

応用ロジスティック回帰
平行性検定、実験の反復
第 11 回 高橋セミナー

高橋 行雄

表紙裏

改訂の記録

2002年11月2日，新規作成

目 次

1. はじめに	1
2. 相対力価	2
2.1. 相対力価とその 95%信頼区間	2
2.2. Mathematica による式の展開	4
ED50 の場合	4
2 次方程式の根, Mathematica 入門	5
Filler の式により分散共分散を用いた ED50 の信頼区間	5
Filler の式による相対力価の 95%信頼区間	6
2.3. 粉虫に対する殺虫剤の毒性 (Colett Ex. 4.3)	7
JMP による解析	9
SAS による分散共分散の計算	11
Filler の式, 2 次方程式の根, 数値計算	13
Mathematica を用いて DDT と併用剤の効力比の信頼区間の計算	14
Excel による効力比の計算	15
3. 自然反応	16
3.1. 殺虫剤実験での自然死	16
3.2. 最大反応がある場合	18
3.3. 小麦粉の害虫に対する殺虫剤の毒性 (Collett 例 4.4)	19
3.4. SAS による解析結果	21
PROBIT プロシジャによる解析結果	21
GENMOD プロシジャに解析	22
4. 非線型ロジスティック回帰モデル	26
4.1. ロジスティック直線が当てはまらない場合	26
4.2. 小麦粉の害虫 (confused flour beetles、こくぞう虫?) の死亡率 (Collett 事例 4.5)	

4.3.	2次曲線のあてはめ.....	28
	JMPによる2次式のロジスティック回帰.....	30
	SASのGENMODプロシジャによるあてはめ.....	31
	元の薬物濃度のままでの解析.....	31
	薬物濃度の平均値を引いた2次の項を用いたGENMODプロシジャ.....	32
	SASのLOGISTICプロシジャによるあてはめ.....	33
4.4.	非線形モデル.....	33
	JMPの非線形モデル.....	34
5.	補2 重対数の適用例.....	36
5.1.	希釈法.....	37
5.2.	4グループ化された生存時間解析.....	39
5.3.	血清学的テスト.....	39
5.4.	マラリアについての血清学的テスト（事例 4.6）.....	39
	年間の感染率.....	39
	SASのGENMODプロシジャによる解析.....	41
	年齢階層ごとのあてはめ.....	41
6.	過大発散（Overdispersion）.....	43
6.1.	過大発散の潜在的な原因（Potential Causes of Overdispersion）.....	43
	例題 6.1 抗肺炎（ <i>pneumococcus</i> ）血清, Collett, Ex 6.1.....	44
	JMPによるLOFの解析.....	45
	反応確率間の変動 Variation between the response probabilities.....	46
	SASのGENMODプロシジャによる解析.....	48
	2値反応間の相関.....	49
6.2.	反応確率のモデル化変動（Modeling Variability in Response Probabilities）.....	49
6.3.	2値反応間の相関.....	52
6.4.	過大発散データのモデル化.....	52
	が同数の特別な場合.....	54
	SASのGENMODプロシジャでの逸脱度の調整.....	56
	SASのLOGISTICプロシジャによる過大発散の調整.....	57
	SASのPROBITプロシジャによる, ロジスティック回帰およびED50の推定.....	57

事例 6.6 ハマウツボの発芽.....	58
6.6. ベータ 2 項分布モデル.....	58
6.7. ロジスティック回帰でのランダム係数モデル.....	59
SAS の NLMIXED を用いた線形ロジスティックモデルでのランダム効果.....	60
Example 46.2: Probit-Normal Model with Binomial Data	60
6.8 手法の比較.....	66

図 表 目 次

表 2.1 DDT、 γ -BHC および併用スプレーの粉虫にたいする毒性.....	7
表 2.2 ED50 と 95%信頼区間	11
表 2.3 Excel による DDT の 95%信頼区間の算出.....	13
表 2.4 Excel による γ BHC と併用薬の効力比の計算.....	15
表 4.1 二硫化炭素ガスの異なる濃度で曝露された n 匹の死亡した害虫の数 y	27
表 5.1 希釈法のデータ例.....	38
表 5.2.....	39
表 6.1 肺炎による死亡数.....	44
表 6.2 ハマウツボの発芽.....	47
表 6.3 ハマウツボの発芽データにロジスティックモデルのあてはめ後の逸脱度	47
表 6.4 ウィリアムの分散の補正, $\phi=0.0249$, $\pi=0.2$ の場合	51
表 6.5 重み付きロジスティック回帰モデルの逸脱度	53
表 6.6 エチレンオキサイドの曝露に作用した粉虫の数	55
表 6.7 有意な逸脱度がある場合の F 検定による調整.....	56
表 6.8 生存胎児数.....	59
表 6.9 ランダム係数の推定.....	64
図 2.1 等価用量 d_S と d_N におけるロジスティック回帰直線.....	3
図 2.2 DDT、 γ -BHC および併用スプレーの粉虫にたいする毒性.....	7
図 2.3 $\ln(\text{dose})$ に対するロジスティック回帰直線のあてはめ	9
図 3.1 自然反応 $\pi=0.15$ がある場合の用量反応	17
図 3.2 自然反応 10%と最大反応 90%がある場合の用量反応	19

