult}(\boldsymbol{X}\text{の範囲}, \hat{\boldsymbol{\beta}}\text{の範囲})$　$df = 2$

回帰の平方和　$S_R = \sum_{i=1}^{9}(\hat{Y}_i - \bar{Y})^2 = \mathrm{SumSq}(\hat{\boldsymbol{Y}}\text{の範囲} - \bar{Y}) = 117.8182$　$df = 2-1 = 1$

誤差平方和　$S_e = \sum_{i=1}^{9}(Y_i - \hat{Y}_i)^2 = \mathrm{SumSq}(\boldsymbol{Y}\text{の範囲} - \hat{\boldsymbol{Y}}\text{の範囲}) = 18.1818$　$df = 9-2 = 7$

なお, 回帰の平方和 S_R は,

$$S_R = S_T - S_e = 136.0000 - 18.1818 = 117.8182$$

としても計算できるが, 回帰の平方和の定義式による計算が, 回帰分析の本質を理解する上で望ましい.

表 11.2　回帰の平方和の計算（表 4.3 再掲）

i	$\boldsymbol{X}$		Y_i	$\bar{Y}$	$Y_i - \bar{Y}$	Y_i^	Y_i^ $-\bar{Y}$	$Y_i - Y_i$^		$\boldsymbol{\beta}$^
1	1	-1	2	8.00	-6.00	2.55	-5.45	-0.55	β_0^ =	7.4545
2	1	-1	3	8.00	-5.00	2.55	-5.45	0.45	β_1^ =	4.9091
3	1	0	6	8.00	-2.00	7.45	-0.55	-1.45		
4	1	0	7	8.00	-1.00	7.45	-0.55	-0.45		
5	1	0	8	8.00	0.00	7.45	-0.55	0.55		
6	1	0	9	8.00	1.00	7.45	-0.55	1.55		
7	1	1	10	8.00	2.00	12.36	4.36	-2.36		
8	1	1	12	8.00	4.00	12.36	4.36	-0.36		
9	1	1	15	8.00	7.00	12.36	4.36	2.64		
		5.00		8.0000	136.0000		117.8182	18.1818		
		ΣX^2		$\bar{Y}$	S_T		S_R	S_e		
自由度 df			9	1	9-1=8	2	2-1=1	9-2=7		2

計算結果を, 表 11.3 の分散分析表にまとめる. 分散分析表の自由度については, 各種の便宜的な説明が行なわれているが, 自由度の本質が把握しづらいので, ここに示したように偏差平方和をベースにした計算式に対応する自由度の計算法を示す.

表 11.3　回帰に対する分散分析表（表 4.4 再掲）

要因		平方和	自由度	平均平方	F 値	p 値
回帰	S_R	117.8182	2-1=1	117.8182	45.3600	0.0003
誤差	S_e	18.1818	9-2=7	2.5974		
全体	S_T	136.0000	9-1=8			

パラメータの共分散行列

分散分析表の誤差の平均平方（誤差分散 $\hat{\sigma}^2$）を用いて，回帰パラメータ $\hat{\beta}_1$ の分散 $Var(\hat{\beta}_1)$ は，正規方程式の解を用いて，式（4.23）および式（4.22）から

$$Var(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum X_i^2}{n\sum (X_i - \bar{X})^2} = \frac{2.5974 \times 5.00}{9 \times 4.8889} = 0.2952$$

$$Var(\beta_1) = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2} = \frac{2.5974}{4.8889} = 0.5313$$

が得られる．また，式（4.25）から，それぞれの分散は，

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2 = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix} = \begin{bmatrix} \frac{\Sigma X_i^2}{n\Sigma (X_i - \bar{X})^2} & \frac{-\bar{X}}{\Sigma (X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\Sigma (X_i - \bar{X})^2} & \frac{1}{\Sigma (X_i - \bar{X})^2} \end{bmatrix} \hat{\sigma}^2$$

パラメータの共分散行列の対角要素であることも示した．実際の計算は，

$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=$			
	0.1136	-0.0227	2.5974
	-0.0227	**0.2045**	σ^2
	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$		

=		
	0.2952	-0.0590
	-0.0590	**0.5313**
	$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma$^2	

となり，$Var(\hat{\beta}_0) = 0.2952$，$Var(\hat{\beta}_1) = 0.5313$ が得られる．それらの平方根から SE を計算しパラメータに関する t 検定が行なう．この行列計算の方法は，変数の数が増えても同じであり，偏差平方和をベースにした計算手順よりも簡潔である．詳しくは第 12.3 節で示す．

表 11.4　回帰パラメータの推定値（表 4.2 再掲）

項	推定値	分散	SE	t 値	p 値
β_0^	7.4545	0.2952	0.5433	13.72	0.0000
β_1^	4.9091	0.5313	0.7289	6.73	0.0003

スチューデント化残差

通常の回帰分析において，残差の検討の重要性は常に強調されている．代表的なのは，図 11.1 に示す予測値 $\hat{y}_i$ に対する残差 $\hat{\varepsilon}_i = y_i - \hat{y}_i$ のプロットおよびスチューデント化残差プロットである．この残差プロットは，予測値 $\hat{y}_i$ が大きくなるにつれ扇型に広がっている．したがって，推定された回帰直線に対して誤差が均一とはみなせないので，変数変換などで，誤差が均一になるような変換を検討することになる．

スチューデント化残差 $\hat{\varepsilon}_i'$ は，残差 $\hat{\varepsilon}_i$ の標準誤差で割って基準化したものであり，残差自体について統計的な考察ができる．図 11.1 右に示すように，扇型に広がってはいるが残差 $\hat{\varepsilon}_i'$ は，(–2 ~ +2) の範囲に入っており，通常の回帰分析を適用しても差し支えないと判断される．

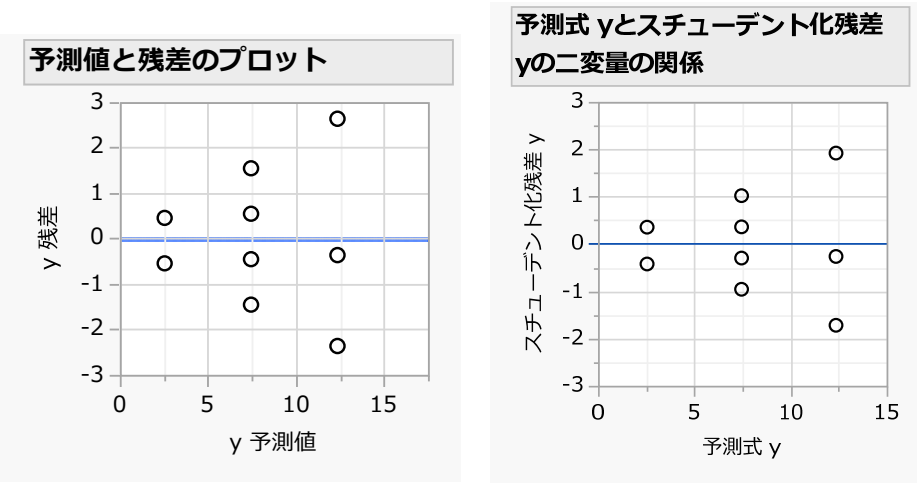


図 11.1　通常の回帰分析における残差およびスチューデント化残差プロット

テコ比・ハット行列

スチューデント化残差 $\hat{\varepsilon}_i'$ は，残差 $\hat{\varepsilon}_i$ の分散 $Var(\hat{\varepsilon}_i) = \hat{\sigma}^2(1-h_{ii})$ を考慮したものである．ここで h_{ii} は，通称テコ比ともいわれており，スチューデント化残差 $\hat{\varepsilon}_i'$ は，ハット行列の対角要素 h_{ii} を用いて基準化したもので

$$\left.\begin{aligned}\hat{\varepsilon}_i' &= \frac{\hat{\varepsilon}_i}{\sqrt{Var(\hat{\varepsilon}_i)}} \\ &= \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}\end{aligned}\right\} \qquad (11.3)$$

として計算されている．ハット行列 $\boldsymbol{H}$ は，$\boldsymbol{Y}$ の推定値を求める式 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ に $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ を代入し，

$$\left.\begin{aligned}\hat{\boldsymbol{Y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\end{aligned}\right\} \qquad (11.4)$$

最後の $\boldsymbol{Y}$ を除いた行列

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \qquad (11.5)$$

で定義されている．実際に計算すると次のような結果となる．まず，$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ の計算結果を示す．次に，$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ を挟んで $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ の行列計算を行う．このようにハット行列 $\boldsymbol{H}$ の計算は，ストレスなく行えるのであるが，その対角要素を列ベクトル化する関数が Excel にないために，この式を使うことに躊躇せざるをえない．

$\boldsymbol{X}^T$								
1	1	1	1	1	1	1	1	1
-1	-1	0	0	0	0	1	1	1

=Transpose($\boldsymbol{X}$の範囲)

(2×9)

$\boldsymbol{X}$	
1	-1
1	-1
1	0
1	0
1	0
1	0
1	1
1	1
1	1

(9×2)

=

$\boldsymbol{X}^T\boldsymbol{X}$	
9	1
1	**5**

(2×2)

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	
0.1136	-0.0227
-0.0227	**0.2045**

=Minverse($\boldsymbol{X}^T\boldsymbol{X}$の範囲)

=Mmult($\boldsymbol{X}^T$の範囲, $\boldsymbol{X}$の範囲)

(2×2)

$\boldsymbol{X}$	
1	-1
1	-1
1	0
1	0
1	0
1	0
1	1
1	1
1	1

(9×2)

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	
0.1136	-0.0227
-0.0227	**0.2045**

(2×2)

$\boldsymbol{X}^T$								
1	1	1	1	1	1	1	1	1
-1	-1	0	0	0	0	1	1	1

(2×9)

=

$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$								
0.36	0.36	0.14	0.14	0.14	0.14	-0.1	-0.1	-0.1
0.36	**0.36**	0.14	0.14	0.14	0.14	-0.1	-0.1	-0.1
0.14	0.14	**0.11**	0.11	0.11	0.11	0.09	0.09	0.09
0.14	0.14	0.11	**0.11**	0.11	0.11	0.09	0.09	0.09
0.14	0.14	0.11	0.11	**0.11**	0.11	0.09	0.09	0.09
0.14	0.14	0.11	0.11	0.11	**0.11**	0.09	0.09	0.09
-0.1	-0.1	0.09	0.09	0.09	0.09	**0.27**	0.27	0.27
-0.1	-0.1	0.09	0.09	0.09	0.09	0.27	**0.27**	0.27
-0.1	-0.1	0.09	0.09	0.09	0.09	0.27	0.27	**0.27**

(9×9)

=Mmult(Mmult($\boldsymbol{X}$の範囲, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$の範囲),$\boldsymbol{X}^T$の範囲)

ハット行列 $\boldsymbol{H}$ 対角要素を h_{ii} としたとき，デザイン行列 $\boldsymbol{X}$ の i 行目のベクトルを $\boldsymbol{x}_i$ を用いることにより，テコ比 h_{ii} は，

$$h_{ii} = \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i^T \qquad (11.6)$$

として求めることができる．この式で，それぞれの h_{ii} を計算することにより，ハット行列 $\boldsymbol{H}$ を計算しなくとも次式で示すように，対角要素をベクトル化してテコ比 h_{ii} を計算することができる．

x_1		$(X^TX)^{-1}$		x_1^T		h_{11}
1	-1	**0.1136**	-0.0227	1	=	**0.3636**
		-0.0227	**0.2045**	-1		
:						
x_8		$(X^TX)^{-1}$		x_8^T		h_{88}
1	1	**0.1136**	-0.0227	1	=	**0.2727**
		-0.0227	**0.2045**	1		

テコ比の活用

単純な誤差のプロットに加えて，テコ比を考慮したスチューデント化残差（標準化残差）による検討も有益である．なお，テコ比とハット行列の意味付けについては，野沢（1992），「テコ比とハット行列」が詳しい．

テコ比 h_{ii} は，回帰の 95%信頼区間を求めるための分散 $Var(\hat{y}_i)$ の計算にも関係している．式（4.36）から，

$$\left.\begin{aligned} Var(\hat{y}_i) &= \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T \\ &= \boldsymbol{x}_i [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2] \boldsymbol{x}_i^T \\ &= h_{ii}\hat{\sigma}^2 \end{aligned}\right\} \qquad (11.7)$$

が導かれる．個々のデータの分散 $\hat{\sigma}^2$ に対しテコ比 h_{ii} は，回帰の推定値の分散 $Var(\hat{y}_i)$ を求めるための割引係数として解される．表 11.5 に Excel でテコ比 h_{ii} を計算し，スチューデント化残差 $\hat{\varepsilon}_i'$ および分散 $Var(\hat{y}_i)$ を計算した結果を示す．

表 11.5　テコ比を用いたスチューデント化残差

					残差	テコ比	ス.化残差	残差 比	分散
i	X		Y	Y^	$\varepsilon = Y - Y$^	h_{ii}	ε'	ε'/ε	$Var(y$^$)$
1	1	-1	2	2.5455	-0.5455	**0.3636**	**-0.4243**	0.7778	**0.9445**
2	1	-1	3	2.5455	0.4545	0.3636	0.3536	0.7778	0.9445
3	1	0	6	7.4545	-1.4545	0.1136	-0.9586	0.6591	0.2952
4	1	0	7	7.4545	-0.4545	0.1136	-0.2996	0.6591	0.2952
5	1	0	8	7.4545	0.5455	0.1136	0.3595	0.6591	0.2952
6	1	0	9	7.4545	1.5455	0.1136	1.0185	0.6591	0.2952
7	1	1	10	12.3636	-2.3636	0.2727	-1.7197	0.7276	0.7084
8	1	1	12	12.3636	-0.3636	0.2727	-0.2646	0.7276	0.7084
9	1	1	15	12.3636	2.6364	0.2727	1.9182	0.7276	0.7084
	9.00	1.00	0.1136	-0.0227	72.0000	β_0=	7.4545	$\varepsilon^T\varepsilon$=	18.1818
	1.00	5.00	-0.0227	0.2045	32.0000	β_1=	4.9091	σ^^{2}=	**2.5974**
	X^TX		$(X^TX)^{-1}$		X^TY	$(X^TX)^{-1}X^TY$			

実際の計算過程を$i=1$の場合について示す．まずテコ比h_{11}を計算し，

$h_{11}=$

1	-1	**0.1136**	-0.0227	1	=	**0.3636**
		-0.0227	**0.2045**	-1		
$\boldsymbol{x}_1$		$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$		$\boldsymbol{x}_1^T$		h_{11}

スチューデント化残差$\hat{\varepsilon}_1'$を求める．

$$\hat{\varepsilon}_1' = \frac{\hat{\varepsilon}_1}{\sqrt{\hat{\sigma}^2(1-h_{11})}}$$
$$= \frac{-0.5455}{\sqrt{2.5974\times(1-0.3636)}} = -0.4243$$

テコ比h_{11}を用いて回帰の推定値$\hat{y}_1$の分散$Var(\hat{y}_1)$の計算もできる．

$$Var(\hat{y}_1) = h_{11}\hat{\sigma}^2$$
$$= 0.3636\times 2.5974 = 0.9445$$

なお，図 11.1 右に示したスチューデント化残差プロットは，JMP のスチューデント化残差プロットに X 軸に予測値を指定できないので，スチューデント化残差をファイルに出力して，別途「二変量の関係」を使って作図したものである．

テコ比は，回帰直線の推定値の分散を誤差分散$\hat{\sigma}^2$に対する割引係数としても理解される．回帰直線の中心部は小さく，外側に向かって大きくなり，推定値の分散が大きくなり 95%信頼区間の幅の変化を示す統計量とも解される．

Excel の「標準残差」に対する使用上の注意

Excel の回帰分析の活用は，分散分析表および回帰パラメータの推定値に関してデザイン行列の計算の煩わしさを軽減するために有益であることを示してきた．さらに，Excel の回帰分析で「残差」に加えて「標準残差」を Excel シートに出力することができる．「標準残差」はスチューデント化残差（標準化残差）と紛らわしいが，テコ比を含まない計算であり，別物である．表 11.5 の場合では，全ての残差$\hat{\varepsilon}_i$に 0.6633 を掛けている．これは，分散$\hat{\sigma}^2$の平方根の逆数 0.6205 に近い値であるが，どのような計算なのかは不明である．いずれにしても，スチューデント化残差（標準化残差）ではないことに注意が必要である．

Excel の統計解析については，正確性が欠けるとの指摘があることは十分に承知しており，「標準残差」もその一例であろう．もっとひどい事例は，第 10.3 節で例示した，折れ線グラフの誤差範囲の設定のいい加減さであり，それを回避する方法を知る必要がある．なお，行列計算などの精度で問題になった経験はないが，計算過程の脆弱性を常に認識し，他のソフトでの検証を怠ってはならない．

11.3.　ポアソン回帰におけるデビアンス・逸脱度

対数尤度を用いた恒等リンクのポアソン回帰は，回帰式を

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim \text{ポアソン分布}$$

としたときに，対数尤度 $\ln L^{\text{回帰(誤差)}}$ を最大にするような回帰パラメータ $\hat{\beta}_0$ と $\hat{\beta}_1$ を次式により，

$$\ln L^{\text{回帰(誤差)}} = \sum_i \ln[\text{Poisson.dist}(y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i, false)]$$

を推定する．$\ln L^{\text{回帰(誤差)}}$ を最大化するために，第 1.4 節および第 5 章で反復重み付き回帰を用いる方法，第 2 章で対数尤度関数の 2 階の偏微分式行列を用いる方法，第 1.9 節で Excel のソルバーにより $\ln L^{\text{回帰(誤差)}}$ が最大になるような $\hat{\beta}_0$ と $\hat{\beta}_1$ を直接求める方法を示してきた．ここでは，簡便な Excel のソルバーを用いる方法を使う．

デビアンス・カイ 2 乗

ポアソン回帰では，第 1.9 節で示したように飽和モデル，完全（最大）モデル，縮小モデル，切片のみの場合には（null モデル）など幾つかの「モデル」が登場する．そして，逸脱度/デビアンスは，飽和モデルと各モデルとのマイナス 2 倍の対数尤度の差で定義されている．

「飽和モデル」の概念が通常の回帰分析にはないので，理解に苦しむことになる．表 11.6 に示すように，飽和モデルの対数尤度 $\ln L^{\text{飽和}}$ は，y_i のポアソン分布の確率を求めるための推定値として $\hat{y}_i = y_i$ のように自分自身 y_i を用いて，

$$\text{飽和モデル：}\ \hat{y}_i = y_i, \quad \begin{cases} \ln L^{\text{飽和}} = \sum_i \ln(\text{Poisson.dist}(y_i, y_i, false)) \\ \qquad = \ln(0.2707) + \ln(0.2240) + \cdots + \ln(0.1024) \\ \qquad = -1.3069 - 1.4959 - \cdots - 2.2785 \\ \qquad = -17.0566 \end{cases}$$

としてポアソン分布の確率を計算している．通常の回帰分析は，偏差平方和の計算で組み立てられているので，無理に飽和モデルを考えても，次のようにゼロなので

$$S^{\text{飽和}} = \sum_i (y_i - y_i)^2 = 0$$

何の役にも立たない．したがって，飽和モデルの概念がない．（完全 or 最大 or 回帰）モデルは，$\hat{\beta}_0$ と $\hat{\beta}_1$ を使った回帰モデルで，$\ln L^{\text{回帰(誤差)}}$ に変えて $\ln L^{\text{完全}}$ とし，

$$\text{完全モデル，}\ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i\ \text{：} \quad \begin{cases} \ln L^{\text{完全}} = \sum_i \ln(\text{Poisson.dist}(y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i, false)) \\ \qquad = \ln(0.2557) + \ln(0.2144) + \cdots + \ln(0.0791) \\ \qquad = -1.3639 - 1.5397 - \cdots - 2.5366 \\ \qquad = -18.0039 \end{cases}$$

となる．これは，表 11.7 に示すように通常の回帰分析の「誤差平方和 S_e」に該当する．

表 11.6　最尤法によるポアソン回帰

			縮小（null）モデル			完全（最大）モデル			飽和モデル		
			$y\hat{}=\beta_0\hat{}$			$y\hat{}=\beta_0\hat{}+\beta_1\hat{}x$			$y\hat{}=y$		
				$\beta_0\hat{}=$	**8.0000**		$\beta_0\hat{}=$	**7.4516**			
							$\beta_1\hat{}=$	**4.9353**			
				$\ln L^{縮小}=$	**-26.2669**		$\ln L^{完全}=$	**-18.0039**		$\ln L^{飽和}=$	**-17.0566**
			$\beta_0\hat{}$	確率	対数尤度	回帰	確率	対数尤度	y	確率	対数尤度
i	x	y	$y\hat{}$	P	$\ln L_i$	$y\hat{}$	P	$\ln L_i$	$y\hat{}$	P	$\ln L_i$
1	-1	2	**8.00**	0.0107	-4.5343	**2.52**	0.2557	-1.3639	**2**	0.2707	-1.3069
2	-1	3	**8.00**	0.0286	-3.5534	**2.52**	0.2144	-1.5397	**3**	0.2240	-1.4959
3	0	6	**8.00**	0.1221	-2.1026	**7.45**	0.1380	-1.9803	**6**	0.1606	-1.8287
4	0	7	**8.00**	0.1396	-1.9691	**7.45**	0.1469	-1.9178	**7**	0.1490	-1.9038
5	0	8	**8.00**	0.1396	-1.9691	**7.45**	0.1369	-1.9888	**8**	0.1396	-1.9691
6	0	9	**8.00**	0.1241	-2.0869	**7.45**	0.1133	-2.1776	**9**	0.1318	-2.0268
7	1	10	**8.00**	0.0993	-2.3100	**12.39**	0.0978	-2.3249	**10**	0.1251	-2.0786
8	1	12	**8.00**	0.0481	-3.0339	**12.39**	0.1137	-2.1744	**12**	0.1144	-2.1683
9	1	15	**8.00**	0.0090	-4.7076	**12.39**	0.0791	-2.5366	**15**	0.1024	-2.2785

（縮小 or 切片 or Null）モデルは，

$$\text{縮小モデル，}\ \hat{y}_i=\hat{\beta}_0\ :\ \begin{cases}\ln L^{縮小}=\sum_i \ln(\text{Poisson.dist}(y_i,\hat{\beta}_0,false))\\ \quad=\ln(0.0107)+\ln(0.0286)+\cdots+\ln(0.0090)\\ \quad=-4.5343-3.5534-\cdots-4.7076\\ \quad=-26.2669\end{cases}$$

として計算される．通常の回帰分析の「総平方和 S_T」に該当する．

回帰の平方和 S_R は，$S_R=S_T-S_e$ で求められると同様に，傾き $\hat{\beta}_1$ に対する 2 倍の対数尤度は，差分

$$\text{差分：}\ \begin{cases}2\ln L^{R:回帰}=2(\ln L^{完全}-\ln L^{縮小})\\ \quad=2\times[-18.0039-(-26.2669)]\\ \quad=16.5260\end{cases}$$

として求められる．この 2 倍の対数尤度に対する検定統計量は，それぞれの自由度の差のカイ 2 乗分布に従うことにより有意差検定が行なえる．完全モデル（$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$）と縮小モデル（$\hat{y}_i=\hat{\beta}_0$）の対数用尤度の差 2 倍なので，$\hat{\beta}_1$ に対する尤度比検定統計量となり，カイ 2 乗検定が行なえる．

通常の回帰分析では，S_R をその自由度 df_R で割った平均平方を，S_e をその自由度 df_e で割った平均平方（誤差分散）との分散比

$$F=\frac{S_R/df_R}{S_e/df_e}$$

が，分母の自由度 df_R，分子の自由度 df_e の F 分布に従うことから有意差検定を行なっている．

JMP によるポアソン回帰の結果を表 11.7 に示す．通常の回帰分析の「分散分析表」と対比して，（差分，完全，縮小）の意味を理解してもらいたい．その後に続く「適合度統計量」は通常の「分散分析表」にはない概念である．しいて言えば，仮定した正規分布からのからの乖離度の統計量であろうか.「適合度統計量」の欄の「デビアンス」の行のカイ 2 乗値＝1.8947 となっているのは，飽和モデルの対数尤度 $\ln L_{飽和}=-17.0566$ と $\ln L_{完全}=-18.0039$ との差の 2 倍で

$$\begin{aligned}\text{デビアンス・カイ2乗} &= 2(\ln L^{飽和}-\ln L^{完全})\\ &= 2\times[-17.0566-(-18.0039)]\\ &= 1.8947\end{aligned}$$

と，デビアンスが計算されている．自由度は，$df^{飽和}=9$ と $df^{完全}=2$ の差から 7 となっている．このデビアンスが，第 9.5 節で示した R 言語の一般化線形モデルで出力されている Residual deviance に対応する．

表 11.7　通常の回帰とポアソン回帰の対比

通常の回帰分析		
要因		自由度
回帰	S_R	2-1=1
誤差	S_e	9-2=7
全体	S_T	9-1=8

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値
差分	8.2630	16.5260	1	<.0001*
完全	18.0039			
縮小	26.2669			

適合度統計量	カイ2乗	自由度	p値
Pearson	1.8944	7	0.9655
デビアンス	1.8947	7	0.9654

AICc
42.0078

Pearson･カイ 2 乗

Pearson の適合度統計量は，反応変数 y_i と推定値 $\hat{y}_i$ の差の 2 乗をその分散 $Var(y_i)=\hat{y}_i$ で割って加えたもので

$$\begin{aligned}\text{Pearson･カイ2乗} &= \sum_i \frac{(y_i-\hat{y}_i)^2}{Var(\hat{y}_i)} = \sum_i \frac{(y_i-\hat{y}_i)^2}{\hat{y}_i}\\ &= \frac{(2-2.52)^2}{2.52}+\frac{(3-2.52)^2}{2.52}+\cdots+\frac{(15-12.39)^2}{12.39}\\ &= 1.8944\end{aligned}$$

として計算されている．これらのデビアンスおよび Peason のカイ 2 乗値を自由度で除したのが過分散の調整パラメータである．この例では，明らかにカイ 2 乗値が自由度 7 より小さいので，過分散が起きていないと判断される．

通常の回帰分析の分散分析表の「要因」の欄の（回帰，誤差，全体）との表記は，ポアソン回帰の「モデル」の表記（差分，完全，縮小）と全く異なるので，それらの対応関係を関連付けることによりによって理解を深めてもらいたい．ただし，「回帰」が「差分」に対応し，「誤差」が「完全」に，「全体」が「縮小」にそれぞれ対応している．このような用語の対応付けを行なったとしても，同義語としては全く認識されないので，それらの対数尤度の定義を理解した上での説明を加えながら注意深く使うことが，理解を深めるための必要と思われる．

AICc

AICc は，修正済み赤池の情報量量基準で，k をパラメータ数 2 とし

$$\left.\begin{aligned}\mathrm{AICc} &= -2\ln L_{\text{完全}} + 2k + \frac{2k(k+1)}{n-k-1} \\ &= -2\times(-18.0039) + 2\times 2 + \frac{2\times 2\times(2+1)}{9-2-1} \\ &= 42.0078\end{aligned}\right\} \qquad (11.8)$$

と計算されている．AICc は，パラメータ数が異なるモデルを比較する際に役に立つ．

表 11.8 のパラメータの関する尤度比検定の「項」の欄の「x」に対する尤度比カイ 2 乗は，16.5260 であり，モデル全体の検定の差分の尤度比カイ 2 乗 16.5260 に一致する．なお，標準誤差は，第 4.2 節で詳しく述べたように対数尤度関数の 2 階の偏微分行列 $\boldsymbol{H}$ の負の逆行列 $(-\boldsymbol{H})^{-1}$ がパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ なので，その対角要素の平方根である．

表 11.8　パラメータに関する尤度比検定

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	7.4516	0.8842	71.0299	<.0001*
x	4.9353	1.0915	16.5260	<.0001*

推定値の共分散

共分散

	切片	x
切片	0.7817	0.4160
x	0.4160	1.1915

「推定値の共分散」の対角要素（0.7817，1.1915）を用いて

$$SE(\text{切片}) = \sqrt{Var(\text{切片})} = \sqrt{0.7817} = 0.8842$$

$$SE(\text{x}) = \sqrt{Var(\text{x})} = \sqrt{1.1915} = 1.0915$$

それぞれの標準誤差が計算される．

11.4. ポアソン回帰における 4 種の残差

ポアソン回帰の場合は，恒等リンクの場合でも対数リンクの場合でも予測値の大きさに比例して分散が大きくなるので，単純な残差プロットによる残差の検討は，不適切である．そのために，ピアソン残差，スチューデント化ピアソン残差，デビアンス残差，あるいは，スチューデント化デビアンス残差などが JMP の一般化線形モデル，SAS の GENMOD プロシジャで提供されている．R では，デビアンス残差が使われている．

デビアンス残差

表 11.7 に示した JMP の「モデル全体の検定」での「適合度統計量」としてのデビアンスは，表 11.6 に示した飽和モデルの $\ln L^{飽和} = -17.0566$ と完全モデル $\ln L^{完全} = -18.0039$ との差の 2 倍で

$$\begin{aligned} デビアンス &= 2(\ln L^{飽和} - \ln L^{完全}) \\ &= 2\times[-17.0566-(-18.0039)], \quad df = 7 \\ &= 1.8947 \end{aligned}$$

であることを示した．これが有意であれば，ポアソン回帰で取り上げた説明変数では説明しきれない誤差変動が残っていることを意味している．他に追加できる変数がなければ，誤差分布がポアソン分布に対し過分散となっていることを意味する．過分散となっている場合は，パラメータの推定値の標準誤差が小さくなり，結果を過大評価することになる．そのために，過分散パラメータで標準誤差を大きくする修正，あるいは，負の 2 項分布（ガンマ・ポアソン分布）などを仮定した解析が必要となる．

デビアンスは，飽和モデルと完全モデルの対数尤度の差の 2 倍であり，それぞれの対数尤度は，個々のデータの対数尤度 $\ln L_i$ の和であることを表 11.6 で示した．デビアンス残差（Deviance Residuals）は，飽和モデルと完全モデルの個々の対数尤度の差の 2 倍の平方根として定義されている．ただし，符号が全てプラスとなり残差とは言えないので，完全モデルの推定値 $\hat{y}_{完全,i}$ と飽和モデルの推定値 $\hat{y}_{飽和,i} = y_i$ の差の符号を付けてデビアンス残差 $\varepsilon_i^{(D)}$ とする．

$$d_i = 2[\ln(L_i^{飽和}) - \ln(L_i^{完全})] \tag{11.9}$$

$$\varepsilon_i^{(D)} = \mathrm{Sign}(\hat{y}_i^{飽和} - \hat{y}_i^{完全})\sqrt{d_i} \tag{11.10}$$

なお，個々のデビアンス d_i は，$\hat{y}_i^{飽和} = y_i$ とし，簡略化した計算公式（11.3）

$$\left.\begin{aligned} d_i &= 2\left[\ln\left(\frac{(\hat{y}_i^{飽和})^{y_i}e^{-\hat{y}_i^{飽和}}}{y_i!}\right) - \ln\left(\frac{(\hat{y}_i^{完全})^{y_i}e^{-\hat{y}_i^{完全}}}{y_i!}\right)\right] \\ &= 2\left[y_i\ln(y_i) - y_i - \ln(y_i!) - y_i\ln(\hat{y}_i^{完全}) + \hat{y}_i^{完全} + \ln(y_i!)\right] \\ &= 2\left[y_i\ln\left(\frac{y_i}{\hat{y}_i^{完全}}\right) - (y_i - \hat{y}_i^{完全})\right] \end{aligned}\right\} \tag{11.11}$$

が，一般的に計算公式として広く用いられている．ただし，この計算式からでは何を意味しているのか，推測しがたい．したがって，元の対数尤度での定義式による理解を勧める．

実際に$i=1$の場合の計算を次に示し，全てのiについての結果を表 11.9 に示す．

$$d_1 = 2[(-1.3069)-(-1.3639)]$$
$$= 2\times 0.0570 = 0.1140$$
$$\mathrm{Sign}(\hat{y}_1^{\text{飽和}} - \hat{y}_1^{\text{完全}}) = \mathrm{Sign}(2.0-2.5163) = \text{マイナス}(-)$$
$$\varepsilon_1^{(D)} = \mathrm{Sign}(\hat{y}_1^{\text{飽和}} - \hat{y}_1^{\text{完全}})\sqrt{d_i}$$
$$= -\sqrt{0.1140} = -0.3377$$

表 11.9　デビアンス残差

				完全(最大)モデル			飽和モデル			デビアンス残差		
					$\hat{\beta}_0=$	**7.4516**						
					$\hat{\beta}_1=$	**4.9353**					平方和 =	**1.8947**
					$\ln L^{\text{完全}}=$	**-18.0039**		$\ln L^{\text{飽和}}=$	**-17.0566**	尤度の差	平方根	残差
i	X		y	$y^{\wedge\text{完全}}$	$P^{\text{完全}}$	$\ln L_i^{\text{完全}}$	$y^{\wedge\text{飽和}}$	$P^{\text{飽和}}$	$\ln L_i^{\text{飽和}}$	d	$\sqrt{(d)}$	$\varepsilon^{(D)}$
1	1	-1	2	2.5163	0.2557	**-1.3639**	2	0.2707	**-1.3069**	**0.1140**	**0.3377**	**-0.3377**
2	1	-1	3	2.5163	0.2144	-1.5397	3	0.2240	-1.4959	0.0875	0.2958	0.2958
3	1	0	6	7.4516	0.1380	-1.9803	6	0.1606	-1.8287	0.3032	0.5506	-0.5506
4	1	0	7	7.4516	0.1469	-1.9178	7	0.1490	-1.9038	0.0279	0.1672	-0.1672
5	1	0	8	7.4516	0.1369	-1.9888	8	0.1396	-1.9691	0.0394	0.1985	0.1985
6	1	0	9	7.4516	0.1133	-2.1776	9	0.1318	-2.0268	0.3015	0.5491	0.5491
7	1	1	10	12.3869	0.0978	-2.3249	10	0.1251	-2.0786	0.4927	0.7019	-0.7019
8	1	1	12	12.3869	0.1137	-2.1744	12	0.1144	-2.1683	0.0122	0.1105	-0.1105
9	1	1	15	12.3869	0.0791	-2.5366	15	0.1024	-2.2785	0.5161	0.7184	0.7184
											平方和＝	**1.8947**

デビアンス残差$\varepsilon_i^{(D)}$の平方和は，

$$\text{デビアンス}=\sum_{i=1}^{9}\left[\varepsilon_i^{(D)}\right]^2$$
$$= (-0.3377)^2 + 0.2958^2 + \cdots + 0.7184^2 = 1.8947$$

となり，デビアンス・カイ 2 乗値に

$$\text{デビアンス・カイ2乗} = 2(\ln L^{\text{飽和}} - \ln L^{\text{完全}})$$
$$= 2\times[-17.0566-(-18.0039)] = 1.8947$$

一致することが確認される．もちろん，計算公式によっても，次のように

$$d_1 = 2\left[y_1 \ln\left(\frac{y_1}{\hat{y}_1^{\text{完全}}}\right) - (y_1 - \hat{y}_1^{\text{完全}})\right]$$
$$= 2\times\left[2\times\ln\left(\frac{2}{2.5163}\right) - (2-2.5163)\right] = 0.1140$$

求められ，一致することが確認できる．

スチューデント化デビアンス残差

スチューデント化デビアンス残差は，反復重み付き回帰での重みを加味したハット行列 $\boldsymbol{H}'$ の対角要素テコ比 h_{ii}' を用いる．恒等リンクの場合の重みは，推定値 $\hat{y}_i$ の逆数であるので，対角要素にそれぞれの重を持つ行列を $\hat{\boldsymbol{W}}$ とする．推定値 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ のパラメータ推定値 $\hat{\boldsymbol{\beta}}$ を重み付き回帰の計算式 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Y}$ を代入すると，

$$\begin{aligned}\hat{\boldsymbol{Y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{Y}\end{aligned} \tag{11.12}$$

となる．ハット行列 $\boldsymbol{H}'$ は，回帰の推定値を求める際に $\hat{\boldsymbol{Y}}$ の式から最後の $\boldsymbol{Y}$ を除いた行列

$$\boldsymbol{H}' = \boldsymbol{X}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}} \tag{11.13}$$

で定義されている．重み行列 $\hat{\boldsymbol{W}}$ の対角要素からなるベクトルを $\hat{\boldsymbol{w}}$ とし，デザイン行列 $\boldsymbol{X}$ の i 行目のベクトルを $\boldsymbol{x}_i$ とすれば，テコ比 h_{ii}' は，

$$h_{ii}' = \boldsymbol{x}_i[(\boldsymbol{X} * \hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1}\boldsymbol{x}_i^T\hat{w}_i \tag{11.14}$$

として求めることができる．スチューデント化デビアンス残差 $\varepsilon_i'^{(D)}$ は，

$$\hat{\varepsilon}_i'^{(D)} = \frac{\hat{\varepsilon}_i^{(D)}}{\sqrt{1-h_{ii}'}} \tag{11.15}$$

として求められる．表 11.10 にスチューデント化デビアンス残差の計算結果を示す．デビアンス残差は，表 11.9 に示した結果を用いる．重み $\hat{w}_i$ を推定値 $\hat{y}_i$ の逆数とし，テコ比を求め，スチューデント化デビアンス残差 $\hat{\varepsilon}_i'^{(D)}$ が計算されている．

表 11.10　スチューデント化デビアンス残差

				完全モデル		飽和モデル						
				$\hat{\beta}_0$ =	**7.4516**			デビアンス残差		スチューデント化		
				$\hat{\beta}_1$ =	**4.9353**			平方和 =	**1.8947**	デビアンス残差		
				$\ln L^{完全}$ =	**-18.0039**	$\ln L^{飽和}$ =	**-17.0566**	尤度の差	残差	重みw^	テコ比	残差
i	$\boldsymbol{X}$		y	y^^{完全}	$\ln L_i{}^{完全}$	y^^{飽和}	$\ln L_i{}^{飽和}$	d	$\varepsilon^{(D)}$	$1/y$^^{完全}	h'_{ii}	$\varepsilon'^{(D)}$
1	1	-1	2	**2.5163**	-1.3639	2.0	-1.3069	0.1140	**-0.3377**	**0.3974**	**0.4510**	**-0.4558**
2	1	-1	3	2.5163	-1.5397	3.0	-1.4959	0.0875	0.2958	0.3974	0.4510	0.3993
3	1	0	6	7.4516	-1.9803	6.0	-1.8287	0.3032	-0.5506	0.1342	0.1049	-0.5820
4	1	0	7	7.4516	-1.9178	7.0	-1.9038	0.0279	-0.1672	0.1342	0.1049	-0.1767
5	1	0	8	7.4516	-1.9888	8.0	-1.9691	0.0394	0.1985	0.1342	0.1049	0.2098
6	1	0	9	7.4516	-2.1776	9.0	-2.0268	0.3015	0.5491	0.1342	0.1049	0.5804
7	1	1	10	12.3869	-2.3249	10.0	-2.0786	0.4927	-0.7019	0.0807	0.2261	-0.7979
8	1	1	12	12.3869	-2.1744	12.0	-2.1683	0.0122	-0.1105	0.0807	0.2261	-0.1256
9	1	1	15	12.3869	-2.5366	15.0	-2.2785	0.5161	0.7184	0.0807	0.2261	0.8167
										0.7817	0.4166	
										0.4166	**1.1863**	
										$[(\boldsymbol{X} * \boldsymbol{w})^T\boldsymbol{X}]^{-1}$		

実際の計算を $i=1$ の場合について示す．スチューデント化デビアンス残差は，

$$\hat{w}_1 = 1/\hat{y}_1^{\text{完全}}$$
$$= 1/2.5163 = 0.3974$$

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = [(\boldsymbol{X} * \hat{\boldsymbol{w}})^T \boldsymbol{X}]^{-1}$$
$$= \text{Minverse(Mmult(Transpose(}\boldsymbol{X}\text{の範囲} * \hat{\boldsymbol{w}}\text{の範囲)}, \boldsymbol{X}\text{の範囲)}$$

$h'_{11} =$

1	-1	**0.7817**	0.4166	1	×	0.3974	=	0.4510
		0.4166	**1.1863**	-1				
$\boldsymbol{x}_1$		$[(\boldsymbol{X} * \boldsymbol{w})^T \boldsymbol{X}]^{-1}$		$\boldsymbol{x}_1{}^T$		w_1		$h_{ii}{}'$

$$\hat{\varepsilon}_i'^{(D)} = \frac{\hat{\varepsilon}_i^{(D)}}{\sqrt{1-h'_{ii}}}$$
$$= \frac{-0.3377}{\sqrt{1-0.4510}} = -0.4558$$

で求められる．

デビアンス残差およびスチューデント化デビアンス残差について，JMP のポアソン回帰で作成した残差プロットを図 11.2 に示す．デビアンス残差に対してスチューデント化デビアンス残差の方が，残差の絶対値が大きめになっていることが確認できる．ただし，実際の解析で複数の残差を用いることは非現実的であり，第 11.5 節を参考にして，選択してほしい．

自ら計算する場合には，手軽に計算できるデビアンス残差であるが，統計的には，スチューデント化デビアンス残差が望ましいと思われる．JMP のディフォルトの残差プロットは，スチューデント化デビアンス残差が使用されているが，SAS の GENMOD プロシジャでは，ユーザの選択に任されている．

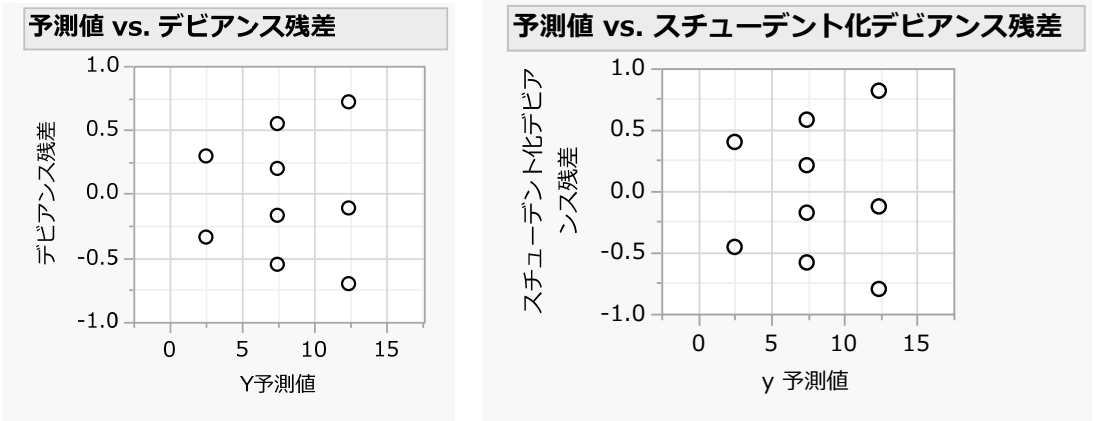


図 11.2　デビアンス残差とスチューデント化デビアンス残差の比較

スチューデント化 Pearson 残差

第 11.2 節で通常の回帰分析の場合に残差 $\hat{\varepsilon}_i = y_i - \hat{y}_i$ に対してテコ比 h_{ii} を用いてスチューデント化残差 $\hat{\varepsilon}_i'$ を求めた．ポアソン回帰の場合，単純な残差 $\hat{\varepsilon}_i$ は，推定値 $\hat{y}_i$ に比例して大きくなるので残差の検討に使えない，そこで，$\hat{y}_i$ の標準誤差で基準化した Pearson 残差 $\hat{\varepsilon}_i^{(P)}$ が使われている．Pearson 残差 $\hat{\varepsilon}_i^{(P)}$ は，

$$\begin{aligned}\hat{\varepsilon}_i^{(P)} &= \frac{y_i - \hat{y}_i}{\sqrt{Var(\hat{y}_i)}} \\ &= \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}\end{aligned} \qquad (11.16)$$

として計算される．Pearson 残差 $\hat{\varepsilon}_i^{(P)}$ は過小評価になりがちなので，テコ比 h_{ii} により調整したスチューデント化 Pearson 残差も使われている．重みは，$\hat{w}_i = 1/\hat{y}_i$ であり，テコ比は，スチューデント化デビアンス残差で計算した表 11.10 と同じ式であり，スチューデント化 Pearson 残差 $\hat{\varepsilon}_i'^{(P)}$ は，スチューデント化デビアンス残差 $\varepsilon_i'^{(D)}$ の場合と同様に

$$\hat{\varepsilon}_1'^{(P)} = \frac{\hat{\varepsilon}_1^{(P)}}{\sqrt{1 - h_{11}'}} \qquad (11.17)$$

となる．表 11.11 に Pearson 残差およびスチューデント化 Pearson 残差についての計算結果を示す．

表 11.11　Pearson 残差およびスチューデント化 Pearson 残差

					Pearson	スチューデント化Pearson残差		
				推定値	残差	重み	テコ比	残差
i	$\boldsymbol{X}$		y	$y\hat{}$	$\varepsilon^{(P)}$	$w\hat{}=1/y\hat{}$	h'_{ii}	$\varepsilon'^{(P)}$
1	1	-1	2	**2.5163**	**-0.3255**	**0.3974**	0.4510	**-0.4393**
2	1	-1	3	2.5163	0.3049	0.3974	0.4510	0.4115
3	1	0	6	7.4516	-0.5318	0.1342	0.1049	-0.5621
4	1	0	7	7.4516	-0.1654	0.1342	0.1049	-0.1749
5	1	0	8	7.4516	0.2009	0.1342	0.1049	0.2123
6	1	0	9	7.4516	0.5672	0.1342	0.1049	0.5995
7	1	1	10	12.3869	-0.6782	0.0807	0.2261	-0.7709
8	1	1	12	12.3869	-0.1099	0.0807	0.2261	-0.1250
9	1	1	15	12.3869	0.7425	0.0807	0.2261	0.8440

推定値は，表 11.10 の完全モデルの推定値，
テコ比は，スチューデント化デビアンス残差でのテコ比に等しい．

実際の計算を $i=1$ の場合について示す．Pearson 残差は，

$$\begin{aligned}\varepsilon_1^{(P)} &= \frac{y_1 - \hat{y}_1}{\sqrt{\hat{y}_1}} \\ &= \frac{2 - 2.5163}{\sqrt{2.5163}} \\ &= -0.3255\end{aligned}$$

スチューデント化デ Pearson 残差は,

$$\hat{w}_1 = 1/\hat{y}_1$$
$$= 1/2.5163 = 0.3974$$

$$h_{11}' = \underbrace{\begin{bmatrix}1 & -1\end{bmatrix}}_{\boldsymbol{x}_1} \underbrace{\begin{bmatrix}\mathbf{0.7817} & 0.4166 \\ 0.4166 & \mathbf{1.1863}\end{bmatrix}}_{[(\boldsymbol{X}*\boldsymbol{w})^T\boldsymbol{X}]^{-1}} \underbrace{\begin{bmatrix}1 \\ -1\end{bmatrix}}_{\boldsymbol{x}_1{}^T} \times \underbrace{0.3974}_{w_1} = \underbrace{0.4510}_{h'_{ii}}$$

$$\hat{\varepsilon}_1^{(P)'} = \frac{\hat{\varepsilon}_i^{(P)}}{\sqrt{1-h_{11}'}}$$
$$= \frac{-0.3255}{\sqrt{1-0.4510}} = -0.4393$$

で求められる.

SAS/GENMOD による各種の残差

SAS の GENMOD プロシジャを用いて，スチューデント化 Pearson 残差を含め，これまでに示したポアソン回帰での 4 種の残差，（デビアンス残差，スチューデント化デビアンス残差，Pearson 残差，スチューデント化 Pearson 残差）を計算し，これまでの Excel による計算結果を検証する．SAS/GENMOD のオプションで「residual」が各種の残差の一括出力指示で，「plots= stdresdev(xbeta)」が，スチューデント化デビアンス残差を推定値 $\hat{y}_i$ に対する残差プロットの作成を指示している．

＜＜SAS/GENMOD によるポアソン回帰＞＞

```
Title "デビアンス残差_a01.sas   2020/01/20  Y.Takahashi" ;
data d01 ;
  input  x y @@ ;
datalines ;
-1  2    -1  3    0  6    0  7    0  8    0  9    1  10    1  12    1  15
;
proc genmod  data=d01  plots=reschi(xbeta)  plots=stdreschi(xbeta)
                       plots=resdev(xbeta)  plots=stdresdev(xbeta) ;
    model   y = X / dist=poisson link= identity  residual ;
    run;
```

表 11.12 に示した SAS の残差の出力で，「標準化」となっているのが「スチューデント化」の意味である「未加工残差」は，$\hat{\varepsilon}_i = y_i - \hat{y}_i$ であるが，右端の「尤度残差」は，スチューデント化デビアンス残差とスチューデント化 Pearson 残差の両方を使って，それらの中間的な残差である［SAS Institute（2016），The GENMOD Procedure :3164-3165］．SAS の残差の出力結果と，Excel で計算した 4 種の残差を照合し，一致することが確認される．

表 11.12　SAS/GENMOD の出力：ポアソン回帰の各種の残差統計量

観測値の統計量						
オブザベーション	未加工残差	Pearson 残差	デビアンス残差	標準化デビアンス残差	標準化 Pearson 残差	尤度残差
1	−0.5163	−0.3255	−0.3377	−0.4558	−0.4393	−0.4484
2	0.4837	0.3049	0.2958	0.3993	0.4115	0.4048
3	−1.4516	−0.5318	−0.5506	−0.5820	−0.5621	−0.5799
:						
9	2.6131	0.7425	0.7184	0.8167	0.8440	0.8229

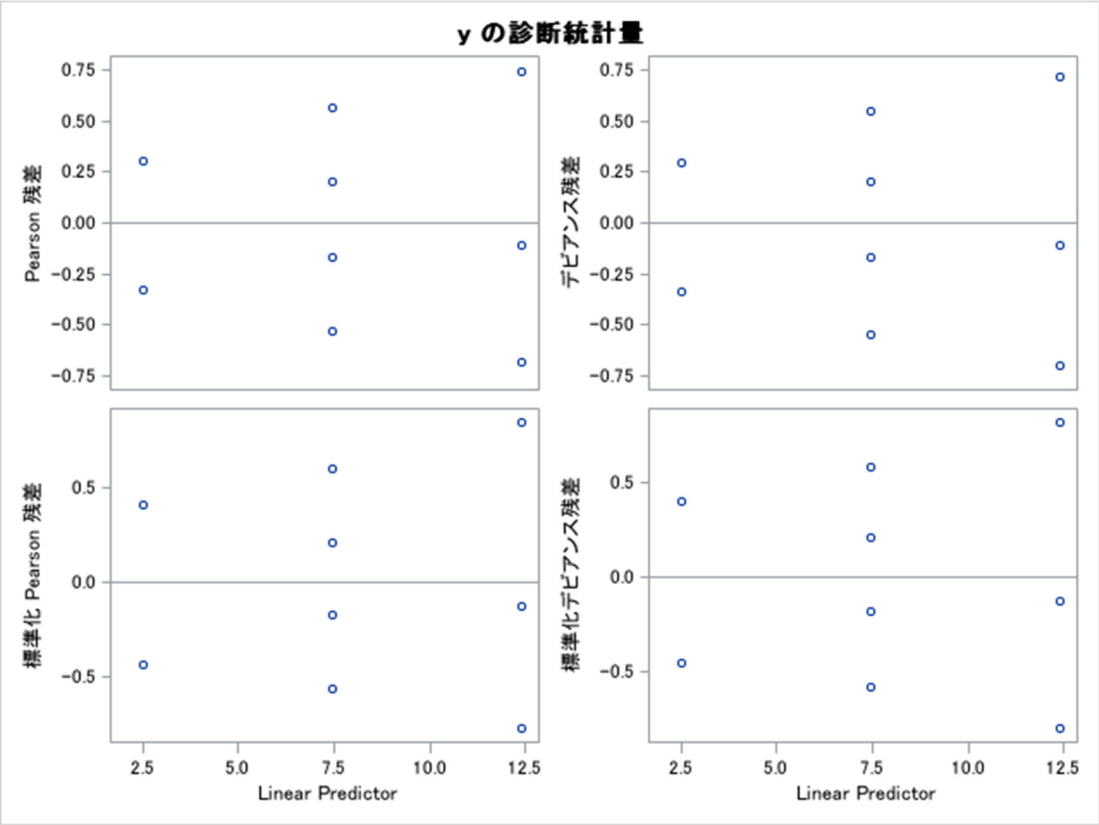


図 11.3　SAS/GENMOD による 4 種の残差プロット

このように沢山の残差があり，どれを使うか選択に窮するのである．最も簡便なのが Pearson 残差であるが，スチューデント化 Pearson 残差に比べて小さめに出るので，残差の大きさを少々過小評価となる．同様にデビアンス残差もスチューデント化デビアンス残差に比べて小さめになっている．

スチューデント化 Pearson 残差か，スチューデント化デビアンス残差かの選択については，この事例ではどちらとも言い切れない．元々の計算過程で，対数尤度を使っているので，飽和モデルと完全モデルの対数尤度の差を用いるデビアンス残差，あるいは，スチューデント化デビアンス残差を使うのが自然の流れのように思われる．

11.5.　カブトガニの事例における 4 種の残差

アグレスティ（2003）のカブトガニのデータについて第 1.13 節で概要を示し，第 7.2 節で探索的な解析の事例として用いた．ここでは，ポアソン回帰で用いられている 4 種の残差を比較検討するために用いる．このデータは，雌のカブトガニに連結する雄のサテライト数（Satellite 数）について 173 匹について，名義尺度（甲羅の色，後体部の棘の状態）の 2 変数，連続尺度（甲羅の幅，体重）の 2 変数，反応変数としてサテライト数が含まれている．

JMP による 4 種の残差の計算

第 1.13 節では，雌のカブトガニに連結する雄のサテライト数を反応変数とし，雌の甲羅の幅を説明変数とした対数リンクでのポアソン回帰を行い，散布図に回帰曲線および 95%信頼区間と予測区間（個別データの 95%信頼区間）を示し，多くのデータが 95%予測区間の外にあることを示した．残差プロットには，JMP の過分散の調整に使われている Pearson 残差について，「予測値 vs. Pearson 残差」を例示した．

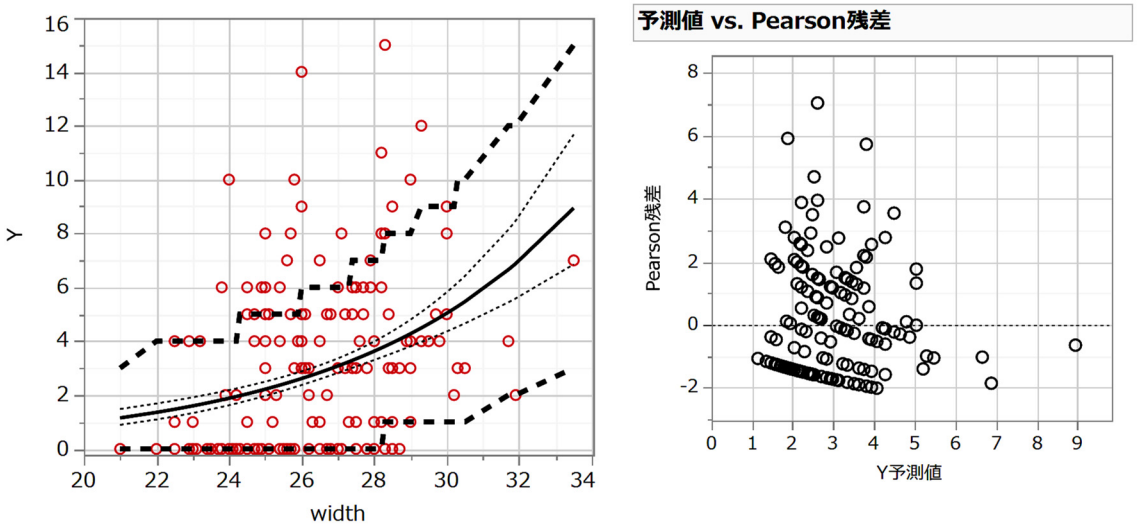


図 11.4　ポアソン回帰に対する 95%信頼区間および pearson 残差（図 1.15 再掲）

Pearson 残差プロットは，プラス側に大きく歪んでいる．これは，推定値に対してポアソン分布がプラス側に裾を引くことによる必然的な現象であること，さらに，ゼロを含む場合にプラス側に大きく裾を引くことも影響している．このような必然的に起きるバイアスを少しでも解消するためにデビアンス残差の使用が望ましい．

JMP で過分散なしの対数リンクでのポアソン回帰分析を行い，4 種の診断プロットを選択し，さらに，4 種の残差を JMP ファイルとして出力して，相互の比較を行う．

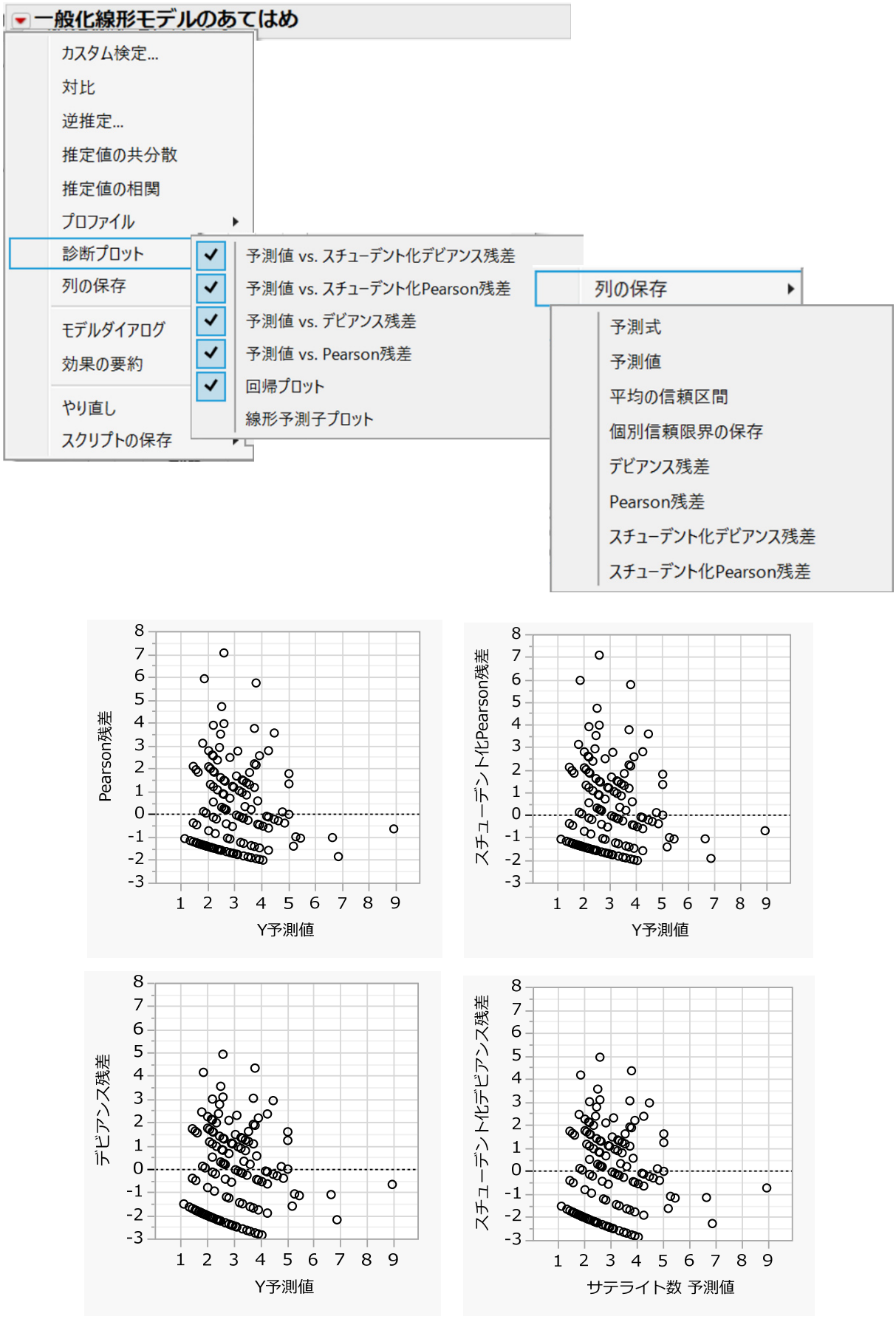


図 11.5　JMP/ポアソン回帰による 4 種の残差プロット

4 種の残差の比較

（Pearson 残差 vs. デビアンス残差）では，正の残差については，デビアンス残差を全体的に小さい方向へシフトし，負の残差の場合は，負の大きい方向にシフトしている．（スチューデント化 Pearson 残差 vs. スチューデント化デビアンス残差）も同様である．

（Pearson 残差 vs. スチューデント化 Pearson 残差）および（デビアンス残差 vs. スチューデント化デビアンス残差）については，式（11.7）および式（11.9）を用いたスチューデント化に用いているテコ比 h_{ii} が 1 以下であることから，必然的に残差ゼロを基準に残差の絶対値を大きい方に広げることになる．言い換えれば，プラスの残差はプラス方向に引き伸ばし，マイナスの残差は，マイナス方向に引き伸ばす．

どの程度の差が実際に起きるのかを実感するために，表 11.13 に JMP ファイルに出力された 4 種の残差を Excel に取り込みデビアンス残差の大きい順に並べて抜粋した結果を示す．スチューデント化した場合の残差の差について計算した結果は，大きいもので小数点以下 2 桁目での差であり，目立った違いではない．

Pearson 残差とデビアンス残差を比較すると，最初の行の No.15 では，Pearson 残差が 7.0448 であるのに対し，デビアンス残差は，4.9221 と明らかにデビアンス残差に縮小効果が表れている．また最後の No. 94 の場合は，-2.0171➔-2.8526 とマイナス方向への明らかな引き延ばしが確認される．

表 11.13　ポアソン回帰の各種の残差統計量

No	甲羅の色	後体部の棘	体重	甲羅の幅	サテライト数	予測式サテライト数	Pearson 残差	スチューデント化 Pearson 残差	Pearson 残差の差	デビアンス残差	スチューデント化デビアンス残差	デビアンス残差の差
15	2	1	2.30	26.0	14	2.6128	7.0448	7.0673	**0.0225**	4.9221	4.9379	**0.0157**
56	2	3	3.00	28.3	15	3.8103	5.7324	5.7608	**0.0284**	4.3279	4.3494	**0.0215**
13	2	3	3.05	28.2	11	3.7483	3.7456	3.7632	**0.0176**	3.0301	3.0443	**0.0142**
165	2	3	2.75	26.5	7	2.8361	2.4725	2.4799	**0.0074**	2.0787	2.0849	**0.0062**
28	2	1	2.70	26.8	5	2.9792	1.1708	1.1743	**0.0035**	1.0659	1.0692	**0.0032**
91	2	1	3.85	29.7	5	4.7941	0.0940	0.0951	**0.0011**	0.0934	0.0945	**0.0011**
124	2	3	1.65	24.2	2	1.9448	0.0396	0.0398	**0.0002**	0.0394	0.0396	**0.0002**
44	2	1	3.30	30.0	5	5.0359	-0.0160	-0.0162	**-0.0002**	-0.0160	-0.0162	**-0.0002**
63	3	1	2.45	27.0	3	3.0786	-0.0448	-0.0449	**-0.0001**	-0.0450	-0.0451	**-0.0001**
50	2	1	3.60	30.3	3	5.2900	-0.9956	-1.0122	**-0.0166**	-1.0848	-1.1028	**-0.0181**
81	3	2	2.25	24.5	0	2.0429	-1.4293	-1.4361	**-0.0068**	-2.0213	-2.0309	**-0.0096**
94	2	1	3.20	28.7	0	4.0688	-2.0171	-2.0297	**-0.0126**	-2.8526	-2.8705	**-0.0178**

全体で 173 サンプルをデビアンス残差の大きい順に並べ，最大値と最小値を残し，デビアンス残差がおおむね等間隔になるように 12 サンプルを抽出した．

図 11.6 に全 173 匹についてスチューデント化した場合の残差の変化について JMP の「対応のあるペア」によって比較した結果を示す．スチューデント化した場合，残差がゼロを起点にプラス方向とマイナス方向に引き伸ばされていることが確認される．

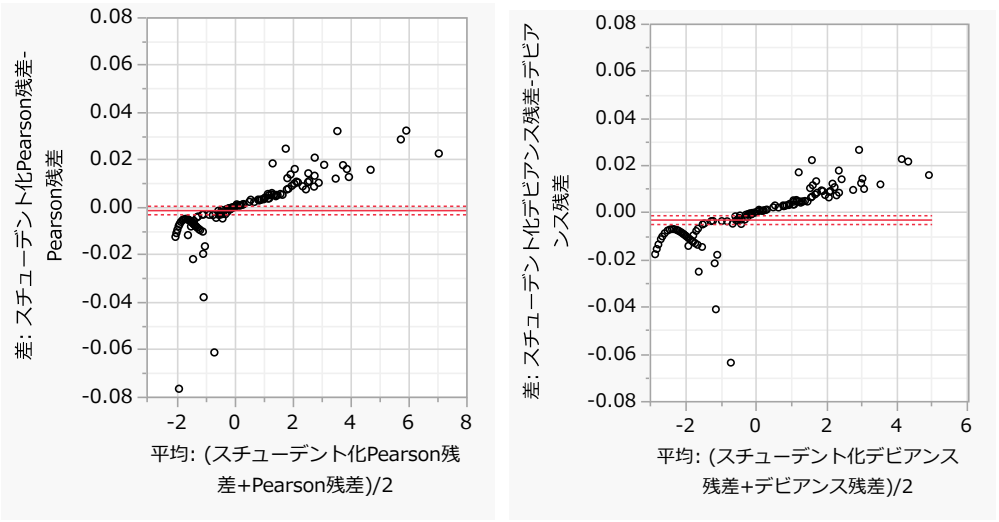


図 11.6　スチューデント化による残差の大きさの比効

図 11.7 に, スチューデント化 Pearson 残差とスチューデント化デビアンス残差の比較を JMP の「二変量の関係」によって作図した結果を示す．デビアンスにした場合に，残差がプラスの場合は圧縮され，マイナスの場合は，引き伸ばされていることが確認される．対数リンクの場合には，観測値が大いい方に裾を引くので，スチューデント化 Pearson 残差は，大きい方に引っ張られる．ビアンス残差あるいはスチューデント化デビアンス残差による補正が行われていることが確認さるれる．したがって，対数リンクの場合には，デビアンス残差あるいはスチューデント化デビアンス残差の使用が望ましい．

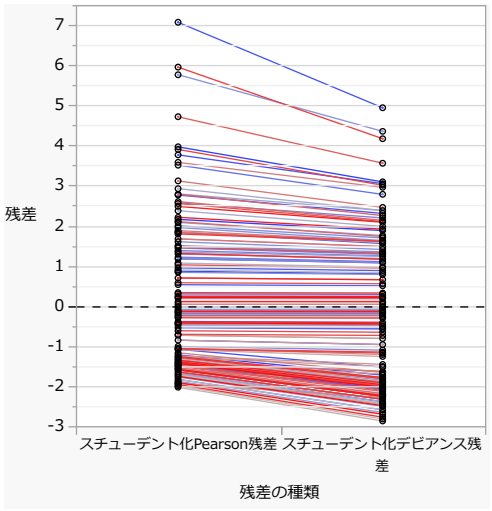


図 11.7　スチューデント化 Pearson 残差とスチューデント化デビアンス残差の比較

12. パラメータの共分散行列の活用

パラメータの共分散行列は，ポアソン回帰のみならず通常の回帰分析における種々の推定値に対する 95%信頼区間を求めるために不可欠な統計量であることを繰返し示してきた．一般的に共分散行列といった場合には，「データの共分散行列」を意味するので，「パラメータの共分散行列」とを明確に区別する必要がある．本章では，最初にデータの共分散行列について Excel による行列計算の入門として取り上げ，パラメータの共分散行列との違いを明確にする．次に，伝統的な偏差平方和行列をベースにした重回帰分析について Excel の行列計算を用いてチャレンジし，デザイン行列をベースにした重回帰分析につなげる．パラメータの共分散行列の活用事例として，2 次曲線の 95%信頼区間の推定を取り上げる．

12.1. データの共分散行列・パラメータの共分散行列

共分散は，身近な統計量であり 2 つの変数の相関係数を算出する過程で使われており，多変量データの関連を概観するための相関行列を算出する際にも共分散行列が使われている．共分散分析も良く知られた統計的方法であるが，質的変数に対する 1 元配置分散分析などに際し「共変量」についての回帰分析を併合した解析として知られており，「共分散行列」を用いた解析ではない．共分散分析については，第 13 章で取り上げるので，ここでは取り扱わない．

第 4 章では，デザイン行列を用いた回帰分析で回帰パラメータの標準誤差を求める際，「パラメータの共分散行列」の対角要素が，パラメータの分散となるので，その平方根が標準誤差として求めることを示した．また，回帰直線の 95%信頼区間および予測区間（個別データの信頼区間）を求める際に，一般的にシグマが用いられている計算式に代え，「パラメータの共分散行列」を活用する方法も示してきた．

多変量ポアソン回帰を Excel によるニュートン・ラフソン法を用いた解析，および，反復重み付き回帰を用いた解析に際し，計算過程の中で「パラメータの共分散行列」を当然のごとく使い，各種の推定値の 95%信頼区間の計算に際しても，「パラメータの共分散行列」を使ってきた．「パラメータの共分散行列」は，多変量ポアソン回帰のみならず，通常の回帰でも重回帰でも共通の存在であるが，日陰の存在であり続けている．理由は，はっきりしている．「パラメータの共分散行列」を統計ソフトの出力結果として得たとしても，それを行列とし

て再定義し，必要となる統計量の分散を計算するためには，行列計算なしでは難儀であるため，避けられ続けていると思われる．

以前より，JMP の「モデルあてはめ」による重回帰分析では，パラメータ間の相関行列を出力することができても，その元となるパラメータの共分散行列が出力できないない状態が現在でも続いている．他方,「モデルのあてはめ」でポアソン回帰を行う一般化線形モデルでは，パラメータの共分散行列および相関行列の出力がサポートされている．なぜなのだろうか．多くのユーザが必要としない機能に対し，統計ソフト・ベンダーが対応していない状況と思われる．

各種の分散分析モデルおよび重回帰分析は，ほとんどすべて一般線形モデルの応用によって解決でき，SAS の GLM プロシジャは，長年にわたり統計ソフトのバージョンアップが行われてきた．その結果として，多くのユーザの必要性に答えてきたことが,「パラメータの共分散行列」を用いて解析したいと思うような事例をほとんど網羅してきたためなのだろうか,「パラメータの共分散行列」を出力するすべがない．ただし，GLM プロシジャでは，例外的に Lsmeans 最小 2 乗平均に関連したパラメータの共分散行列の出力に限定して対応している．他方，重回帰に特化した SAS の REG プロシジャでは，cov オプションでパラメータの共分散行列の出力に対応している．

文献で取り上げられている事例を使った探索的ポアソン回帰をこれまで示したきたのであるが，JMP の一般化線形モデルにおけるポアソン回帰の出力は，回帰パラメータの推定に限定されており，Excel を用いて追加の推定を行う必要があった．その際に，パラメータの共分散行列が，中心的な役割を果すことを実感した．JMP の「予測プロファイル」は，これまでの統計ソフトにはない，素朴ではあるが画期的な出力であり，Excel でパラメータの共分散行列を用いて再現することを通じ，これまで私も真剣に向き合わなかった各種の課題に対し，丁寧な解説を試みる切っ掛けとなった．

回帰分析のパラメータの標準誤差を求める際に，それらの分散を計算する場合にシグマを用いた計算式が示されることが普通であるが，それらをパラメータの共分散行列としてまとめて扱っている成書に遭遇することはまれである．実際，パラメータの共分散行列が得られたとしても，それらを活用した種々の推定を行うためには，行列計算を前提にする必要があるために，意図的に避けているとも思われる．もちろん私も自分自身の理解を深める為に統計ソフトに付随する行列計算言語による解析を行ってきたのであるが，それを推奨することは難儀であることを実感してきた．

共分散行列といえばデータの共分散行列がメジャーであり，パラメータの共分散行列といっても「データの共分散行列のこと？」と思われるに違いない．パラメータの共分散行列を理解し活用するためには，データの共分散行列について Excel の行列関数を使った計算方法を知ることも必要である．データの共分散行列が得られれば，相関行列も定義に従って，行列計算により簡単に求めることができる．

Excel の行列関数を使って統計計算を行う入門としては，2 変数の相関係数の算出は，統計の基礎知識でもあり，Excel の行列計算でなくともシグマ的な計算でも，相関係数を求める Correl() 関数でも容易に求めることができ，多変数の相関行列を作成もイメージしやすい．

多変量データについての相関行列の算出は，データの共分散行列をベースにしているので Excel の行列関数を使った計算方法の入門に適している．これらの行列計算は，偏差平方和ベースの重回帰分析の基礎であり，また，デザイン行列ベースでの重回帰分析の基礎でもある．さらに，反復重み付き回帰によるポアソン回帰への拡張に対しても必須の知識でもある．

第 4 章は，Excel によるデザイン行列を用いた回帰分析の入門としたが，相関行列の作成は，逆行列が含まれないので，回帰分析よりも行列計算の入門として適していると思われる．

12.2.　アイリスデータの共分散行列および相関行列

多変量データとして，表 12.1 に示すようにフィッシャーのアイリスデータからバーシカラー種の 50 個のデータを抜き出して用いる．このアイリスのデータは，多変量解析の代表的な事例であり，Web 上で沢山の解説記事が見いだされ，データを手軽にダウンロードすることができる．どんな統計ソフトでも，多変量データの共分散行列および相関行列の計算は標準的にサポートされているので，実用的には Excel で計算する必要性は全くない．しかし，理論を学習し応用力を養うためには，Excel の行列計算などにより，各種の統計計算を実際に行う経験を積むことが，理論を確実なものにすると期待される．多変量データの相関行列をいかにスマートに計算するかは，行列計算の最初の課題として適している．

Excel の行列関数を用いた相関行列の算出

表 12.1 に示すデータは，50 行 4 列データの矩形状の集まりであり，行列 $\boldsymbol{X}$ とする．行列 $\boldsymbol{X}$ の 1 列目を列ベクトル $\boldsymbol{X}_1$（50 行×1 列）とし，順次 $\boldsymbol{X}_2$，$\boldsymbol{X}_3$，$\boldsymbol{X}_4$ とし，行方向は，行ベクトル $\boldsymbol{x}_1$（1 行×4 列）とし，順次 $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_{50}$ とする．

表 12.1　アイリスデータのバーシカラー種の相関行列の計算

	種類: versicolor								
i	がくの長さ x_1	がくの幅 x_2	花弁の長さ x_3	花弁の幅 x_4		がくの長さ x_1	がくの幅 x_2	花弁の長さ x_3	花弁の幅 x_4
1	7.0	3.2	4.7	1.4	平均 $\bar{\boldsymbol{x}}$=	5.9360	2.7700	4.2600	1.3260
2	6.4	3.2	4.5	1.5					
3	6.9	3.1	4.9	1.5	共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$=	**0.2664**	0.0852	0.1829	0.0558
4	5.5	2.3	4.0	1.3	（データの）	0.0852	**0.0985**	0.0827	0.0412
5	6.5	2.8	4.6	1.5		0.1829	0.0827	**0.2208**	0.0731
6	5.7	2.8	4.5	1.3		0.0558	0.0412	0.0731	**0.0391**
7	6.3	3.3	4.7	1.6					
8	4.9	2.4	3.3	1.0	分散 $\boldsymbol{\sigma}^2$=	0.2664	0.0985	0.2208	0.0391
9	6.6	2.9	4.6	1.3					
10	5.2	2.7	3.9	1.4	相関行列 $\boldsymbol{R}(\boldsymbol{x})$=	**1**	0.5259	0.7540	0.5465
11	5.0	2.0	3.5	1.0	（データの）	0.5259	**1**	0.5605	0.6640
12	5.9	3.0	4.2	1.5		0.7540	0.5605	**1**	0.7867
13	6.0	2.2	4.0	1.0		0.5465	0.6640	0.7867	**1**
:									
48	6.2	2.9	4.3	1.3					
49	5.1	2.5	3.0	1.1					
50	5.7	2.8	4.1	1.3					

手順 1） $\boldsymbol{X}_1$ の平均を $\bar{x}_1$ = Average($\boldsymbol{X}_1$の範囲）により計算し，右方向にフィルハンドルで計算式をコピーし，4 個の平均をベクトル $\bar{\boldsymbol{x}}$ する．

	がくの長さ x_1	がくの幅 x_2	花弁の長さ x_3	花弁の幅 x_4
平均 $\bar{\boldsymbol{x}}$ =	5.9360	2.7700	4.2600	1.3260

手順 2）偏差は［($\boldsymbol{X}$の範囲)－($\bar{\boldsymbol{x}}$の範囲)］の行列の引き算として計算する．行列計算では，50×4 の行列と 1×4 のベクトルとの差は．50×4 の行列となり，平均からの偏差が計算される．行列の積 Mmult() 関数で一気に 4×4 のデータの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$ を作成する．なお，49 は，自由度である

$\boldsymbol{\Sigma}(\boldsymbol{x})$=Mmult(Transpose(($\boldsymbol{X}$の範囲)－($\bar{\boldsymbol{x}}$の範囲))，(($\boldsymbol{X}$の範囲)－($\bar{\boldsymbol{x}}$の範囲)))/49

	Trannspose(($\boldsymbol{X}$の範囲) - ($\bar{\boldsymbol{x}}$の範囲))						(($\boldsymbol{X}$の範囲) - ($\bar{\boldsymbol{x}}$の範囲))				
	1	2	3	…	50		x_1	x_2	x_3	x_4	
x_1	1.0640	0.4640	0.9640	…	-0.2360		1.0640	0.4300	0.4400	0.0740	1
x_2	0.4300	0.4300	0.3300		0.0300		0.4640	0.4300	0.2400	0.1740	2
x_3	0.4400	0.2400	0.6400		-0.1600		0.9640	0.3300	0.6400	0.1740	3
x_4	0.0740	0.1740	0.1740		-0.0260		-0.4360	-0.4700	-0.2600	-0.0260	4
							:				:
							-0.2360	0.0300	-0.1600	-0.0260	50

共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$=	**0.2664**	0.0852	0.1829	0.0558
（データの）	0.0852	**0.0985**	0.0827	0.0412
	0.1829	0.0827	**0.2208**	0.0731
	0.0558	0.0412	0.0731	**0.0391**

手順 3）データの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$ の対角要素である変数 x_1 の分散を別途

$$Var(x_1) = \text{SumSq}((\boldsymbol{X}_1\text{の範囲}) - \bar{x}_1) / 49$$

あるいは， $Var(x_1)$ = Var.S($\boldsymbol{X}_1$の範囲)

で計算し，計算式をフィルハンドルでコピーし，4 個の分散を計算する．

分散 $\boldsymbol{\sigma}^2$=	0.2664	0.0985	0.2208	0.0391

手順 4）相関行列 $\boldsymbol{R}(\boldsymbol{x})$ を計算する

$$\boldsymbol{R}(\boldsymbol{x}) = (\boldsymbol{\Sigma}(\boldsymbol{x})\text{の範囲}) \,/\, \text{Sqrt}(\boldsymbol{\sigma}^2\text{ の範囲}) \,/\, \text{Sqrt}(\text{Transpose}(\boldsymbol{\sigma}^2\text{ の範囲}))$$

	データの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$				SD^T		データの相関行列 $\boldsymbol{R}(\boldsymbol{x})$			
相関行列 $\boldsymbol{R}(\boldsymbol{x})$=	**0.2664**	0.0852	0.1829	0.0558	0.5162	=	**1**	0.5259	0.7540	0.5465
	0.0852	**0.0985**	0.0827	0.0412	0.3138		0.5259	**1**	0.5605	0.6640
	0.1829	0.0827	**0.2208**	0.0731	0.4699		0.7540	0.5605	**1**	0.7867
	0.0558	0.0412	0.0731	**0.0391**	0.1978		0.5465	0.6640	0.7867	**1**
				対応する行で除す						
SD=	0.5162	0.3138	0.4699	0.1978						
	対応する列で除す									

JMP の「多変量の相関」で計算した結果を表 12.2 に示す．Excel での計算結果と一致することが確認される．

表 12.2　JMP の「多変量の相関」によるデータの共分散行列および相関行列

共分散行列

	がくの長さ	がくの幅	花弁の長さ	花弁の幅
がくの長さ	0.2664	0.0852	0.1829	0.0558
がくの幅	0.0852	0.0985	0.0827	0.0412
花弁の長さ	0.1829	0.0827	0.2208	0.0731
花弁の幅	0.0558	0.0412	0.0731	0.0391

相関

	がくの長さ	がくの幅	花弁の長さ	花弁の幅
がくの長さ	1.0000	0.5259	0.7540	0.5465
がくの幅	0.5259	1.0000	0.5605	0.6640
花弁の長さ	0.7540	0.5605	1.0000	0.7867
花弁の幅	0.5465	0.6640	0.7867	1.0000

相関はリストワイズ法によって推定されました。

データの相関行列と標準偏差 SD を用いてデータの共分散行列を逆に求めることもできる．

手順 5）相関行列 $\boldsymbol{R}(\boldsymbol{x})$ と分散ベクトルから共分散行列を計算する．

$$\boldsymbol{\Sigma}(\boldsymbol{x}) = (\boldsymbol{R}(\boldsymbol{x})\text{の範囲}) * \text{Sqrt}(\boldsymbol{\sigma}^2\text{ の範囲}) * \text{Sqrt}(\text{Transpose}(\boldsymbol{\sigma}^2\text{ の範囲}))$$

	データの相関行列 $\boldsymbol{R}(\boldsymbol{x})$				SD^T		データの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$			
共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$=	**1**	0.5259	0.7540	0.5465	0.5162	=	**0.2664**	0.0852	0.1829	0.0558
	0.5259	**1**	0.5605	0.6640	0.3138		0.0852	**0.0985**	0.0827	0.0412
	0.7540	0.5605	**1**	0.7867	0.4699		0.1829	0.0827	**0.2208**	0.0731
	0.5465	0.6640	0.7867	**1**	0.1978		0.0558	0.0412	0.0731	**0.0391**
				対応する行で掛ける						
SD=	0.5162	0.3138	0.4699	0.1978						
	対応する行で掛ける									

こに示した行列計算は，中間的な計算結果を示していないので，Excel の行列計算に不慣れな場合には，Ecxcel シート上に中間結果を書き出すことを薦める．「手順 2）偏差は $(\boldsymbol{X}$の範囲$)-(\bar{\boldsymbol{x}}$の範囲$)$で計算する」などは，練習用に小さな行列を用いることから始めるとよい．なお，行列計算の詳細は，第 4 章で丁寧に説明しているので，参考にしてもらいたい．

逆行列の練習には，相関行列から偏相関行列を求める課題もあり，Excel による行列計算の入門に含めることもできるが，割愛する（添付の Excel シートには含まれている）．

分析ツールを使う場合

Excel の分析ツールで，データの共分散行列および相関行列を計算することができる．ただし，表 12.3 に示すようにデータの共分散行列は，母集団を仮定した場合であり，標本を仮定した場合ではないので，表 12.2 で示した結果とは，微妙に異なる．なお，データ相関行列は，どちらを仮定した場合でも一致する．

表 12.3　Excel の分析ツールによる共分散行列および相関行列

Excel データの共分散行列(母集団)				
	x1	x2	x3	x4
がくの長さ x1	**0.2611**			
がくの幅 x2	0.0835	**0.0965**		
花弁の長さ x3	0.1792	0.0810	**0.2164**	
花弁の幅 x4	0.0547	0.0404	0.0716	**0.0383**

Excel データの相関行列				
	x1	x2	x3	x4
がくの長さ x1	**1**			
がくの幅 x2	0.5259	**1**		
花弁の長さ x3	0.7540	0.5605	**1**	
花弁の幅 x4	0.5465	0.6640	0.7867	**1**