図 3.3	自然死を 10%とした場合のロジスティック曲線	20
図 4.1	死亡率とロジット変換後のプロット	27
図 4.2	ロジット変換値での直線および 2 次曲線のあてはめ	28
図 4.3	死亡率での 2 次曲線のあてはめ	29
図 4.4	指数関数のあてはめ	34
図 4.5	2 次シグモイドおよび指数シグモイド曲線のあてはめ	35
図 5.1	$\alpha=0$ 、 $\kappa=1$ の場合のガンベル確率密度関数	36
図 5.2	希釈法での μ および $\log(\mu)$	38
図 5.3	出現率と補二重対数変換のグラフ	38
図 5.4		40
図 6.1	ロジスティック直線のあてはめ	44
図 6.2	対数用量に対するロジスティック回帰直線	45
図 6.3	ハマウツボの各バッチの発芽割合と経験ロジット	47
図 6.4	ウィリアムの補正後の CV	52
図 6.5	ロジスティック直線のあてはめ	55
図 6.6	群ごとの生存胎児の割合	59
図 6.7	リッターサイズ m と生存胎児数の確率楕円	61
図 6.8	生存胎児数の割合と経験ロジットでの平均値の比較	61
図 6.9	ランダム効果を調整した推定値 $\hat{\eta}_i$	65
図 6.10	リッターサイズを含めたランダム効果を調整した胎児の生存率	65

1. はじめに

前回は、ロジスティック回帰の基礎、プロビット法とのロジット法の比較など生物検定法の基礎について述べた。今回は、さらに複数の化合物の相対力価、自然反応がある場合、直線があてはまらない場合、実験に反復がある場合などについて述べる。

第1回目セミナーでも相対力価の課題はとりあげたのであるが、JMP でどこまで対応可能なのか、についての解析の結果を主体にした (<http://www.yukms.com/biostat/>からダウンロードできる)。これまでは、使用する統計ソフトは JMP を主体にしていたのであるが、今回のテーマは、JMP では取り扱えない課題も含まれるので、SAS の GENMOD および LOGISTIC プロシジャも使用することにした。プログラムリストを添付するので参考にしてもらいたい。

今回の講義と実習も、D. Collett (1991)、*Modeling Binary Data* をベースにする。この本は、統計モデルを用いた豊富な薬理試験、毒性試験などの生物系の実験データの解析事例を用いつつ、2 値データの解析の理論的な背景が丁寧に解説されている。生物検定法は第4章 *Bioassay and some applications* で取り上げられている。同一条件で複数のバッチに対して実験をしたような場合、あるいは繁殖試験での同一母体で胎児の異常数を扱うような場合については、第6章の *Overdispersion* で取り上げられている。

2. 相対力価

2.1. 相対力価とその 95%信頼区間

生物検定法の中で、しばしば異なるデータセットの比較のためにモデルのあてはめが必要となる。考え方を固めるための例示をしよう。新たに合成されたある化合物 (N) が、同じ実験条件下で標準物 (S) と比較されるとしよう。それぞれの化合物について、用量の増加ごとの個体の反応の比率が記録されるとしよう。2つの化合物の比較のために、2つの異なるロジスティック回帰直線、平行な直線、および共通の直線が、あてはめられ比較される。もしも、用量反応が、それぞれの化合物で異なるならば、2つの化合物の効果の違いは、一定ではなく、相対的な効果での簡単な要約ができない。2つの化合物の用量反応関係が、ロジットのスケールで平行であるとき、2つの化合物の相対力価を評価することができる。実際、2つの刺激物の力価の比較のための分析計画は、しばしば平行線検定法として言われている。

新しい化合物が、標準物より一定の高い反応を与えること期待されて、また、対数用量が説明変数であるとしたとき、その平行なロジスティック回帰直線は、

$$\text{化合物 N : } \text{logit}(p) = \alpha_N + \beta \log(\text{dose})$$

$$\text{化合物 S : } \text{logit}(p) = \alpha_S + \beta \log(\text{dose})$$

ここで、 $\alpha_N > \alpha_S$

となる。2つの化合物の相対力価は、同じ効果での用量の比として定義される。相対力価の計算式を推論してみよう。新薬の用量を d_N 、標準薬の用量を d_S が、同じ反応を引き起こしたとしたとき、2つの用量が等しい力価をもつようなロジットのスケール上を λ としよう。標準品に対する相対的な新化合物の力価は、 $r_{NS} = d_N / d_S$ としたときに、新化合物は、標準品よりも r_{NS} 倍効果があるということである。この状況は、図 2.1 に示す。 $\text{logit}(p) = \lambda$ のとき、 $\alpha_N + \beta \log(d_N) = \alpha_S + \beta \log(d_S)$ であるので、Sに対するNの力価は、

$$r_{NS} = \frac{d_S}{d_N} = \exp\left\{\frac{\alpha_N - \alpha_S}{\beta}\right\}$$