手順 5）によりデータの相関行列から共分散行列を作成することを示したが，そのためには，上三角行列を代入文で埋める必要があり，手作業的な操作となるので省略する．

共分散関数を使う場合

変数間の共分散については，Excel の Covariance.S()関数を使うと標本に対する共分散が得られる．ただし，2 変数間なので，共分散行列にするためには一工夫する必要がある．

表 12.4　Excel の Covariance.S() 関数による共分散行列

Covariance.S() 関数				
	x1	x2	x3	x4
がくの長さ x1	**0.2664**	0.0852	0.1829	0.0558
がくの幅 x2	0.0852	**0.0985**	0.0827	0.0412
花弁の長さ x3	0.1829	0.0827	**0.2208**	0.0731
花弁の幅 x4	0.0558	0.0412	0.0731	**0.0391**

共分散行列の列 1 と行 1 で

=Covariance.S(C5 : C54,　C$5 : C$54)

（$\boldsymbol{X}_1$ に固定，$\boldsymbol{X}_1$ の行方向を固定）

のように Covariance.S() 関数をセットし，フィルハンドルで列方向に計算式をコピーする．次に列 2 と行 2 で

=Covariance.S(D5 : D54,　D$5 : D$54)

（$\boldsymbol{X}_2$ に固定，$\boldsymbol{X}_2$ の行方向を固定）

のように関数をセットし，フィルハンドルで列の左右に計算式をコピーする．これを繰り返す．このように，いくつかあるの計算手段を知ったうえで，目的に応じて簡便で汎用性の高い方法を選択することを勧める．

Correl() 関数を使い相関係数行列を直接作成することもできるが，Covariance.S() 関数と同様なので割愛する．

12.3.　偏差平方和ベースの重回帰分析

先人たちによって，計算手段が限られていた時代に様々な工夫により，偏差平方和を主体にした単回帰分析を拡張し重回帰分析の定式化がなされてきた．多くの多変量解析および重回帰分析の書物が出版されている中で，版を重ね読み継がれてきたのは，奥野・久米・芳賀・吉沢著（1981)，「多変量解析法 改訂版」である．

奥野ら(1981）は，重回帰分析について行列計算を用いずに，シグマを用いた偏差平方和をベースにした方法で多様な事例を丁寧に示している．しかし，それらの式の意味を理解し習得するために，Excel でシグマを用いた計算を行うことは絶望的ですらある．そこで，シグマで示されている計算式を Excel の行列関数を用いて行うことにし，デザイン行列を活用した計算方法と対比する．なお，Excel の計算式についての説明は省略するので，添付の Eceel シートを参照されたい．

奥野ら(1981）の「第 4 章 偏回帰係数の解釈」の「表 4.1 材料，工数と生産量の関係」を用いて，行列計算により偏差平方和ベースの重回帰分析の手順を示す．あるガラス加工工程で，投入材料 x_1，使用工数 x_2 と生産量 y の関係を調べたところ，表 12.5 に示すデータが得られた．

表 12.5　材料, 工数と生産量の関係［奥野ら(1981), 表 4.1］

	材料	工数	生産量		材料	工数	生産量
No.	x_1(m^2)	x_2(hr)	y (m^2)	No.	x_1(m^2)	x_2(hr)	y (m^2)
1	54	29	50	12	82	50	73
2	61	39	51	13	75	39	74
3	52	26	52	14	92	60	78
4	70	48	54	15	96	62	82
5	63	42	53	16	92	61	80
6	79	62	60	17	91	50	87
7	68	45	59	18	85	43	84
8	65	30	65	19	106	72	88
9	79	51	67	20	96	52	92
10	76	44	70	計	1,553	941	1,389
11	71	36	70	平均	77.65	47.05	69.45

手順 1）材料 x_1，工数 x_2 の 20 行分のデータを行列 $\boldsymbol{X}$ （20×2）とし，平均値をベクトル $\bar{\boldsymbol{x}}$ （2×1）とし，次式で偏差平方和行列 $\boldsymbol{S}_{xx}$ を計算する．

$$\boldsymbol{S}_{xx} = (\boldsymbol{X} - \bar{\boldsymbol{x}})^T (\boldsymbol{X} - \bar{\boldsymbol{x}}) = \begin{bmatrix} \mathbf{4218.55} & 3009.35 \\ 3009.35 & \mathbf{2856.95} \end{bmatrix}$$

手順 2）材料 x_1，工数 x_2 と生産量 y との偏差平方和を計算する．20 行分の生産量 y をベクトル $\boldsymbol{y}$，その平均を $\bar{y}$ とし，次式で偏差平方和ベクトル $\boldsymbol{S}_{xy}$ を計算する．

$$\boldsymbol{S}_{xy} = (\boldsymbol{X} - \bar{\boldsymbol{x}})^T (\boldsymbol{y} - \bar{y}) = \begin{array}{|r|}\hline 3499.15 \\ 1860.55 \\ \hline \end{array}$$

手順 3） $\boldsymbol{S}_{xx}$ の逆行列 $\boldsymbol{S}^{xx}$ を計算する．

$$\boldsymbol{S}^{xx} = (\boldsymbol{S}_{xx})^{-1} = \begin{array}{|rr|}\hline \textbf{0.000954} & -0.001004 \\ -0.001004 & \textbf{0.001408} \\ \hline \end{array}$$

手順 4）推定値 $\hat{\beta}_1$， $\hat{\beta}_2$ のベクトル $\hat{\boldsymbol{\beta}}$ （2×1）を次式で計算する．

$$\hat{\boldsymbol{\beta}} = \boldsymbol{S}^{xx}\boldsymbol{S}_{xy} = \begin{array}{|r|}\hline \textbf{1.4679} \\ \textbf{-0.8950} \\ \hline \end{array}$$

手順 5）切片 $\hat{\beta}_0$ を次式で計算する．

$$\hat{\beta}_0 = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}} = \boxed{\textbf{-2.4245}}$$

手順 6）以上の計算から回帰式が得られる．

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -2.4245 + 1.4679 x_1 - 0.8950 x_2$$

奥野ら（1981）に「この式で奇妙なことは，x_2 の係数がマイナスになっていることである．これをそのまま解釈すれば，工数を減らせば生産量（絶対量）が増加するということだから，こんなうまい話はない．はたしてそうであろうか？」このような疑問が示されている．

手順 7）重相関係数 R^2 を次式で計算する．.

$$R^2 = (\hat{\boldsymbol{\beta}}^T \boldsymbol{S}_{xy}) / S_{yy} = \boxed{0.9904}$$

ただし， $S_{yy} = (\boldsymbol{y} - \bar{y})^T (\boldsymbol{y} - \bar{y}) = \boxed{3504.95}$

手順 8）誤差分散 $\hat{\sigma}^2$ を次式で計算する．.

$$\hat{\sigma}^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}}) / (n-3) = \boxed{\textbf{1.9803}}$$

ただし， $\hat{\boldsymbol{y}} = \hat{\beta}_0 + \boldsymbol{X}\hat{\boldsymbol{\beta}}$ ， 自由度： $(n-3) = 20 - 3 = 17$

手順 9）推定値 $\hat{\beta}_1$， $\hat{\beta}_2$ の分散 $Var(\hat{\beta}_1)$， $Var(\hat{\beta}_2)$ を次式で計算する．

$$Var(\hat{\beta}_1) = \boldsymbol{S}^{11}\hat{\sigma}^2 = \boxed{0.001888}$$

$$Var(\hat{\beta}_2) = \boldsymbol{S}^{22}\hat{\sigma}^2 = \boxed{0.002788}$$

ただし， $\boldsymbol{S}^{ii}$ は， $\boldsymbol{S}^{xx}$ の対角要素

手順 10）推定値 $\hat{\beta}_0$ の分散 $Var(\hat{\beta}_0)$ を次式で計算する．

$$Var(\hat{\beta}_0)=\left(\frac{1}{n}+\bar{\boldsymbol{x}}\boldsymbol{S}^{xx}\bar{\boldsymbol{x}}^T\right)\hat{\sigma}^2 = \boxed{3.1236}$$

手順 11）分散分析表のための平方和 S_T， S_R， S_e を次式で計算する．

$$S_T = S_{yy} = \boxed{\mathbf{3504.9500}}$$

$$S_R = (\boldsymbol{S}_{xy})^T\hat{\boldsymbol{\beta}} = \boxed{\mathbf{3471.2842}}$$

$$S_e = S_{yy} - S_R = \boxed{\mathbf{33.6658}}$$

手順 12）以上の Excel での計算シートを表 12.6 示す．その結果は，表 12.7 に示す Excel の回帰分析による分散分析表およびパラメータの推定値と一致することが確認される．

表 12.6　偏差平方和ベースの重回帰の Excel の計算シート

	$(\boldsymbol{X}-(\boldsymbol{x}^-))^T(\boldsymbol{X}-(\boldsymbol{x}^-))$			$(\boldsymbol{X}-(\boldsymbol{x}^-))^T(\boldsymbol{y}-(y^-))$			$(\boldsymbol{y}-(y^-))^T(\boldsymbol{y}-(y^-))$	
$\boldsymbol{S}_{xx}=$	**4218.55**	3009.35	$\boldsymbol{S}_{xy}=$	3499.15		$S_{yy}=$	3504.95	
	3009.35	**2856.95**		1860.55				
	$(\boldsymbol{S}_{xx})^{-1}$			$\boldsymbol{S}^{xx}\boldsymbol{S}_{xx}$			$\boldsymbol{S}^{11}\boldsymbol{\sigma}\hat{}^2$	*SE*
$\boldsymbol{S}^{xx}=$	**0.000954**	-0.001004	$\boldsymbol{\beta}\hat{}=$	**1.4679**		$Var(\beta_1\hat{})=$	0.001888	**0.0435**
	-0.001004	**0.001408**		**-0.8950**			$\boldsymbol{S}^{22}\boldsymbol{\sigma}\hat{}^2$	
						$Var(\beta_2\hat{})=$	0.002788	**0.0528**
	$(\boldsymbol{\beta}\hat{}^t\boldsymbol{S}_{xx})/\boldsymbol{S}_{xy}$			$(y^-)-(\boldsymbol{x}^-)\boldsymbol{\beta}\hat{}$			$(1/n+(\boldsymbol{x}^-)\boldsymbol{S}_{xx}(\boldsymbol{x}^-)^T)\sigma\hat{}^2$	
$R^2=$	0.9904		$\beta_0\hat{}=$	**-2.4245**		$Var(\beta_0\hat{})=$	3.1236	**1.7674**
	$(\boldsymbol{y}-\boldsymbol{y}\hat{})^T(\boldsymbol{y}-\boldsymbol{y}\hat{})/(n-3)$							
$\boldsymbol{\sigma}\hat{}^2=$	**1.9803**		$S_T=$	**3504.9500**	S_{yy}			
	$\boldsymbol{y}\hat{}=\beta_0\hat{}+\boldsymbol{X}\boldsymbol{\beta}\hat{}$		$S_R=$	**3471.2842**	$(S_{xy})^T\boldsymbol{\beta}\hat{}$			
			$S_e=$	**33.6658**	$S_{yy}-S_R$			

表 12.7　Excel の回帰分析による分散分析表およびパラメータの推定値

分散分析表					
	自由度	変動	分散	分散比	有意 *F*
回帰	2	3471.2842	1735.6421	876.4369	0.0000
残差	17	33.6658	1.9803		
合計	19	3504.9500			
	係数	標準誤差	*t*	*P*-値	
切片	-2.4245	1.7674	-1.3718	0.1880	
材料 x1	1.4679	0.0435	33.7791	0.0000	
工数 x2	-0.8950	0.0528	-16.9485	0.0000	

注）Excel の回帰分析で簡単にできることを，細々と計算するのは，私にとってもストレスであるが，基本を身に付けなければ，第 12.4 節で示す 2 次式の回帰曲線の 95%信頼区間を描くこととすらできない．先人たちの様々な工夫を学ぶことも，新たな応用力を付けるために必要である．

手順 13）回帰の推定値 $\hat{y}_i$ および分散 $Var(\hat{y}_i)$ を次式で求める．

$\boldsymbol{x}_1$=	54	29

$$\hat{y}_1 = \hat{\beta}_0 + \boldsymbol{x}_1\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,1} + \hat{\beta}_2 x_{2,1} = \boxed{50.8883}$$

$$Var(\hat{y}_1) = \left[\frac{1}{n} + (\boldsymbol{x}_1 - \bar{\boldsymbol{x}})\boldsymbol{S}^{xx}(\boldsymbol{x}_1 - \bar{\boldsymbol{x}})^T\right]\hat{\sigma}^2 = \boxed{0.3655}$$

全ての $\hat{y}_i$ と $Var(\hat{y}_i)$ を求め，結果を表 12.8 に示す．

表 12.8　偏差平方和ベースの重回帰の推定値と分散

	材料	工数	生産量	推定値	分散
No.	x_1	x_2	y	y^	$Var(y$^$)$
1	**54**	**29**	50	**50.89**	**0.3655**
2	61	39	51	52.21	0.2700
3	52	26	52	50.64	0.4290
4	70	48	54	57.37	0.2410
5	63	42	53	52.46	0.2811
:					
18	85	43	84	83.86	0.3652
19	106	72	88	88.74	0.5386
20	96	52	92	91.96	0.4419
計	1,553	941	1,389		
平均	77.65	47.05	69.45		

奥野ら（1981）の重回帰分析の計算手順は，単回帰分析で一般的となっている偏差平方和に基づく計算手順を重回帰分析に拡張したものである．計算手段が乏しかった時代の標準的の方法であり，生物統計の名著であるスネデカー・コクラン（1972）でも，医学統計の名著であるアーミテージら（2001）でも説明変数の偏差平方和に基づく重回帰分析の手順として示されていることから，世界的に標準的な方法として普及してきたと理解される．ただし，手順 11）に示されている回帰の平方和 $S_R = (\boldsymbol{S}_{xy})^T\hat{\boldsymbol{\beta}}$ の計算方法，誤差平方和 $S_e = S_{yy} - S_R$ の計算は，計算量を減らすために計算公式であり，表 12.9 に示されている平方和の定義式による計算法が，回帰分析の理解に欠かせない．

このような偏差平方和ベースの重回帰分析の解析手順は，計算手段が乏しく有効数字の桁数が 7 程度であった単精度実数の時代の標準的な手順として理解できる．それぞれの平均を差し引いた後に，標準偏差を計算し基準化して計算精度を確保することは，数値計算の常識でもある．現在の Excel は倍精度実数での計算が標準であり，平均値を引かなければ計算精度が保てないことはなくなった．ただし，多項式回帰を行う際に，べき乗の項については，桁数のインフレーションを防ぐ何らかの手立てをすることが必要である．

12.4.　デザイン行列ベースの重回帰分析

Excel によるデザイン行列ベースの重回帰分析

「第 4.5 節 デザイン行列を用いた回帰分析の実際」では，単回帰分析に対しても偏差平方和に基づく方法ではなく，切片を含むデザイン行列を用いた行列計算による回帰分析の解析法を示した．この方法は，切片の推定値を別途計算するのではなく，説明変数のパラメータ推定も併せて行うので，手順としてはスマートであり，ポアソン回帰などへの拡張が容易である．表 12.5 に示した奥野ら（1981）のデータを表 12.9 に示すように切片 x_0 を加えたデザイン行列 $\boldsymbol{X}$ に対し，Excel シート上で重回帰を行ったので，偏差平方和ベースの重回帰分析の方法と比較してもらいたい．

表 12.9　Excel によるデザイン行列ベースの重回帰分析

	デザイン行列 $\boldsymbol{X}$				推定値	分散					
No.	x_0	x_1	x_2	y	y^	$Var(y$^$)$					
1	1	54	29	50	**50.89**	**0.3655**	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$=	**1.5773**	-0.0268	0.0117	
2	1	61	39	51	52.21	0.2700		-0.0268	**0.0010**	-0.0010	
3	1	52	26	52	50.64	0.4290		0.0117	-0.0010	**0.0014**	
4	1	70	48	54	57.37	0.2410					
5	1	63	42	53	52.46	0.2811	$\boldsymbol{\beta}$^=	**-2.4245**	: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$		
6	1	79	62	60	58.05	0.6454		**1.4679**			
7	1	68	45	59	57.12	0.2079		**-0.8950**			
8	1	65	30	65	66.14	0.3538					
9	1	79	51	67	67.90	0.1248	$\boldsymbol{y}$^= $\boldsymbol{X\beta}$^				
10	1	76	44	70	69.76	0.1101					
11	1	71	36	70	69.58	0.2307	σ^2=	1.9803	: $(\boldsymbol{y}-\boldsymbol{y}$^$)^T(\boldsymbol{y}-\boldsymbol{y}$^$)/17$		
12	1	82	50	73	73.20	0.1080					
13	1	75	39	74	72.76	0.2081	$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)$=	**3.1236**	-0.0530	0.0233	: $Var(\beta_0$^$)$
14	1	92	60	78	78.92	0.2162		-0.0530	**0.0019**	-0.0020	: $Var(\beta_1$^$)$
15	1	96	62	82	83.01	0.2667		0.0233	-0.0020	**0.0028**	: $Var(\beta_2$^$)$
16	1	92	61	80	78.03	0.2341					
17	1	91	50	87	86.41	0.3032	$Va\boldsymbol{x}_i\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)\boldsymbol{x}_i{}^T$				
18	1	85	43	84	83.86	0.3652					重相関
19	1	106	72	88	88.74	0.5386	S_T=	**3504.9500**	: SumSq($\boldsymbol{y}-\bar{y}$)		$\boldsymbol{R}^2=S_R/S_T$
20	1	96	52	92	91.96	0.4419	S_R=	**3471.2842**	: SumSq($\boldsymbol{y}$^$-\bar{y}$)		0.9904
平均	1.00	77.65	47.05	69.45			S_e=	**33.6658**	: SumSq($\boldsymbol{y}-\boldsymbol{y}$^$)$		

手順 a）切片 x_0，材料 x_1，工数 x_2 のデザイン行列 $\boldsymbol{X}$（20×3）に対し積和行列を求め，その逆行列を計算する．

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$=	**1.5773**	-0.0268	0.0117
	-0.0268	**0.0010**	-0.0010
	0.0117	-0.0010	**0.0014**

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ = Minverse(Mmult(Transpose($\boldsymbol{X}$の範囲), $\boldsymbol{X}$の範囲))

以下，Excel での計算式は省略する．

手順 b）推定値 $[\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_3]^T$ のベクトル $\hat{\boldsymbol{\beta}}$ （3×1）を次式で計算する．

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{array}{|r|}\hline \textbf{-2.4245} \\ \textbf{1.4679} \\ \textbf{-0.8950} \\ \hline \end{array}$$

手順 c）推定された $\hat{\boldsymbol{\beta}}$ を用いて推定値 $\hat{\boldsymbol{y}}$ を計算する．

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$\begin{aligned}\hat{y}_1 &= \hat{\beta}_0 x_{0,1} + \hat{\beta}_1 x_{1,1} + \hat{\beta}_2 x_{2,1} \\ &= -2.4245\times 1 + 1.4679\times 54 - 0.8950\times 29 \\ &= 50.89\end{aligned}$$

手順 d）誤差分散 $\hat{\sigma}^2$ を次式で計算する．

$$\begin{aligned}\hat{\sigma}^2 &= (\boldsymbol{y}-\hat{\boldsymbol{y}})^T(\boldsymbol{y}-\hat{\boldsymbol{y}})/(n-3) \\ &= \text{SumSq}(\boldsymbol{y}\text{の範囲} - \hat{\boldsymbol{y}}\text{の範囲})/17 \\ &= 1.9803\end{aligned}$$

手順 e）パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を計算する．対角要素がパラメータの分散になる．

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2 = \begin{array}{|r|r|r|l|}\hline \textbf{3.1236} & -0.0530 & 0.0233 & : Var(\hat{\beta_0}) \\ -0.0530 & \textbf{0.0019} & -0.0020 & : Var(\hat{\beta_1}) \\ 0.0233 & -0.0020 & \textbf{0.0028} & : Var(\hat{\beta_2}) \\ \hline \end{array}$$

手順 f）分散分析表のための平方和 S_T， S_R， S_e を次式で計算する．

$$S_T = \text{SumSq}(\boldsymbol{y}\text{の範囲} - \bar{y}) = 3504.9500$$

$$S_R = \text{SumSq}(\hat{\boldsymbol{y}}\text{の範囲} - \bar{y}) = 3471.2842$$

$$S_e = \text{SumSq}(\boldsymbol{y}\text{の範囲} - \hat{\boldsymbol{y}}\text{の範囲}) = 33.6658$$

手順 g）重相関係数 R^2 を次式で計算する..

$$R^2 = S_R / S_T = 0.9904$$

手順 h）回帰の推定値 $\hat{y}_1$ の分散 $Var(\hat{y}_1)$ を次式で求め，

$\boldsymbol{x}_1$=	1	54	29

$$Var(\hat{y}_1) = \boldsymbol{x}_1\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_1^T$$

	$\boldsymbol{x}_1$				$\boldsymbol{\Sigma}(\boldsymbol{\beta}\hat{})$				$\boldsymbol{x}_i^T$		
=	1	54	29		**3.1236**	-0.0530	0.0233		1	=	0.3655
					-0.0530	**0.0019**	-0.0020		54		
					0.0233	-0.0020	**0.0028**		29		

フィルハンドルを用いて計算式をコピーして全ての $Var(\hat{y}_i)$ を求める．

2 変量の重回帰分析について，奥野ら（1981）に忠実に偏差平方和をベースにした解析を Excel の行列関数を用いて行い，それと対比する形でデザイン行列をベースにした解析法を対

比した．手順数は 13 から 8 に減少し，数式も大幅に簡素化された．これは，偏差平方和をベースにした場合に，常に平均値のベクトルを考慮した式となり，さらに，切片の推定値 $\hat{\beta}_0$ を別計算で求める煩雑さが付きまとっている．しかし，慣れ親しんだ 1 変量の回帰分析の手順を多変量に拡張する手順が示されていることは，興味深い．ただし，ポアソン回帰を含んだ一般化線形モデルへの拡張を視野に入れた場合には，切片を含めたデザイン行列をベースにした回帰分析の方法が理解の助けになる．

等高線図

手順 6）および手順 c）で推定された回帰式，

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -2.4245 + 1.4679 x_1 - 0.8950 x_2$$

に対し，x_2 のパラメータがマイナスとなっており，そのまま解釈すれば，工数を減らせば生産量（絶対量）が増加するということだから，こんなうまい話はない．「はたしてそうであろうか」との疑問が起きる．

この疑問に答えるために，2 変量の等高線図を用いると理解しやすい．図 12.1 に X 軸に材料 x_1 を，Y 軸に工数 x_2 とし，生産量 y の値を散布図として示す．工数 x_2 を 50 hr に固定し，材料 x_1 の 70 m^2～90 m^2 の等高線を読むと生産量 y は，60，70，80 m^2 と増加している．次に，

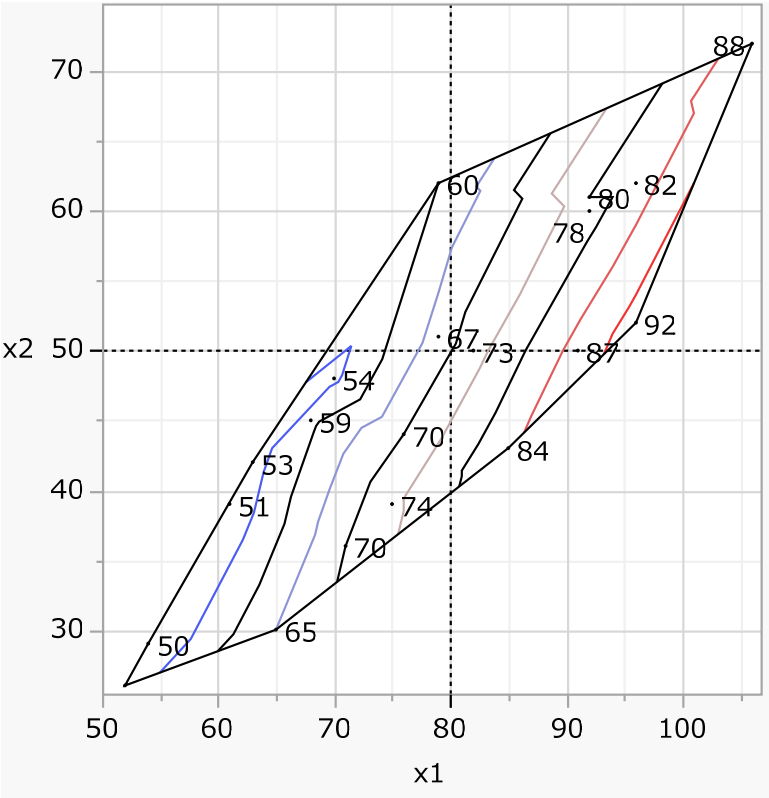


図 12.1 材料 x_1 と工数 x_2 に対する生産量 y の JMP による等高線図

材料 x_1 を 80 m^2 に固定し，工数 x_2 の 40 hr～60 hr の等高線を読むと生産量 y は，おおよそ 80, 70，60 m^2 と減少している．

等高線から，工数 x_2 を増やすことなく材料 x_1 を増やせば，生産量 y を増やすこと読み取れる．材料 x_1 が同じでも工数 x_2 を減らすことにより生産量をわずかであるが増やすことができる．材料 x_1 が同じでも工数 x_2 を増加させると生産量が落ちることは，何か別の測定していない変数の影響が潜んでいることが伺われる．

予測プロファイル

等高線図を用いた検討方法をさらに細かく「予測プロファイル」を作成し検討する．表 12.10 に示すように，等高線図を参考にして材料 x_1 を 80 m^2 に固定し，工数 x_2 を（30，40, . . . , 70 hr）と変化させて，生産量 y を推定する．材料 x_1 を 80 m^2，工数 x_2 を 30 hr とした場合に，

$$\begin{aligned}\hat{y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 x_{1,1} + \hat{\beta}_2 x_{2,1} \\ &= -2.4245 + 1.4679 \times 80 - 0.8950 \times 30 \\ &= 88.16\end{aligned}$$

が推定される．分散は，手順 h）で示した方法で，

$$Var(\hat{y}_1) = \boldsymbol{x}_1 \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_1^T =$$

1	80	30	**3.1236**	-0.0530	0.0233	1	=	1.0795
			-0.0530	**0.0019**	-0.0020	80		
			0.0233	-0.0020	**0.0028**	30		

にて計算する．95%信頼区間は，

$$\begin{aligned}(L95\%, U95\%) &= \hat{y}_1 \pm t(0.05,17)\sqrt{Vay(\hat{y}_1)} \\ &= 80.82 \pm 2.1098\sqrt{1.0795} \\ &= (85.97, 90.35)\end{aligned}$$

表 12.10　予測プロファイルのための予測値と 95%信頼区間の計算

	切片	材料	工数	推定値	分散	95%信頼区間					
i	x_0	x_1	x_2	y^	$Var(y$^$)$	$L95\%$	$U95\%$				
1	1	**80**	30	88.16	1.0795	85.97	90.35	$\boldsymbol{\beta}$^=	-2.4245		
2	1	**80**	40	79.21	0.3139	78.03	80.39		1.4679		
3	1	**80**	50	70.26	0.1061	69.57	70.95		-0.8950		
4	1	**80**	60	61.31	0.4560	59.88	62.73				
5	1	**80**	70	52.36	1.3636	49.90	54.82	$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)$=	**3.1236**	-0.0530	0.0233
6	1	60	**50**	40.90	0.9187	38.88	42.92		-0.0530	**0.0019**	-0.0020
7	1	70	**50**	55.58	0.3236	54.38	56.78		0.0233	-0.0020	**0.0028**
8	1	80	**50**	70.26	0.1061	69.57	70.95				
9	1	90	**50**	84.94	0.2664	83.85	86.03	$Var(y_i$^$)$= $\boldsymbol{x}_i \boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)\boldsymbol{x}_i^T$			
10	1	100	**50**	99.62	0.8043	97.73	101.51				

となる．さらに，工数 x_2 を 50 hr に固定し，材料 x_1 を（60，70, . . . 100）と変化させて，生産量 $\hat{y}$ と 95%信頼区間を計算する．

表 12.10 で計算した推定値と 95%信頼区間を，Excel の散布図を用いて図 12.2 に「予測プロファイル」を作成する．表 12.10 の x_1 および x_2 を変化させると，図 12.2 の予測プロファイル」も連動して変化する．

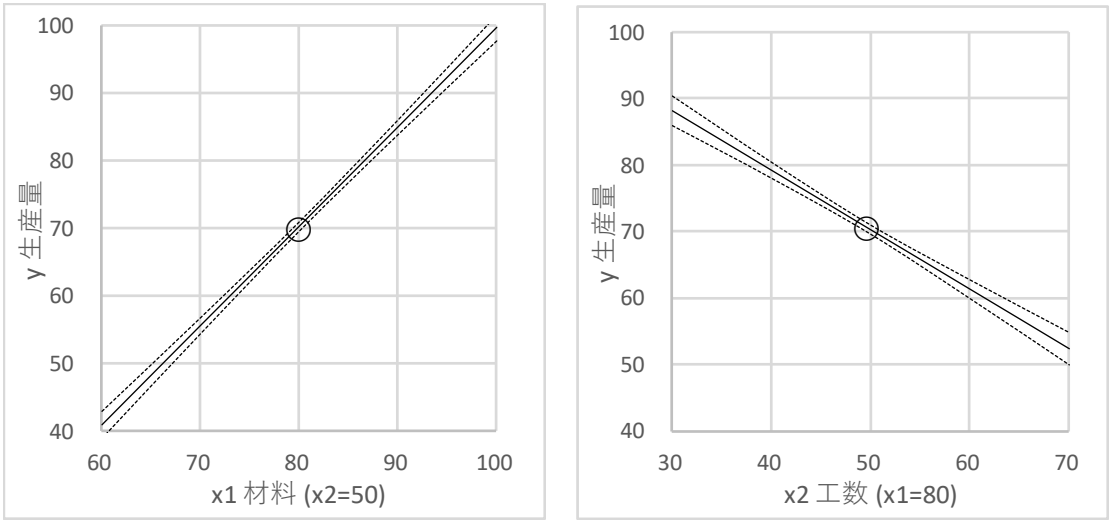


図 12.2　予測プロファイル
図左：工数 x_2 を 50 hr に固定し材料 x_1 を変化せた場合の生産量の変化
図右：材料 x_1 を 80 m^2 に固定し工数 x_2 を変化せた場合の生産量の変化

偏差平方和ベース vs. デザイン行列ベース

ドレーパー・スミス（1968）では，単回帰分析について，偏差平方和をベースにした解析を示した後に，重回帰分析への導入を意図し，同じデータを用いてデザイン行列ベースの単回帰分析を示している．Excel の分析ツールの「回帰分析」により，誰にでも手軽に重回帰分析が行えるようになったのは，素晴らししいことと思う．しかし，Excel の「回帰分析」に欠けているのは，パラメータの共分散行列の出力がないことである．単回帰分析の 95%信頼区間をグラフに示したいと思った場合には，説明変数 x について偏差平方和 S_{xx} を別途計算し，分散の公式を使って，何とかできる範囲ではある．さて，説明変数が複数ある場合には，どうしたらよいのだろうかと，途方に暮れることになる．

偏差平方和をベースした場合には，計算手順の中で，偏差平方和の逆行列を

手順 3） $\boldsymbol{S}_{xx}$ の逆行列 $\boldsymbol{S}^{xx}$ を計算する．

$$\boldsymbol{S}^{xx} = (\boldsymbol{S}_{xx})^{-1} = \begin{array}{|c|c|}\hline \mathbf{0.000954} & -0.001004 \\ \hline -0.001004 & \mathbf{0.001408} \\ \hline \end{array}$$

とあり，これに $\hat{\sigma}^2 = 1.9803$ を掛ければ，

	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$
$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) =$			
		0.0019	-0.0020
		-0.0020	**0.0028**

のように，切片を含まないパラメータについての共分散行列となる．さらに，手順 10）で示した切片の分散

$$Var(\hat{\beta}_0) = \left(\frac{1}{n} + \bar{\boldsymbol{x}} \boldsymbol{S}^{xx} \bar{\boldsymbol{x}}^T \right) \hat{\sigma}^2 = 3.1236$$

の結果を加えても，

	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$
$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) =$	**3.1236**	未	未
	未	**0.0019**	-0.0020
	未	-0.0020	**0.0028**

のように，切片と他のパラメータの共分散が欠けた共分散行列しかできない．

現実的な対応として，Excel の「回帰分析」を使い，パラメータの推定値と誤差分散 $\hat{\sigma}^2$ 求める．パラメータに関する共分散行列は，説明変数に切片を加えたデザイン行列 $\boldsymbol{X}$ を設定し，これまでも示してきた計算式

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \hat{\sigma}^2$$
$$= \text{Minverse}(\ \text{Mmult}(\text{Transpose}(\boldsymbol{X}\text{の範囲}), \boldsymbol{X}\text{の範囲}\ )\ ) * \hat{\sigma}^2$$

$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) =$	**3.1236**	-0.0530	0.0233
	-0.0530	**0.0019**	-0.0020
	0.0233	-0.0020	**0.0028**

によって，簡単に求めることができる．いずれにしても，切片を含むデザイン行列 $\boldsymbol{X}$ を使うことが，各種の推定値に 95%信頼区間を計算する際に必須の対応法である．

統計教育の現場での葛藤

統計教育に携わってきた人達から，「シグマを使うと嫌われ，ましてや行列を出すとそっぽを向かれる」を散々聞かされてきたが，Excel で行列を扱い始めると，シグマを使った計算式は，冗長で見たくも計算したくもないと思うようになるのではないだろうか．私も久々に奥野ら（1981）のシグマを用いた計算式に遭遇してめまいを感じた．救いは，サマリーとして行

列表記が適宜挿入されていて，これにより偏差平方和ベースの計算を Excel の行列計算で容易に行うことができた．この経験により，切片を含むデザイン行列ベースの回帰分析の簡潔な計算方法の優位性を認識しつつ，先人たちの苦悩を再認識した．

私は，幸いなことに切片を含むデザイン行列を用いた方法に慣れ親しんできたので，奥野ら(1981）の偏差平方和ベースによる重回帰分析の解析法には，ほとんど関心がなかった．あらためて Eecel での計算をし，計算手段が乏しい時代の先人たちの苦労を垣間見た．他方，Excel の分析ツールの回帰分析（重回帰分析）のみならず，全ての統計ソフトでの重回帰分析は，切片を含まない変数の指定を前提にしている．これは，ごく自然のことと思うのであるが，解析結果に含まれる切片について認識不足にもなる原因である．

デザイン行列ベースの重回帰の変遷

偏差平方和ベースおよびザイン行列ベースの重回帰について Excel を用いて例示してきたのであるが, 詳しくは, 無料公開されている新村(1983a,b) の「行列表現による重回帰分析 (1) および (2)」を参照されたい．この論文は，応答変数 y と説明変数 x_1 から x_4 までの 4 個の説明変数からなる 7 個の観測データを用いて行列表現による重回帰分析について詳しく解説されている．

また，偏差平方和ベースの重回帰分析について,「規準化データによる重回帰」として例示と文献の引用もあり，時代的な背景の理解に役に立つ．データの共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{x})$，推定値 $\hat{\boldsymbol{\beta}}$ の分散行列（パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ の明示的な使い分けなどについても示唆された．さらに，テコ比，スチューデント化残差の具体的な計算事例についても，Excel での計算を行う際に参考にした．

この時代の重回帰分析は，逆行列を求めるために掃き出し計算が用いられており，奥野ら（1981）にも丁寧な解説がある．本書では，Excel の Minverse() 関数を使うことを前提にし，逆行列の計算はブラック・ボックスのままにしてきた．詳しくは，新村(1983c)，「重回帰分析における掃き出し演算子」を参照のこと．なお，私にとっても掃き出し計算による逆行列の計算は, Fortran を使っていたころ慣れ親しんできたので, なつかしく思うのであるが, Excel のソルバーを含む基本の計算機能だけでは逆行列の計算は実現できなかったので，深入りはしない．

12.5.　2 次曲線の 95%信頼区間

通常の回帰分析において，回帰直線の 95%信頼区間の計算式，95%予測区間（個別データの 95%信頼区間）の計算式は，ほとんどの統計の教科書で示されていて，統計ソフトでもこれらの信頼区間のグラフ表示も標準的にサポートされている．反応が直線でなく曲線となるような場合に，2 次式あるいは 3 次式のあてはめを検討することも一般的に薦められている．

ところで，2 次式の 95%信頼区間の計算はどのようにしたらよいのだろうか．パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を使って，デザイン行列 $\boldsymbol{X}$ の行ベクトル $\boldsymbol{x}_i$ から，分散 $Var(\hat{y}_i)$ を次式

$$Var(\hat{y}_i) = \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T$$

で計算し

$$95\%CL = \hat{y}_i \pm t(0.05, df)\sqrt{Var(\hat{y}_i)}$$

によって 95%信頼区間を計算すればよい．本書でも，各種の事例で推定値の 95%信頼区間を計算しグラフ表示をしてきたのであるが，初心に戻り 2 次式の 95%信頼区間を実際に計算し，Excel で散布図上に描いてみる．

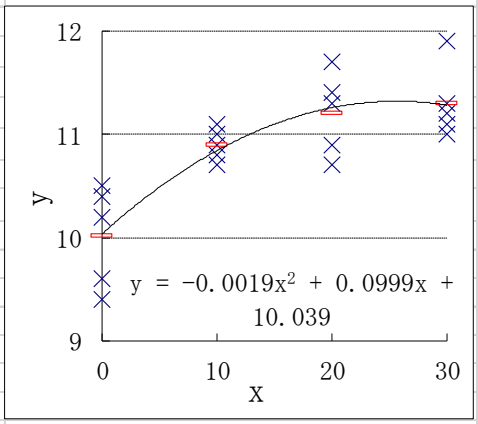
芳賀の事例

事例としては，表 12.11 に芳賀(2009)，「医薬品開発のための統計解析　第 2 部　実験計画法」の第 2.2 節「非線形の関係」で使われているデータを示す．なお，芳賀(2009）に掲載されている Excel の図表類は，https://scientist-press.com/download/（2020 年 5 月 4 日アクセス）から得られる．

表 12.11　2 次曲線となるデータ［芳賀(2009)，表示 2.2.1 再掲］

			データ				
水準	n	平均	1	2	3	4	5
0	5	10.02	10.5	9.6	10.4	10.2	9.4
10	5	10.90	10.8	10.7	11.1	10.9	11.0
20	5	11.20	11.4	10.7	10.9	11.3	11.7
30	5	11.30	11.9	11.2	11.0	11.1	11.3
全体	20	10.855					
水準数	4						

			残差				
水準	標準偏差	効果	1	2	3	4	5
0	0.49	−0.84	0.48	−0.42	0.38	0.18	−0.62
10	0.16	0.04	−0.10	−0.20	0.20	0.00	0.10
20	0.40	0.34	0.20	−0.50	−0.30	0.10	0.50
30	0.35	0.45	0.60	−0.10	−0.30	−0.20	0.00



分散分析表

要因	平方和	自由度	平均平方	F 比	p値
水準間	5.082	3	1.694	12.27	0.0002
残差	2.208	16	0.138	1.00	
全体	7.290	19			
(検算)	7.290	19			

高橋の注）
この分散分析表は，
4水準の一元配置の分散分析表であり，
2次式のあてはめた場合の分散分析表ではない．

芳賀(2009)は，2 次式の回帰曲線のパラメータを求めるために Excel の LinEst() 関数を用いており，詳しい解説がなされている．2 次式の回帰曲線に対する 95%信頼区間は，JMP の「二変量の関係」によるグラフで示しているが，その計算方法については示されていない．そこで，Excel の「分析ツール：回帰分析」を用いて回帰パラメータおよび誤差分散を計算し，パラメータの共分散行列を Excel の行列関数を用いて算出し，2 次式の回帰曲線の 95%信頼区間をグラフ化する手順を示す．

Excel による 2 次式のあてはめ

表 12.11 のデータを表 12.12 に示すようにデザイン行列の形に整え，Excel の分析ツールの回帰分析で得られた分散分析表と回帰パラメータの推定値を示す．分散分析表の「残差」の行の分散の列が誤差分散 $\hat{\sigma}^2 = 0.1320$ となる．得られた列ベクトルの回帰パラメータの推定値を

$$\hat{\boldsymbol{\beta}} = [\,10.0390 \quad 0.0999 \quad -0.0020\,]^T$$

のように行ベクトルに転置記号「T」を付けて示す．デザイン行列 $\boldsymbol{X}$ の転置行列 $\boldsymbol{X}^T$ を Transpose() 関数で求め，$\boldsymbol{X}^T$ と $\boldsymbol{X}$ の積を Mmult()関数で計算する．

表 12.12　2 次式のあてはめ

	デザイン行列 $\boldsymbol{X}$					Excel 分析ツール 回帰分析			
i	切片	x	x^2	y		分散分析表			
1	1	0	0	10.5			自由度	変動	分散
2	1	0	0	9.6		回帰	2	5.0454	2.5227
3	1	0	0	10.4		残差	17	2.2441	**0.1320**
4	1	0	0	10.2		合計	19	7.2895	
5	1	0	0	9.4			係数	標準誤差	
6	1	10	100	10.8		切片	**10.0390**	0.1584	
7	1	10	100	10.7		x	**0.0999**	0.0254	
8	1	10	100	11.1		x2	**-0.0020**	0.0008	
9	1	10	100	10.9					
10	1	10	100	11.0		行列計算			
11	1	20	400	11.4		$\boldsymbol{X}^T\boldsymbol{X}=$	**20**	300	7000
12	1	20	400	10.7			300	**7000**	180000
13	1	20	400	10.9			7000	180000	**4900000**
14	1	20	400	11.3					
15	1	20	400	11.7		$(\boldsymbol{X}^T\boldsymbol{X})^{-1}=$	**0.1900**	-0.0210	0.0005
16	1	30	900	11.9			-0.0210	**0.0049**	-0.0002
17	1	30	900	11.2			0.0005	-0.0002	**5.00E-06**
18	1	30	900	11.0					
19	1	30	900	11.1		$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma\text{^}2=$	**2.508E-02**	-2.772E-03	6.600E-05
20	1	30	900	11.3		$\boldsymbol{\Sigma}(\boldsymbol{\beta}\text{^})$	-2.772E-03	**6.468E-04**	-1.980E-05
							6.600E-05	-1.980E-05	**6.600E-07**

							X						
	X^T						切片	x	x^2		X^TX		
$X^TX=$	1	1	1	・・・	1		1	0	0	=	**20**	300	7000
	0	0	0		30		1	0	0		300	**7000**	180000
	0	0	0		900		1	0	0		7000	180000	**4900000**
							:						
							1	30	900				
					=Mmult(Transpose(Xの範囲), Xの範囲)								

X^TX の逆行列を Minverse() 関数で求め，

		$(X^TX)^{-1}$		
$(X^TX)^{-1}$ =		**0.1900**	-0.0210	0.0005
		-0.0210	**0.0049**	-0.0002
		0.0005	-0.0002	**5.00E-06**
		= Minverse(X^TXの範囲)		

Excel の分散分析表の「分散」列の「残差」の行の誤差分散 $\hat{\sigma}^2 = 0.1320$ を掛けて，パラメータの共分散行列 $\Sigma(\hat{\beta})$

	$(X^TX)^{-1}$			$\sigma\hat{}^2$		$\Sigma(\beta\hat{})$		
$(X^TX)^{-1}\sigma\hat{}^2=$	**0.1900**	-0.0210	0.0005	0.1320	=	**2.508E-02**	-2.772E-03	6.600E-05
	-0.0210	**0.0049**	-0.0002			-2.772E-03	**6.468E-04**	-1.980E-05
	0.0005	-0.0002	**5.00E-06**			6.600E-05	-1.980E-05	**6.600E-07**

を計算する．パラメータの共分散行列 $\Sigma(\hat{\beta})$ の対角要素の平方根が，回帰パラメータの標準誤差となっている．

$Ver(\beta\hat{})=$	0.025081118	$SE(\beta\hat{})=$	0.1584
	0.000646829		0.0254
	0.000000660		0.0008

段階を追って計算式を示したが，

$$\Sigma(\hat{\beta}) = \text{Minverse}(\ \text{Mmult}(\text{Transpose}(X\text{の範囲}), X\text{の範囲})\ ) * \hat{\sigma}^2$$

のようにまとめて計算することができる．

推定値の 95%信頼区間

Excel で推定された回帰パラメータを用いて推定値 $\hat{y} = X\hat{\beta}$ を計算する．デザイン行列 X の i 行目のベクトル x_i を用いて $\hat{y}_i = x_i\hat{\beta}$ として計算し，フィルハンドルを用いて計算式をコピーする方法でも良い．

	X'			$\beta\hat{}$		$y\hat{}$
$X'\beta\hat{}=$	1	0	0	10.0390	=	10.04
	1	10	100	0.0999		10.84
	1	20	400	-0.0020		11.26
	1	30	900			11.28
	=Mmult(Xの範囲，$\beta\hat{}$の範囲)					

表 12.13 に示すように，推定値 $\hat{y}_i$ の分散 $Var(\hat{y}_i)$ は， $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を挟む $\boldsymbol{x}_i$ の 2 次形式 $\boldsymbol{x}_i\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i^T$ によって計算することができる．

$Var(\hat{y}''_5) = \boldsymbol{x}''_5\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}''^T_5 =$

$\boldsymbol{x}''_5$			$\boldsymbol{\Sigma}(\boldsymbol{\beta}\text{^})$			$\boldsymbol{x}''^T_5$		$Var(y''_5\text{^})$
1	10	100	**2.51E-02**	-2.77E-03	6.60E-05	1	=	0.0145
			-2.77E-03	**6.47E-04**	-1.98E-05	10		
			6.60E-05	-1.98E-05	**6.60E-07**	100		

推定値の $Var(\hat{y}_i)$ 分散の平方根を用いて，95%信頼区間および 95%予測区間を

$$\text{信頼区間}= \hat{y}''_5 \pm t(0.05,17)\sqrt{Var(\hat{y}''_5)} = (10.5888, 11.0972)$$

$$\text{予測区間}= \hat{y}''_5 \pm t(0.05,17)\sqrt{Var(\hat{y}''_5) + \hat{\sigma}^2} = (10.0354, 11.6506)$$

として計算することができる．回帰曲線の滑かな線グラフにするために，x の範囲を広げ（–10, –5, . . . , 40）について，推定値 $\hat{y}$，95%信頼区間，95%予測区間を計算する．

表 12.13　散布図に上書きする 2 次曲線と信頼区間の計算シート

			—— $\boldsymbol{X}''$ ——					信頼区間		予測区間	
i	x	y	切片	x''	x''^2	y''^	$Var(y''\text{^})$	L95%	U95%	L95%	U95%
1	0	10.5	1	-10	100	8.85	0.2046	7.8907	9.7993	7.6209	10.0691
2	0	9.6	1	-5	25	9.49	0.0776	8.9029	10.0786	8.5247	10.4568
3	0	10.4	1	0	0	10.04	0.0251	9.7049	10.3731	9.2028	10.8752
4	0	10.2	1	5	25	10.49	0.0123	10.2558	10.7237	9.6883	11.2912
5	0	9.4	1	10	100	10.84	**0.0145**	**10.5888**	**11.0972**	**10.0354**	**11.6506**
6	10	10.8	1	15	225	11.10	0.0169	10.8244	11.3731	10.2846	11.9129
7	10	10.7	1	20	400	11.26	0.0145	11.0028	11.5112	10.4494	12.0646
8	10	11.1	1	25	625	11.32	0.0123	11.0838	11.5517	10.5163	12.1192
9	10	10.9	1	30	900	11.28	0.0251	10.9469	11.6151	10.4448	12.1172
10	10	11.0	1	35	1225	11.15	0.0776	10.5589	11.7346	10.1807	12.1128
11	20	11.4	1	40	1600	10.92	0.2046	9.9607	11.8693	9.6909	12.1391
12	20	10.7									
13	20	10.9				$\boldsymbol{\beta}$^		$\boldsymbol{\Sigma}(\boldsymbol{\beta}\text{^}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$			
14	20	11.3			切片	**10.0390**		**2.51E-02**	-2.77E-03	6.60E-05	
15	20	11.7			x	**0.0999**		-2.77E-03	**6.47E-04**	-1.98E-05	
16	30	11.9			x^2	**-0.0020**		6.60E-05	-1.98E-05	**6.60E-07**	
17	30	11.2									
18	30	11.0					σ^2		$t(0.05, 17)$		
19	30	11.1					**0.1320**		**2.1098**		
20	30	11.3									

Excel の「散布図」による 2 次の回帰曲線のグラフを図 12.3 に示す．最初に x と y の散布図を描き，その上に「データの選択」機能を用いて，（$\hat{y}$，信頼区間の下限・上限，予測区間下限・上限）を上書きし，「データ行列の書式設定」機能を用いて整える．

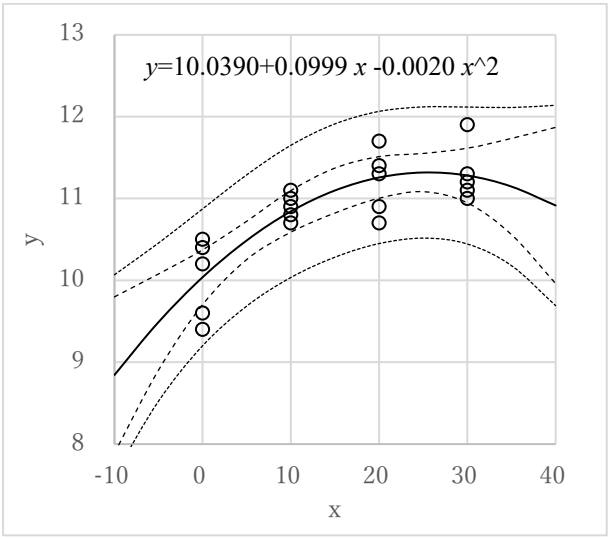


図 12.3　Excel の散布図を用いた 2 次曲線の 95%信頼区間および 95%予測区間

JMP の「二変量の関係」による 2 次曲線のあてはめ

JMP の「二変量の関係」には，多項式に対する 95%信頼区間を上書きする機能があるので，結果を示す．芳賀（2009）には，この JMP で作成した図が示されているが，計算方法は示されていない．

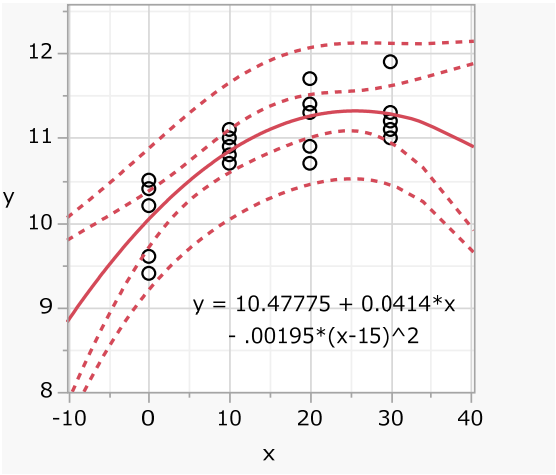
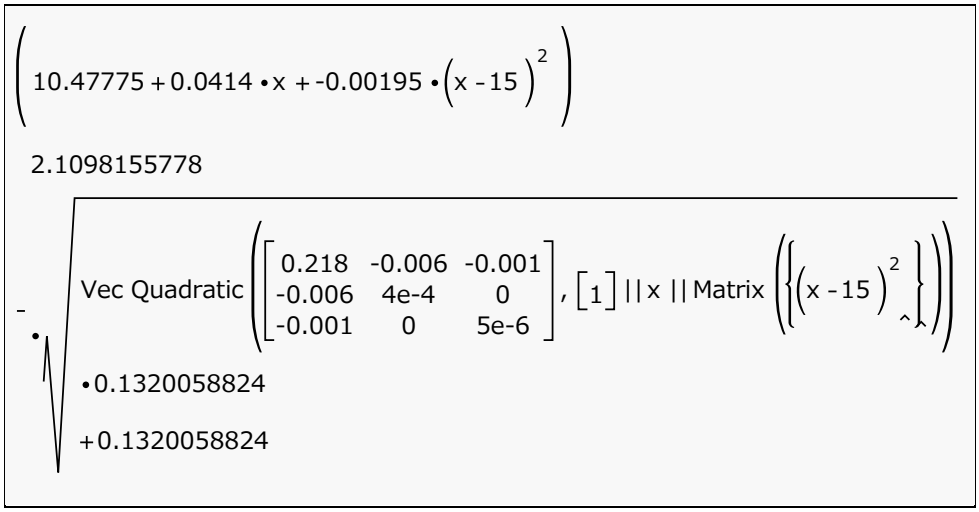


図 12.4　JMP による 2 次曲線の 95%信頼区間および 95%予測区間

JMP には，95%信頼区間および個別データの 95%信頼区間の計算式を JMP ファイルに出力する機能があり，これを用いて JMP の内部でどのような計算をしているかを可視化することができる．表 12.14 に示すように，推定値 $\hat{y}$ の計算が最初に示されているが，2 次の項は，$(x-15)^2$ となっており，x から平均値 15 を引くことにより，2 乗の桁数が増えないように「多項式の中心化」が行われている．次に 95%信頼区間の計算のために自由度が 17 の両側 5%の t 値が 2.1098 と示され，標準誤差の計算のために分散の平方根が現われ，その中で 2 次形式

表 12.14　JMP で生成された個別データの下側 95%信頼区間雄の計算式



の VecQuadratic()関数の第 1 引数とし積和行列 $(\boldsymbol{X}-15)^T(\boldsymbol{X}-15)$ の逆行列，第 2 引数として [1 x　$(x\text{-}15)^2$]が示されている．これに誤差分散 $\hat{\sigma}^2=0.1320$ を掛け推定値の回帰の分散 $Var(\hat{y})$ を求めている．さらに，誤差分散 $\hat{\sigma}^2=0.1320$ を加え，個別データの分散としている．

$$\sqrt{Var(個別\hat{y})}=\sqrt{\text{VecQuadratic}\{[(\boldsymbol{X}-15)^T(\boldsymbol{X}-15)]^{-1},[1\ \ x\ \ (x-15)^2]\}\hat{\sigma}^2+\hat{\sigma}^2}$$

なお，$(x-15)^2$ と中心化した場合の $[(\boldsymbol{X}-15)^T(\boldsymbol{X}-15)]^{-1}$ を Excel で別途計算すると，

$[(\boldsymbol{X}-15)^T(\boldsymbol{X}-15)]^{-1}=$	**0.218125**	-0.006000	-0.000625
	-0.006000	**0.000400**	0.000000
	-0.000625	0.000000	**0.000005**

が得られ，JMP の計算式に一致することが確認される．このように，統計ソフトの内部では，行列計算が行われている．通常，ユーザが目にすることはできないが，JMP では，表 12.14 にその一部に示すように可視化することができるようになっており，統計ソフトを通じての学習効果が期待される．

2 次式の 95%信頼区間の計算式を行列表記でなくシグマで表わすことも，分散共分散の要素を使って表わすこともできるが，冗長であり示すこともためらわれ，さらに，それらの式を使った計算を例示することもためらわれる．パラメータの共分散行列を使った 2 次形式で示したら読者からそっぽを向かれるにちがいない．このような事情により，2 次式の 95%信頼区間の計算方法が，ブラック・ボックス化している要因となっていると思われる．いずれにしても，容易かつ可視化に優れる Excel の行列計算が，読者のブレイクスルーのための立役者となることを期待したい．

「自然科学の統計学」での事例

東京大学教養学部統計学教室編（1992),「基礎統計学 III 自然科学の統計学」の第 2 章の表 2.3 に「2 次多項式のデータ：液体のある成分と曇り点の関係」が示されている．このデータは，ある溶液の成分（I-8）の比率とその溶液の曇り点（透明な溶液が温度の変化によって曇りを生じさせる温度）の関係である．元のデータを x の小さい順に並び替え，切片および 2 乗項を付け加え，Excel の「回帰分析」の結果を加えたものを表 12.15 に示す．さらに，分散分析表の残差の分散から $\hat{\sigma}^2 = 0.155412$ とし，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$ を Excel の行列関数で計算した結果を加えてある．

表 12.15　2 次多項式のデータ：液体のある成分と曇り点の関係

	——$\boldsymbol{X}$——			曇り点
No.	切片	x	x^2	y
1	1	0	0	21.9
2	1	0	0	22.1
3	1	0	0	22.8
4	1	1	1	24.5
5	1	2	4	26.0
6	1	2	4	26.1
7	1	3	9	26.8
8	1	3	9	27.3
9	1	4	16	28.2
10	1	4	16	28.5
11	1	5	25	28.9
12	1	6	36	29.8
13	1	6	36	30.0
14	1	6	36	30.3
15	1	7	49	30.4
16	1	8	64	31.4
17	1	8	64	31.5
18	1	9	81	31.8
19	1	10	100	33.1

分散分析表			
	自由度	変動	分散
回帰	3	15064.41	5021.47
残差	16	2.486599	**0.155412**
合計	19	15066.90	

	係数	標準誤差	t
切片	**22.5612**	0.1984	113.6984
x	**1.6680**	0.0990	16.8568
x2	**-0.0680**	0.0103	-6.5911

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

0.253356	-0.097147	0.007903
-0.097147	**0.063004**	-0.006266
0.007903	-0.006266	**0.000684**

$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$

0.039375	-0.015098	0.001228
-0.015098	**0.009792**	-0.000974
0.001228	-0.000974	**0.000106**

「自然科学の統計学」では，デザイン行列 $\boldsymbol{X}$ を用いて $\boldsymbol{X}^T\boldsymbol{X}$，$\boldsymbol{X}^T\boldsymbol{Y}$，$(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ の計算過程が示され，2 次式のパラメータの推定値 $\hat{\boldsymbol{\beta}}$ も $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ の計算結果として示されている．残念なのは，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ についての記載がないことである．そのためであろうか，2 次曲線の推定値の分散 $Var(\hat{y})$ についても例示がない．もちろん，一般論としてのパラメータの共分散行列については丁寧な説明があるが，それを用いた数値例は示されていない．

2 次式の推定値に対する分散（4 次式）を実際に求めるためには，行列計算なしには計算事例として示し難いことは確かである．他方，2 変数以上の回帰分析における推定値の分散は，直観的な意味付けができにくいのであるが，2 次式を含めて多項式回帰の推定値の分散を計

算し，その 95%信頼区間を求め，グラフ化することとの意義は明確であり，また入門的でもある．表 12.15 の結果を用いて，表 12.16 に示すように 2 次曲線の 95%信頼区間および個別データの 95%信頼区間の計算を行い，その結果を図 12.5 に示す．

表 12.16　2 次多項式の 95%信頼区間および予測区間

	I-8(%)	曇り点		——X'——					信頼区間		予測区間	
No.	x	y		切片	x'	x'^2	y'^	$Var(y'$^$)$	L95%	U95%	L95%	U95%
1	0	21.9		**1**	**-1**	**1**	**20.83**	**0.0839**	**20.21**	**21.44**	**19.79**	**21.86**
2	0	22.1		1	0	0	22.56	0.0394	22.14	22.98	21.63	23.50
3	0	22.8		1	1	1	24.16	0.0196	23.86	24.46	23.27	25.05
4	1	24.5		1	2	4	25.63	0.0141	25.37	25.88	24.75	26.50
5	2	26.0		1	3	9	26.95	0.0150	26.69	27.21	26.08	27.83
6	2	26.1		1	4	16	28.15	0.0171	27.87	28.42	27.27	29.03
7	3	26.8		1	5	25	29.20	0.0176	28.92	29.48	28.32	30.08
8	3	27.3		1	6	36	30.12	0.0162	29.85	30.39	29.24	31.00
9	4	28.2		1	7	49	30.91	0.0153	30.64	31.17	30.03	31.78
10	4	28.5		1	8	64	31.56	0.0199	31.26	31.85	30.67	32.44
11	5	28.9		1	9	81	32.07	0.0373	31.66	32.48	31.14	33.00
12	6	29.8		1	10	100	32.45	0.0775	31.86	33.04	31.42	33.47
13	6	30.0		1	11	121	32.69	0.1532	31.86	33.52	31.51	33.86
14	6	30.3										
15	7	30.4				$\hat{\boldsymbol{\beta}}$			$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma$^^{2}			
16	8	31.4			切片	**22.5612**	β_0		**0.0394**	-0.0151	0.0012	
17	8	31.5			x	**1.6680**	β_1		-0.0151	**0.0098**	-0.0010	
18	9	31.8			x^2	**-0.0680**	β_2		0.0012	-0.0010	**0.0001**	
19	10	33.1			$t(0.05,16)$=	**2.1199**	σ^^{2}=		**0.1554**			

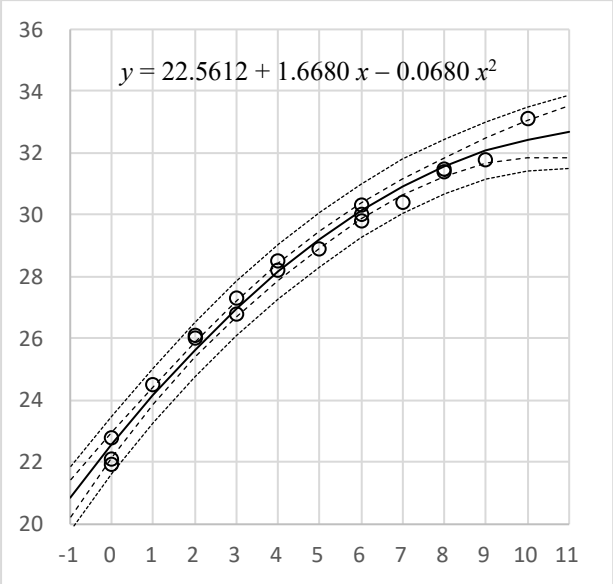


図 12.5　Excel の散布図を用いた 2 次曲線の推定曲線と 95%信頼区間および予測区間

「自然科学の統計学」の 38 ページに行列計算の結果が示されている．Excel の行列計算の計算式および結果を示すので，実際に手を動かして計算することが，各種の統計モデルに対する 95%信頼区間を求めるための練習となる．

X^TX			β		X^TY
19.0	84.0	550.0	β_0	=	531.4
84.0	**550.0**	4068.0	β_1		2536.1
550.0	4068.0	**32374.0**	β_2		16994.1

X^TX = Mmult(Transpose(Xの範囲), Xの範囲)

X^TY = Mmult(Transpose(Xの範囲), Yの範囲)

$(X^TX)^{-1}$		
0.253356	-0.097147	0.007903
-0.097147	**0.063004**	-0.006266
0.007903	-0.006266	**0.000684**

$(X^TX)^{-1}$ = Minverse(X^TXの範囲)

β^	$(X^TX)^{-1}$			X^TY
22.5612 =	**0.253356**	-0.097147	0.007903	531.4
1.6680	-0.097147	**0.063004**	-0.006266	2536.1
-0.0680	0.007903	-0.006266	**0.000684**	16994.1

$\hat{\beta}$ = Mmult((X^TX)$^{-1}$の範囲, X^TYの範囲)

σ^2
0.155412

$\hat{\sigma}^2$ = SumSq(Yの範囲 − Mmult(Xの範囲, $\hat{\beta}$の範囲)) / (19 − 3)

—— $\hat{y}$ の計算——

$\Sigma(\beta$^$)=(X^TX)^{-1}\sigma$^2		
0.039375	-0.015098	0.001228
-0.015098	**0.009792**	-0.000974
0.001228	-0.000974	**0.000106**

$\Sigma(\hat{\beta})$ = ((X^TX)$^{-1}$の範囲) * $\hat{\sigma}^2$

$Var(y'$^$)$		切片	x'	x'^2	$\Sigma(\beta$^$)=(X^TX)^{-1}\sigma$^2			$x'_1{}^T$
0.0839	=	1	-1	1	**0.039375**	-0.015098	0.001228	1
					-0.015098	**0.009792**	-0.000974	-1
					0.001228	-0.000974	**0.000106**	1

$Var(\hat{y}'_1)$ = Mmult(Mmult(x'_1の範囲, $\Sigma(\hat{\beta})$の範囲), Transpose(x'_1の範囲))

以下 略 （フィルハンドルでコピー）

95%信頼区間 = $\hat{y}'_1 \pm t(0.05,16)\sqrt{Var(\hat{y}'_1)}$ = **20.21 21.44**

95%予測区間 = $\hat{y}'_1 \pm t(0.05,16)\sqrt{Var(\hat{y}'_1)+\hat{\sigma}^2}$ = **19.79 21.86**

以下 略 （フィルハンドルでコピー）

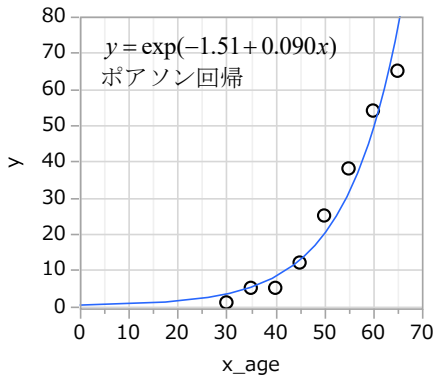
12.6.　対数リンクでのポアソン回帰の 95%信頼区間

第 1.5 節でオフセットを考慮した「冠動脈心疾患の死亡者数」の事例に対し，第 2.6 節では，オフセットを考慮した対数リンクのポアソン回帰について Excel によるニュートン・ラフソン法による最尤法を例示した．第 5.2 節では，同じ事例について反復重み付き回帰によるオフセット無しで対数リンクによるポアソン回帰を適用し，95%信頼区間を含むグラフを示し，さらに 2 次式によるポアソン回帰も例示した．どちらも Excel による例示となっている．

ここでは，オフセット無しの対数リンクに対するポアソン回帰として取り上げる．回帰パラメータは，統計ソフトでの計算結果を用いることを想定し，Excel でパラメータの共分散行列を計算し，95%信頼区間および 95%予測区間のグラフ表示の方法について示す．表 12.17 に示すように第 5.3 節で取り上げた冠動脈心疾患の死亡者数［ドブソン(2008)］を，オフセットなしの対数リンクの事例として取り上げる．

表 12.17　オーストラリアのある地方の冠動脈心疾患の死亡者数（表 5.8 再掲）

	年齢層	死亡者数
No.	x	y
1	30	1
2	35	5
3	40	5
4	45	12
5	50	25
6	55	38
7	60	54
8	65	65



第 5.2 節では，対数リンクに対する反復重み付き回帰の入門とし，ポアソン回帰曲線の 95%信頼区間の求め方についても示したのであるが，ここでは，ポアソン回帰のパラメータがすでに得られていることを前提にし，事後的に 95%信頼区間を求める方法に焦点をあてる．

表 12.18 に示すように，ポアソン回帰係数 $\hat{\beta}_0 = -1.5078$，$\hat{\beta}_1 = 0.0899$ がパラメータの推定値として得られているとする．指数曲線の推定値 $\hat{y}_1$ は，

$$\begin{aligned}\hat{y}_1 &= \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1) \\ &= \exp(-1.5078 + 0.0899 x_1) \\ &= 3.2831\end{aligned}$$

として，計算されている．

表 12.18　冠動脈心疾患の死亡者数（対数リンク）

	デザイン行列		死亡	回帰	対数		対数	対数		
	$\boldsymbol{X}$		者数	推定値	尤度	重み	推定値	分散	95%信頼区間	
i	$\boldsymbol{X}_0$	$\boldsymbol{X}_1$	$\boldsymbol{y}$	$\boldsymbol{y}$^	$\ln L_i$	$\boldsymbol{w}$^=$\boldsymbol{y}$^	$\ln \boldsymbol{y}$^	$Var(\ln \boldsymbol{y}$^)	L95%	U95%
1	1	30	1	3.2830	-2.0943	3.2830	1.1888	**0.0532**	2.0887	5.1603
2	1	35	5	5.1458	-1.7424	5.1458	1.6382	0.0372	3.5268	7.5081
3	1	40	5	8.0656	-2.4151	8.0656	2.0876	0.0243	5.9418	10.9486
4	1	45	12	12.6422	-2.1849	12.6422	2.5370	0.0147	9.9675	16.0346
5	1	50	25	19.8154	-3.1575	19.8154	2.9865	0.0083	16.5665	23.7014
6	1	55	38	31.0589	-3.4635	31.0589	3.4359	0.0052	26.9557	35.7866
7	1	60	54	48.6819	-3.1954	48.6819	3.8853	0.0053	42.1830	56.1820
8	1	65	65	76.3044	-3.8895	76.3044	4.3347	0.0087	63.5534	91.6138
	1	70		119.6002			4.7842	0.0153	93.8501	152.415
			β_0^=	**-1.5078**	**-22.1425**	$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^)=	**0.2178**	-0.0037		
			β_1^=	**0.0899**	$\ln L$		-0.0037	**0.0001**		
							$[(\boldsymbol{X}*\boldsymbol{w}$^$)^T\boldsymbol{X}]^{-1}$			

このパラメータの推定値は，Excel の関数 Poisson.dist() 関数で対数尤度 $\ln L$ を Excel のソルバーで最大化して求めたものであるが，後の手順では，パラメータの推定値のみを使う．

$$\ln L = \sum_i \ln(\text{Poisson.dist}(y_i\ \hat{y}_i, false)) = -22.1425$$

第 5 章の反復重み付き回帰で示したように，対数リンクの場合にパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を求めるためには，重み $\hat{w}_i = \hat{y}_i$ を対角要素とした行列 $\boldsymbol{W}$ としたデザイン行列の積和行列の逆数を計算して得られる．

1	1	1	1	1	1	1	1
30	35	40	45	50	55	60	65
			$\boldsymbol{X}^T$				

3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	**5.1**	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	**8.1**	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	**12.6**	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	**19.8**	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	**31.1**	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	**48.7**	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	**76.3**
			$\boldsymbol{W}$				

1	30
1	35
1	40
1	45
1	50
1	55
1	60
1	65
$\boldsymbol{X}$	

=

205.0	11750.1
11750.1	**688912.4**
$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$	

0.2178	-0.0037
-0.0037	**0.0001**
$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)=(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	

簡便的には，表 12.18 に示したように $[(\boldsymbol{X}*\hat{\boldsymbol{w}})^T\boldsymbol{X}]^{-1}$ として計算することができる．実際に確認して見ると，次に示すように同じ結果が得られる．このような技巧的な計算は Excel の行列関数にベクトルを対角化する関数，逆に行列の対角要素をベクトル化する関数がないためである．

1	30	*	3.2831	=	3.28	98.49
1	35		5.1460		5.15	180.11
1	40		8.0658		8.07	322.63
1	45		12.6425		12.64	568.91
1	50		19.8159		19.82	990.79
1	55		31.0596		31.06	1708.28
1	60		48.6830		48.68	2920.98
1	65		76.3062		76.31	4959.90
X			***w***		***X*w***	

3.28	5.15	8.07	12.64	19.82	31.06	48.68	76.31
98.49	180.11	322.63	568.91	990.79	1708.28	2920.98	4959.90
			$(\boldsymbol{X}*\boldsymbol{w})^T$				

1	30	=	**205.0** 11750.1
1	35		11750.1 **688912.4**
1	40		$(\boldsymbol{X}*\boldsymbol{w})^T\boldsymbol{X}$
1	45		
1	50		**0.2178** -0.0037
1	55		-0.0037 **0.0001**
1	60		$[(\boldsymbol{X}*\boldsymbol{w})^T\boldsymbol{X}]^{-1}$
1	65		$\boldsymbol{\Sigma}(\boldsymbol{\beta}\text{^})$
X			

パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いて，回帰直線の 95%信頼区間を求める．デザイン行列 $\boldsymbol{X}$ それぞれの i 行目ごとに，$\boldsymbol{x}_i = [x_{0,i} \quad x_{1,i}]$ としたとき $\hat{y}_i$ の分散は，次の 2 次形式で求められるので，

$$Var(\ln \hat{y}_i) = \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T$$

$\hat{y}_i$ の 95%信頼区間は，

$$(L95\%, U95\%) = \exp\left[\ln \hat{y}_i \pm 1.96\sqrt{Var(\ln \hat{y}_i)} \right]$$

で求められる．元のスケールは，対数の 95%信頼区間について指数を取って計算したものである．

共分散行列を用いた計算の実例を，$i=1$ の場合について示す．まず，対数についての $Var(\hat{y}_i)$ は，2 次形式の計算法で，0.0532 が得られる．

$Var(y_1\text{^})$=	1	30	**0.2178**	-0.0037	1	=	**0.0532**
	$\boldsymbol{x}_{1,i}$		-0.0037	**0.0001**	30		
			$\boldsymbol{\Sigma}(\boldsymbol{\beta}\text{^}) = [(\boldsymbol{X}*\boldsymbol{w}\text{^})^T\boldsymbol{X}]^{-1}$		$\boldsymbol{x}_{1,i}^T$		

信頼区間の計算は，

$$\ln \hat{y}_1 \pm 1.96\sqrt{Var(\ln \hat{y}_1)} = 1.1888 \pm 1.96\sqrt{0.0532} = (0.7365, 1.6408)$$

で得られる．指数を取って元のスケールでは，

$$\hat{y}_1 = \exp(\ln \hat{y}_1) = \exp(1.1888) = 3.283$$

$$L95\% = \exp(0.7365) = 2.089$$

$$U95\% = \exp(1.6408) = 5.160$$

となる．これらの計算結果より，図 12.6 に 95%信頼区間を示す．

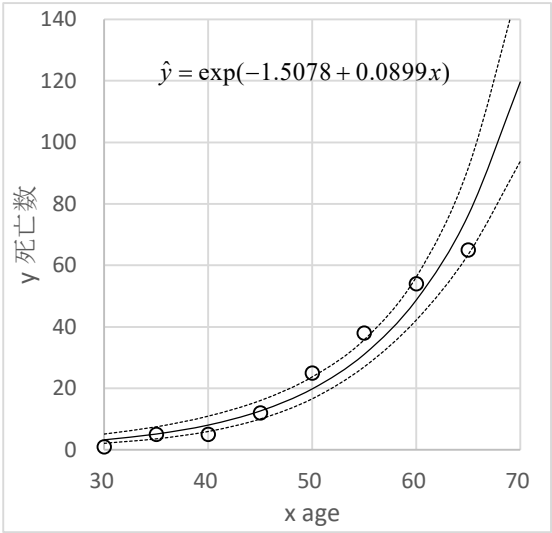


図 12.6　死亡数の 95%信頼区間

大変な計算と思われるかもしれなので，段階的な計算手順を示す．

手順 1）表 12.17 に示す冠動脈疾患の死亡者データに対し，使い慣れた統計ソフトで対数リンクのポアソン回帰を行い，回帰パラメータ $\hat{\beta}_0 = -1.5078$，$\hat{\beta}_1 = 0.0899$ を得る．

手順 2）推定値 $\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_0 x_i$ を計算し，その指数を計算し，表 12.17 右に示す散布図に指数の線グラフを上書きする．Excel でも統計ソフトでも，このグラフを書けることが必須である．

手順 3）統計ソフトでパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ が得られない場合は，表 12.19 に示すように Excel で計算する．

共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta}$^$)$	
0.2178	-0.0037
-0.0037	**0.0001**

手順 4）推定値 $\ln \hat{y}_1$ の分散 $Var(\ln \hat{y}_1)$ を計算し，計算式をコピーする．

手順 5）95%信頼区間を計算する．

手順 6）95%信頼区間を散布図に上書きし，形式を整える．

表 12.19　対数リンクでの 95%信頼区間の計算

デザイン		回帰		推定値	重み=推定値			分散	95%信頼区間	
行列		パラメータ		$\ln \boldsymbol{y}$^	$\boldsymbol{y}$^=$\boldsymbol{w}$^	共分散行列 $\boldsymbol{\Sigma}(\boldsymbol{\beta}$^)		$Var(\ln \boldsymbol{y}$^)	L95%	U95%
1	30	**-1.5078**	=	1.1888	3.2831	**0.2178**	-0.0037	0.0532	2.0888	5.1604
1	35	**0.0899**		1.6382	5.1460	-0.0037	**0.0001**	0.0372	3.5269	7.5082
1	40			2.0876	8.0658			0.0243	5.9420	10.9489
1	45			2.5371	12.6425			0.0147	9.9677	16.0349
1	50			2.9865	19.8159			0.0083	16.5670	23.7019
1	55			3.4359	31.0596			0.0052	26.9563	35.7874
1	60			3.8853	48.6830			0.0053	42.1841	56.1832
1	65			4.3348	76.3062			0.0087	63.5550	91.6157
$\boldsymbol{X}$		$\boldsymbol{\beta}$		=$\boldsymbol{X\beta}$	=exp($\ln \boldsymbol{y}$^)	$\boldsymbol{\Sigma}(\boldsymbol{\beta}$^)=		=$\boldsymbol{x}_i \boldsymbol{\Sigma} \boldsymbol{x}_i^T$	=exp($\ln \boldsymbol{y}$^±1.96*	
						$[(\boldsymbol{X}^* \boldsymbol{w}$^$)^T \boldsymbol{X}]^{-1}$			sqrt($Var(\ln \boldsymbol{y}$^))	

Excel での計算は以下の通り．

$$\ln(\hat{\boldsymbol{y}}) = \text{Mmult}(\boldsymbol{X}\text{の範囲}, \hat{\boldsymbol{\beta}}\text{の範囲})$$

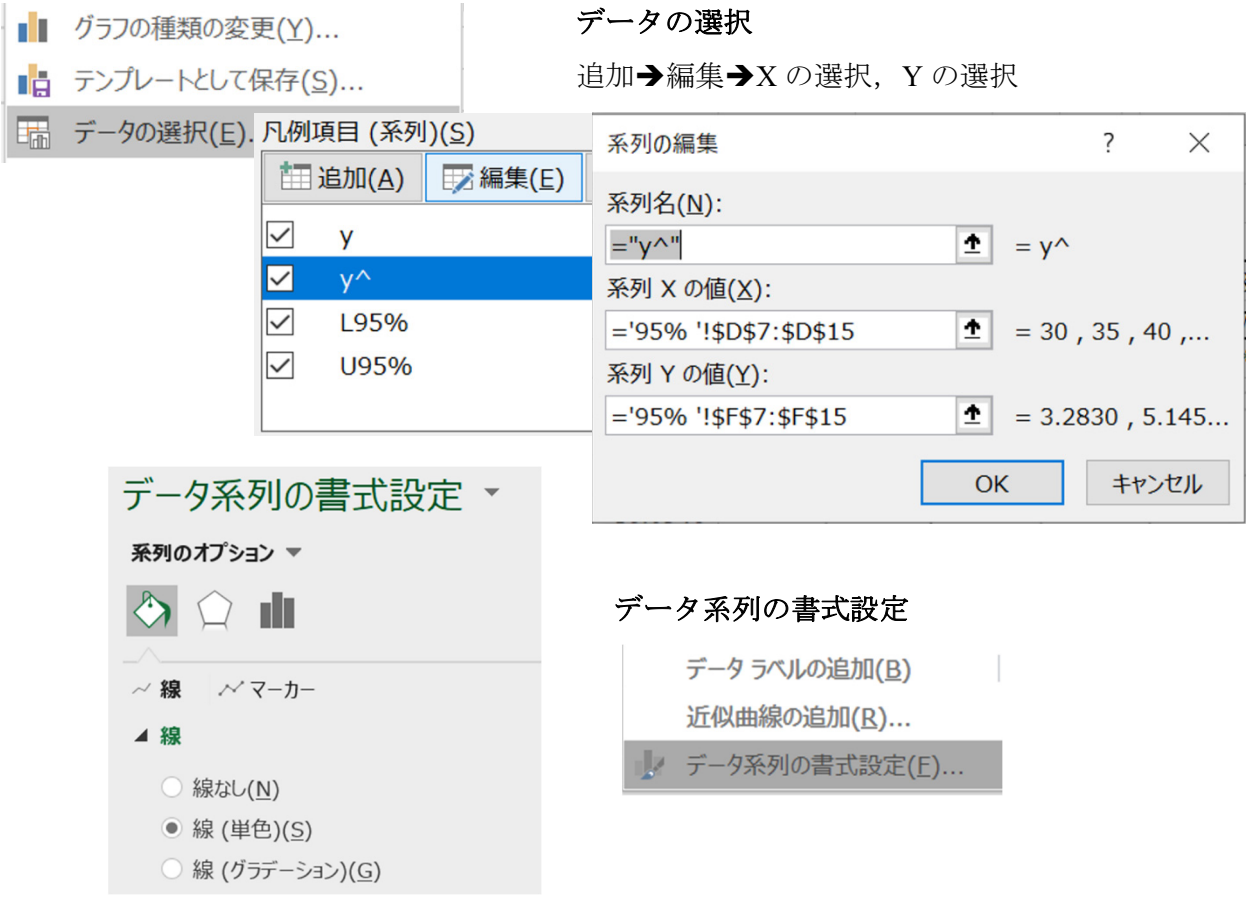
$$\hat{\boldsymbol{y}} = \exp(\ln(\hat{\boldsymbol{y}})\text{の範囲})$$

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = \text{Minverse}(\text{Mmult}(\text{Transpose}(\boldsymbol{X}\text{の範囲} * \hat{\boldsymbol{w}}\text{の範囲}), \boldsymbol{X}\text{の範囲}))$$

$$Var(\ln \hat{y}_i) = \text{Mmult}(\text{Mmult}(\boldsymbol{x}_i\text{の範囲}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\text{の範囲}), \text{Transpose}(\boldsymbol{x}_i))$$

$$(L95\%, U95\%) = \exp(\ln \hat{y}_i \pm 1.96 * \text{sqrt}(Var(\ln \hat{y}_i)))$$

Excel の散布図の活用のヒント



データの選択

追加➔編集➔X の選択，Y の選択

データ系列の書式設定

12.7.　オフセットを含むポアソン回帰の各種の 95%信頼区間

第 3.6 節では，喫煙習慣による年齢階層ごとの冠動脈疾患による死亡データについて，各種の統計モデルをあてはめ，尤度比検定を用いて探索的な解析の手順を示した．そのために，パラメータの共分散行列を用いたが，95%信頼区間については示さなかった．データを表 12.20 に再掲する［ドブソン(2008)］.

表 12.20　年齢階層毎の喫煙習慣による冠動脈心疾患による死亡数（表 3.32 再掲）

年齢		非喫煙者（$x_{smoke}=0$）			喫煙者（$x_{smoke}=1$）		
範囲	歳	死亡	人年	10万人比	死亡	人年	10万人比
35-44	40	2	18,790	10.6	32	52,407	61.1
45-54	50	12	10,673	112.4	104	43,248	240.5
55-64	60	28	5,710	490.4	206	28,612	720.0
65-74	70	28	2,585	1083.2	186	12,663	1468.8
75-84	80	31	1,462	2120.4	102	5,317	1918.4

第 3.6 節では，年齢階層を無視した（非喫煙・喫煙）のポアソン回帰による 2 群間比較に引き続き，年齢をモデルに加えたモデル，年齢の 2 乗の項を加えたモデル，さらに，年齢と喫煙習慣の交互作用，年齢の 2 乗と喫煙習慣の交互作用をモデルに加え，尤度比検定によるモデル選択を行った.

2 次式のあてはめ

基本の主効果モデル

$$y_i = n_i \exp(\beta_0 + \beta_1 x_{smoke} + \beta_2 x_{age}) + \varepsilon_i$$

に対し，年齢の 2 乗と年齢と喫煙習慣の交互作用を加えたモデル

$$y_i = n_i \exp(\beta_0 + \beta_1 x_{smoke} + \beta_2 x_{age} + \beta_3 x^2_{(age/10)} + \beta_4 x_{smoke\times age}) + \varepsilon_i$$

が尤度比検定によって選択された.

図 12.7 左の基本モデルのあてはめは，年齢が高くなるにつれて，死亡者数が指数関数的に増大し続けることになり，モデルとしては不適切である．さらに，図 12.7 右の対数目盛での直線に対し，プロットされた点は，上に凸となっており，あてはめは支持されない.

直線のあてはめが支持されない場合は，便宜的な方法ではあるが，年齢について 2 乗の項を加えて 2 次式をあてはめて様子をみることになる．図 12.8 に示すように，×印の喫煙者へのあてはまりは良くなったが，〇印の非喫煙者の 80 歳代に対しては，過小評価となっていてモデルとしては不十分である.