となる。図 2.1 で、 $\log(r_{NS}) = \log(d_S) - \beta \log(d_N)$ は、平行な回帰直線の水平な分割線であり、 $r_{NS} = DE50_S / ED50_N$ は、標準品と新化合物のそれぞれのED50の比である。ロジット・スケール上での平行線モデルが、観測された用量反応データのあてはまっている認められた後に、相対力価は、

$$\hat{r}_{NS} = \exp\left(\frac{\hat{\alpha}_N - \hat{\alpha}_S}{\hat{\beta}}\right) = \frac{D\hat{E}50_S}{E\hat{D}50_N}$$

をもちいたパラメータの推定値から推定される。

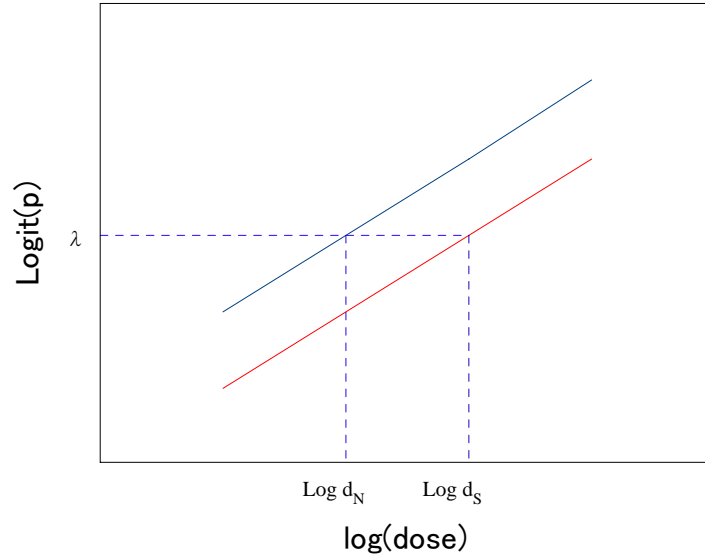


図 2.1 等価用量 d_S と d_N におけるロジスティック回帰直線

相対力価の信頼区間は、フィラーの定理を用いて得ることができる。 $DE50$ についてフィラーの定理で 95% 信頼区間を計算したと同様の計算手順を示す。 $\rho = (\alpha_N - \alpha_S) / \beta$ に対する信頼区間は、最初に関数 $\psi = \hat{\alpha}_N - \hat{\alpha}_S - \rho\hat{\beta}$ を考えることから得られる。 ψ の期待値はゼロであるが、その分散は次式によって与えられる。

$$V = \text{Var}(\hat{\alpha}_N) + \text{Var}(\hat{\alpha}_S) + \rho^2 \text{Var}(\hat{\beta}) - 2\text{Cov}(\hat{\alpha}_N, \hat{\alpha}_S) - 2\rho\text{Cov}(\hat{\alpha}_N, \beta) + 2\rho\text{Cov}(\hat{\alpha}_S, \beta)$$

まえと同様に、

$$\frac{\alpha_N - \alpha_S - \rho\beta}{\sqrt{V}}$$

は、標準正規分布となり、相対力価の対数についての $100(1 - \alpha_S)\%$ 信頼区間は、次式の ρ により求められる。

$$|\hat{\alpha}_N - \hat{\alpha}_S - \rho\hat{\beta}| = z_{\alpha/2} \sqrt{V}$$

両辺の平方し、 V を代入 (substituting for V) することにより、 ρ の 2 次式が導かれる。

最終的に、この式の根は、相対力価自身の信頼区間を与えるために指数化される。

Mathematika で解いた次の結果より,

$$\text{quad} = b_0^2 - 2b_0b_1 + b_1^2 - v_{00}z^2 + 2v_{01}z^2 - v_{11}z^2 + r(-2b_0b_2 + 2b_1b_2 + 2v_{02}z^2 - 2v_{12}z^2) + r^2(b_2^2 - v_{22}z^2)$$

効力比の 95%信頼区間は, r に $-\rho$ を代入した次の 2 次式の根により求められる.

$$\{\hat{\alpha}_S^2 - 2\hat{\alpha}_S\hat{\alpha}_N + \hat{\alpha}_N^2 + \text{Var}(\hat{\alpha}_S)z^2 + 2\text{Cov}(\hat{\alpha}_N, \hat{\alpha}_S)z^2 - \text{Var}(\hat{\alpha}_N)z^2\} + \{2(\hat{\alpha}_S\hat{\beta} - \hat{\alpha}_N\hat{\beta} - \text{Cov}(\hat{\alpha}_S, \hat{\beta})z^2 + \text{Cov}(\hat{\alpha}_N, \hat{\beta})z^2)\rho + \{\hat{\beta}^2 - \text{Var}(\hat{\beta})z^2\}\rho^2 = 0$$

2.2. Mathematica による式の展開

JMP では、ED50 についてはフィラーの定理を用いて、その 95%信頼区間を計算できるが、相対力価の 95%信頼区間については逆推定のパラメータの設定画面が不調で、求めることができない。SAS の LIGOSTIC および GENMOD プロシジャは、逆推定はサポートされていない。このために、数値計算を正しく行なうために Mathematica によりフィラーの定理を解き、数値計算を行なうことにした。

ED50 の場合

ED50 による二次方程式は次のようであった。

$$(\hat{\beta}_1^2 - z_{\alpha/2}^2 \hat{v}_{11})ED50^2 - (2\hat{v}_{01}z_{\alpha/2}^2 - 2\hat{\beta}_0\hat{\beta}_1)ED50 + (\hat{\beta}_0^2 - \hat{v}_{00}z_{\alpha/2}^2) = 0$$