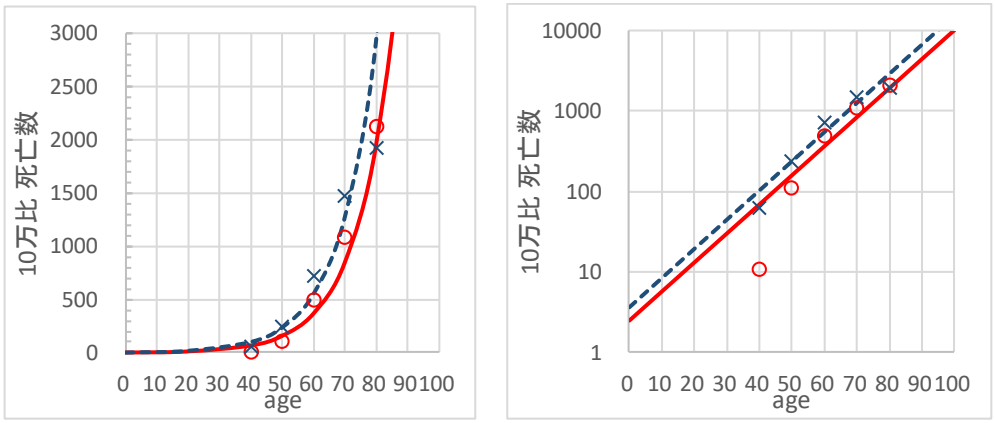


図 12.7　2 本のポアソン回帰直線（図 3.4 再掲）

○非喫煙：$\hat{y}_i = 100{,}000 \times \exp(-10.6260 + 0.4064 \times 0 + 0.0836 \times x_{age})$

×喫煙：　$\hat{y}_i = 100{,}000 \times \exp(-10.6260 + 0.4064 \times 1 + 0.0836 \times x_{age})$

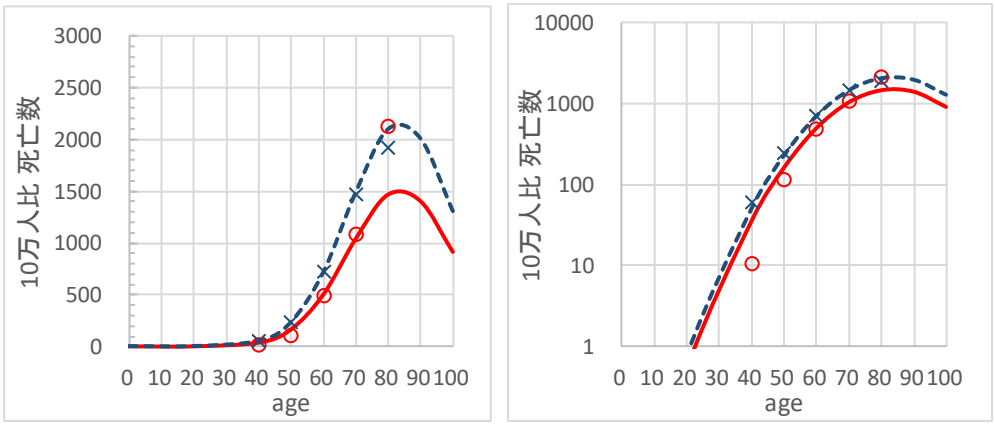


図 12.8　2 本の 2 次曲線（図 3.5 再掲）

○非喫煙：$\hat{y}_i = 100{,}000 \times \exp(-17.8583 + 0.3548 \times 0 + 0.3285 x_{age} - 0.1942 x^2_{(age/10)})$

×喫煙：　$\hat{y}_i = 100{,}000 \times \exp(-17.8583 + 0.3548 \times 1 + 0.3285 x_{age} - 0.1942 x^2_{(age/10)})$

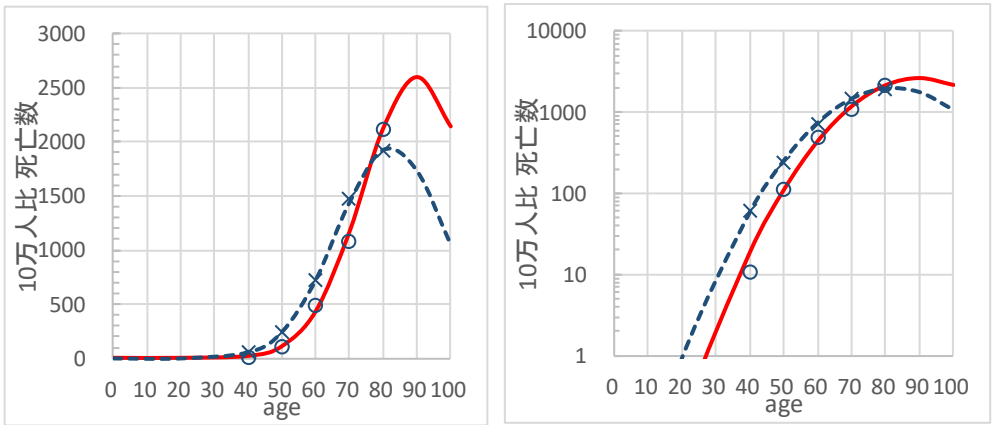


図 12.9　交互作用を含む 2 本の 2 次曲線（図 3.6 再掲）

○非喫煙：$\hat{y}_i = 100{,}000 \times \exp(-19.7003 + 2.3636 \times 0 + 0.3563 x_{age} - 0.1977 x^2_{(age/10)} - 0.0308 \times 0)$

×喫煙：　$\hat{y}_i = 100{,}000 \times \exp(-19.7003 + 2.3636 \times 1 + 0.3563 x_{age} - 0.1977 x^2_{(age/10)} - 0.0308 \times 1 \times x_{age})$

図 12.9 に示すように，喫煙習慣と年齢の交互作用を加えたモデルは，良くあてはまっていると判断される．喫煙習慣と年齢の交互作用を含むオフセット付き対数リンクのポアソン回帰の結果を表 12.21 に示す．推定値 $\hat{y}_i$ は，

$$\hat{y}_i = n_i \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{smoke} + \hat{\beta}_2 x_{age} + \hat{\beta}_3 x^2_{(age/10)} + \hat{\beta}_4 x_{smoke\times age})$$

として求められている．対数尤度 $\ln L_i$ は，

$$\ln L_i = \ln\left[\text{Poisson.dist}(y_i, \hat{y}_i, false)\right]$$

Excel のポアソン関数で計算し，対数尤度 $\ln L = \Sigma_i \ln L_i$ を最大化するようにパラメータ $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，$\hat{\beta}_3$，$\hat{\beta}_4$ に適当な初期値を入れて，Excel のソルバーで変化させた結果を示す．JMP での結果と微妙に異なるが，$\ln L = -28.3517$ と小数点以下 4 桁までで同じ結果となっている．JMP の方が小数点以下 5 桁目で大きくなっているので，JMP で推定したパラメータを以後使うことにする．

表 12.21　交互作用モデル（推定値は JMP による）

	切片	喫煙	年齢	年齢^2	交互.	死亡	人年		推定値	対数尤度		
i	x_0	x_1	x_2	x_3	$x_1 \times x_2$	y	n	10万y	y^=w^	$\ln Li$		JMP
1	1	0	40	16	0	2	18,790	10.6	3.4	-1.65171	$\hat{\beta}_0$=	**-19.7003**
2	1	0	50	25	0	12	10,673	112.4	11.5	-2.17732	$\hat{\beta}_1$=	**2.3636**
3	1	0	60	36	0	28	5,710	490.4	24.7	-2.79350	$\hat{\beta}_2$=	**0.3563**
4	1	0	70	49	0	28	2,585	1,083.2	30.2	-2.67231	$\hat{\beta}_3$=	**-0.1977**
5	1	0	80	64	0	31	1,462	2,120.4	31.1	-2.63870	$\hat{\beta}_4$=	**-0.0308**
6	1	1	40	16	40	32	52,407	61.1	29.6	-2.75042		Excel
7	1	1	50	25	50	104	43,248	240.5	106.8	-3.27928		**-19.6978**
8	1	1	60	36	60	206	28,612	720.0	208.2	-3.59493		**2.3677**
9	1	1	70	49	70	186	12,663	1,468.8	182.8	-3.55962		**0.3561**
10	1	1	80	64	80	102	5,317	1,918.4	102.6	-3.23387		**-0.1975**
			$x_3=(x_2/10)^2$						$\ln L$ =	**-28.35166**		**-0.0308**

2 次式の 95%信頼区間

図 12.9 に示した喫煙習慣と年齢の交互作用を含む 10 万人比での推定死亡曲線に，95%信頼区間を重ね書きするために，パラメータの共分散行列を計算する．対数リンクの場合の重み $\hat{w}_i$ は，推定値 $\hat{y}_i$ に等しいことは，前節で示した．求めるパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = [(\boldsymbol{X} * \hat{\boldsymbol{w}})^T \boldsymbol{X}]^{-1}$$

として得られ，表 12.22 に結果を示す．

得られたパラメータの推定値 $\hat{\boldsymbol{\beta}}$ を用いて，表 12.23 に示すように 0 歳から 100 歳までの 1 人あたりの対数死亡者の推定値 $\ln \hat{y}_i$ を

$$\ln \hat{y}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}$$

表 12.22　パラメータの共分散行列

		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=$	$\hat{\beta}_0$	**1.5712**	-0.4352	-0.0439	0.0300	0.0064
	$\hat{\beta}_1$	-0.4352	**0.4306**	0.0074	-0.0016	-0.0063
	$\hat{\beta}_2$	-0.0439	0.0074	**0.0013**	-0.0010	-0.0001
	$\hat{\beta}_3$	0.0300	-0.0016	-0.0010	**0.0007**	0.0000
	$\hat{\beta}_4$	0.0064	-0.0063	-0.0001	0.0000	**0.0001**

=Minverse(Mmult(Transpose($\boldsymbol{X}$の範囲*$\boldsymbol{w}$への範囲), $\boldsymbol{X}$の範囲))

で求める．その分散 $Var(\ln \hat{y}_i)$ は，

$$Var(\ln \hat{y}_i) = \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T$$

である．95%信頼区間は，

$$(\mathrm{L95\%}, \mathrm{U95\%}) = \ln \hat{y}_i \pm 1.96\sqrt{Var(\ln \hat{y}_i)}$$

で求められる．これらを 10 万人比にするために，

$$\hat{y}_i = 100{,}000 \exp(\ln \hat{y}_i)$$

$$L95\% = 100{,}000 \exp(\ln L95\%)$$

$$U95\% = 100{,}000 \exp(\ln U95\%)$$

で換算する．

表 12.23　10 万人比における 95%信頼区間

切片	喫煙	年齢	年齢^2	交互	——— 1人当たりの推定値 対数———				———10万人比———		
x_0	x_1	x_2	x_3	$x_1 \times x_2$	$\ln y$^	$Var(\ln y$^$)$	$L95\%$	$U95\%$	y^	$L95\%$	$U95\%$
1	0	0	0	0	-19.7003	1.5712	-22.1571	-17.2435	0.0	0.0	0.0
1	0	20	4	0	-13.3659	0.4393	-14.6650	-12.0668	0.2	0.0	0.6
1	0	40	16	0	-8.6130	0.0851	-9.1847	-8.0412	18.2	10.3	32.2
1	0	50	25	0	-6.8295	0.0342	-7.1920	-6.4670	108.1	75.3	155.4
1	0	60	36	0	-5.4414	0.0152	-5.6828	-5.2001	433.3	340.4	551.6
1	0	70	49	0	-4.4487	0.0112	-4.6564	-4.2409	1169.4	950.0	1439.4
1	0	80	64	0	-3.8513	0.0237	-4.1527	-3.5499	2125.2	1572.2	2872.9
1	0	90	81	0	-3.6492	0.0716	-4.1738	-3.1247	2601.1	1539.4	4395.0
1	0	100	100	0	-3.8426	0.1923	-4.7021	-2.9830	2143.9	907.6	5064.2
1	1	0	0	0	-17.3367	1.1314	-19.4215	-15.2518	0.0	0.0	0.0
1	1	20	4	20	-11.6174	0.2408	-12.5791	-10.6557	0.9	0.3	2.4
1	1	40	16	40	-7.4795	0.0204	-7.7592	-7.1999	56.5	42.7	74.7
1	1	50	25	50	-6.0036	0.0042	-6.1302	-5.8771	247.0	217.6	280.3
1	1	60	36	60	-4.9231	0.0027	-5.0249	-4.8213	727.7	657.2	805.6
1	1	70	49	70	-4.2379	0.0024	-4.3348	-4.1410	1443.8	1310.4	1590.8
1	1	80	64	80	-3.9481	0.0079	-4.1221	-3.7740	1929.2	1621.0	2296.0
1	1	90	81	90	-4.0536	0.0415	-4.4527	-3.6544	1736.0	1164.7	2587.6
1	1	100	100	100	-4.5544	0.1436	-5.2972	-3.8117	1052.0	500.6	2211.2

表 12.21 で求めた 10 万人比 y_i，表 12.23 で求めた 10 万人比の $\hat{y}_i$，$L95\%$，$U95\%$を非喫煙者に対しては図 12.10 に，喫煙者に対しては図 12.11 に示す．

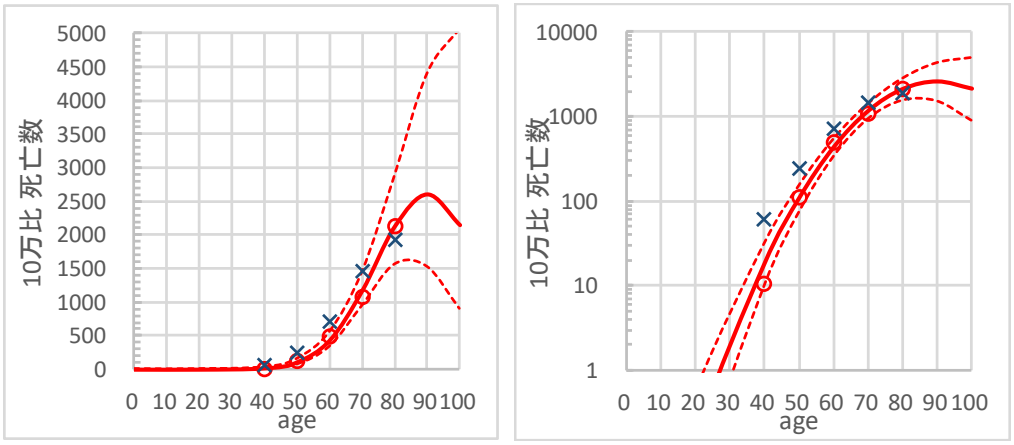


図 12.10　非喫煙者に対する 10 万人比での 95%信頼区間

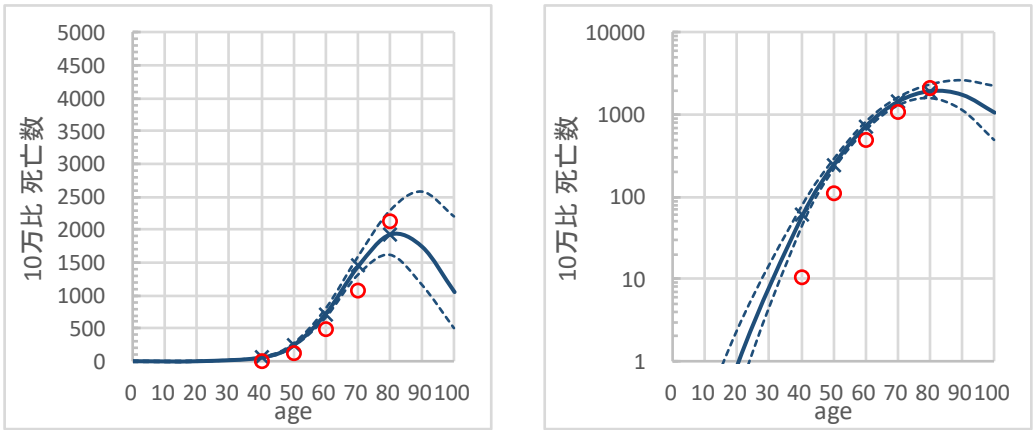


図 12.11　喫煙者に対する 10 万人比での 95%信頼区間

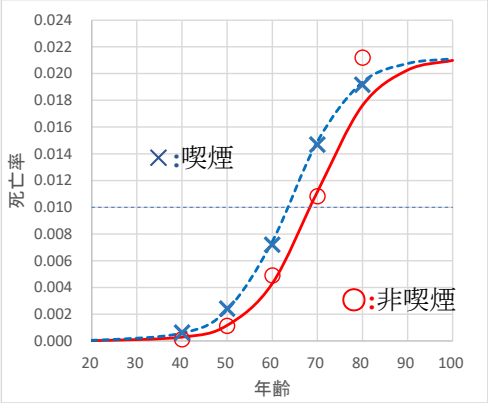
回帰分析の推定値および 95%信頼区間について，あえて外挿を行い対数リンクでの 2 次曲線のあてはめの形状の認識を行った．2 次曲線のあてはめは，対数目盛上で行っているので，指数を取って元に戻した結果も示した．対数目盛上で 80 歳以上の外挿の落ち込み具合は，少ないように見えても，元の目盛り上では，落ち込みが激しい．逆に，40 歳以下の場合は，指数を取ることにより，きれいに 0 人に収束していることが読み取れる．

上限がある場合のシグモイド曲線のあてはめ

2 次曲線のあてはめは，1 次直線のあてはめが適しているかの検討のためにであって，統計モデルとしては，便宜的な方法である．第 2.6 節では，死亡率について上限を新たな変数としたロジスティック回帰でシグモイド曲線をあてはめる方法を示した．この事例であれば，喫煙者と非喫煙者に共通の死亡率を新たな変数とし，形状が同じで，位置が異なるシグモイド曲線をあてはめ，喫煙者の位置パラメータと非喫煙者の位置パラメータの違いを検討するのが本質的な解析方法である．

表 12.24　死亡率に上限を持つ 2 本のロジスティック回帰の同時あてはめ

					$\hat{\beta}_0$=	**-10.2556**		
					$\hat{\beta}_1$=	**0.7578**		
					$\hat{\beta}_2$=	**0.1480**		対数尤度
					Umax=	**0.0212**	ln*L*=	**-29.7838**
切片	喫煙	年齢	死亡	人年	死亡率	推定値	二項分布	対数尤度
x_0	x_1	x_2	y	n	p	$\hat{\pi}$	P	lnL_i
1	0	0				0.0000		
1	0	20				0.0000		
1	0	40	2	18,790	0.0001	0.0003	0.0770	-2.5640
1	0	50	12	10,673	0.0011	0.0012	0.1140	-2.1715
1	0	60	28	5,710	0.0049	0.0043	0.0585	-2.8389
1	0	70	28	2,585	0.0108	0.0112	0.0747	-2.5948
1	0	80	31	1,462	0.0212	0.0176	0.0432	-3.1428
1	0	90				0.0203		
1	0	100				0.0210		
1	1	0				0.0000		
1	1	20				0.0000		
1	1	40	32	52,407	0.0006	0.0006	0.0667	-2.7072
1	1	50	104	43,248	0.0024	0.0023	0.0364	-3.3122
1	1	60	206	28,612	0.0072	0.0074	0.0252	-3.6792
1	1	70	186	12,663	0.0147	0.0149	0.0288	-3.5465
1	1	80	102	5,317	0.0192	0.0194	0.0397	-3.2267
1	1	90				0.0208		
1	1	100				0.0211		



死亡率が π_0 となる $\hat{x}_0$ の推定

$$\pi_0 = U_{max} \cdot \frac{\exp(\boldsymbol{x}\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}\hat{\boldsymbol{\beta}})}$$

$$logit_0 = \ln\left(\frac{\pi_0 / U_{max}}{1-\pi_0 / U_{max}}\right)$$

$$logit_0 = \boldsymbol{x}\hat{\boldsymbol{\beta}}$$

$$\hat{x}_0^{\text{喫煙}} = (logit_0 - \hat{\beta}_0 - \hat{\beta}_1)/\hat{\beta}_2$$

$$\hat{x}_0^{\text{非喫煙}} = (logit_0 - \hat{\beta}_0)/\hat{\beta}_2$$

図 12.12　喫煙者と非喫煙者に対する上限があるシグモイド曲線の同時あてはめ

喫煙者と非喫煙者の共通の死亡率の上限は，$U_{max} = 2.12\%$ と推定される．詳細は省くが，死亡率が 1%となる年齢は，$logit_0 = \ln[(\pi_0 / U_{max})/(1-\pi_0 / U_{max})] = -0.1151$ としたときに，次式から

$$\hat{x}_0^{\text{喫煙}} = (logit_0 - \hat{\beta}_0 - \hat{\beta}_1)/\hat{\beta}_2 = [-0.1151-(-10.2556)-0.7578]/0.1480 = 68.53$$

$$\hat{x}_0^{\text{非喫煙}} = (logit_0 - \hat{\beta}_0)/\hat{\beta}_2 = [-0.1151-(-10.2556)]/0.1480 = 63.41$$

と推定される．したがって，喫煙者は，非喫煙者に比べて 5.12 歳ほど早死にすると推定される．

13. 最小 2 乗平均の謎を予測プロファイルで解く

最小 2 乗法は，統計的用語として広く使われているが，「最小 2 乗平均」は，竹内ら（1989），「統計学辞典」の索引にも載っていない方言みたいなものである．「最小 2 乗平均」は，SAS ユーザであれば，Lsmeans で出力される「調整平均」と広く認識されており，SAS を使った解析結果には，方言であることの認識なく広く使われている．高橋・大橋・芳賀（1989），「SAS による実験データの解析」の第 15 章では，最小 2 乗平均について丁寧に説明している．残念ながら，「最小 2 乗平均」は，SAS ユーザの方言のまま現在に至っている．JMP では，「予測プロファイル」が新たな方言として加わっており，探索的な解析結果を可視化するために斬新な統計的な方法であることをこれまで示してきた．「最小 2 乗平均」は，「予測プロファイル」に包含される統計量の一種であり，「予測プロファイル」は，探索的データ解析に際し，データの内部構造を描き出すことに優れていることを示してきた．

13.1. 最小 2 乗平均（Lsmeans）とは

予測プロファイルについては，第 7.2 節「カブトガニのサテライト数に対する探索的解析」で，メインの解析方法として例示してきた．この事例は，ポアソン回帰での適用であるが，予測プロファイルの算出方法は，共分散分析，重回帰分析などの場合も同様に適用できる汎用的な方法である．特に，質的因子と量的因子が混在し，交互作用が無視できないような探索的な解析に際し，結果の考察に役に立つ．

JMP の「予測プロファイル」は，「最小 2 乗平均」の概念を拡張し，GUI を兼ね備えた動的なグラフ表示となっていて，JMP ユーザにとって身近な存在である．ただ，どのような計算方法なのか，多くのユーザにとってブラック・ボックス的な存在でもある．使い勝手の良い方法であっても，その統計的な性質が適切に示され Excel などで計算が誰にでもでも再現できなければ，しょせん「方言」の域を出ない．第 7.2 節では，JMP で作成した予測プロファイルを Excel により JMP と同様のグラフとして再現した．その際に必要なのは，パラメータの共分散行列であり，その活用方法について繰返し言及してきた．

重回帰分析では，多重共線性の把握にパラメータの相関行列が広く活用されているが，パラメータの共分散行列は軽視されている．たとえば，JMP の「モデルのあてはめ」においてディフォルトの最小 2 乗法による解析を選択した場合に，パラメータの相関行列は得ること

ができるが，パラメータの共分散行列を得ることができない．これは，多くの統計の教科書に「パラメータの共分散行列」の活用事例が見いだされないことの反映でもある．

JMP の「予測プロファイル」は，最小 2 乗平均の考え方を拡張し，その結果を GUI による探索的な要素を持った動的なグラフとして提供している．これにより，従来は，パラメータの推定値による隔靴掻痒的な結果の解釈に対し，直接的に複雑なデータの構造を切り出し，データの内部構造の可視化に貢献している．

とはいえ，多くの JMP ユーザにとって「予測プロファイル」が，ブラック・ボックス的であれば，方言のままとなる．幸い，現代の算盤である Excel を用いて自己完結的に「予測プロファイル」を作成できることを示してきた．ただし，主体はポアソン回帰についであり，通常の回帰については断片的な取り扱いであった．

そこで，豊富で魅力的な事例に富んだ奥野ら（1981）で取り上げられている重回帰分析の事例に着目した．第 12.3 節では，2 変量の重回帰の事例として「材料，工数と生産量の関係」について奥野ら（1981）で用いられている「偏差平方和ベース」に対し「デザイン行列ベース」の解析アプローチの違いを比較検討するために用いた．さらに，奥野ら（1981）から層別因子を含む回帰分析の事例，共変量が 2 変数の共分散分析の事例，守屋ら（2018）の繰り返しが不揃いな 2 因子の共分散分析のデータについて最小 2 乗平均（Lsmeans）の謎を Excel によって解き明かす．

13.2. 交互作用を考慮した共分散分析

奥野ら(1981）の事例 7.1 は，4 水準の層別因子を含む回帰分析の事例である．表 13.1 に示すように，亜硫酸ガス回収塔の洗浄用水の温度 x と，回収液の濃度 y との関係を調べたい．しかし, 測定回数が少ないので, 長期間にわたってとられたデータを使わなければならない．その間に季節の変化ばかりでなく，操業条件も変わっているので，データを 4 つ季節に分けてみることにした.

共分散分析の拡張

厳密な定義による共分散分析は, 1 元配置型などの実験で制御することはできないが, 結果に影響を与える量的変数を解析モデルに共変量として含める解析方法である．伝統的な解析法は, 1 元配置型などの分散分析に回帰分析を複合的に付け加えており, 一般的には難解な方法として認識されている.

奥野ら(1981）のタイトルは,「層別因子を含む回帰分析」であり，回帰分析を主体とするが結果に影響がある質的因子（層別因子）を付加的に解析モデルに組み込もうとしている．表 13.1 に示すように層別因子として操業月を 4 区分の季節とし，解析モデル的には，共分散分析と同じであるが，厳密な定義での共分散分析ではなく，回帰分析を主体にした探索的な解析である.

表 13.1　季節ごとの洗浄用水の温度 x と回収液の濃度 y［奥野（1981），表 7.2］

	A_1(7月・8月)		A_2(9月・10月)		A_3(11月・12月)		A_4(1月・2月)	
No.	x	y	x	y	x	y	x	y
1	30	8.0	34	9.1	27	15.0	16	23.4
2	28	10.5	19	19.4	20	21.0	16	28.2
3	30	8.2	22	20.5	14	24.2	20	29.5
4	29	13.0	25	14.2	18	15.3	16	22.2
5	28	10.1	22	11.0	12	17.3	6	40.2
6	17	16.2	25	19.1	18	24.7	16	36.6
7	24	16.1	22	16.0	13	23.9	12	35.5
8	22	13.4	23	11.1	21	20.2	11	42.1
9	30	13.0	28	12.0	24	13.6	9	42.5
10	30	7.3	24	17.1			22	25.8
11	25	14.0	30	13.3				
12			24	8.9				
平均	26.6364	11.8000	24.8333	14.3083	18.5556	19.4667	14.4000	32.6000
						総平均	21.4762	19.1119

データのグラフ表示

得られたデータの関連を概観するために，洗浄用水の温度 x と回収液の濃度 y の散布図上に季節別の回帰直線を上書きした結果を図 13.1 に示す．季節が A_4（1 月・2 月）の場合に多

重比較などをするまでもなく明らかに他の季節に比べ洗浄水の温度が低い方に分布し，他の季節と共通の洗浄水の温度 x が 20 度で比べた場合に，回収液の濃度 y が高めになっている．季節に関連する操業条件などの他の因子の影響があるのかも知れない．

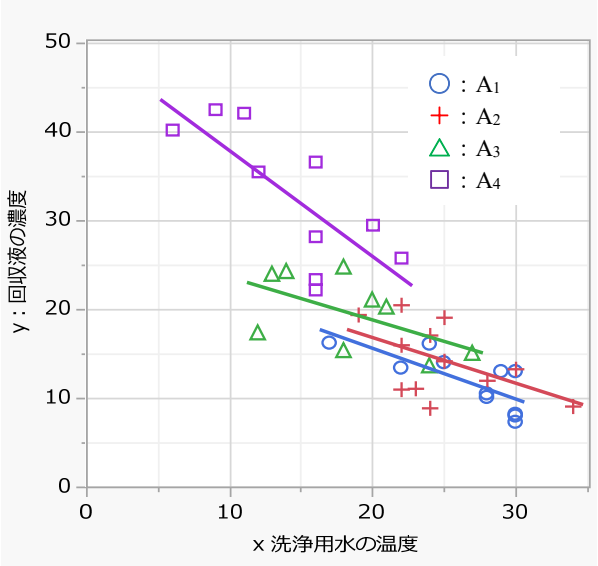


図 13.1　JMP のグラフ・ビルダーによる 4 本の回帰直線のあてはめ

伝統的な共分散分析の考え方

奥野ら（1981）には，共分散分析における伝統的な交互作用の解析方法が示されている．生物統計の名著であるスネデカー・コクラン（1972）でも，医療統計の名著であるアーミテージら（2001）でも，同様の解析方法が示されていることもあり，解析法として定着している．

1 元配置モデルでの解析に際し，反応 y に影響を与えることがはっきりしている変数があるが，実験に際し一定に保つことがどうしてもできない．この変数を「共変量」として，共変量の影響を除去して，因子 A の水準間の比較をしたいとの要望に対処するための方法として厳密な定義での共分散分析が定着している．そのために，共変量の影響が均一であることが条件であり，均一でなければ共分散分析が適用できないとの共通認識となっている．

この節で取り上げる観察データに対する「層別因子を含む回帰分析」は，データの構造および解析の考え方は，厳密な定義での「共分散分析」と全く同一である．ただし，後付けで設定した層ごとの共変量 x が同一の範囲に入り，反応 y に対する回帰直線の傾きが層ごとに共通であるとは限らない．季節ごとの共変量 x の平均値には，統計的方法を用いなくても散布図から明らかと言い切ったのであるが，p 値による判断を好む人たちもいるので，共変量 x について季節で 1 元配置分散分析をすると F=16.2716，p=0.00000058 (df_1=3, df_2=38) となる．し

たがって，「交互作用を考慮した共分散分析」というのではなく，「質的変数と量的変数を含む重回帰分析における交互作用解析」というべきであろう．

質的変数と量的変数を含む重回帰分析における交互作用解析

あえて本節のタイトルを「交互作用を考慮した共分散分析」としたのは，厳密な定義による共分散分析の解析の考え方が普及していて，交互作用があったら共分散分析ができないとの頑なな迷信が蔓延している．これら事柄については，高橋（2019b)，「投与前値がある場合の解析のレビュー」に歴史的な経緯がまとめられている．このために，共分散分析と同様のデータ構造をもつ観察データの探索的な解析に際しても「迷信」にとらわれてしまいがちになる．そこで，あえて，厳密な共分散分析の適用がためらわれる事例について，交互作用が否定できにくい場合の探索的な解析方法を示すことにした．交互作用の探索的な解析には，最小 2 乗平均および予測プロファイルによるアプローチが有益であることも示す．

伝統的な共分散分析の解析方法での交互作用の検出方法は，因子 A の水準ごとに反応 y に対する共変量 x を説明変数とする単回帰分析を行い，偏差平方和（$S_{xy}^{(1)}$， $S_{xy}^{(2)}$, . . . ， $S_{xy}^{(k)}$)，（$S_{xx}^{(1)}$， $S_{xx}^{(2)}$, . . . ， $S_{xx}^{(k)}$）を求めて，共通の切片 $\hat{\beta}_1^{(共通)}$ として，

$$\hat{\beta}_1^{(共通)} = \frac{\sum_{i=1}^{k} S_{xy}^{(k)}}{\sum_{i=1}^{k} S_{xx}^{(k)}}$$

を計算し，別々の直線をあてはめた場合の平方和，共通の傾きを持つ回帰直線をあてはめた場合の誤差平方和を計算し，それらの誤差平方和の差から，交互作用の検討を行っている．

この方法は，コンピュータによる重回帰分析が手軽にできない時代の計算手順であり，「共分散分析」というと，この伝統的な解析手順にとらわれてしまいがちになる．正確には，質的変数と量的変数，さらにそれらの交互作用を含むデザイン行列を活用した探索的な重回帰分析と認識すべきある．反応がカウント・データデータであれば，ポアソン回帰を使った解析となり，2 値反応ならば，ロジスティック回帰を使った解析となる．第 3.5 節の「2 本の回帰直線に対する各種のデザイン行列」は，ポアソン回帰での結果を示しているが，重回帰分析の場合でも，まったく共通する考え方である．

統計ソフト JMP を用いた共分散分析

JMP の最小 2 乗法による回帰分析を用い，層別因子 A と洗浄水の温度 x との交互作用 A×x を含む重回帰分析を適用し，予測プロファイルを用いて比較検討する．解析モデルは，4 水準の季節 A，洗浄水の温度 x，それらの交互作用 A×x を含める．表 13.2 に示すような分散分析表が得られる．

表 13.2　季節 A と洗浄水の温度 x の交互作用を含む分散分析表

分散分析

要因	自由度	平方和	平均平方	F値
モデル	7	3145.7754	449.3965	30.0158
誤差	34	509.0487	14.9720	p値(Prob>F)
全体(修正済み)	41	3654.8240		<.0001*

効果の検定

要因	自由度	平方和	平均平方	F値	p値(Prob>F)
A	3	182.8749	60.9583	4.0715	0.0142*
x	1	369.0901	369.0901	24.6520	<.0001*
A*x	3	71.2462	23.7487	1.5862	0.2107

交互作用 A×x が有意ではないので，解析モデルから除いて，季節 A に共通の傾きをあてはめたくなる．だが，季節により洗浄水の温度 x の範囲が明らかに異なるので，ためらわれる．このような観察データでは，厳密な定義での共分散分析では起こりえないような状況が発生する．したがって，交互作用を含めた“共分散分析”とし，季節に共通の洗浄水の温度である 20 度での季節間の回収液の温度 y の推定値の比較をしてみよう．

予測プロファイル

JMP の予測プロファイルを用い，図 13.2 に示すように，洗浄水の温度xを 20 度に固定し回収液の濃度 y を季節 A_4（1 月・2 月）と季節 A_3（11 月・12 月）の推定値と 95%信頼区間を求める．季節 A_4 の回収液の濃度の推定値 $\hat{y}$ は，25.9338 で 95%信頼区間は（22.0423，29.8254），季節 A_3 の場合は，18.7719（16.0328，21.5109）であり，95%信頼区間が互いに重ならないので有意な差があると判断される．

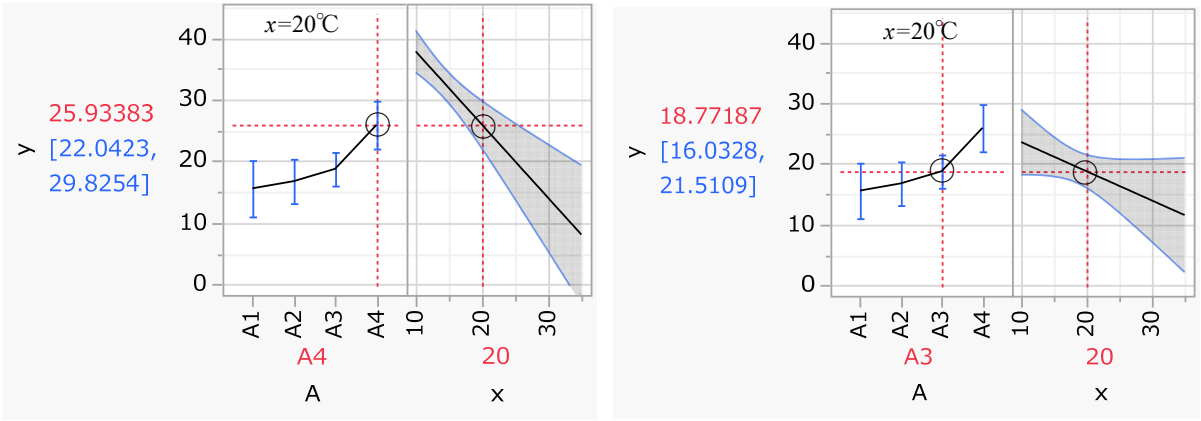


図 13.2　A_4 および A_3 を選択し場合の予測プロファイル

予測プロファイルには，季節を A_4（1 月・2 月）とした場合，洗浄水の温度 x に対する回収液の濃度 y について回帰直線および 95%信頼区間が示されている．マウスで「季節 A_3」を選択すると図 13.2 右の傾きが異なる回帰直線が示される．これは，解析モデルに交互作用を含めたからである．洗浄水の温度 x を 20 度に設定してあるが，ディフォルトでは，総平均の 21.4762 度である．他の温度も自由に設定することができ，図 13.3 に示すように温度を 10 度と 30 度に設定すると，交互作用を含めているために季節 A ごとの水準平均と 95%信頼区間ががらりと変わることが示されている．

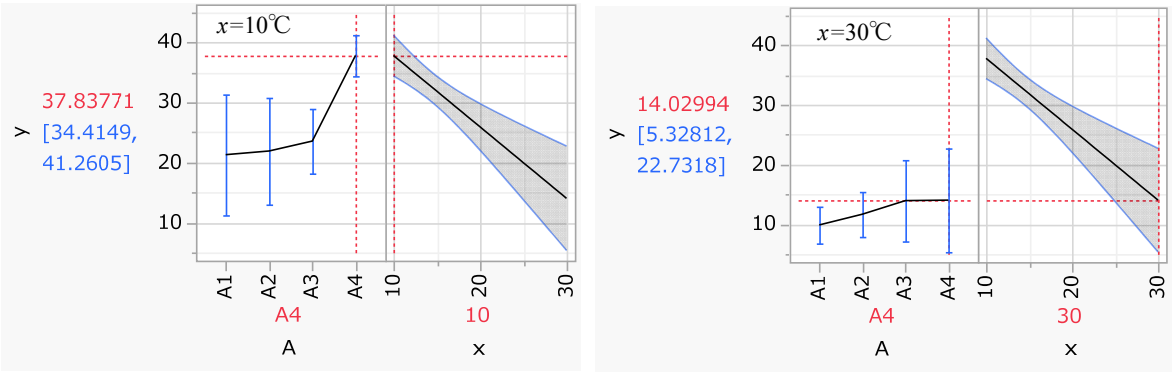


図 13.3　洗浄水の温度を変えた場合の季節 A の各水準の推定値と 95%信頼区間

対比による水準間の差の推定

JMP には，対比の機能を使って任意の水準間の差について統計量を算出する機能が備わっている．表 13.3 に示すように，洗浄水の温度を 20 度に固定し，季節 A_3 を+1，季節 A_4 を –1 とする対比の指定により，図 13.2 に示した A_3 の推定値の 18.7719 と A_4 の推定値 25.9338 との差が –7.1620, 標準誤差が 2.3417 と推定されている. 95%信頼区間の出力はないが, t=3.0585

表 13.3　洗浄水の温度 20 度の場合の季節 A_4 と季節 A_3 の差の統計量

対比

対比の指定

温度を 20℃に設定

A		x	
A1	0	x	20
A2	0		
A3	1		
A4	-1		

+または-をクリックして対比値を作成。

検定の詳細

A1	0.0000
A2	0.0000
A3	1.0000
A4	-1.0000
推定値	-7.1620
標準誤差	2.3417
t値	-3.0585
p値(Prob>|t|)	0.0043
平方和	140.053

パラメータ関数

パラメータ	
切片	0
A[A1]	1
A[A2]	1
A[A3]	2
x	0
x*A[A1]	20
x*A[A2]	20
x*A[A3]	40

平方和	分子自由度	分母自由度	F値	p値(Prob>F)
140.053	1.0000	34.0000	9.3543	0.0043

x	20

と有意な差であることが分かる．JMP での対比の設定は，直観的で季節（＋A_3－A_4）となっているが，JMP の内部での設定は，表 13.3 右の「パラメータ関数」で示すように（＋A_1＋A_2＋2A_3）が使われている．これは，JMP が（1，-1）対比型のデザイン変数をディフォルトにしていることに関係している．表 13.4 に示す対比型デザイン変数から明らかなように，（＋A_3－A_4）は，パラメータ関数で示されているように，内部では（＋a_1＋a_2＋2a_3）と置き換えられている．

表 13.4　対比型デザイン変数（＋A_3－A_4）

A	a_1	a_2	a_3
A_1	**1**	0	0
A_2	0	**1**	0
A_3	0	0	**1**
A_4	**-1**	**-1**	**-1**
A_3-A_4	**1**	**1**	**2**

パラメータ関数

パラメータ	
切片	0
A[A1]	1
A[A2]	1
A[A3]	2

なお，SAS の GLM プロシジャなどの場合は，Estimete ステートメントあるいは Contrust ステートメントで同様の推定を行うことができる．一般化線形モデルに対する SAS の GENMOD プロシジャでも同様の推定ができる．

Excel による交互作用を含む解析

JMP の予測プロファイルは，探索的な解析のための有用なツールであることを多くの事例で示したきたが，さらに，交互作用を含む事例について Excel で予測プロファイルを再現することにより，さらに理解を深めたい．交互作用を含むので表 13.6 に示すようにデザイン行列 $\boldsymbol{X}$ は，（42×8）と大きくなるが，これまでと同様に Excel の分析ツールの回帰分析を適用することができる．

季節 A は 4 水準なので，表 13.5 に示すように（1，-1）対比型デザイン変数(a_1，a_2，a_3)を与える．洗浄水の温度 x との交互作用は，(a_1x，a_2x，a_3x) のように積で与える．

表 13.5　対比型デザインでの交互作用に対するデザイン変数

A	a_1	a_2	a_3
A_1	**1**	0	0
A_2	0	**1**	0
A_3	0	0	**1**
A_4	**-1**	**-1**	**-1**

A	a_1x	a_2x	a_3x
A_1	$\boldsymbol{x}$	0	0
A_2	0	$\boldsymbol{x}$	0
A_3	0	0	$\boldsymbol{x}$
A_4	$\boldsymbol{-x}$	$\boldsymbol{-x}$	$\boldsymbol{-x}$