この 2 方程式を標準的な手順により解き、ED50 値の $100(1-\alpha)\%$ の信頼限界の計算ために次の式を得る。

$$ED50 = \frac{-\left(\hat{\rho} - g \frac{\hat{v}_{01}}{\hat{v}_{11}}\right) \pm \frac{z_{\alpha/2}}{\hat{\beta}} \left\{ \hat{v}_{00} - 2\hat{\rho}\hat{v}_{01} + \hat{\rho}^2\hat{v}_{11} - g \left(\hat{v}_{00} - \frac{\hat{v}_{01}^2}{\hat{v}_{11}} \right) \right\}^{1/2}}{1 - g}$$

ここで、 $\hat{\rho} = \hat{\beta}_0 / \hat{\beta}_1$ 、 $g = z_{\alpha/2}^2 \hat{v}_{11} / \hat{\beta}_1^2$ である。

Mathematica によって 2 次方程式を求め、その根を求めた結果を次に示す。

2次方程式の根, Mathematica入門

プログラム

```
Solve[c + b x + a x^2 == 0, x]
```

実行結果

$$\left\{ \left\{ x \rightarrow \frac{-b - \sqrt{b^2 - 4ac}}{2a} \right\}, \left\{ x \rightarrow \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right\} \right\}$$

2 次方程式の根の公式が, **Solve** 関数によって出力される. Mathematica は, 数値計算だけでなく, 文字変数についての解を出すソフトである.

Filler の式により分散共分散を用いたED50の信頼区間

ED50 の 95%信頼区間は, 最終的に 2 次式の根として得られるので, Mathematica により計算してみる.

プログラム

```
v = v00 + r^2 v11 - 2 r v01
```

```
y = (b0 - r b1)^2 - z^2 v
```

```
r = -ED50
```

```
quad = Collect[y, ED50]
```

```
cl = Solve[y == 0, ED50]
```

実行結果

```
v = v00 + 2 ED50 v01 + ED50^2 v11
```

```
y = (b0 + b1 ED50)^2 - (v00 + 2 ED50 v01 + ED50^2 v11) z^2
```

```
r = -ED50
```

```
quad = b0^2 - v00 z^2 + ED50 (2 b0 b1 - 2 v01 z^2) + ED50^2 (b1^2 - v11 z^2)
```

```
cl = { { ED50 -> \frac{1}{b1^2 - v11 z^2} \left( -b0 b1 + v01 z^2 - \right. \\ \left. z \sqrt{b1^2 v00 - 2 b0 b1 v01 + b0^2 v11 + v01^2 z^2 - v00 v11 z^2} \right) }, \\ \left\{ ED50 -> \frac{1}{b1^2 - v11 z^2} \left( -b0 b1 + v01 z^2 + \right. \\ \left. z \sqrt{b1^2 v00 - 2 b0 b1 v01 + b0^2 v11 + v01^2 z^2 - v00 v11 z^2} \right) \right\} }
```

Fillerの式による相対力価の95%信頼区間

プログラム

マトリックス表現で作成した.

```

r=.
c1=.
z=.
a = {1, -1, -r}
MatrixForm[a]
V = {{v00, v01, v02}, {v01, v11, v12}, {v02, v12, v22}}
MatrixForm[V]
aVa = a. V. a
y = (b0 - b1 - r b2)^2 - z^2 aVa
quad = Collect[y, r]
r = -c1
Solve[Y == 0, c1]

```

実行結果

$$a = \begin{pmatrix} 1 \\ -1 \\ -r \end{pmatrix}$$

$$V = \begin{pmatrix} v00 & v01 & v02 \\ v01 & v11 & v12 \\ v02 & v12 & v22 \end{pmatrix}$$

$$aVa = v00 - 2 v01 - r v02 + v11 + r v12 - r (v02 - v12 - r v22)$$

$$y = (b0 - b1 - b2 r)^2 - (v00 - 2 v01 - r v02 + v11 + r v12 - r (v02 - v12 - r v22)) z^2$$

$$\text{quad} = b0^2 - 2 b0 b1 + b1^2 - v00 z^2 + 2 v01 z^2 - v11 z^2 + r (-2 b0 b2 + 2 b1 b2 + 2 v02 z^2 - 2 v12 z^2) + r^2 (b2^2 - v22 z^2)$$

$$c1 = \left\{ \left\{ c1 \rightarrow \frac{1}{2 (b2^2 - v22 z^2)} \left(-2 b0 b2 + 2 b1 b2 + 2 v02 z^2 - 2 v12 z^2 - \sqrt{\left((2 b0 b2 - 2 b1 b2 - 2 v02 z^2 + 2 v12 z^2)^2 - 4 (b0^2 - 2 b0 b1 + b1^2 - v00 z^2 + 2 v01 z^2 - v11 z^2) (b2^2 - v22 z^2) \right)} \right) \right\}, \right. \\ \left. \left\{ c1 \rightarrow \frac{1}{2 (b2^2 - v22 z^2)} \left(-2 b0 b2 + 2 b1 b2 + 2 v02 z^2 - 2 v12 z^2 + \sqrt{\left((2 b0 b2 - 2 b1 b2 - 2 v02 z^2 + 2 v12 z^2)^2 - 4 (b0^2 - 2 b0 b1 + b1^2 - v00 z^2 + 2 v01 z^2 - v11 z^2) (b2^2 - v22 z^2) \right)} \right) \right\} \right\}$$

2.3. 粉虫に対する殺虫剤の毒性 (Colett Ex. 4.3)

Hewlett と Plackett (1950) によって報告された殺虫試験に、粉虫、*Tribolium castaneum*、は、Shell油 P31 に溶解された 3 つ異なる殺虫剤の一つについて噴霧された。使われた 3 種の殺虫剤は、2.0% w/v の dichlorodiphenyltrichloroethane (DDT)、1.5% w/v の γ -benzene hexachloride (γ -BHC)、および、2 つの混合物であった。実験は、約 50 匹のバッチが、噴霧器の varying deposits で曝露され、 $\text{mg}/10\text{cm}^2$ 単位に換算された。6 日後に殺された害虫の割合が実験結果のデータであり、表 2.1 に示す。

これらのデータのモデル化で、殺虫剤の蓄積量の対数は、線形ロジスティックモデルで説明変数として用いられる。最初のステップは、殺虫剤間違いについて、死亡の確率と蓄積量の関係に基づいて、その広がりを見ることである。

表 2.1 DDT、 γ -BHC および併用スプレートの粉虫にたいする毒性

Dose $\text{mg}/10\text{cm}^2$	DDT	γ -BHC	DDT+ γ -BHC
2.00	3 / 50	2 / 50	28 / 50
2.64	5 / 49	14 / 49	37 / 50
3.48	19 / 47	20 / 50	46 / 50
4.59	19 / 50	27 / 50	48 / 50
6.06	24 / 49	41 / 50	48 / 50
8.00	35 / 50	40 / 50	50 / 50

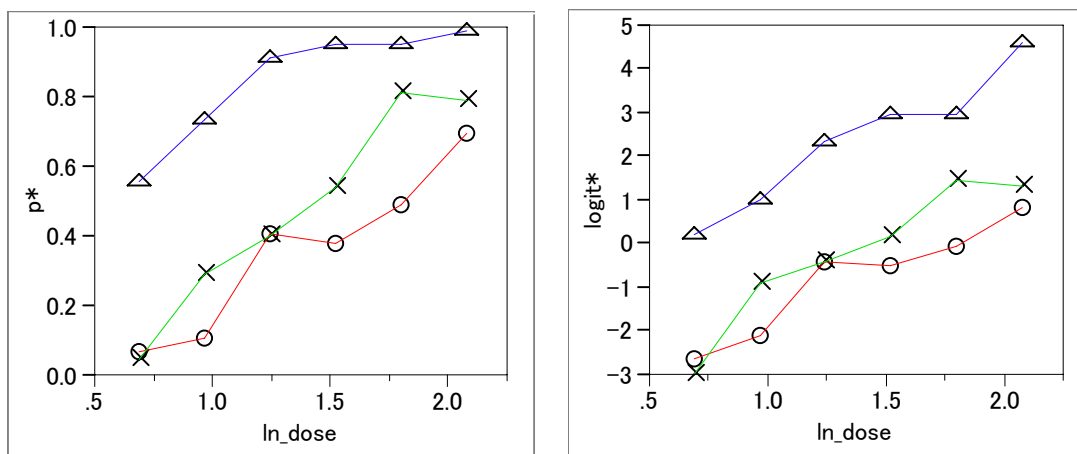


図 2.2 DDT、 γ -BHC および併用スプレートの粉虫にたいする毒性
DDT(—○—)、 γ BHC (—×—)、併用 (—△—)

ロジスティック回帰モデルは、(i) 異なる切片と傾き、(ii) 異なる切片と共通の傾き、(iii) 共通の切片と傾き、のように系統的にモデルあてはめを行う。それぞれの 3 つの

モデルの逸脱度は次ぎのように与えられる。

- (i) 異なる直線 : Deviance = 17.89, d.f. = 12
- (ii) 平行な直線 : Deviance = 21.28, d.f. = 14
- (iii) 共通の直線 : Deviance = 246.83, d.f. = 16

平行線であるようなロジスティック直線のあてはめようとしたときの逸脱度の増加は、 $21.28 - 17.89 = 3.39$, d.f. = 2 であり、有意ではない ($P=0.184$)。それゆえに、共通の傾きもつを回帰直線が、3つの別々の直線と同じようにデータにあてはまると結論する。用量の増加に反応のパターンが、3つの殺虫剤で似ていることを意味する。同様な切片を持つ平行線をあてはめによる逸脱度の増加は、高度に有意 ($P<0.001$) であり、それが、殺虫剤間に明瞭な違いがある。さらに、平行線モデルに対する逸脱度は、受け入れ難くはない大きさであり、このモデルは、3つの殺虫剤の効果を比較する基盤となる。