表 13.6 に交互作用モデルに対するデザイン行列に対して，Excel の回帰分析による分散分析表およびパラメータの推定結果を示す．その下に，Excel の行列関数を用いて計算したパラメータの共分散行列の計算結果を示す．

表 13.6　交互作用モデルに対する Excel によるパラメータの共分散行列の推定

		—— デザイン行列 $\boldsymbol{X}$ ——								
No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	y
1	A_1	1	1	0	0	30	30	0	0	8.0
2		1	1	0	0	28	28	0	0	10.5
3		1	1	0	0	30	30	0	0	8.2
4		1	1	0	0	29	29	0	0	13.0
5		1	1	0	0	28	28	0	0	10.1
6		1	1	0	0	17	17	0	0	16.2
7		1	1	0	0	24	24	0	0	16.1
8		1	1	0	0	22	22	0	0	13.4
9		1	1	0	0	30	30	0	0	13.0
10		1	1	0	0	30	30	0	0	7.3
11		1	1	0	0	25	25	0	0	14.0
12	A_2	1	0	1	0	34	0	34	0	9.1
13		1	0	1	0	19	0	19	0	19.4
14		1	0	1	0	22	0	22	0	20.5
15		1	0	1	0	25	0	25	0	14.2
16		1	0	1	0	22	0	22	0	11.0
17		1	0	1	0	25	0	25	0	19.1
18		1	0	1	0	22	0	22	0	16.0
19		1	0	1	0	23	0	23	0	11.1
20		1	0	1	0	28	0	28	0	12.0
21		1	0	1	0	24	0	24	0	17.1
22		1	0	1	0	30	0	30	0	13.3
23		1	0	1	0	24	0	24	0	8.9
24	A_3	1	0	0	1	27	0	0	27	15.0
25		1	0	0	1	20	0	0	20	21.0
26		1	0	0	1	14	0	0	14	24.2
27		1	0	0	1	18	0	0	18	15.3
28		1	0	0	1	12	0	0	12	17.3
29		1	0	0	1	18	0	0	18	24.7
30		1	0	0	1	13	0	0	13	23.9
31		1	0	0	1	21	0	0	21	20.2
32		1	0	0	1	24	0	0	24	13.6
33	A_4	1	-1	-1	-1	16	-16	-16	-16	23.4
34		1	-1	-1	-1	16	-16	-16	-16	28.2
35		1	-1	-1	-1	20	-20	-20	-20	29.5
36		1	-1	-1	-1	16	-16	-16	-16	22.2
37		1	-1	-1	-1	6	-6	-6	-6	40.2
38		1	-1	-1	-1	16	-16	-16	-16	36.6
39		1	-1	-1	-1	12	-12	-12	-12	35.5
40		1	-1	-1	-1	11	-11	-11	-11	42.1
41		1	-1	-1	-1	9	-9	-9	-9	42.5
42		1	-1	-1	-1	22	-22	-22	-22	25.8
総平均					21.4762				19.1119	

Excelによる回帰分析（x_0〜a_3x 8列）

分散分析表　「定数に0を使用」を on とする

	自由度	変動	分散	
回帰	8	18487	2310.9	
残差	34	509.0	**14.972**	$\sigma^{\wedge 2}$
合計	42	18996		

	係数	SE	t	P-値	
切片	0	#N/A	#N/A	#N/A	
x0	**33.04**	3.11	10.61	0.00	β_0
a1	**-6.03**	6.33	-0.95	0.35	β_1
a2	**-6.02**	5.95	-1.01	0.32	β_2
a3	**-4.65**	4.81	-0.97	0.34	β_3
x	**-0.69**	0.14	-4.97	0.00	β_4
a1 x	**0.12**	0.25	0.47	0.64	β_5
a2 x	**0.18**	0.24	0.72	0.48	β_6
a3 x	**0.21**	0.24	0.88	0.39	β_7

パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{\wedge 2}$

	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x
β_0	**9.70**	5.52	3.18	-2.97	-0.41	-0.14	-0.09	0.07
β_1	5.52	**40.12**	-18.40	-12.25	-0.14	-1.53	0.65	0.49
β_2	3.18	-18.40	**35.45**	-9.91	-0.09	0.65	-1.43	0.43
β_3	-2.97	-12.25	-9.91	**23.15**	0.07	0.49	0.43	-1.09
β_4	-0.41	-0.14	-0.09	0.07	**0.02**	0.00	0.00	0.00
β_5	-0.14	-1.53	0.65	0.49	0.00	**0.06**	-0.02	-0.02
β_6	-0.09	0.65	-1.43	0.43	0.00	-0.02	**0.06**	-0.02
β_7	0.07	0.49	0.43	-1.09	0.00	-0.02	-0.02	**0.06**

対比型デザイン変数

A	a_1	a_2	a_3
A_1	**1**	0	0
A_2	0	**1**	0
A_3	0	0	**1**
A_4	**-1**	**-1**	**-1**

交互作用のデザイン変数

A	a_1x	a_2x	a_3x
A_1	***x***	0	0
A_2	0	***x***	0
A_3	0	0	***x***
A_4	***-x***	***-x***	***-x***

$t(0.05,34)=$ **2.0322**

=T.Inv.2T(0.05,34)

パラメータに関する共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，デザイン行列と分散分析表の誤差分散=14.972 を用いて

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$$

$$= \text{Minverse(Mmult(Transpose(}\boldsymbol{X}\text{の範囲), }\boldsymbol{X}\text{の範囲)) *14.972}$$

=

	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x
β_0	**9.70**	5.52	3.18	-2.97	-0.41	-0.14	-0.09	0.07
β_1	5.52	**40.12**	-18.40	-12.25	-0.14	-1.53	0.65	0.49
β_2	3.18	-18.40	**35.45**	-9.91	-0.09	0.65	-1.43	0.43
β_3	-2.97	-12.25	-9.91	**23.15**	0.07	0.49	0.43	-1.09
β_4	-0.41	-0.14	-0.09	0.07	**0.02**	0.00	0.00	0.00
β_5	-0.14	-1.53	0.65	0.49	0.00	**0.06**	-0.02	-0.02
β_6	-0.09	0.65	-1.43	0.43	0.00	-0.02	**0.06**	-0.02
β_7	0.07	0.49	0.43	-1.09	0.00	-0.02	-0.02	**0.06**

のように計算されている．なお，これら行列計算の詳細は，第 4 章を参照のこと．

4 本の回帰直線の推定

季節 A_1 の回帰直線は，対比型のデザインであることから，次のように求めることができる．図 13.1 を参照して，推定結果が適切であることが確認できる．

$$A_1: \begin{cases} y_{A1} = (\hat{\beta}_0 + \hat{\beta}_1)\times 1 + (\hat{\beta}_4 + \hat{\beta}_5)x \\ \quad = 33.04 - 6.03 + (-0.69 + 0.12)x \\ \quad = 27.02 - 0.57x \end{cases}$$

$$A_2: \begin{cases} y_{A2} = (\hat{\beta}_0 + \hat{\beta}_2)\times 1 + (\hat{\beta}_4 + \hat{\beta}_6)x \\ \quad = 27.03 - 0.51x \end{cases}$$

$$A_3: \begin{cases} y_{A3} = (\hat{\beta}_0 + \hat{\beta}_3)\times 1 + (\hat{\beta}_4 + \hat{\beta}_7)x \\ \quad = 28.39 - 0.48x \end{cases}$$

$$A_4: \begin{cases} y_{A4} = [\hat{\beta}_0 - (\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)]\times 1 + [\hat{\beta}_4 - (\hat{\beta}_5 + \hat{\beta}_6 + \hat{\beta}_7)]x \\ \quad = 33.04 - (-16.70) + (-0.69 - 0.50)x \\ \quad = 49.74 - 1.19x \end{cases}$$

分散分析表

表 13.6 で示した Excel での分散分析表は，切片を含むデザイン変数すべてを用いたので，「回帰」の自由度が 8 で（回帰の変動=18487）となり，表 13.2 に示した JMP の結果（モデルの平方和=3145.7754）とは異なるが，以後の計算に必要な（残差の分散=14.972）は，JMP の（誤差の平均平方=14.9720）と同じである．Excel による回帰分析では，因子 A，温度 x，交互作用 A×x に関する平方和の分解がないので，別途対応する必要がある．

交互作用 A×x の分散（平均平方）は，表 13.7(右）に示すように交互作用を含まない主効果モデルでの分散分析表を作成し（S_{A+x}：回帰の変動=3047.5291）を得る．表 13.7(左）に示

すように交互作用を含めたモデルで（$S_{A+x+(A\times x)}$：回帰の変動=3145.7754）を得る．交互作用 $S_{(A\times x)}$ の変動（平方和）は，それらの差

$$\begin{aligned} S_{(A\times x)} &= S_{A+x+(A\times x)} - S_{A+x} \\ &= 3145.7754 - 3074.5291 \\ &= 71.2462 \end{aligned}$$

となり，表 13.2 に示した JMP での交互作用 $A\times x$ の平方和 $S_{(A\times x)} = 71.2462$ が得られる．

表 13.7　交互作用モデル vs. 主効果モデル

交互作用モデル：$y = A + x + (A\times x)$			
	自由度	変動	分散
回帰	7	**3145.7754**	449.3965
残差	34	509.0487	14.9720
合計	41	3654.8240	

a_1 a_2 a_3 x a_1x a_2x a_3x を選択

主効果モデル：$y = A + x$			
	自由度	変動	分散
回帰	4	**3074.5291**	768.6323
残差	37	580.2949	15.6836
合計	41	3654.8240	

a_1 a_2 a_3 x を選択

「定数に 0 を使用」を off に設定

表 13.8 に示すように，主効果 A および x の平方和は，Type III の平方和と言われて，因子 A ごとの x の平均が等しいと仮定するような平方和で，Excel の回帰分析では計算することができない．表 13.2 の交互作用を含む場合に対して，因子 A の平方和は 128.8749➔551.4944 と大幅に増加し，因子 x の平方和は 369.0901➔389.1143 と微妙に異なる．表 13.8 に示した平方和は，Type II の平方和と言われていて，Excel の回帰分析で求めることができる．

表 13.8　JMP での主効果モデルでの効果の検定

分散分析

要因	自由度	平方和	平均平方	F値
モデル	4	3074.5291	768.6323	49.0085
誤差	37	580.2949	15.6836	p値(Prob>F)
全体(修正済み)	41	3654.8240		<.0001*

効果の検定

要因	自由度	平方和	平均平方	F値	p値(Prob>F)
A	3	551.4944	183.8315	11.7212	<.0001*
x	1	389.1143	389.1143	24.8102	<.0001*

表 13.9 に示すように $(y = A+x)$ モデルでの平方和は $S_{A+x} = 3074.5291$，因子 A のみの $(y = A)$ モデルでの平方和は $S_A = 2685.4149$ であり，この差

$$\begin{aligned} S_x &= S_{A+x} - S_A \\ &= 3074.5291 - 2685.4149 = 389.1143 \end{aligned}$$

が，因子 A の存在下で因子 x を加えた場合の平方和の増分となっている．同様に，因子 x の

みの $(y=x)$ モデルでの平方和は $S_x = 2523.0347$ であり，この差

$$S_A = S_{A+x} - S_x$$
$$= 3074.5291 - 2523.0347 = 551.4944$$

が，因子 x の存在下で因子 A を加えた場合の平方和の増分となっている．

表 13.9　3 種の主効果モデル

主効果モデル: $y = A + x$

	自由度	変動
回帰	4	**3074.5291**
残差	37	580.2949
合計	41	3654.8240

a_1 a_2 a_3 x を選択

主効果モデル: $y = A$

	自由度	変動
回帰	3	**2685.4149**
残差	38	969.4092
合計	41	3654.8240

a_1 a_2 a_3 を選択

主効果モデル: $y = x$

	自由度	変動
回帰	1	**2523.0347**
残差	40	1131.7893
合計	41	3654.8240

x を選択

交互作用がモデルに含まれている場合の主効果の平方和の計算方法には，幾つかの考え方があり，そもそも，交互作用がある場合の主効果の検定にどんな意味があるのかとも言われている．Excel で追試ができる範囲の Type II の平方和を基準にして対応することを勧める．これらの平方和の扱いについては，芳賀（2009），「医薬品開発のための統計解析 第 2 部 実験計画法 初版」の第 4 章「共分散分析」に詳しく説明されている．また，同書について「じっくり勉強すれば身につく統計入門」シリーズ第 3 回で，杉本・橘田（2011）が「共分散分析の基礎・医薬品開発における共分散分析の例」で丁寧に解析しているので参照してもらいたい．

Excel による予測プロファイル

JMP の予測プロファイルを用いた推定を Excel で行う．洗浄水の温度 x を季節に共通の 20 度に固定し回収液の濃度の推定値 $\hat{y}_i$ と分散 $Var(\hat{y}_i)$ を推定する．表 13.10 に示すように，デザイン行列の変数をセットし，

$$\hat{y}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}$$
$$Var(\hat{y}_i) = \boldsymbol{x}_i \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^T$$
$$(L95\%, U95\%) = \hat{y}_i \pm t(0.05, 34)\sqrt{Var(\hat{y}_i)}$$

による計算を行う．季節 A_1 の場合は，

	x_1								β^		
y^$_{A1}$=	1	1	0	0	20	20	0	0	33.0450	=	15.5912
									-6.0281		
									-6.0157		
									-4.6528		
									-0.6887		
									0.1175		
									0.1765		
									0.2077		

$Var(y\hat{}_{A1})$=					$\boldsymbol{x}_1$				$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$								$\boldsymbol{x}_1^T$		
	1	1	0	0	20	20	0	0	**9.70**	5.52	3.18	-2.97	-0.41	-0.14	-0.09	0.07	1	=	5.0542
									5.52	**40.12**	-18.40	-12.25	-0.14	-1.53	0.65	0.49	1		
									3.18	-18.40	**35.45**	-9.91	-0.09	0.65	-1.43	0.43	0		
									-2.97	-12.25	-9.91	**23.15**	0.07	0.49	0.43	-1.09	0		
									-0.41	-0.14	-0.09	0.07	**0.02**	0.00	0.00	0.00	20		
									-0.14	-1.53	0.65	0.49	0.00	**0.06**	-0.02	-0.02	20		
									-0.09	0.65	-1.43	0.43	0.00	-0.02	**0.06**	-0.02	0		
									0.07	0.49	0.43	-1.09	0.00	-0.02	-0.02	**0.06**	0		

$$
\begin{aligned}
(L95\%, U95\%) &= \hat{y}_{A1} \pm t(0.05, 34)\sqrt{Var(\hat{y}_{A1})} \\
&= 15.5912 \pm 2.0322\sqrt{5.0542} \\
&= (11.0224, 20.1600)
\end{aligned}
$$

が得られる．他の季節に対しても同様に計算する．季節 A_4 の推定値 $\hat{y}_{A4}$ は，25.9338 で 95%信頼区間は（22.0423，29.8254）であり，季節 A_3 の場合は，18.7719（16.0328，21.5109）となり図 13.2 に示した予測プロアイルの推定値に一致する．

表 13.10　交互作用を考慮した季節別の推定値と 95%信頼区間

No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	$y\hat{}$	$Var(y\hat{})$	SE	$L95\%$	$U95\%$
1	A_1	1	1	0	0	20	20	0	0	15.5912	5.0542	2.2482	11.0224	20.1600
2	A_2	1	0	1	0	20	0	20	0	16.7842	3.1520	1.7754	13.1762	20.3922
3	**A_3**	1	0	0	1	20	0	0	20	**18.7719**	1.8165	1.3478	**16.0328**	**21.5109**
4	**A_4**	1	-1	-1	-1	20	-20	-20	-20	**25.9338**	3.6669	1.9149	**22.0423**	**29.8254**

この結果を Excel の「折れ線」グラフで作成した結果を図 13.4 に示す．Excel での信頼区間の幅を付ける際には，推定値からの長さでの設定になっているので $t(0.05, df) \times SE$ による髭の長さを別途計算する必要がある．

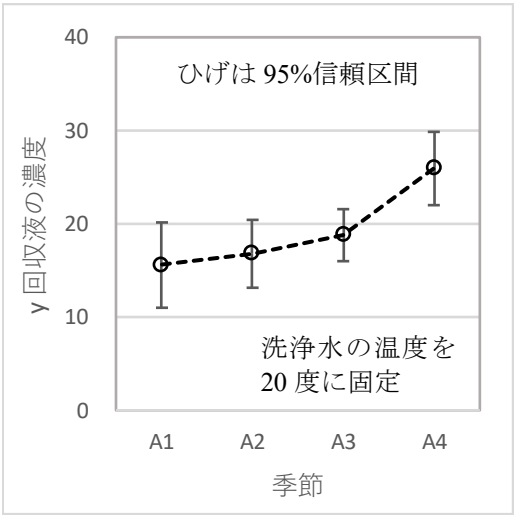


図 13.4　Excel の「折れ線」による予測プロファイル

Excel での「折れ線グラフ」の髭の長さの設定は，かなり深いところにあり，

グラフの要素➔誤差範囲➔その他のオプション➔ユーザの設定➔
値の設定➔正の誤差の値（範囲）➔負の誤差の値（範囲）

のように慎重に設定する．気を抜くと尤もらしいまがい物の誤差範囲となってしまう．

水準間の差の予測プロファイル

回収液の濃度が最も高い季節 A_4 を基準とし，温度が 20 度の場合の他の季節との差の推定値と 95%信頼区間を求めたい．季節 A_3 と A_4 の差の推定値は，表 13.11 に示すように，デザイン変数の差を計算する．推定値は，

$$\begin{aligned}\hat{y}_{(A_3-A_4)} &= \hat{y}_{A_3} - \hat{y}_{A_4} \\ &= 18.7719 - 25.9338 = -7.1620\end{aligned}$$

として求められる．この差に対する分散の推定のためのベクトル $\boldsymbol{d}_{(A_3-A_4)}$ は，季節 A_3 のベクトル $\boldsymbol{x}_{A_3}$ と季節 A_4 のベクトル $\boldsymbol{x}_{A_4}$ の差

$$\begin{aligned}\boldsymbol{x}_{A_3} &= [1 \quad 0 \quad 0 \quad 1 \quad 20 \quad 0 \quad 0 \quad 20] \\ -) \quad \boldsymbol{x}_{A_4} &= [1 \quad -1 \quad -1 \quad -1 \quad 20 \quad -20 \quad -20 \quad -20] \\ \hline \boldsymbol{d}_{(A_3-A_4)} = \boldsymbol{x}_{A_3} - \boldsymbol{x}_{A_4} &= [0 \quad 1 \quad 1 \quad 2 \quad 0 \quad 20 \quad 20 \quad 40]\end{aligned}$$

であり，$\hat{y}_{(A_3-A_4)}$ の分散 $Var[\hat{y}_{(A_3-A_4)}]$ は，

$$\begin{aligned}Var[\hat{y}_{(A_3-A_4)}] &= \boldsymbol{d}_{(A_3-A_4)} \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) \boldsymbol{d}^T_{(A_3-A_4)} \\ &= 5.4834\end{aligned}$$

として求められ．この結果は，表 13.3 に示した「パラメータ関数」に示された結果に一致する．差 –7.1620 の 95%信頼区間が（–11.9208，–2.4031）と推定されゼロを含まないので，有意な差であることが示されている．

表 13.11　2 水準間の差のデザイン変数と推定値

No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	y^	$Var(y$^$)$	SE	L95%	U95%
3	A_3	1	0	0	1	20	0	0	20	18.7719				
4	A_4	1	-1	-1	-1	20	-20	-20	-20	25.9338				
差	d	0	1	1	2	0	20	20	40	**-7.1620**	5.4834	2.3417	**-11.9208**	**-2.4031**

同様に季節 A_4 を基準とし，他の季節の差を表 13.12 に示す．差のデザイン行列については，季節 A_3 と季節 A_4 の場合と同様にする．差の分散についてもそれぞれの差のデザイン行列の場合と同様に推定し．95%信頼区間を計算する．JMP にも表 13.3 で示したように，ある温度設定に対し，因子 A の水準間の差の推定ができる．

表 13.12　季節 A_4 を基準とした 2 水準間の差の推定

No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	y^	$Var(y$^)	SE	L95%	U95%
5	A_1-A_4	0	2	1	1	0	40	20	20	**-10.3426**	8.7211	2.9532	**-16.3441**	**-4.3411**
6	A_2-A_4	0	1	2	1	0	20	40	20	**-9.1496**	6.8189	2.6113	**-14.4564**	**-3.8428**
7	A_3-A_4	0	1	1	2	0	20	20	40	**-7.1620**	5.4834	2.3417	**-11.9208**	**-2.4031**
8	A_4-A_4	0	0	0	0	0	0	0	0	0.0000	0.0000	0.0000	0.0000	0.0000

図 13.5 に季節 A_4 を基準とした差の推定値および95%信頼区間の予測プロファイルを示す．この図は，洗浄水の温度を 20 度に固定した場合であり，温度が低くなれば季節 A_4 との差は広がり，温度が高くなれば差は縮まる．

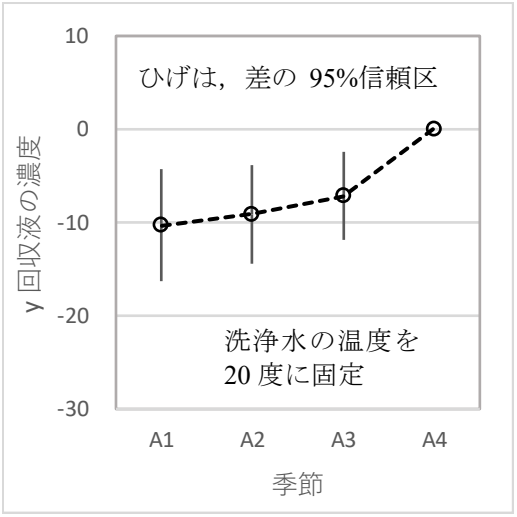


図 13.5　Excel の「折れ線」による差の予測プロファイル

洗浄水の温度に関する予測プロファイル

季節 A_4 と季節 A_3 について洗浄水の温度を変化させた場合の予測プロファイルを作成する．表 13.13 に示すように，季節 A_4 のデザイン行列変数のベクトル [−1　−1　−1]，洗浄水の温度 xを 5 度から 25 度まで変化させ，さらに [−1　−1　−1] と x の交互作用を加えたベクトルとする．

季節 A_3 のベクトル [0　0　1] に対しても同様に，洗浄水の温度 x を 5 度から 25 度まで変化させ，さらに [0　0　1] と x の交互作用を加えたベクトルを生成する．これらのベクトルに対し，パラメータの推定値 $\hat{\boldsymbol{\beta}}$ およびパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いて予測値 $\hat{y}$ と分散を推定し，95%信頼区間を求める．

表 13.13　洗浄水の温度を変えた場合の回収液の濃度

No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	y^	$Var(y$^$)$	SE	L95%	U95%
9	A_4	1	-1	-1	-1	5	-5	-5	-5	43.7896	7.6105	2.7587	38.1833	49.3960
10	A_4	1	-1	-1	-1	10	-10	-10	-10	37.8377	2.8367	1.6842	34.4149	41.2605
11	A_4	1	-1	-1	-1	15	-15	-15	-15	31.8858	1.5221	1.2337	29.3785	34.3930
12	**A_4**	1	-1	-1	-1	**20**	-20	-20	-20	**25.9338**	3.6669	1.9149	**22.0423**	**29.8254**
13	A_4	1	-1	-1	-1	25	-25	-25	-25	19.9819	9.2710	3.0448	13.7940	26.1697
14	A_3	1	0	0	1	10	0	0	10	23.5820	7.0298	2.6514	18.1937	28.9703
15	A_3	1	0	0	1	15	0	0	15	21.1769	2.5904	1.6095	17.9061	24.4478
16	**A_3**	1	0	0	1	**20**	0	0	20	**18.7719**	1.8165	1.3478	**16.0328**	**21.5109**
17	A_3	1	0	0	1	25	0	0	25	16.3668	4.7083	2.1699	11.9571	20.7765
18	A_3	1	0	0	1	30	0	0	30	13.9618	11.2657	3.3564	7.1407	20.7829

表 13.13 で計算された季節 A_4 と季節 A_3 の予測値と 95%信頼区間を図 13.6 に示すような予測プロファイルを作成する．季節 A_4 の洗浄水の温度 x が 10 度の場合は，回収液の濃度 $\hat{y}$ は，37.8377 と季節 A_3 の 23.5820 と 14.2557 の差がある．洗浄水の温度 x が 10 度の場合は，差が 7.1620 と縮小するが，95%信頼区間が互いに重なっていないので統計的な差があると判断される．

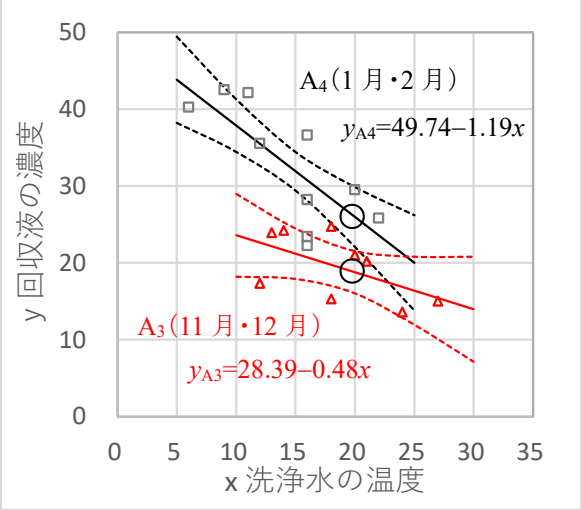


図 13.6　洗浄水の温度に対する季節 A_4 と季節 A_3 の回収液の濃度

回収液の濃度の差ついての予測プロファイル

洗浄水の温度を変化させた場合に季節 A_4 と季節 A_3 の回収液の濃度の差の予測値と 95%信頼区間について定量的に検討したい．季節 A_3 のデザイン変数のベクトル $\boldsymbol{x}_{A_3}$ と季節 A_4 のベクトル $\boldsymbol{x}_{A_4}$ の差は，

$$
\begin{array}{rcccccccc}
 & x_0 & a_1 & a_2 & a_3 & x & a_1x & a_2x & a_3x \\
\boldsymbol{x}_{A_3} = & [1 & 0 & 0 & 1 & x & 0 & 0 & x] \\
-) \quad \boldsymbol{x}_{A_4} = & [1 & -1 & -1 & -1 & x & -x & -x & -x] \\
\hline
\boldsymbol{d}_{(A_3-A_4)} = \boldsymbol{x}_{A_3} - \boldsymbol{x}_{A_4} = & [0 & 1 & 1 & 2 & 0 & x & x & 2x]
\end{array}
$$

となるので，表 13.14 に洗浄水の温度 x を変化させたときの予測値と 95%信頼区間を計算した結果を示す．図 13.7 には，差の予測値と 95%信頼区間を示したものである．95%信頼区間の上限（U95%）がゼロを横切るのは，おおよそ 22 度と推測される．正確には，95%上側信頼区間が 0.0 となるように温度 x をソルバーで変化させると 21.99 度が得られる．同等に，推定値 $\hat{y}$ が 0.0 となるのは，30.10 度と推定される．

表 13.14　A_4 と季節 A_3 の回収液の濃度の差の 95%信頼区間

No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	y^	$Var(y$^$)$	SE	L95%	U95%
19	A_3-A_4	0	1	1	2	0	5	5	10	-17.8026	22.7455	4.7692	-27.4948	-8.1104
20	"	0	1	1	2	0	10	10	20	-14.2557	9.8665	3.1411	-20.6392	-7.8722
21	"	0	1	1	2	0	15	15	30	-10.7088	4.1125	2.0279	-14.8301	-6.5876
22	"	0	1	1	2	0	20	20	40	-7.1620	5.4834	2.3417	-11.9208	-2.4031
23	"	0	1	1	2	0	25	25	50	-3.6151	13.9793	3.7389	-11.2134	3.9833
24	"	0	1	1	2	0	30	30	60	-0.0682	29.6002	5.4406	-11.1248	10.9884
ソルバー		0	1	1	2	0	**21.99**	21.99	43.98	-5.7513	8.0090	2.8300	-11.5026	**0.0000**
ソルバー		0	1	1	2	0	**30.10**	30.10	60.19	**0.0000**	29.9703	5.4745	-11.1256	11.1256

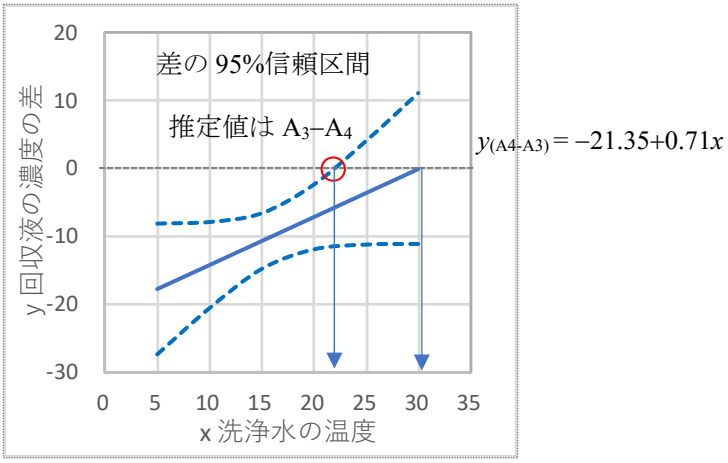


図 13.7　洗浄水の温度に対する季節 A_4 と季節 A_3 の回収液の濃度の差

最小 2 乗平均（Lsmeans）

表 13.2 に示した JMP による「分散分析表・効果の検定」，に引き続き表 13.3 で季節 A の水準間の差の検討結果を示したが，季節 A についての最小 2 乗平均の結果を示さなかった．これは，簡単に説明しにくいからで，予測プロファイルのある特定の場合として，最小 2 乗平均を説明することが理解しやすいと思われる．

表 13.15 に示すように，季節 A について最小 2 乗平均，標準誤差，95%信頼区間，平均が出力されている．平均は表 13.1 に示した回収液の濃度の季節 A の各種水準の濃度の算術平均である．季節 A_4 の濃度 y の最小 2 乗平均は，24.1766，A_3 の場合は，18.0618 と表示されている．

これは，洗浄水の温度xの総平均 21.4762 に固定した場合の季節 A の各種水準の濃度 y の推定値で，図 13.6 で示した洗浄水の温度を $x=20.0$ から $x=21.4762$ にずらした場合の濃度 y の推定値に相当する．

表 13.15　JMP による季節 A の各水準に対する最小 2 乗平均と 95%信頼区間

最小2乗平均表

水準	最小2乗平均	標準誤差	下側95%	上側95%	平均
A1	14.7479	1.8958	10.8953	18.6006	11.8000
A2	16.0280	1.4719	13.0368	19.0192	14.3083
A3	18.0618	1.5129	14.9872	21.1364	19.4667
A4	24.1766	2.2275	19.6499	28.7033	32.6000

JMP の予測プロファイルを使って確認する．図 13.8 に示すように，洗浄水の温度を $x=21.4762$ とした場合の，季節 A_4 の濃度 y の推定値は，24.1766（19.6498，28.7033），季節 A_3 の場合は，18.0618（14.9872，19.4667）のように，表 13.15 に示す最小 2 乗平均に一致する．

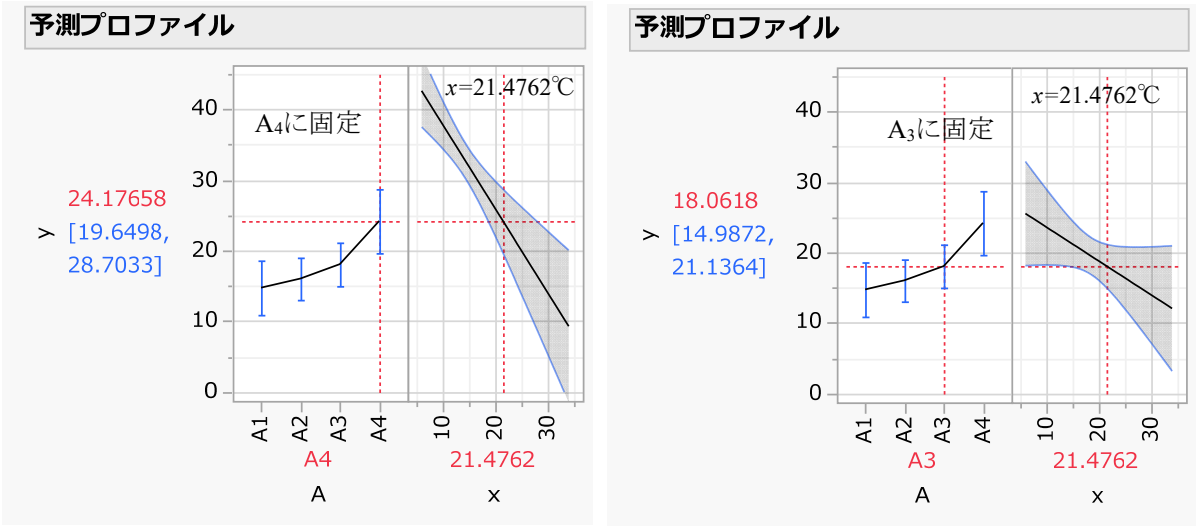


図 13.8　洗浄水の温度 21.476 に対する季節 A_4 および季節 A_3 の予測プロファイル

なぜ，洗浄水温度 x の総平均を各季節の予測値の計算のために使用したのだろうか．これは，最小 2 乗法による回帰分析に際し，y の推定値の分散が最小になるのが x の総平均であるためと理解される．このような，設定方法であると理解しても，各水準の標準誤差，95%信頼区間は，どのようにして求めているかのかの説明は，簡単ではない．説明のためには，これまで繰り返し使ってきたパラメータの共分散行列を使わずにはできない．

表 13.10 は，洗浄水の温度を $x=20.0$ に固定した場合の季節 A の各水準での濃度 y の推定値であるが，総平均 $x=21.4762$ に置き換えれば，「最小 2 乗平均」を推定することができる．表

13.16 に Excel の行列計算による「最小 2 乗平均」の推定と「95%信頼区間」の推定結果を示す．JMP で出力した表 13.15 と完全に一致していることが確かめられる．

表 13.16　季節 A の各水準に対する最小 2 乗平均と 95%信頼区間

No.	A	x_0	a_1	a_2	a_3	x	a_1x	a_2x	a_3x	y^	$Var(y$^$)$	$L95\%$	$U95\%$
1	A_1	1	1	0	0	**21.4762**	21.48	0	0	14.7479	3.5939	10.8952	18.6006
2	A_2	1	0	1	0	**21.4762**	0	21.48	0	16.0280	2.1664	13.0368	19.0192
3	**A_3**	1	0	0	1	**21.4762**	0	0	21.48	**18.0618**	2.2889	**14.9872**	**21.1364**
4	**A_4**	1	-1	-1	-1	**21.4762**	-21.48	-21.48	-21.48	**24.1766**	4.9616	**19.6498**	**28.7033**

Excel による探索的な交互作用解析

Excel を主体にし，交互作用を考慮した解析方法を詳細に示し，JMP および SAS の GLM プロシジャで同様の解析を試みたのであるが，表 13.14 で示したような，共変量 x がどのくらいから有意な差となるかとの推定は，実現できなかった．本質的に交互作用がある場合の“共分散分析”の解析方法として，図 13.7 に示した結果の表示が容易にできたことは，Excel による探索的な解析の賜物であり，また，驚きでもあった．

交互作用があるか否かは，伝統的に分散分析表によっての判断することで対応がなされてきた．交互作用があった場合には，私自身も「適切なグラフ作成して考察する」ことで対応してきた．これは，95%信頼区間が含まれないような図 13.1 の回帰直線をみて「適当」に考察するような態度であり，「統計的」な判断を放棄したことに等しい．少なくとも図 13.6 に示すように，2 本の回帰直線に 95%信頼区間を上書きすることにより「統計的」判断を行うことができるようになる．

先進的な JMP でも，図 13.6 のように 2 本の回帰直線に 95%信頼区間を上書きしようと試みたことはある．ただし，細かなファイル操作などを経て，JMP でのグラフ作成機能では，図 13.6 と同等なもの作成は，やってできないことはないが，生産性が著しく低く，まったく薦められない．これは，Excel の「散布図」のきめ細かさが JMP のスペックより勝っているためである．

13.3. 共変量が 2 変量の場合の探索的な共分散分析

共分散分析は，質的因子を対象にした 1 元配置などの実験に際し，反応変数 y に何らかの影響を与えることが事前に分かっていて測定値を得ることができる量的因子があるが，実験に際して制御し難い変数を「共変量」として解析モデルに組み込む方法である．したがって，共分散分析は，「共分散・分析」ではなく「共（変量を含む）・分散分析」の意味である．

表 13.17 に示すのは，奥野ら(1981）の第 7 章の事例 7.2 であり，ある電気部品の製造工程で，熱処理前の部品寸法 x_1，x_2 が事前に測定され，4 つの炉で熱処理をして電気特性 y を計測した結果である．このデータは，4 つの炉で熱処理した電気部品の電気特性 y に対し，炉によるバイアスが疑われたのであろう．そこで，幾つかの試験片を作成し，炉による熱処理を行い，電気特性 y を測定したと思われる．質的因子である 4 つの炉に対し，さまざま寸法の電気部品を集めて熱処理後の電気特性 y に対し，部品寸法 x_1，x_2 が，特徴的な共変量として認識される．

表 13.17　電気部品と電気特性［奥野(1981)，表 7.6］

	1 号 炉			2 号 炉			3 号 炉			4 号 炉		
i	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
1	31	24	14	29	15	14	11	36	0	16	15	15
2	25	31	24	29	34	21	26	32	13	44	21	29
3	37	22	23	36	25	22	26	26	16	29	41	22
4	20	35	9	16	34	3	18	13	4	18	49	22
5	11	28	0	35	16	12	32	34	9	21	33	23
6	21	39	2	30	26	16	31	26	20	31	7	17
7	9	28	1	28	21	19	30	49	19	21	37	10
8				29	27	19	19	40	7	29	12	22
9				21	11	14				21	52	18
10				22	37	17						
平均	22.0	29.6	10.4	27.5	24.6	15.7	24.1	32.0	11.0	25.6	29.7	19.8
									総平均	25.06	28.71	14.59

共変量の効き方

通常の共分散分析は，制御できる質的因子に対して 1 つの共変量に対する解析法として知られている．この事例は，2 つの共変量があるので，探索的な解析が必要となる．まず，部品寸法 x_1，x_2 のどちらが，共変量として効いているのか，両方が効いているとした場合に，互いに独立なのか，互いに補完し合っているのか，などを最初に明らかにする必要がある．

図 13.9 に JMP の「二変量の関係」で作成した（x_1 と y），（x_2 と y）の散布図上に炉ごとに 50%の確率楕円をあてはめた結果を示す．このような層別確率楕円図は，炉ごとの説明変数

と反応変数 y の関係を一目で把握することができる．なお，探索的な解析の場合に，経験的に50%程度の緩い確率楕円が，探索的な解析では見通しがよいようである．

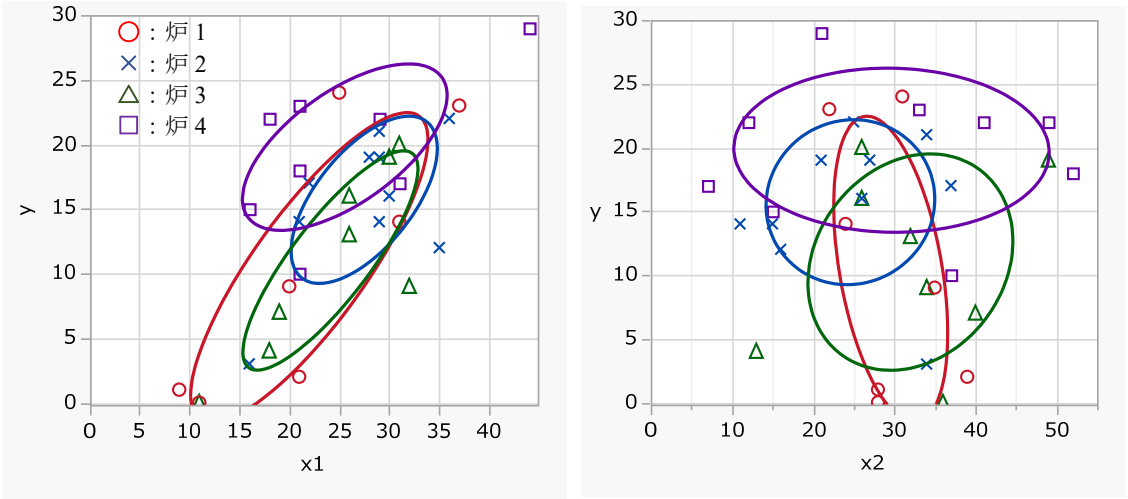


図 13.9　寸法 x_1 と x_2 別の電気特性 y について炉ごとの 50% 確率楕円

寸法 x_1 と y の関係は，どの炉でも同様の正の関連がみられる．炉の 1～3 は重なっており，炉 4 が他に比べて高い電気特性を持っているようである．寸法 x_2 と y の関係は，ほとんど関連がなく，電気的性能に及ぼす影響が単独ではないと判断される．

2 つの特性が互いに関係しているかを調べるためには，交互作用をモデルに入れることで検討できる．また，共変量と y の関係が 1 次的であることの確認のために，解析モデルに x_1^2，x_2^2 および $x_1 \times x_2$ を加えた応答局面法による解析も有益である．ここでは，詳細は省くが JMP の応答局面プロファイルの結果を図 13.10 に示す．