3つの殺虫剤の線形ロジスティックモデルの回帰式は次に示す。

$$\begin{aligned} \text{DDT} : & \quad \text{logit}(p) = -4.555 + 2.696 \log(\text{dose}) \\ \gamma\text{-BHC} : & \quad \text{logit}(p) = -3.842 + 2.696 \log(\text{dose}) \\ \text{DDT} + \gamma\text{-BHC} : & \quad \text{logit}(p) = -1.425 + 2.696 \log(\text{dose}) \end{aligned}$$

図 2.3 に、経験ロジットが、 $\log(\text{dose})$ に対してプロットされて、あてはめられた平行なロジスティック回帰直線が、重ね書きされている。

3つの殺虫剤に対する ED_{50} 値は、すでにモデルのあてはめた場合のパラメータ推定から得られていて、

$$\text{DDT については } \exp\{-(-4.555/2.696)\} = 5.417 \text{ mg/10cm}^2,$$

$$\gamma\text{-BHC については } \exp\{-(-3.842/2.696)\} = 4.158,$$

$$\text{DDT と } \gamma\text{-BHC の混合物については } \exp\{-(-1.425/2.696)\} = 1.696$$

である。2つの元の殺虫剤に対する混合物の相対力価は、推定された ED_{50} 値の適切な比から推定される。それゆえに、

DDT 単独の場合に対する

$$\text{DDT と } \gamma\text{-BHC の混合物の相対力価は、} 5.417/1.696 = 3.19$$

であり、同時に、

$\gamma\text{-BHC}$ に対する

$$\text{混合物の相対力価は、} 4.158/1.696 = 2.45$$

である。それゆえに、DDT と $\gamma\text{-BHC}$ の混合物は、DDT 単独の 3 倍以上の力価であり、

γ -BHC 単独の 2.5 倍より少し低い力価である。二つの基本的な毒素については、 γ -BHC が粉の中の害虫に対してより力価があるのは、同様に明らかである。

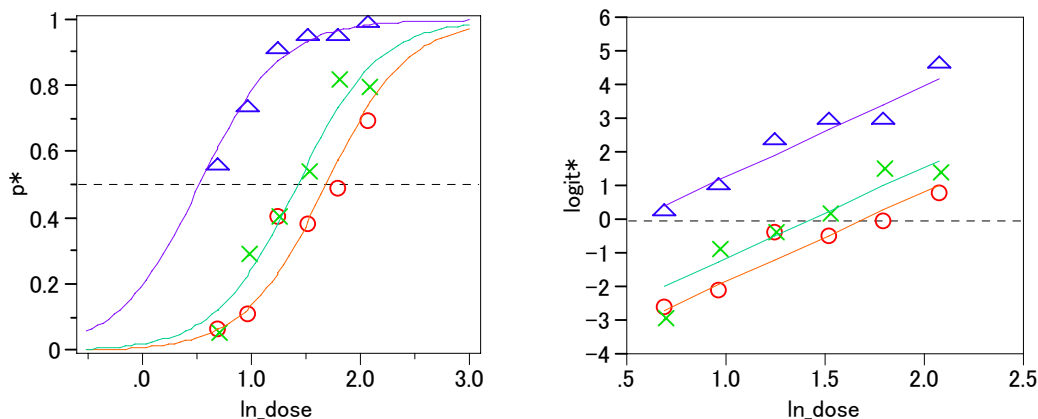


図 2.3 $\ln(\text{dose})$ に対するロジスティック回帰直線のあてはめ

DDT(—○—)、 γ BHC (—×—)、併用 (—△—)

JMPによる解析

JMP により、(i) 異なる直線 のあてはめは、次に示すようにモデルに交互作用を入れることにより実現できる。(ii) 平行な直線 をあてはめた場合の逸脱度との差 3.39 は、交互作用 $\text{Compound} * \ln_dose$ の尤度比カイ 2 乗で得られる。

効果の尤度比検定

要因	パラメータ数	自由度	尤度比カイ2乗	p値(Prob>ChiSq)
Compound	2	2	214.335133	0.0000
\ln_dose	1	1	202.676215	0.0000
$\text{Compound} * \ln_dose$	2	2	3.3922133	0.1834

逸脱度は、あてはまりの悪さ (LOF) のカイ 2 乗値 17.8897 である。

あてはまりの悪さ(LOF)

要因	自由度	(-1)*対数尤度	カイ2乗
あてはまりの悪さ(LOF)	12	8.94485	17.8897
飽和モデル	17	405.03954	p値(Prob>ChiSq)
あてはめたモデル	5	413.98439	0.1191

JMP により平行線を求めるために 3 水準の薬剤名 Compound を使うと内部で生成される計算用のダミー変数が対比型となり、それぞれの回帰式を直接求められない。

パラメータ推定値

項	推定値	標準誤差	カイ2乗	p値(Prob>ChiSq)
切片	-3.2740619	0.2995361	119.47	<.0001
Compound[1:DDT]	-1.2810541	0.128659	99.14	<.0001
Compound[2:γ-BHC]	-0.5682847	0.1193807	22.66	<.0001
ln_dose	2.69568401	0.2156831	156.21	<.0001

推定値は次の対数オッズに対するものです: 0/1

そこで、インディケータタイプのダミー変数 C1、C2、C3 を別途用意しておき、切片なしの回帰モデルにすると直接3本の回帰係数を求めることができる。

パラメータ推定値

項	推定値	標準誤差	カイ2乗	p値(Prob>ChiSq)
C1	-4.555116	0.3611378	159.09	<.0001
C2	-3.8423466	0.3326627	133.41	<.0001
C3	-1.424723	0.2850544	24.98	<.0001
ln_dose	2.69568401	0.2156831	156.21	<.0001

推定値は次の対数オッズに対するものです: 0/1

C1、C2、C3 に対する推定値がそれぞれの殺虫剤のロジスティック回帰直線の切片となり、ln_dose が共通の傾きとなっている。このモデルでの逸脱度は、あてはまりの悪さの出力から 21.28191 が得られる。

あてはまりの悪さ(LOF)

要因	自由度	(-1)*対数尤度	カイ2乗
あてはまりの悪さ(LOF)	14	10.64096	21.28191
飽和モデル	17	405.03954	p値(Prob>ChiSq)
あてはめたモデル	3	415.68050	0.0946

ED50 と 95%信頼区間の計算は、逆推定の機能を用いて計算する。ダミー変数が対比型の場合は、次のように入力する。ln_dose の欄に推定したい変数としては「？」を入力する。

薬剤	切片	d1	d2	ln_dose
DDT	1	1	0	?
γ BHC	1	0	1	?
併用	1	-1	-1	?

DDT の ED50

逆推定

確率	予測値 ln_dose	下限	上限	1-Alpha
0.50000000	1.68978114	1.58681107	1.80025800	0.9500
X値				
1	1	0	.	

γ BHC の ED50

逆推定					
	確率	予測値 ln_dose	下限	上限	1-Alpha
	0.50000000	1.42536982	1.32550595	1.52613470	0.9500
X値	1	0	1	.	

併用剤の ED50

逆推定					
	確率	予測値 ln_dose	下限	上限	1-Alpha
	0.50000000	0.528520030	0.361537139	0.670012947	0.9500
X値	1	-1	-1	.	

JMP での推定の結果は、自然対数で表されているので元の用量に直した結果を次に示す。

表 2.2 ED50 と 95%信頼区間

薬剤	ln(ED50)	ln(L95%)	ln(U95%)	ED50	L95%	U95%
DDT	1.690	1.587	1.800	5.418	4.888	6.051
γ BHC	1.425	1.326	1.526	4.159	3.764	4.600
併用	0.529	0.362	0.670	1.696	1.436	1.954

DDT と併用薬， γ BHC と併用薬の効力比はそれぞれ， $5.417/1.696 = 3.19$ 、 $4.158/1.696 = 2.45$ であるが，JMP では現在のところ推定できない。それぞれの薬剤の ED50 から効力比から点推定値から手計算できるが，95%信頼区間は，分散共分散行列が現在のところ出力できないので，SAS の GENMOD または LOGISTIC の力を借りて，分散共分散行列を求め Mathematica および Excel による効力比の 95%信頼区間の計算結果を示す。

SASによる分散共分散の計算

SAS Program Ex403-1 r/n 形式でのデータセット、インディケータ変数の生成

```
Title1 'C_ex403a.sas 2002-6-28 Y.Takahashi' ;
```

```
data C401 ;
  input dose @@ ;
  ln_dose = log(dose) ;
  do drug = '1:DDT', '2:BHC', '3:mix' ;
    input r n @@;
    D1 = (drug='1:DDT') ;
    D2 = (drug='2:BHC') ;
    D3 = (drug='3:mix') ;
  output ;
end ;
```

```

datalines ;
2.00 3 50 2 50 28 50
2.64 5 49 14 49 37 50
3.48 19 47 20 50 46 50
4.59 19 50 27 50 48 50
6.06 24 49 41 50 48 50
8.00 35 50 40 50 50 50
;
Title2 '<<< Parallel lines (class) >>>';
proc genmod data=C401 ;
class drug ;
model r/n = drug ln_dose / dist=binomial link=logit noint covb;
run ;

```

SAS Output Ex403-2 共通の傾きを持つ回帰直線、逸脱度=21.28

Criteria For Assessing Goodness of Fit							
Criterion	DF	Value	Value/DF				
Deviance	14	21.2819	1.5201				

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
drug 1:DDT	1	-4.5553	0.3611	-5.2631	-3.8474	159.10	<.0001
drug 2:BHC	1	-3.8425	0.3327	-4.4945	-3.1904	133.41	<.0001
drug 3:mix	1	-1.4248	0.2851	-1.9835	-0.8661	24.98	<.0001
ln_dose	1	2.6958	0.2157	2.2730	3.1185	156.21	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Parameter Information		
Parameter	Effect	drug
Prm1	Intercept	
Prm2	drug	1:DDT
Prm3	drug	2:BHC
Prm4	drug	3:mix
Prm5	ln_dose	

Estimated Covariance Matrix 分散共分散行列				
	Prm2	Prm3	Prm4	Prm5
Prm2	0.13043	0.10093	0.07404	-0.07170
Prm3	0.10093	0.11067	0.06762	-0.06548
Prm4	0.07404	0.06762	0.08126	-0.04804
Prm5	-0.07170	-0.06548	-0.04804	0.04652