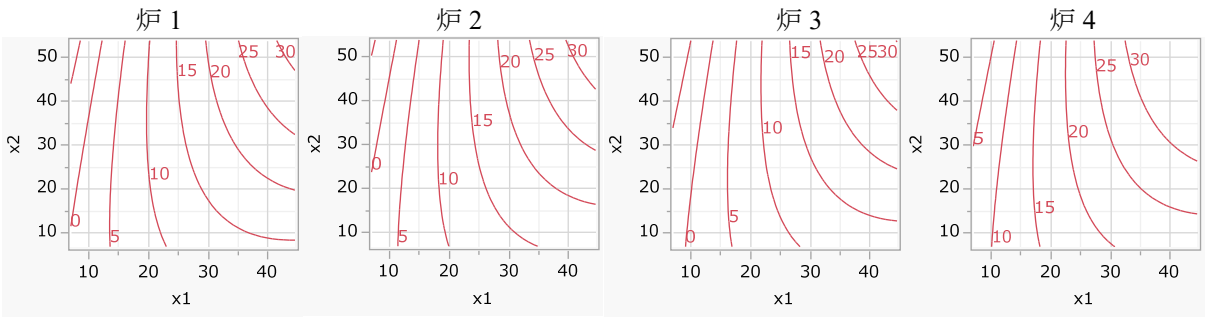


図 13.10　JMP の応答局面解析で作成した等高線プロファイル

応答局面解析の結果から寸法 x_1，x_2 間で明確な交互作用は見出されず，寸法 x_1 の 1 次成分のみが明確な縦縞の等高線として得られている．寸法 x_2 に関しては，寸法 x_1 を固定した場合に電気特性 y に対する影響は，ほとんど見いだせない．

Excel による 2 変量の共分散分析

実際の解析では，寸法 x_1 のみを共変量とした共分散分析の問題に帰着するが，寸法 x_1， x_2 が互いに独立に，電気特性 y に対し直線関係であるとして，2 つの共変量とする共分散分析の事例とする．表 13.18 に Excel シート上に展開したデザイン行列，分析ツールの回帰分析での結果，行列関数で計算したパラメータの共分散行列を示す．寸法 x_2 の p 値は，0.2129 と有意でないことが示されている．

表 13.18　分析ツールの回帰分析による結果およびパラメータの共分散行列

		—— デザイン行列 ——						
			—— 炉 ——			— 寸法 —		反応
i	炉	x_0	a_1	a_2	a_3	x_1	x_2	y
1	A1	1	1	0	0	31	24	14
2		1	1	0	0	25	31	24
3		1	1	0	0	37	22	23
4		1	1	0	0	20	35	9
5		1	1	0	0	11	28	0
6		1	1	0	0	21	39	2
7		1	1	0	0	9	28	1
8	A2	1	0	1	0	29	15	14
9		1	0	1	0	29	34	21
10		1	0	1	0	36	25	22
11		1	0	1	0	16	34	3
12		1	0	1	0	35	16	12
13		1	0	1	0	30	26	16
14		1	0	1	0	28	21	19
15		1	0	1	0	29	27	19
16		1	0	1	0	21	11	14
17		1	0	1	0	22	37	17
18	A3	1	0	0	1	11	36	0
19		1	0	0	1	26	32	13
20		1	0	0	1	26	26	16
21		1	0	0	1	18	13	4
22		1	0	0	1	32	34	9
23		1	0	0	1	31	26	20
24		1	0	0	1	30	49	19
25		1	0	0	1	19	40	7
26	A4	1	-1	-1	-1	16	15	15
27		1	-1	-1	-1	44	21	29
28		1	-1	-1	-1	29	41	22
29		1	-1	-1	-1	18	49	22
30		1	-1	-1	-1	21	33	23
31		1	-1	-1	-1	31	7	17
32		1	-1	-1	-1	21	37	10
33		1	-1	-1	-1	29	12	22
34		1	-1	-1	-1	21	52	18

Excelによる回帰分析

分散分析表　「定数に0を使用」をonとする

	自由度	変動	分散	分散比	有意 F
回帰	6	8554.77	1425.80	60.38	0.0000
残差	28	661.23	**23.62**	$\sigma^{\wedge 2}$	
合計	34	9216			

	係数	標準誤差	t	P-値	
切片	0	#N/A	#N/A	#N/A	
x0	**-5.4994**	4.1387	-1.3288	0.1946	β_0
a1	**-1.9702**	1.5777	-1.2487	0.2221	β_1
a2	**0.0893**	1.4326	0.0623	0.9507	β_2
a3	**-3.0839**	1.4964	-2.0609	0.0487	β_3
x1	**0.6763**	0.1134	5.9624	0.0000	β_4
x2	**0.1021**	0.0801	1.2747	0.2129	β_5

パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{\wedge 2}$

	x_0	a_1	a_2	a_3	x_1	x_2
β_0	**17.1288**	-0.7848	-0.1288	0.5044	-0.3821	-0.2399
β_1	-0.7848	**2.4892**	-0.8111	-0.8578	0.0346	0.0022
β_2	-0.1288	-0.8111	**2.0524**	-0.7055	-0.0253	0.0221
β_3	0.5044	-0.8578	-0.7055	**2.2391**	0.0020	-0.0181
β_4	-0.3821	0.0346	-0.0253	0.0020	**0.0129**	0.0022
β_5	-0.2399	0.0022	0.0221	-0.0181	0.0022	**0.0064**

対比型デザイン変数

A	a_1	a_2	a_3
A_1	**1**	0	0
A_2	0	**1**	0
A_3	0	0	**1**
A_4	**-1**	**-1**	**-1**

$t(0.08,28)$= **2.0484**

平均x_1= **25.0588**

平均x_2= **28.7059**

回帰パラメータの推定値を用いた回帰式は，

$$\hat{y} = \hat{\beta}_0 x_0 + \hat{\beta}_1 a_1 + \hat{\beta}_2 a_2 + \hat{\beta}_3 a_3 + \hat{\beta}_4 x_1 + \hat{\beta}_5 x_2$$
$$= -5.4994 - 1.9702 a_1 + 0.0893 a_2 - 3.0839 a_3 + 0.6763 x_1 + 0.1021 x_2$$

である．$\hat{\beta}_0 = -5.4994$ は，$a_1 = a_2 = a_3 = x_1 = x_2 = 0$ の場合でなので，寸法 x_1 と x_2 の共通の Y 切片であり，また，炉 A に関しては効果の平均の推定値となっている．$\hat{\beta}_0 + \hat{\beta}_1 = -5.4994 - 1.9702$ は，寸法 x_1 と x_2 の切片における A_1 の推定値であり，$\hat{\beta}_4 = 0.6763$ は，寸法 x_1 が切片 0 から 1 単位増加した時の増分である．このような回帰パラメータの推定値の解釈は，結果を説明する際に，現実から遊離していていて理解が得られない．

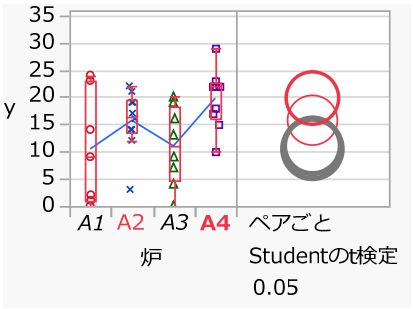
最小 2 乗平均

まず，表 13.19 に示すように炉別の基本統計量を示し，ボックス・プロットなどを用いて，共変量を含まない比較を行う．炉 A_4 が熱処理後の電気特性に対する性能が高く，A_2 は平均が 2 番目に高いが，飛び離れ値の存在が気になり，炉 A_1 と炉 A_3 が，低めになっている．

表 13.19　炉の番号別の基本統計量（単純平均）

平均と標準偏差

水準	数	平均	標準偏差	標準誤差	下側95%	上側95%
A1	7	10.43	10.21	3.86	0.98	19.87
A2	10	15.70	5.50	1.74	11.77	19.63
A3	8	11.00	7.21	2.55	4.97	17.03
A4	9	19.78	5.47	1.82	15.57	23.98



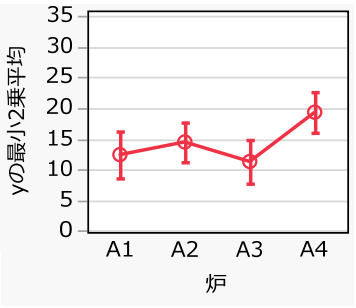
炉によって，試験片の寸法が不揃いであり，図 13.9 に示したように，寸法 x_1 が共変量なので，共分散分析によって，試験片の平均値を統計的に調整し，調整済み平均値と 95%信頼区間を示したい．さらに，ある炉を基準とした調整済み平均値の 95%信頼区間も示し，総合的に評価を行いたい．

SAS および JMP ユーザならば，迷わず最小 2 乗平均 Lsmeans を出力して評価をするに違いない．JMP を使えば，表 13.20 に示すように炉の番号別の最小 2 乗平均と 95%信頼区間，さらにグラフが標示される．さて，さまざまな関係者から「最小 2 乗平均」とは何ですか，どのように計算されたものですか，との質問にどのように答えたら良いのであろうか．表 13.20 に炉の番号別の最小 2 乗平均と 95%信頼区間を示す．表の右端に「算術平均」が示されており，最小 2 乗平均とは明らかに異なる．

表 13.20　炉 A の番号別の最小 2 乗平均と 95%信頼区間（2 つの共変量で調整済み）

最小2乗平均表

水準	最小2乗平均	標準誤差	下側95%	上側95%	平均
A1	12.41	1.87	8.58	16.23	10.43
A2	14.47	1.58	11.23	17.71	15.70
A3	11.30	1.74	7.74	14.85	11.00
A4	19.34	1.62	16.02	22.67	19.78



回帰パラメータの推定値を用いた推論の場合には，$x_1=0$，$x_2=0$ の場合の共通の Y 切片上での炉の番号別の推定値となり，現実データと遊離しているので，説明がややこしくなる．そのために，「最小 2 乗平均」は，炉の推定値の分散が最小になると期待される x_1 の総平均 20.0588 と x_2 の総平均 28.7059 としたときの炉 A の水準ごとの反応 y の推定値を「最小 2 乗平均」としている．炉 A_1 の場合は，

$$\hat{y}_{A_1}=\hat{\beta}_0x_0+\hat{\beta}_1a_1+\hat{\beta}_2a_2+\hat{\beta}_3a_3+\hat{\beta}_4x_1+\hat{\beta}_5x_2$$

	x_0	a_1	a_2	a_3	x_1	x_2	$\hat{\beta}$		
y^A1=	1	1	0	0	25.0588	28.7059	-5.4994	=	12.4088
							-1.9702		
							0.0893		
							-3.0839		
							0.6763		
							0.1021		

と推定される．表 13.21 に A_1 以外の炉について最小 2 乗平均を推定した結果を示す．JMP で求めた表 13.20 と完全に一致していることが確認される．分散は，これまでと同様に，

$$Var(\hat{y}_{A_1})=\boldsymbol{x}_{A1}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{A1}{}^T=3.49$$

によって計算されている．

表 13.21　Excel による炉 A の番号別の最小 2 乗平均と 95%信頼区間

			—— 炉 ——			寸法		推定値	分散	—誤差 —		95%信頼区間	
	炉	x_0	a_1	a_2	a_3	x_1	x_2	y^	$Var(y$^$)$	SE	$t_\alpha SE$	L95%	U95%
						25.06	**28.71**						
1	A_1	1	1	0	0	25.06	28.71	**12.41**	3.49	1.87	3.83	8.58	16.23
2	A_2	1	0	1	0	25.06	28.71	**14.47**	2.50	1.58	3.24	11.23	17.71
3	A_3	1	0	0	1	25.06	28.71	**11.30**	3.02	1.74	3.56	7.74	14.85
4	A_4	1	-1	-1	-1	25.06	28.71	**19.34**	2.64	1.62	3.33	16.02	22.67

注）　$t_\alpha SE$ は，Excel で 95%信頼区間のひげを付けるために計算している．

JMP および SAS での「最小 2 乗平均」，は，合理的ではあるが，出力された結果について，手計算で計算結果を検証することは，なかなか困難である．ここに示したように Excel の行列計算を用いることにより，「最小 2 乗平均」の内部を可視化することができる．表 13.21 で，$x_1 = 25$，$x_2 = 30$ のように Excel シート上のデータを置き換えることにより，切れの良い数字での推定値が得られ，結果の説明も容易になる．

水準間の差の推定

炉の水準ごとの推定値に対し，炉 A_4 を基準にし，他の炉との差および 95%信頼区間を推定したい．また，炉 A_1，A_2，A_3 水準は，互いに 95%信頼区間が推定値に互いにかぶり合い，統計的には差がないことが明らかであるが，炉 A_4 との差も考察のために求めたいとする．水準間の差とその 95%信頼区間を求めたい場合には，比較したい水準のデザイン変数の差を計算すればよい．炉 A_3 と炉 A_4 の差の場合であれば，

	炉	x_0	a_1	a_2	a_3	x_1	x_2
	A_3	1	0	0	1	25.0	30.0
−)	A_4	1	-1	-1	-1	25.0	30.0
=	A_3-A_4	0	1	1	2	0	0

とする．推定値も分散の計算も，表 13.22 に示すように，これまでと同様に計算すればよい．炉 A の水準間の推定に際しては，共変量の大きさが反映されるが，交互作用がない主効果モ

表 13.22　Excel による炉 A_4 との差の最小 2 乗平均と 95%信頼区間

			—— 炉 ——			寸法		推定値	分散	—誤差 —		95%信頼区間	
	炉	x_0	a_1	a_2	a_3	x_1	x_2	$y_{差}$^	$Var(y_{差}$^)	SE	$t_\alpha SE$	L95%	U95%
	A_4	**1**	**-1**	**-1**	**-1**	**25.0**	**30.0**						
A_1	−A_4	0	2	1	1	0.0	0.0	**-6.93**	6.16	2.48	5.08	-12.02	-1.85
A_2	−A_4	0	1	2	1	0.0	0.0	**-4.88**	5.16	2.27	4.65	-9.53	-0.22
A_3	−A_4	0	1	1	2	0.0	0.0	**-8.05**	5.62	2.37	4.86	-12.91	-3.19
A_4	−A_4	0	0	0	0	0.0	0.0	**0.00**	0.00	0.00	0.00	0.00	0.00

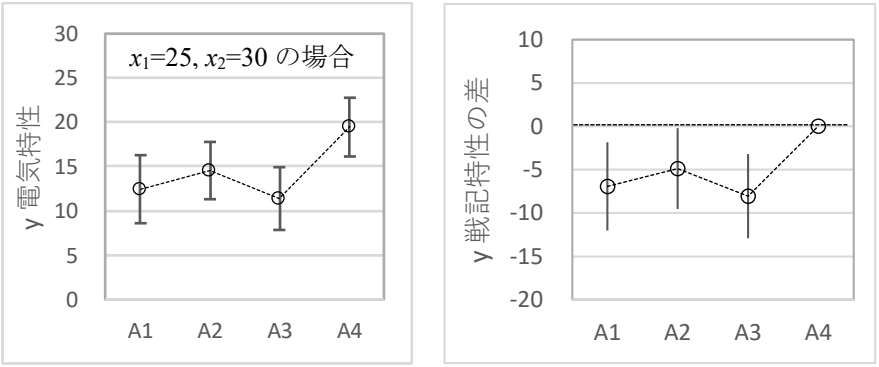


図 13.11　炉 A の水準間の最小 2 乗平均および炉 A_4 との差の 95%信頼区間

デルの場合での水準間の差における共変量の大きさはゼロとなっており，共変量x_1およびx_2の大きさに関わらず差の推定値も差の分散も同じである．

デザイン変数の活用による A_4 との差

ここに示した共分散分析は，共変量が 2 つ，炉 A が 4 水準で，複雑な共分散分析の事例となっている．一般的な共分散分析は，共変量が一つで，2 群に対して平行な直線をあてはめる場面設定であり，通常の解析手順は，分散分析表を主体にした解析結果が示されている．推定された 2 本の回帰直線間の統計的な判断は，分散分析表の群間差の p 値に基づいている．

推定された 2 つの Y 切片の差の 95%信頼区間を示し，結果を考察するような手順とはなっていないが，デザイン変数を，工夫することにより表 13.22 に示したと同等の結果を得ることができる．Excel の回帰分析で，表 13.23 に示すような炉 A_4 を基準とする（1，0）型のデザイン変数にすることにより，表 13.24 に示すように A_1 と A_4 の差，A_2 と A_4 の差，A_3 と A_4 の差についての *SE* および 95%信頼区間も同時推定ができる．

表 13.23　炉 A_4 を基準とするデザイン変数

A	x_0	a_1	a_2	a_3	推定値
A_1	1	1	0	0	A1-A4
A_2	1	0	1	0	A2-A4
A_3	1	0	0	1	A3-A4
A_4	1	0	0	0	**基準**

表 13.24 に示した a_3 の行の「係数」欄には，炉 A_3 と炉 A_4 の差の推定値 –8.0468（–12.9057, –3.1914）が推定されていて，表 13.22 の結果に一致する．

表 13.24　Excel の回帰分析による炉 A_4 を基準とする他の水準との差と 95%信頼区間

	自由度	変動	分散	分散比	有意 F
回帰	6	8554.77	1425.80	60.38	0.0000
残差	28	661.23	**23.62**		
合計	34	9216			

		係数	標準誤差	t	P-値	下限 95%	上限 95%
	切片	0	#N/A	#N/A	#N/A	#N/A	#N/A
β_0	x0	-0.5347	4.4698	-0.1196	0.9056	-9.6907	8.6213
β_1	a1	**-6.9349**	**2.4823**	-2.7938	0.0093	-12.0196	-1.8502
β_2	a2	**-4.8754**	**2.2707**	-2.1471	0.0406	-9.5266	-0.2242
β_3	a3	**-8.0486**	**2.3712**	-3.3943	0.0021	-12.9057	-3.1914
β_4	x1	0.6763	0.1134	5.9624	0.0000	0.4439	0.9086
β_5	x2	0.1021	0.0801	1.2747	0.2129	-0.0620	0.2662

予測プロファイルによる共変量の影響の可視化

共変量の効き方は，表 13.18 の回帰係数の推定値と標準誤差によって評価できるが，共変量の変化によって電気特性 y に対しどのような状況かを可視化したい．図 13.12 に JMP の予測プロファイルで（炉 A=A$_4$，$x_1=25$，$x_2=30$）を選択した結果を示す．部品寸法 x_1 に関するプロファイルは，炉を A$_4$ に固定し，部品寸法 x_2 を 30 に固定した場合であり，明らかな正の関連が読み取れる．

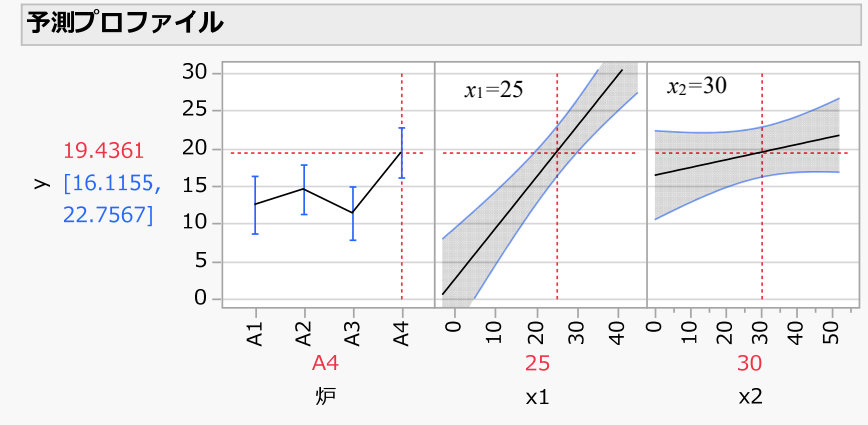


図 13.12　炉 A$_4$ の場合の共変量の影響

Excel を用いて内部での計算方法を表 13.25 に示す. 基本は, 表 13.21 と同様であり, 表 13.24 に示すように炉 A$_4$ と $x_2=30$ に固定し x_1 を (0, 10, . . . , 50) と変化させた場合の炉 A$_4$ と $x_1=25$

表 13.25　炉 A$_4$ に対する共変量の変化に対する電気特性 y に対する影響

			—— 炉 ——			寸法		推定値	分散	誤差	95%信頼区間	
	炉	x_0	a_1	a_2	a_3	x_1	x_2	y^	$Var(y$^$)$	SE	$L95\%$	$U95\%$
x_1=25	A$_1$	1	1	0	0	25.0	30.0	12.50	3.50	1.87	8.67	16.33
x_2=30	A$_2$	1	0	1	0	25.0	30.0	14.56	2.57	1.60	11.28	17.84
	A$_3$	1	0	0	1	25.0	30.0	11.39	2.98	1.73	7.85	14.92
	A$_4$	1	**-1**	**-1**	**-1**	25.0	30.0	**19.44**	2.63	1.62	16.12	22.76
x_2=30	A4	1	-1	-1	-1	**0.00**	30.00	2.53	10.99	3.32	-4.26	9.32
	A4	1	-1	-1	-1	**10.00**	30.00	9.29	5.72	2.39	4.39	14.19
	A4	1	-1	-1	-1	**25.00**	30.00	**19.44**	2.63	1.62	16.12	22.76
	A4	1	-1	-1	-1	**30.00**	30.00	22.82	2.89	1.70	19.34	26.30
	A4	1	-1	-1	-1	**40.00**	30.00	29.58	5.33	2.31	24.85	34.31
	A4	1	-1	-1	-1	**50.00**	30.00	36.34	10.35	3.22	29.75	42.93
x_1=25	A4	1	-1	-1	-1	25.00	**0.00**	16.37	8.35	2.89	10.45	22.29
	A4	1	-1	-1	-1	25.00	**10.00**	17.39	5.16	2.27	12.74	22.05
	A4	1	-1	-1	-1	25.00	**20.00**	18.41	3.25	1.80	14.72	22.11
	A4	1	-1	-1	-1	25.00	**30.00**	**19.44**	2.63	1.62	16.12	22.76
	A4	1	-1	-1	-1	25.00	**40.00**	20.46	3.29	1.81	16.74	24.17
	A4	1	-1	-1	-1	25.00	**50.00**	21.48	5.23	2.29	16.79	26.16

を固定し x_2 を（0，10, . . . , 50）と変化させた場合に電気特性 y に対する推定値，分散，*SE*，および，95%信頼区間の計算をしている．こ推定結果を Excel の折れ線グラフ，および，散布図でグラフ化した結果を図 13.13 に示す．

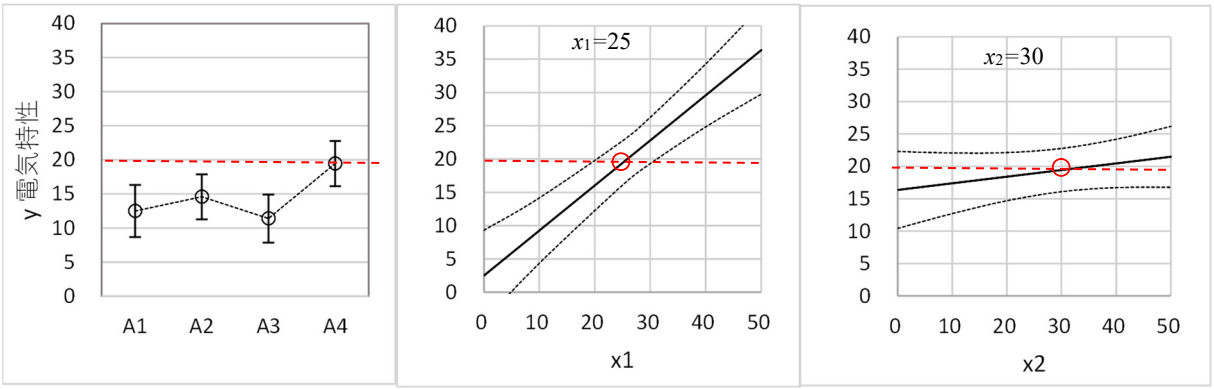


図 13.13　Excel による炉 A_4 に対応する共変量の予測プロファイル

炉 A の予測プロファイルは，部品寸法を（$x_1 = 25$，$x_2 = 30$）に固定した場合，部品寸法 x_1 の予測プロファイルは，（炉 A=A_4，$x_2 = 30$）に固定した場合，部品寸法 x_2 の予測プロファイルは，（炉 A=A_4，$x_1 = 25$）に固定した場合の推定値であり，互いに関連している．炉 A=A_4，$x_1 = 25$，$x_2 = 30$ とそれぞれが一致する場合の予測値は 19.44 と揃っていることが確認される．

JMP が使えるならば，Excel でわざわざ作成する必要はないが，「予測プロファイル」は統計の一般用語ではないことを認識し，他者に対する資料として作成する場合には，何らかの解説を付ける必要がある．他の統計ソフトでは提供されていないので，なおさら，Excel による計算方法と作図法を示す必要がある．

Excel による作図のヒントは，本章以外でも第 7.2 節，第 9.2 節，第 10.3 節，第 12.3 節で例示しているので参考にしてもらいたい．

13.4.　繰返しが不揃いな 2 因子の共分散分析における最小 2 乗平均

検索エンジンを使って,「最小 2 乗平均」について書かれている論文を探したところ, 守屋・広岡（2018）の「R パッケージを用いた最小 2 乗分散分析と最小 2 乗平均値の算出」が見いだされた．彼らは，データ数が不揃いの実験データの解析のために SAS の GLM プロシジャに代えて，R の lsmeans パッケージの使用法について論じている．

守屋らは,「各要因のグループ（水準）内のデータ数が等しい釣り合い型データ（balanced data）に対しては通常の生物統計学の教科書で紹介されている分散分析法が適用でき, EXCEL のような広く普及している表計算ソフトでも対応できる．一方，データ数が等しくない不釣り合い型データ（unbalanced data）や分析に共変量を含むケースは通常の分散分析の手法ではなく最小 2 乗分散分析法を用いる必要がある」ことを提示している．

そのために,「SAS の GLM プロシジャで実行されることが多かった．しかし，SAS は有料でしかも高価であったため，実行できる環境は非常に限定されていた」と述べ，最近，R 環境においても lsmeans パッケージが提供されたので, SAS の GLM プロシジャでの Lsmeans と比較検討した結果を報告している．論文には，例示として，（因子 A，因子 B，共変量 x，反応変数 y）についての 15 個のデータが示されている．

表 13.26　守屋らの 2 因子共分散データ

因子A	因子B	共変量 x	反応 y
A_1	B_1	350	970
		400	1000
		360	980
	B_2	350	980
		340	970
A_2	B_1	390	990
		340	950
	B_2	410	980
		430	990
		390	980
A_3	B_1	400	990
		320	940
	B_2	330	930
		390	1000
		420	1000
	平均	374.67	976.67
	反応 y :1 日当たりの増体量		

解析モデルとして，交互作用を含むモデル

反応 y = A + B + A×B + 共変量 x

についての最小 2 乗平均を提示し，R の lsmeans パッケージの結果が，SAS の GLM プロシジャの Lsmeans の結果と一致することを報告している．

対比型デザイン変数を用いた場合の最小 2 乗平均

表 13.26 のデータに対して，対比型のデザイン変数を表 13.27 に示す．交互作用は，因子 A と因子 B の水準の組み合わせで，A×B に対し（a_1b_1， a_2b_1）のようにデザイン変数間の積で定義する．

表 13.27　（1，-1）対比型のデザイン変数

A	a_1	a_2
A_1	**1**	0
A_2	0	**1**
A_3	**-1**	**-1**

B	b_1
B_1	**1**
B_2	**-1**

A×B		a_1b_1	a_2b_1
A_1	B_1	**1**	0
	B_2	**-1**	0
A_2	B_1	0	**1**
	B_2	0	**-1**
A_3	B_1	**-1**	**-1**
	B_2	**1**	**1**

第 13.2 節では，質的因子が 1 つで共変量が 1 つの場合であり，第 13.3 節では質的因子が 1 つで共変量が 2 つの場合であった．ここでは，質的因子が 2 つで交互作用を含め共変量が 1 つの場合であり，最小 2 乗平均はどのように求められているのだろうか．表 13.28 にデザイン変数を組み込み，Excel の回帰分析を適用した結果を示す．

表 13.28　対比型のデザイン変数用いた Excel による回帰係数の推定

			——— デザイン行列 ***X*** ———							
				A		B	A×B		共変量	反応
i	A	B	x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x	y
1	A_1	B_1	1	1	0	1	1	0	350	970
2			1	1	0	1	1	0	400	1000
3			1	1	0	1	1	0	360	980
4		B_2	1	1	0	-1	-1	0	350	980
5			1	1	0	-1	-1	0	340	970
6	A_2	B_1	1	0	1	1	0	1	390	990
7			1	0	1	1	0	1	340	950
8		B_2	1	0	1	-1	0	-1	410	980
9			1	0	1	-1	0	-1	430	990
10			1	0	1	-1	0	-1	390	980
11	A_3	B_1	1	-1	-1	1	-1	-1	400	990
12			1	-1	-1	1	-1	-1	320	940
13		B_2	1	-1	-1	-1	1	1	330	930
14			1	-1	-1	-1	1	1	390	1000
15			1	-1	-1	-1	1	1	420	1000

Excelによる回帰分析

分散分析表（「定数に0を使用」をon）

	自由度	変動	分散	
回帰	7	14314030	2044861	
残差	8	670.4128	**83.8016**	$\sigma^{\wedge 2}$
合計	15	14314700		

	係数	標準誤差	t	P-値
切片	0	#N/A	#N/A	#N/A
x0	**718.1167**	32.6778	21.9757	0.0000
a1	**13.4238**	3.6307	3.6973	0.0061
a2	**-9.8560**	3.6833	-2.6759	0.0281
b1	**1.8400**	2.4822	0.7413	0.4797
a1b1	**-6.3315**	3.8031	-1.6648	0.1345
a2b1	**7.0782**	3.6833	1.9217	0.0909
x	**0.6927**	0.0877	7.8996	0.0000

パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ は，デザイン行列 $\boldsymbol{X}$ と誤差分散 $\hat{\sigma}^2$ を用いて

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$$

$$= \text{Minverse}(\text{Mmult}(\text{Transpose}(\boldsymbol{X}\text{の範囲}), \boldsymbol{X}\text{の範囲}))*\hat{\sigma}^2$$

として表 13.29 に示すように計算される．

表 13.29　対比型デザイン行列に対するパラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$

		—— A ——		B	—— A×B ——		共変量
	x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x
β_0	**1067.84**	-40.4806	45.2430	-18.6617	53.2160	-44.4671	-2.8575
β_1	-40.4806	**13.1821**	-7.5441	-0.8258	-2.8635	2.1125	0.1089
β_2	45.2430	-7.5441	**13.5665**	-0.0356	2.7211	-0.3755	-0.1217
β_3	-18.6617	-0.8258	-0.0356	**6.1613**	-0.9824	0.8115	0.0513
β_4	53.2160	-2.8635	2.7211	-0.9824	**14.4635**	-8.1527	-0.1474
β_5	-44.4671	2.1125	-0.3755	0.8115	-8.1527	**13.5665**	0.1217
β_6	-2.8575	0.1089	-0.1217	0.0513	-0.1474	0.1217	**0.0077**
		パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{\wedge 2}$					

表 13.30 に対比型デザイン変数を用いた因子 A，因子 B および交互作用 A×B の各水準の Lsmeans を示す．この計算は，Lsmeans のデザイン変数を $\boldsymbol{x}_i$ ベクトルとし，推定されたパラメータ $\hat{\boldsymbol{\beta}}$ としたときに，因子 A の第 1 水準 A_1 は，$a_1=1$，それ以外は 0，共変量 x については，総平均 374.67 とし，991.06 が Lsmeans として推定されている．

$\hat{y}_{A1} = \boldsymbol{x}_{16}\hat{\boldsymbol{\beta}} =$

x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x	$\hat{\boldsymbol{\beta}}$		
1	1	0	0	0	0	374.67	718.1167	=	**991.06**
							13.4238		
							-9.8560		$y^{\wedge}_{A1}$
							1.8400		
							-6.3315		
							7.0782		
							0.6927		

表 13.30　最小 2 乗平均・Lsmeans の計算結果

			——— デザイン行列 $\boldsymbol{X}$ ———							反	最小2乗		標準		
				— A —		B	A×B		共変量	応	平均	分散	誤差	95%信頼区間	
i	A	B	x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x	y	$y^{\wedge}$	$Var(y^{\wedge})$	SE	$L95\%$	$U95\%$
16	**A_1**		1	1	0	0	0	0	374.67	-	**991.06**	**19.7243**	**4.4412**	980.82	1001.30
17	A_2		1	0	1	0	0	0	374.67	-	967.78	18.7249	4.3272	957.80	977.76
18	A_3		1	-1	-1	0	0	0	374.67	-	974.07	17.6261	4.1983	964.38	983.75
19	B_1		1	0	0	1	0	0	374.67	-	979.47	13.1335	3.6240	971.12	987.83
20	B_2		1	0	0	-1	0	0	374.67	-	975.79	10.9665	3.3116	968.16	983.43
21	A_1	B_1	1	1	0	1	1	0	374.67	-	**986.57**	**28.1013**	5.3011	**974.34**	**998.79**
22		B_2	1	1	0	-1	-1	0	374.67	-	995.55	48.6673	6.9762	979.46	1011.64
23	A_2	B_1	1	0	1	1	0	1	374.67	-	976.70	42.6192	6.5283	961.64	991.75
24		B_2	1	0	1	-1	0	-1	374.67	-	958.86	37.5322	6.1264	944.73	972.99
25	A_3	B_1	1	-1	-1	1	-1	-1	374.67	-	975.16	43.5546	6.5996	959.94	990.38
26		B_2	1	-1	-1	-1	1	1	374.67	-	972.97	28.1526	5.3059	960.74	985.21

さらに分散と *SE* は，パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$ を用いて，

$$Var(\hat{y}_{A_1}) = \boldsymbol{x}_{16}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{16}^{T}$$

$=$

x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x
1	1	0	0	0	0	374.67

1067.8	-40.48	45.24	-18.66	53.22	-44.47	-2.86	1
-40.5	**13.18**	-7.54	-0.83	-2.86	2.11	0.11	1
45.2	-7.54	**13.57**	-0.04	2.72	-0.38	-0.12	0
-18.7	-0.83	-0.04	**6.16**	-0.98	0.81	0.05	0
53.2	-2.86	2.72	-0.98	**14.46**	-8.15	-0.15	0
-44.5	2.11	-0.38	0.81	-8.15	**13.57**	0.12	0
-2.9	0.11	-0.12	0.05	-0.15	0.12	**0.01**	375

パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^{\wedge 2}$

	$Var(y^{\wedge}_{A1})$	
=	**19.7243**	
	SE	
	4.4412	

として計算されている．表 13.31 に守屋ら（2018）の R の lsmeans パッケージによる最小 2 乗平均の算出例を示す．Excel で求めた Lsmeans と *SE* が，一致することが確かめられる．

表 13.31　R の lsmeans パッケージによる最小 2 乗平均の算出例

FactorA	lsmean	SE	df	lower.CL	upper.CL	.group
A1: a	**991.0573**	**4.441210**	8	980.8159	1001.2988	2
A2: b	967.7775	4.327225	8	957.7989	977.7561	1
A3: c	974.0657	4.198345	8	964.3843	983.7472	12

守屋ら（2018）の表 3 の一部を抜粋・編集

探索的な解析では，何らかの結果のグラフ表示が欠かせない．JMP の予測プロファイルよるグラフ表示，グラフ・ビルダーによる層別散布図を示そう．

JMP による 2 因子交互作用に共変量を含む最小 2 乗法での解析で「予測プロファイル」を選択すると，図 13.14 に示すような図が表示される．マウスで因子 A の A_1 を選択し，因子 B の B_1 を選択した結果が示されている．共変量 x は，最小 2 乗平均の推定のために 374.67 が自動的にセットされている．この条件下で反応 y の推定値が，986.67（974.34，998.79）と表示され，

表 13.30 に示されている推定値 $\hat{y}$ と分散 $Var(\hat{y})$ を使い 95%信頼区間は

$$\begin{aligned}(L95\%, U95\%) &= \hat{y}_{A_1B_1} \pm t(0.05,8)\sqrt{Var(\hat{y}_{A_1B_1})} \\ &= 986.57 \pm 2.3060\sqrt{28.1013} \\ &= (974.34,\ \ 998.79)\end{aligned}$$

となり，図 13.14 の上段の結果に一致する．因子 A は A_1 水準のまま，因子 B を B_2 水準に変更したのが図 13.14 の下段である．因子 A のプロファイルの上段と下段で明らかに異なるのは，交互作用を解析モデルに含めたからである．

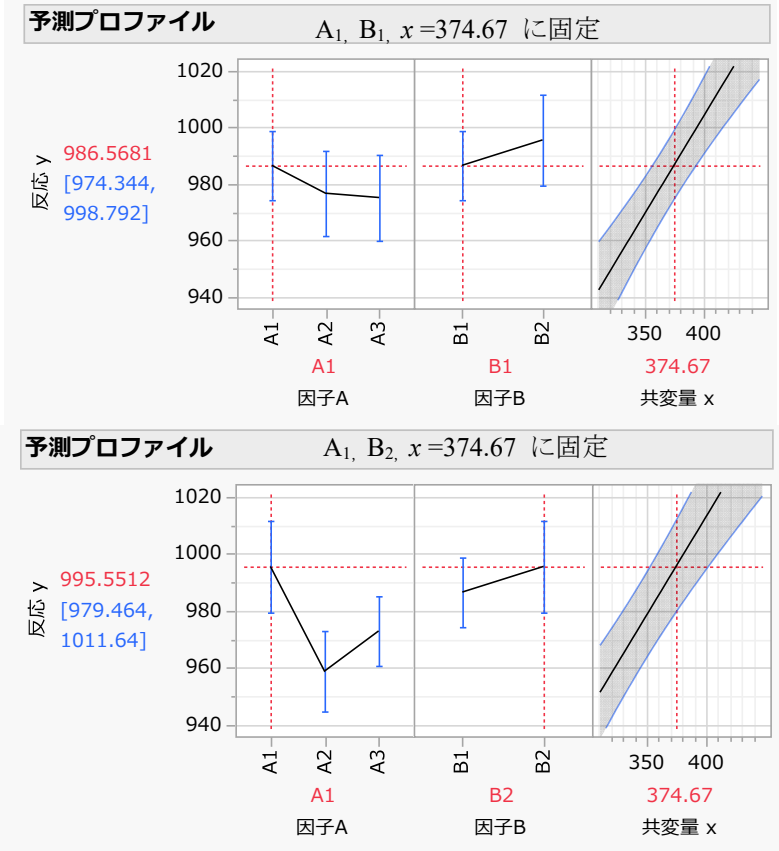


図 13.14　交互作用モデルに対する最小 2 乗平均による予測プロファイル

探索的な解析に際し，因子 A と因子 B を組み合わせた層別散布図により，データの全体像を把握することの必要性をこれまでも示してきた．図 13.15 に JMP のグラフ・ビルダーでの結果を示す．最小 2 乗平均の図 13.14 と比べることにより，探索的な解析結果についての理解が深まると思われる．

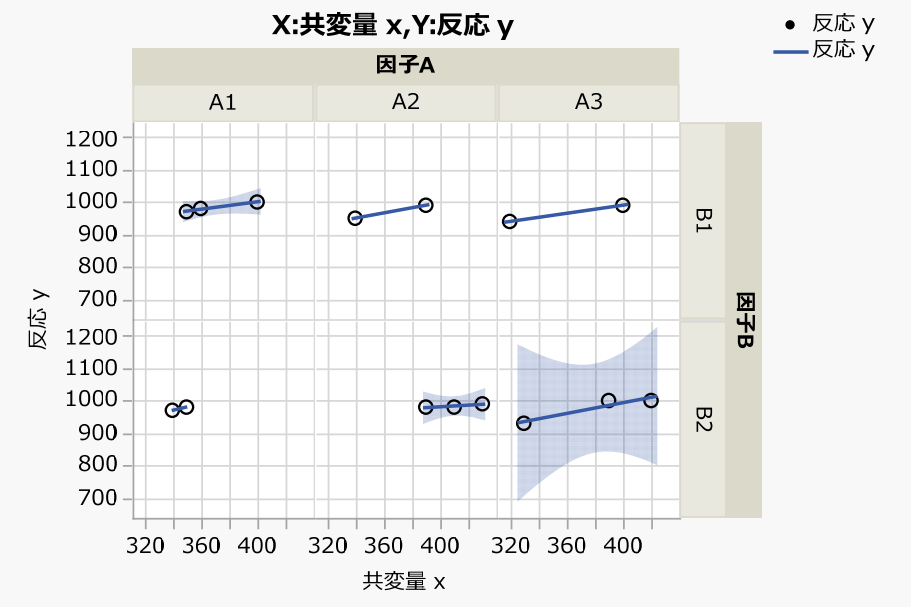


図 13.15　グラフ・ビルダーによる層別散布図によるデータの俯瞰

SAS の GLM プロシジャでのデザイン変数

これまでは，(1，–1）対比型のデザイン変数を用いた Lsmeans を Excel で再現してきたのであるが，SAS の GLM プロシジャでのデザイン変数は，各因子の最後の水準を基準としているので，表 13.32 で示す（1，0）型デザイン変数を使って Lsmeans を再現する．

表 13.32　最後の水準を基準とした（1，0)型デザイン変数

A	a_1	a_2
A_1	**1**	0
A_2	0	**1**
A_3	0	0
平均	1/3	1/3

B	b_1
B_1	**1**
B_2	0
平均	1/2

A×B		a_1b_1	a_2b_1
A_1	B_1	**1**	0
	B_2	0	0
A_2	B_1	0	**1**
	B_2	0	0
A_3	B_1	0	0
	B_2	0	0

表 13.33 に示すように，因子 A の各水準の Lsmeans を求めるためには，因子 A のデザイン変数 a_1 と a_2 を使い，因子 B の b_1 には，デザイン変数の平均としての (1 / 2) を用いて，交互作用は，Lsmeans は，$\hat{y}_{1.16} = \boldsymbol{x}_{1.16}\hat{\boldsymbol{\beta}}$ として計算されている．もちろん，
表 13.30 に示した lsmeans パッケージの結果に一致している．因子 B の場合は，因子 A のデザイン変数 a_1 と a_2 には，3 水準の平均 (1 / 3) を用いる．交互作用については，因子 A と因子 B のデザイン変数の組み合わせによる積のままでよい．

表 13.33　GLM 式デザイン行列に対する Lsmeans を求めるためのデザイン変数

			———— デザイン変数 ————								最小2乗		標準
				— A —		B	A×B		共変量	反応	平均	分散	誤差
i	A	B	x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x	y	y^	$Var(y$^$)$	SE
1.16	A_1		1	1	0	**0.50**	**0.50**	**0**	374.67	-	991.06	19.724	4.441
1.17	A_2		1	0	1	**0.50**	**0**	**0.50**	374.67	-	967.78	18.725	4.327
1.18	A_3		1	0	0	**0.50**	**0**	**0**	374.67	-	974.07	17.626	4.198
1.19	B_1		1	**0.33**	**0.33**	1	**0.33**	**0.33**	374.67	-	979.47	13.133	3.624
1.20	B_2		1	**0.33**	**0.33**	0	**0**	**0**	374.67	-	975.79	10.967	3.312
1.21	A_1	B_1	1	1	0	1	1	0	374.67	-	986.57	28.101	5.301
1.22		B_2	1	1	0	0	0	0	374.67	-	995.55	48.667	6.976
1.23	A_2	B_1	1	0	1	1	0	1	374.67	-	976.70	42.619	6.528
1.24		B_2	1	0	1	0	0	0	374.67	-	958.86	37.532	6.126
1.25	A_3	B_1	1	0	0	1	0	0	374.67	-	975.16	43.555	6.600
1.26		B_2	1	0	0	0	0	0	374.67	-	972.97	28.153	5.306

	β_0^	β_1^	β_2^	β_3^	β_4^	β_5^	β_6^
$\boldsymbol{\beta}^{\wedge T}$=	713.4557	22.5765	-14.1131	2.1865	-11.1697	15.6498	0.6927
	x_0	a_1	a_2	b_1	a_1b_1	a_2b_1	x

表 13.27 に示した対比型のデザイン行列の場合は，それぞれのデザイン変数の合計および平均はゼロなので，表 13.32 に示した最後の水準を基準とした SAS の GLM プロシジャと同

様の考え方が必要なのであるが，他の因子に対する調整は必要とない．R のパッケージの場合は，最初の水準を基準としているので，本質的には SAS の GLM プロシジャと同様のデザイン変数なので，Lsmeans の推定には，求めたい因子以外の因子のデザイン変数について，該当する因子の水準平均を用いる必要がある．

詳細は，高橋ら（1989）の 15 章「4 種の平方和と LSMESN」および第 16 章「GLM プロシジャの計算方式」，魚住（2014）の「LS-Means 再考－GLM と PLM によるモデル推定後のプロセスー」，Littell ら（2002）の「SAS for Linear Medels, Chapter 6 Understanding Linear Models Concepts」を参照のこと．

R 言語などでの最初の水準を基準にする場合のデザイン変数

最初の水準を基準にする場合について，SAS の GLM プロシジャでの最後の水準を基準とする場合を参考にして，Lsmeans の計算のためのデザイン変数を表 13.34 に示し，表 13.35 にこれらを反映した（0，1）型デザイン行列を用いて Excel の回帰分析の結果を示す．

表 13.34　最初の水準を基準とした（0，1）型デザイン変数

A	a_2	a_3
A_1	0	0
A_2	**1**	0
A_3	0	**1**
平均	1/3	1/3

B	b_2
B_1	0
B_2	**1**
平均	1/2

A×B		a_2b_2	a_3b_2
A_1	B_1	0	0
	B_2	0	0
A_2	B_1	0	0
	B_2	**1**	0
A_3	B_1	0	0
	B_2	0	**1**

表 13.35　最初の水準を基準とする Excel による回帰係数の推定

			―――― デザイン行列 ――――													
				― A－B			A×B		共変量	反応	Excelによる回帰分析					
i	A	B	x_0	a_2	a_3	b_2	a_2b_2	a_3b_2	x	y	分散分析表(「定数に0を使用」をon)					
2.01	A_1	B_1	1	0	0	0	0	0	350	970		自由度	変動	分散		
2.02			1	0	0	0	0	0	400	1000	回帰	7	14314030	2044861		
2.03			1	0	0	0	0	0	360	980	残差	8	670.41	**83.8016**		
2.04		B_2	1	0	0	1	0	0	350	980	合計	15	14314700			
2.05			1	0	0	1	0	0	340	970						
2.06	A_2	B_1	1	1	0	0	0	0	390	990		係数	標準誤差	t	P-値	
2.07			1	1	0	0	0	0	340	950	切片	0	#N/A	#N/A	#N/A	
2.08		B_2	1	1	0	1	1	0	410	980	x0	**727.0489**	32.8702	22.1188	0.0000	β_0
2.09			1	1	0	1	1	0	430	990	a2	**-9.8700**	8.3682	-1.1795	0.2721	β_1
2.10			1	1	0	1	1	0	390	980	a3	**-11.4067**	8.4026	-1.3575	0.2117	β_2
2.11	A_3	B_1	1	0	1	0	0	0	400	990	b2	**8.9832**	8.6394	1.0398	0.3288	β_3
2.12			1	0	1	0	0	0	320	940	a2b2	**-26.8196**	13.3170	-2.0139	0.0788	β_4
2.13		B_2	1	0	1	1	0	1	330	930	a3b2	**-11.1697**	12.4595	-0.8965	0.3962	β_5
2.14			1	0	1	1	0	1	390	1000	x	**0.6927**	0.0877	7.8996	0.0000	β_6
2.15			1	0	1	1	0	1	420	1000						

デザイン行列は，表 13.28 と表 13.35 は異なるので，パラメータの推定値も異なる．表 13.36 で示すパラメータの共分散行列は，表 13.29 と比較すると共変量 x の分散 0.0077 は一致するが，他は全く異なる．

表 13.36　最初の水準を基準とした場合のパラメータの共分散行列

	パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2$						
	切片	― A ―		B	― A×B ―		共変量
	x_0	a_2	a_3	b_2	a_2b_2	a_3b_2	x
β_0	**1080.4513**	-42.1571	-56.3803	-99.0499	227.0588	155.9427	-2.8446
β_1	-42.1571	**70.0269**	28.3183	28.8949	-72.5255	-29.6637	0.0384
β_2	-56.3803	28.3183	**70.6035**	29.8559	-33.3156	-73.2944	0.0769
β_3	-99.0499	28.8949	29.8559	**74.6398**	-83.2891	-78.4839	0.1922
β_4	227.0588	-72.5255	-33.3156	-83.2891	**177.3416**	94.0526	-0.5382
β_5	155.9427	-29.6637	-73.2944	-78.4839	94.0526	**155.2380**	-0.3460
β_6	-2.8446	0.0384	0.0769	0.1922	-0.5382	-0.3460	**0.0077**

表 13.37 に最初の水準を基準とした共分散分析に対して，Lsmeans を求めるためのデザイン変数を示す．因子 A の Lsmeans を求めるために他の因子 B にはデザイン変数 b_2 の平均とし，共変量 x はデータの平均とする．交互作用 A×B は，デザイン変数の掛け算で求める．因子 A の各水準の Lsmeans は，$\hat{y}_{A_1}=\boldsymbol{x}_{2.16}\hat{\boldsymbol{\beta}}=991.06$ として計算され，表 13.31 に示した lsmeasn パッケージの結果に一致する．因子 A_1 の標準誤差は，$Var(\hat{y}_{A_1})=\boldsymbol{x}_{2.16}\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{2.16}^T=19.7243$ の平方根で 4.4412 と推定されていて，表 13.31 の結果に一致する．

表 13.37　最初の水準を基準とした場合の Lsmeans を求めるためのデザイン変数

			―――― デザイン変数 ――――								最小2乗		標準
				― A ―		B	A×B		共変量	反応	平均	分散	誤差
i	A	B	x_0	a_2	a_3	b_2	a_2b_2	a_3b_2	x	y	y^	$Var(y$^$)$	SE
2.16	A_1		1	0	0	**0.50**	0	0	374.67	-	**991.06**	19.7243	4.4412
2.17	A_2		1	1	0	**0.50**	**0.50**	**0**	374.67	-	967.78	18.7249	4.3272
2.18	A_3		1	0	1	**0.50**	**0**	**0.50**	374.67	-	974.07	17.6261	4.1983
2.19	B_1		1	**0.33**	**0.33**	0	**0**	**0**	374.67	-	979.47	13.1335	3.6240
2.20	B_2		1	**0.33**	**0.33**	1	**0.33**	**0.33**	374.67	-	975.79	10.9665	3.3116
2.21	A_1	B_1	1	0	0	0	0	0	374.67	-	986.57	28.1013	5.3011
2.22		B_2	1	0	0	1	0	0	374.67	-	995.55	48.6673	6.9762
2.23	A_2	B_1	1	1	0	0	0	0	374.67	-	976.70	42.6192	6.5283
2.24		B_2	1	1	0	1	1	0	374.67	-	958.86	37.5322	6.1264
2.25	A_3	B_1	1	0	1	0	0	0	374.67	-	975.16	43.5546	6.5996
2.26		B_2	1	0	1	1	0	1	374.67	-	972.97	28.1526	5.3059

因子 B の各水準の Lsmeans は，因子 A のデザイン変数（a_2，a_3）の平均値 1/3=0.3333 とし，交互作用 A×B は，デザイン変数の掛け算で求める．このように謎めいた Lsmeans も Excel の行列計算の活用により，身近な統計量として使われるようになることを期待したい．

13.5. ポアソン回帰における最小２乗平均（Lsmeans）

これまでの探索的なポアソン回帰の例示では，「最小２乗平均」ではなく「予測プロファイル」によって結果を示してきた．これは，最小２乗平均は，説明変数に質的変数 A と量的変数 x が混合するような場合に，量的変数 x の平均を $\bar{X}$ としたときの質的変数 A の水準平均を推定する方法であり，量的変数については適用されないからである．JMP の予測プロファイルは，質的変数 A のある水準にセットした場合に量的変数のプロファイルを図示する機能が含まれていて探索的な解析の結果の解釈に役立つことを示してきた．

第 1.13 節で取り上げた「雌のカブトガニに連結する雄の数」は，質的変数２因子，量的変数２変数の対数リンクでのポアソン回帰の事例で，過分散となっていることを示した［アグレスティ（2003)］．第 7.2 節では，「カブトガニのサテライト数に対する探索的解析」において予測プロファイルの使い方を詳細に示した．表 13.38 に示すように質的変数として（甲羅の色，後体部の棘)，量的変数として（甲羅の幅，体重）があり，反応変数はサテライト数である．

表 13.38　カブトガニのータのリスト

No	甲羅の色	後体部の棘	甲羅の幅	体重	サテライト数
1	2:中ぐらい	3:両方破損	28.3	3.05	8
2	3:やや暗い	3:両方破損	22.5	1.55	0
3	1:やや明るい	1:正常	26.0	2.30	9
4	3:やや暗い	3:両方破損	24.8	2.10	0
5	3:やや暗い	3:両方破損	26.0	2.60	4
:					
171	1:やや明るい	1:正常	28.0	2.63	0
172	4:暗い	3:両方破損	27.0	2.63	0
173	2:中ぐらい	2:一方破損	24.5	2.00	0
		平均	26.2988	2.4372	2.9191

カブトガニのデータは，観察研究で得られたデータなので，表 13.39 に示すように，２つの質的変数のクロス表のセルは，不揃いである．

表 13.39　甲羅の色，後体部の棘の状態のクロス表

	後体部の棘			
甲羅の色	1:正常	2:一方破損	3:両方破損	計
1:やや明るい	9	2	1	12
2:中ぐらい	24	8	63	95
3:やや暗い	3	4	37	44
4:暗い	1	1	20	22
計	37	15	121	173

2 つの量的変数（甲羅の幅，体重）の内，体重がサテライト数に対してより説明力があることが結果として示されているので，サテライト数を反応変数とし，甲羅の色，後体部の棘，および，体重を説明変数としたポアソン回帰を行い，最小 2 乗平均を求める．デザイン変数は，表 13.40 に示すように JMP のディフォルトの対比型である．

表 13.40　（1，–1）対比型デザイン変数

甲羅の色	a_1	a_2	a_3	後体部の棘	b_1	b_2
1:やや明るい	1	0	0	1:正常	1	0
2:中ぐらい	0	1	0	2:一方破損	0	1
3:やや暗い	0	0	1	3:両方破損	-1	-1
4:暗い	-1	-1	-1			

JMP によるポアソン回帰の結果を表 13.41 に示す．適合度統計量から過分散であるが，ここでは，Lsmeans の例示として用いるので無視する．

表 13.41　JMP によるポアソン回帰（対数リンク，過分散なし）

モデル全体の検定

モデル	(-1)*対数尤度	尤度比カイ2乗	自由度	p値(Prob>ChiSq)
差分	41.5446	83.0892	6	<.0001*
完全	452.5001			
縮小	494.0447			

適合度統計量	カイ2乗	自由度	p値(Prob>ChiSq)
Pearson	533.4818	166	<.0001*
デビアンス	549.7025	166	<.0001*

AICc
919.6789

効果の検定

要因	自由度	尤度比カイ2乗	p値
甲羅の色	3	9.7851	0.0205*
後体部の棘	2	2.1025	0.3495
体重	1	52.4907	<.0001*

パラメータ推定値

項	推定値	標準誤差	尤度比カイ2乗	p値
切片	-0.3979	0.1990	4.0206	0.0449*
甲羅の色[1:やや明るい]	0.3321	0.1313	6.0106	0.0142*
甲羅の色[2:中ぐらい]	0.0644	0.0735	0.7721	0.3796
甲羅の色[3:やや暗い]	-0.1888	0.0961	3.9353	0.0473*
後体部の棘[1:正常]	0.0233	0.0945	0.0607	0.8054
後体部の棘[2:一方破損]	-0.1374	0.1292	1.1923	0.2749
体重	0.5476	0.0732	52.4907	<.0001*

予測プロファイルを図 13.16 に示す．甲羅の色を 1:明るい，後体部の棘を 1:正常，体重を平均値の 2.4372 kg としたときの推定値が，3.6396（2.7211，4.8681）と推定されている．

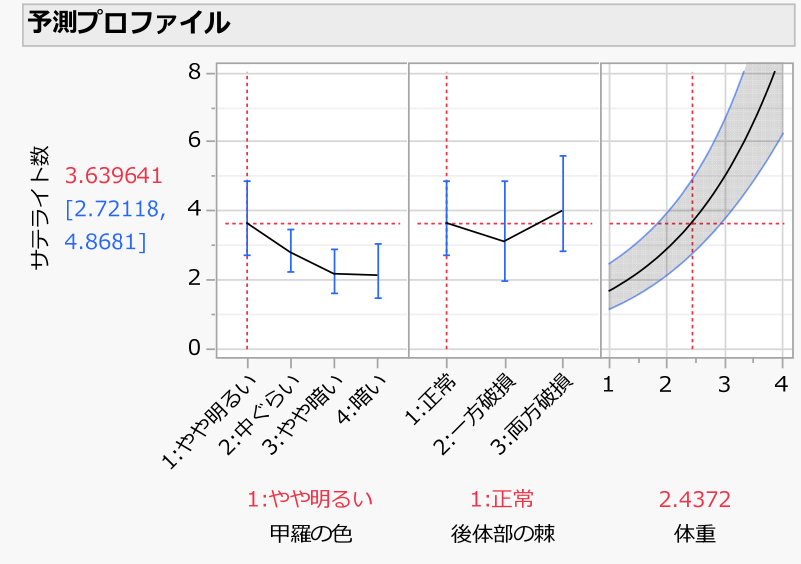


図 13.16　予測プロファイルによる結果の表示

表 13.42 に, JMP によって得られた推定値とパラメータの共分散行列を Excel に取り込んだ結果を示す.

表 13.42　JMP によるポアソン回帰の推定値とパラメータの共分散行列

			パラメータの共分散行列 $\Sigma(\hat{\beta})$						
	項	推定値	切片	1:やや明るい	2:中ぐらい	3:やや暗い	1:正常	2:一方破損	体重
	切片	-0.3979	**0.0396**	-0.0005	0.0006	-0.0007	0.0031	0.0015	-0.0134
甲羅	1:やや明るい	0.3321	-0.0005	**0.0172**	-0.0028	-0.0056	-0.0034	-0.0001	0.0004
の色	2:中ぐらい	0.0644	0.0006	-0.0028	**0.0054**	0.0003	0.0006	-0.0001	-0.0012
	3:やや暗い	-0.1888	-0.0007	-0.0056	0.0003	**0.0092**	0.0018	-0.0007	0.0000
後体部	1:正常	0.0233	0.0031	-0.0034	0.0006	0.0018	**0.0089**	-0.0095	-0.0022
の棘	2:一方破損	-0.1374	0.0015	-0.0001	-0.0001	-0.0007	-0.0095	**0.0167**	0.0019
	体重	0.5476	-0.0134	0.0004	-0.0012	0.0000	-0.0022	0.0019	**0.0054**

表 13.43 に，図 13.16 の予測プロファイルでの推定値と 95%信頼区間をされた表 13.42 の結果を元に計算した結果を示す．甲羅の色を 1:明るい，後体部の棘を 1:正常，体重を平均値の 2.4372 kg としたときの推定値が，3.6396（2.7211，4.8681）が得られている．JMP の予測プロファイルは，ここに示したように各因子の水準の組み合わせた場合を基本としている

表 13.43　因子の組み合わせ推定値と 95%信頼区間

		x_0	a_1	a_2	a_3	b_1	b_2	x	$\ln y\hat{}$	$Var(y\hat{})$	$y\hat{}$	$L95\%$	$U95\%$
甲羅の色													
	1:やや明るい	1	**1**	0	0	**1**	0	2.4372	1.2919	0.0220	**3.6396**	**2.7212**	**4.8681**
	2:中ぐらい	1	0	**1**	0	**1**	0	2.4372	1.0242	0.0125	2.7849	2.2367	3.4674
	3:やや暗い	1	0	0	**1**	**1**	0	2.4372	0.7710	0.0221	2.1620	1.6156	2.8931
	4:暗い	1	**-1**	**-1**	**-1**	**1**	0	2.4372	0.7522	0.0342	2.1217	1.4768	3.0482
後体部の棘													
	1:正常	1	**1**	0	0	**1**	0	2.4372	1.2919	0.0220	**3.6396**	**2.7212**	**4.8681**
	2:一方破損	1	**1**	0	0	0	**1**	2.4372	1.1312	0.0530	3.0994	1.9739	4.8669
	3:両方破損	1	**1**	0	0	**-1**	**-1**	2.4372	1.3827	0.0301	3.9858	2.8375	5.5989

表 13.44 に互いの因子についてセル平均ゼロとした最小 2 乗平均を示す．表 13.43 と大きく異なるのは，ほとんどの分散が小さくなることである．これは，名目のサンプル数が増加することによる影響である．JMP でのポアソン回帰の場合には，この「最小 2 乗平均」は，現在は求めることができないが，SAS の GENMOD プロシジャでは，ポアソン回帰の場合でも Lsmeans ステートメントによって求めることができる．

表 13.44　甲羅の色および体部の棘についての最小 2 乗平均の推定値と 95%信頼区間

		x_0	a_1	a_2	a_3	b_1	b_2	x	$\ln y\hat{}$	$Var(y\hat{})$	$y\hat{}$	$L95\%$	$U95\%$
甲羅の色													
	1:やや明るい	1	**1**	0	0	0	0	2.4372	1.2686	**0.0243**	3.5559	2.6198	4.8265
	2:中ぐらい	1	0	**1**	0	0	0	2.4372	1.0009	**0.0068**	2.7208	2.3140	3.1992
	3:やや暗い	1	0	0	**1**	0	0	2.4372	0.7477	**0.0140**	2.1122	1.6754	2.6630
	4:暗い	1	**-1**	**-1**	**-1**	0	0	2.4372	0.7290	**0.0274**	2.0729	1.4981	2.8682
後体部の棘													
	1:正常	1	0	0	0	**1**	0	2.4372	0.9598	**0.0108**	2.6112	2.1300	3.2012
	2:一方破損	1	0	0	0	0	**1**	2.4372	0.7992	**0.0351**	2.2237	1.5406	3.2098
	3:両方破損	1	0	0	0	**-1**	**-1**	2.4372	1.0507	**0.0050**	2.8596	2.4880	3.2867

いずれにしても，「最小 2 乗平均」は SAS および JMP ユーザの方言であり，「予測プロファイル」は，JMP ユーザの方言ではあるが，探索的な解析における結果の表示方法として優れた推定方法である．

これまでも示してきたように「予測プロファイル」は，「最小 2 乗平均」包含する概念であり，探索的な解析のための基本ツールとの認識し，JMP と同様の予測プロファイルを Excel で作成する手順を示してきた．

最尤法によるポアソン回帰分析入門

文献・索引 目次

偶数ページ

文　献

Web アクセスは全て 2020 年 4 月中旬

－あ－

1） アグレスティ著, 渡邉裕之, 菅波秀規, 吉田光弘, 角野修二, 寒水孝司, 松永信人 訳（2003）, カテゴリカルデータ解析入門, 110-127, 169-179, サイエンティスト社.

2） Agresti A.（2019）, An Introduction to Categorical Data Analyssis 3rd ed., Wiley.

3） Agresti A.（2013）, Categorical Data Analyssis 3rd ed., 75-77, 552-555, Wiley.

4） アーミテージ, ベリー著, 椿美智子, 椿広計 訳（2001）, 医学研究のための統計的方法, 原著第 3 版, 282-302, 377-389, サイエンティスト社.

5） Armitage P., Berry G. and Matthews J.N.S.（2002）, Statistical Methods in Medical Reserch, 4th ed. Blakwell.

6） アルトマン著, 木船義久, 佐久間昭 訳（1999）, 医学研究における実用統計学, 56-58, サイエンティスト社.

7） 岩崎学（2010）, カウントデータの統計解析, 168-189, 朝倉書店.

8） 魚住龍史（2014）, LS-Means 再考－GLM と PLM によるモデル推定後のプロセス－, SAS ユーザー総会論文集:449-463.
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/sugj2014.pdf

9） 大和田章一（2010）, 線形モデルと非線形モデルの基本的な考え方－逆推定の解析, 標準誤差と信頼限界－, 第 2 期医薬安全研究会 第 7 回定例会.
https://biostat.jp/archive_teireikai_2_download.php?id=19

10）奥野忠一, 久米均, 芳賀敏郎, 吉沢正著（1981）, 多変量解析法 改訂版:49-64, 112-123, 123-128, 日科技連出版社.

－か－

11） Cameron A.C. and Trivedi P. K.（1998）, Regression Analysis of Count Data, 76-80, Cambridge University Press.

12） 橘田久美子, 福島慎二 （2013）, 効力比の推定, じっくり勉強すれば身につく統計入門 第 6 回. https://scientist-press.com/wp/wp-content/uploads/2019/07/seminar6.pdf

13） 久保拓弥（2012）, データ解析のための統計モデリング入門 一般化線形モデル・階層ベイズモデル・MCMC, 39-65, 岩波書店.

14） 久保拓弥 訳, Murrell P. 著（2009）, R グラフィックス －R で思いどおりのグラフを作図するためにー, 6-6, 125-147, 共立出版.

15） Collett D.（2003）, Modellng Binary Data 2nd Edition, 103-128, Chapman & Hall.

－さ－

16） 佐久間昭（1977），薬効評価－計画と解析-I，285-295，312-323，東大出版会．

17） 佐久間昭 著，五所正彦，酒井弘憲，佐藤泰憲，竹内久朗 編（2017），新版 薬効評価，232-244，255-276，東大出版会．

18） SAS Institute（2016），SAS/STAT® 14.2 User's Guide，The GENMOD Procedure，3164-65．http://support.sas.com/documentation/onlinedoc/stat/142/genmod.pdf

19） 下野嘉子（2010），R を用いた一般化線形モデル（回帰係数編）:カウントデータを例に，雑草研究，Vol.55(4):287-94．https://www.jstage.jst.go.jp/article/weed/55/4/55_4_287/_pdf

20） 杉本典子，橘田久美子（2011），共分散分析の基礎・医薬品開発における共分散分析の例，じっくり勉強すれば身につく統計入門 第 4 回．https://scientist-press.com/wp/wp-content/uploads/2019/07/seminar4.pdf

21） 杉本典夫（），統計学入門，13.3 節 勾配比検定法．http://www.snap-tck.com/room04/c01/stat/stat13/stat1303.html

22） 杉本典夫（），統計学入門，13.2 節 平行線検定法．http://www.snap-tck.com/room04/c01/stat/stat13/stat1302.html

23） スネデカー，コクラ ン著，畑村又好，奥野忠一，津村善郎 訳（1972），統計的方法，原著第 6 版，213-216，岩波書店．

24） Snedecor G.W., Cochran W.G.（1989），Statistical Methods, 8th ed.，Iowa State Press．

25） 新村秀一（1983a），行列表現による重回帰分析（1），オペレーションズ・リサーチ，vol.28：439-445． http://orsj.or.jp/~archive/pdf/bul/Vol.28_09_439.pdf

26） 新村秀一（1983b），行列表現による重回帰分析（2），オペレーションズ・リサーチ，vol.28：506-628． http://www.orsj.or.jp/~archive/pdf/bul/Vol.28_10_506.pdf

27） 新村秀一（1983c），重回帰分析における掃き出し演算子，オペレーションズ・リサーチ，vol.28：565-669． http://orsj.or.jp/~archive/pdf/bul/Vol.28_11_565.pdf

－た－

28） 高波洋平，舟尾暢男（2016），SAS Stadio によるやさしい統計データ分析，オーム社．

29） 高橋行雄，大橋靖雄，芳賀敏郎（1989），SAS による実験データの解析，307-333，東京大学出版会．

30） 高橋行雄（2002），GENMOD プロシジャによる計数データの解析，SAS ユーザー総会論文集：193-202． https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/sugj2002.pdf

31） 高橋行雄（2004），各種の効力比の統計を支える非線形最小 2 乗法入門，SAS ユーザー総会論文集：3-22． https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/sugj2004.pdf

32） 高橋行雄（2006），SAS ユーザのための S-Plus 活用術.
http://www.msi.co.jp/splus/usersCase/medical/pdf/06takappt.pdf
33） 高橋行雄（2011），JMP による各種分割実験入門 －変量効果モデルの基礎－，続・高橋セミナー第 1 回. https://www.yukms.com/biostat/takahasi2/rec/001.htm
34） 高橋行雄（2013a），回帰分析・再入門 －統計ソフトが対応していない生物統計の各種の課題を Excel でサクサク解こう－，じっくり勉強すれば身につく統計入門 第 7 回.
https://scientist-press.com/wp/wp-content/uploads/2019/07/seminar7.pdf
35） 高橋行雄（2013b），応用回帰分析 1 －各種の重み付き回帰における逆推定－，続・高橋セミナー第 3 回. https://www.yukms.com/biostat/takahasi2/rec/003.htm
36） 高橋行雄（2015），寿命試験データの統計解析，続・高橋セミナー第 4 回.
https://www.yukms.com/biostat/takahasi2/rec/004.htm
37） 高橋行雄（2017），一般化線形モデルを Excel で極め活用する－プロビット法・ロジット法・補 2 重対数法－，続・高橋セミナー第 6 回.
http://www.yukms.com/biostat/takahasi2/rec/006.htm
38） 高橋行雄（2018），正規分布を仮定した打ち切りデータを含む回帰分析入門，続・高橋セミナー第 7 回. https://www.yukms.com/biostat/takahasi2/rec/007.htm
39） 高橋行雄（2019a），最尤法による探索的ポアソン回帰，続・高橋セミナー第 8 回.
https://www.yukms.com/biostat/takahasi2/rec/008.htm
40） 高橋行雄（2019b），投与前値がある場合の解析のレビュー，第 2 期 医薬安全性研究会，第 24 回定例会. https://biostat.jp/archive_teireikai_2_download.php?id=164
41） 竹内啓ら（1989），統計学辞典，1135，東洋経済.
42） 竹内啓（1979），数理統計学，308-317，東洋経済.
43） 東京大学教養学部統計学教室編（1972），自然科学の統計学，31-40，東大出版会.
44） ドブソン 著，田中豊，森川義彦，山中竹春，富田誠 訳（2008），一般化線形モデル入門，原著 第 2 版，60-63，76-79，168-189，共立出版.
45） Dobson A.J. and Barnett A.G.（2018），An Introduction to Generalized Linear Models 4th ed., CRC Press.
46） 富山茂巳，杉本忠則（2004），複数の物質の変異原性の強さの比較，医薬安全性研究会会報，Vol.49：43-53. https://biostat.jp/archive_kaihou/ANZ_KIH_49_2004_01.pdf.
47） ドレーパ，スミス著，中村慶一 訳（1968），応用回帰分析，47-87，216-302，森北出版.
48） Draper N.R. and Smith H.（1998），Applied Regression Analysis,3rd.ed. A Wiley-Interscience Publication.

－な－

49） 中西展大（2010），非線形回帰を用いた逆推定の基礎，じっくり勉強すれば身につく統計入門 第 12 回. https://scientist-press.com/tokei-nyumon/

50） 野沢昌弘（1992），テコ比とハット行列，応用統計学，Vol. 21, N03 : 165-166. https://www.jstage.jst.go.jp/article/jappstat1971/21/3/21_3_165/_pdf/-char/ja

－は－

51） 芳賀敏郎（2004），最小 2 乗法，最尤法，線形モデル，非線形モデル. http://www.yukms.com/biostat/haga/download/archive/likelihood/Likelihood.pdf

52） 芳賀敏郎（2009)，医薬品開発のための統計解析 第 2 部 実験計画法 初版，72-81，サイエンティスト社.

53） 芳賀敏郎（2010），医薬品開発のための統計解析 第 3 部 非線形モデル，13-17，サイエンティスト社.

54） 原田淳（2017），平行線検定を利用した薬物の効力比較，日薬理誌，Vol 150：16-22. https://www.jstage.jst.go.jp/article/fpj/150/1/150_16/_pdf/-char/ja

55） 原田淳，吉池通晴（2017），平行線検定（直線及びシグモイド曲線）による効力比較，第 2 期医薬安全性研究会：第 21 回定例会. https://biostat.jp/archive_teireikai_2_download.php?id=140

56） Finny D.J.（1971），Probit Analisys 3rd ed.：50-80，Cambrige University Press.

57） Finny D.J.（1978），Statistical Method in Biological Assay 3rd ed.，39-68，Charles Griffin.

－ま－

58） McCullagh P. and Nelder J.A.（1989），Generalized Linear Models：204-208, Chapman Hall.

59） 南美穂子，Lennert-Cody C.E.（2013），ゼロの多いデータの解析：負の 2 項回帰モデルによる傾向の過大推定，統計数理，第 61 巻第 2 号：71–87. https://www.ism.ac.jp/editsec/toukei/pdf/61-2-271.pdf

60） 蓑谷千凰彦（2010），統計分布ハンドブック，608-610，朝倉書店.

61） 蓑谷千凰彦（2013），一般化線形モデルと生存時間分析，214-224，朝倉書店.

62） 守屋和幸，広岡博之（2018），R パッケージを用いた最小 2 乗分散分析と最小 2 乗平均値の算出，日畜会報 Vol.89：1-6. https://www.jstage.jst.go.jp/article/chikusan/89/1/89_1/.

－や－

63） 吉村功，大橋靖雄 責任編集（1992），毒性試験データの統計解析，147-66，地人書館.

－ら－

64） Little R.C., Stroup W.W. and Freund R.j.（2020），SAS for Linear Models 4th ed.：163-227，SAS Institute.

65） 臨床評価研究会（ACE）基礎解析分科会（2017），新版 実用 SAS 生物統計ハンドブック，サイエンティスト社. http://www.ace-jp.org/book/favor.html

文 献 索 引

た 高橋(2015) - 寿命試験データの統計解析 70
高橋(2017) - 一般化線形モデルをExcelで極め活用するープロビット法・ロジット法・補2重対数法ー 2, 70, 176, 201
高橋(2018) - 正規分布を仮定した打ち切りデータを含む回帰分析入門 70
高橋(2019a) - 最尤法による探索的ポアソン回帰 2, 224, 243
高橋(2019b) - 投与前値がある場合の解析のレビューー 425
竹内(1979) - 数理統計学 163
竹内ら(1989) - 統計学辞典 421
東京大学教養学部統計学教室編(1992) - 自然科学の統計学 160, 407
ドブソン著, 田中・森川・山中・富田 訳(2008) - 一般化線形モデル入門, 原著 第2版 2, 16, 23, 49, 78, 88, 125, 136, 180, 186, 195, 361, 410, 415
富山・杉本(2004) - 細菌を用いた用量反応試験データ 36, 119, 284
ドレーパ・スミス著, 中村訳(1968) - 応用回帰分析 135, 146, 174, 398
Draper and Smith(1998) - Applied Regrettion Anarysis 3rd ed. 135
な 中西(2016) - 非線形最小2乗法の基本的な考え方 174
野沢昌弘(1992) - テコ比とハット行列 366
は 芳賀(2004) - 最小2乗法, 最尤法, 線形モデル, 非線形モデル 3
芳賀(2009) - 医薬品開発のための統計解析 第2部 実験計画法 401, 432
芳賀(2010) - 医薬品開発のための統計学, 第3部 非線形モデル 163, 174
原田(2017) - 平行線検定を利用した薬物の効力比較 278
原田・吉池(2017) - 平行線検定(直線及びシグモイド曲線)による効力比較 278
Finney(1971) - Probit Analisis 3rd ed. 201
Finney(1978) - Statistical Metod in Biological Assay 3rd ed. 201
ま McCullagh and Nelder(1989) - Generalized Linear Models 2nded 323
南・Cheridy(2013) - ゼロの多いデータの解析:負の2 項回帰モデルによる傾向の過大推定 314
蓑谷(2010) - 統計分布ハンドブック 増補版 212
蓑谷(2013) - 一般線形モデルと生存時間解析 258
守屋・広岡(2018) - Rパッケージを用いた最小2乗分散分析と最小2乗平均値の算出 6, 422, 449
や 吉村・大橋 責任編集(1992) - 毒性試験データの統計解析 32, 109, 237
ら Littleら(2002) - SAS for Linear Medels 455
臨床評価研究会(ACE)基礎解析分科会(2017) - 新版 実用SAS生物統計ハンドブック 293, 354

索　引

解析用ファイル 一 覧

ファイル名
第1章01_ポアソン確率.xlsx
第1章02_ポアソン_2項分布.xlsx
第1章03_種子数.jmp
第1章03_種子数.xlsx
第1章04_人工データ.jmp
第1章04_人工データ.xlsx
第1章05_冠動脈疾患.jmp
第1章05_冠動脈疾患.xlsx
第1章05_冠動脈疾患01反応.jmp
第1章05_冠動脈疾患01反応グラフ.jmp
第1章06_満月新月.jmp
第1章06_満月新月.xlsx
第1章06_満月新月_SAS.txt
第1章07_細菌2x2.jmp
第1章07_細菌2x2.xlsx
第1章08_変異原性試験.jmp
第1章08_変異原性試験.xlsx
第1章08_変異原性試験_SAS.txt
第1章09_久保_種子.jmp
第1章09_久保_種子.xlsx
第1章09_久保_種子_Cグラフ化.jmp
第1章10_軍人_癌.jmp
第1章10_軍人_癌.xlsx
第1章11_タバコと冠動脈疾患.jmp
第1章11_タバコと冠動脈疾患.xlsx
第1章12_通院回数.jmp
第1章12_通院回数.xlsx
第1章13_カブトガニ.jmp
第1章13_カブトガニ.xlsx
第1章13_カブトガニ_探索.jmp
第2章01_ポアソン確率.jmp
第2章01_種子数_尤度関数.xlsx
第2章02_種子数_ソルバー.xlsx
第2章03_種子数_ニュートン.xlsx
第2章03_種子数_偏微分.jmp
第2章04_人工データ.xlsx
第2章04_人工データ_偏微分式.jmp
第2章05_久保_種子C群.xlsx
第2章05_久保_種子_グラフ化.jmp
第2章05_久保_種子_偏微分式.jmp
第2章06_冠動脈疾患_2項分布.jmp
第2章06_冠動脈疾患_オフセット.xlsx
第2章06_冠動脈疾患_偏微分式.jmp
第2章06_冠動脈疾患グラフ.jmp

ファイル名
第3章01_満月新月_01.jmp
第3章01_満月新月_01.xlsx
第3章02_満月新月_01ロジット.jmp
第3章02_満月新月_01ロジット.xlsx
第3章03_満月新月_ポアソン.jmp
第3章03_満月新月_ポアソン.xlsx
第3章04_細菌_01型.jmp
第3章04_細菌_2x2.xlsx
第3章05_Ames_用量反.xlsx
第3章05_Ames_用量反応.jmp
第3章06_タバコと冠動脈疾患.jmp
第3章06_タバコと冠動脈疾患.xlsx
第3章06_タバコと冠動脈疾患b.jmp
第4章01_回帰_入門.xlsx
第4章02_回帰_正規方程式.xlsx
第4章03_回帰_逆行列.xlsx
第4章05_回帰_デザイン行列.xlsx
第4章06_回帰_逆推定.xlsx
第4章07_回帰_JMP.xlsx
第4章07_回帰_逆推定.jmp
第5章01_対数リンク.xlsx
第5章02_重み_計算式.xlsx
第5章03_重み_恒等.xlsx
第5章04_重み_2次式.jmp
第5章04_重み_2次式.xlsx
第5章04_重み_対数リンク.jmp
第5章04_重み_対数リンク.xlsx
第5章05_オフセット.jmp
第5章05_オフセット.xlsx
第5章05_オフセット2次.jmp
第5章06_プロビット.jmp
第5章06_プロビット.xlsx
第6章01_負の二項分布.xlsx
第6章01_負の二項分布_ポアソン.jmp
第6章02_ガンマポアソン.xlsx
第6章02_事故_ポアソン.jmp
第6章03_通院回数.jmp
第6章03_通院回数.xlsx
第6章03_通院回数グラフ.jmp
第6章04_ゼロ_ポアソン.jmp
第6章04_ゼロ_ポアソン.xlsx
第6章05_ゼロ_ガンマ・ポアソン.jmp
第6章05_ゼロ_ガンマ・ポアソン.xlsx

ファイル名

第6章06_カブトガニ_ガンマポアソン.xlsx
第6章06_過大分散比較.jmp
第6章07_カブトガニ_zero過剰ガンマポアソン.xlsx

第7章01_細菌2x2.jmp
第7章01_細菌2x2.xlsx
第7章02a_カブトガニ_クロス表.jmp
第7章02b_カブトガニ_回帰.jmp
第7章02c_カブトガニ_甲羅色_中.jmp
第7章02d_カブトガニ_グラフ・ビルダー.jmp
第7章02__カブトガニ_.xlsx
第7章02__カブトガニ_プロファイル.xlsx
第7章03_被害者.jmp
第7章03_被害者.xlsx
第7章03_被害者_SAS.txt

第8章01_細菌の増殖_勾配比.jmp
第8章01_細菌の増殖_勾配比.xlsx
第8章02_ヒスタミン平行線.jmp
第8章02_ヒスタミン平行線.xlsx
第8章03_Ames_ポアソン.xlsx
第8章03_Ames_回帰_別々ポアソン.jmp

第9章01_花数.jmp
第9章01_花数.xlsx
第9章02_花数_交互作用.jmp
第9章02_花数_交互作用.xlsx
第9章03_花数_SAS.txt
第9章04_花数_オフセット.jmp
第9章04_花数_オフセット.xlsx
第9章04_花数_負の2項回帰_SAS.txt
第9章05_花数_負の2項.xlsx

第10章01_Ship Damage_データ.jmp
第10章01_Ship Damage_データ.xlsx
第10章02_Ship Damage_主効果.jmp
第10章02_Ship Damage_主効果.xlsx
第10章03_Ship Damage_予測プロファイル.xlsx
第10章03_Ship Damage_共分散.jmp
第10章04_Ship Damage_交互作用.jmp

ファイル名

第10章04_Ship Damage_交互作用.xlsx
第10章05_Ship Damage_主効果_予測.jmp
第10章05_Ship Damage_主効果_予測値.xlsx
第10章06_Ship Damage_ソルバー.xlsx
第10章06_Ship Damage_主効果.jmp
第10章07_Ship Damage.xlsx
第10章07_Ship Damage_SAS.txt

第11章02_人工データ_線形.jmp
第11章02_人工データ_線形.xlsx
第11章02_人工データ_線形.jmp
第11章03_ポアソン回帰_デビアンス残差.xlsx
第11章04_人工データ_4種の残差.xlsx
第11章04_人工データ_4種の残差_SAS.txt
第11章05_カブトガニ_4種の残差.jmp
第11章05_カブトガニ_4種残差.xlsx
第11章05_カブトガニ_転置_グラフ作成.jmp

第12章02_iris.jmp
第12章02_iris_相関行列.xlsx
第12章03_ガラス工程_偏差平方和ベース.xlsx
第12章04_ガラス工程.jmp
第12章04_ガラス工程_デザイン行列ベース.xlsx
第12章05_2次回帰_自然科学の統計学.xlsx
第12章05_芳賀2次回帰一DE改2- 1 因子(量).xlsm
第12章05_芳賀_2次回帰.xlsx
第12章05_芳賀_2次式.jmp
第12章06_冠動脈心疾患.jmp
第12章06_冠動脈心疾患.xlsx
第12章07_タバコと冠動脈心疾患.jmp
第12章07_タバコと冠動脈心疾患.xlsx

第13章02_層別共分散.jmp
第13章02_層別共分散.xlsx
第13章03_2変量共分散.jmp
第13章03_2変量共分散.xlsx
第13章04_守屋_2因子共変量.jmp
第13章04_守屋_2因子共変量.xlsx
第13章05_カブトガニ.jmp
第13章05_カブトガニ_プロファイル.xlsx

著 者 紹 介

高 橋 行 雄（たかはし ゆきお）

1971 年　中央大学理工学部管理工学科終了
　　富士通電算機専門学校研究科終了
同　年　日本ロシュ株式会社
　　研究所 研究統計課
　　前臨床および臨床試験の統計解析に従事
2002 年　中外製薬株式会社 統計解析部
2011 年　同社 退社
同　年　BioStat 研究所（株）設立　現在に至る

著書　毒性・薬効データの統計解析, サイエンティスト社, 1987
SAS による実験データの解析，東京大学出版会，1989
毒性試験データの統計解析, 地人書館, 1992

第 9 回 続高橋セミナー

最尤法によるポアソン回帰分析入門

BioStat 研究所（株）

〒105-0014 東京都 港区 芝 1-12－3 の 1005

2021 年 1 月　　高橋 行雄

takahashi.stat@nifty.com ， FAX：03-342-8035