	1:DDT	2:BHC	3:mix	ln_dose
1:DDT	0.13043	0.10093	0.07404	-0.07170
2:BHC	0.10093	0.11067	0.06762	-0.06548
3:mix	0.07404	0.06762	0.08126	-0.04804
ln_dose	-0.07170	-0.06548	-0.04804	0.04652

DDTのED50 および 95%信頼区間をExcelの計算シートを用いて計算し、JMPの出力と対比してみよう。計算に持ちいるのは、DDTの切片 $\hat{\alpha}_{\text{DDT}} = -4.5553$ 、共通の傾き $\hat{\beta} = 2.6958$ 、およびそれらの分散と共分散 $\text{Var}(\hat{\alpha}_{\text{DDT}}) = 0.13043$ 、 $\text{Var}(\hat{\beta}) = 0.04652$ 、 $\text{Cov}(\hat{\alpha}_{\text{DDT}}, \hat{\beta}) = -0.07170$ であるので、これらをシートに入力して、ED50=5.418、L95%=4.889、U95%=6.051 を得る。これは、表 2.2 の結果に一致する。

表 2.3 Excel による DDT の 95%信頼区間の算出

ロジットの回帰係数からEDxx%の推定、信頼区間の計算

x	ln(dose)	Var [^] Cov [^]	0	1
切片 beta0 [^] =	-4.5553	0	0.13043	
傾き beta1 [^] =	2.6958	1	-0.0717	0.04652

y% =	50				
ln(ED(y%)) [^] =	1.689777		ED(y%) [^] = 5.418271		
Q_a=	7.088626	ln_L95% [^] =	1.58691	L95% [^] =	4.88862
Q_b=	-24.0095	ln_U95% [^] =	1.800131	U95% [^] =	6.050442
Q_c=	20.2497				

Fillerの式, 2次方程式の根, 数値計算

DDT の回帰係数および分散共分散の推定値から Mathematica を用いて計算してみよう.

```
r = .
V = v00 + r ^ 2 v11 - 2 r v01 /.
  {v00 -> 0.13043, v01 -> -0.0717, v11 -> 0.04652}
y = (b0 - r b1) ^ 2 - z ^ 2 V /. {b0 -> -4.5553, b1 -> 2.6958, z -> 1.96}
r = -ED50
quad = Collect[y, ED50]
cl = Solve[y == 0, ED50]
```

実行結果

```
V= 0.13043 + 0.1434 r + 0.04652 r2
y= (-4.5553 - 2.6958 r)2 - 3.8416 (0.13043 + 0.1434 r + 0.04652 r2)
quad= 20.2497 - 24.0095 ED50 + 7.08863 ED502
cl= {{ED50 -> 1.58691}, {ED50 -> 1.80013}}
```

95%信頼区間の推定値は対数で, (1.58691, 1.80013) が得られる. これは, 次の効力比計算のための練習問題として示した.

Mathematica を用いて DDTと併用剤の効力比の信頼区間の計算

DDT と併用薬の効力比とその 95%信頼区間を Mathematica を用いて計算してみよう。この計算のためには、さらに併用薬の切片 $\hat{\alpha}_{\text{mix}} = -1.4248$ ，およびそれらの分散と共分散 $\text{Var}(\hat{\alpha}_{\text{mix}}) = 0.08126$ ， $\text{Cov}(\hat{\alpha}_{\text{DDT}}, \hat{\alpha}_{\text{mix}}) = -0.08126$ ， $\text{Cov}(\hat{\alpha}_{\text{mix}}, \beta) = -0.04804$ が必要となる。

```

a = {1, -1, -r}
MatrixForm[a]
v = {{v00, v01, v02}, {v01, v11, v12}, {v02, v12, v22}} /.
  {v00 → 0.13043, v01 → 0.07404, v11 → 0.08126, v02 → -0.0717,
   v12 → -0.04804, v22 → 0.04652}
MatrixForm[v]
V = a. v. a
y = (b0 - b1 - r b2) ^ 2 - z ^ 2 V /.
  {b0 → -4.5553, b1 → -1.4248, b2 → 2.6958, z → 1.96}
r = -c1
Y = Collect[y, c1]
ratio = (b1 - b0) / b2 /. {b0 → -4.5553, b1 → -1.4248, b2 → 2.6958}
Solve[Y == 0, c1]

```

実行結果

```

a=  $\begin{pmatrix} 1 \\ -1 \\ -r \end{pmatrix}$ 
V=  $\begin{pmatrix} 0.13043 & 0.07404 & -0.0717 \\ 0.07404 & 0.08126 & -0.04804 \\ -0.0717 & -0.04804 & 0.04652 \end{pmatrix}$ 
aVa= 0.06361 + 0.02366 r - (-0.02366 - 0.04652 r) r
y= (-3.1305 - 2.6958 r) ^ 2 -
  3.8416 (0.06361 + 0.02366 r - (-0.02366 - 0.04652 r) r)
quad(c1)= 9.55567 - 16.6966 c1 + 7.08863 c1 ^ 2
ratio= 1.16125
95%cl= {{c1 → 0.98032}, {c1 → 1.37509}}

```

Excelによる効力比の計算

効力比の計算式を Excel で作成した. 実際の計算は, Excel が便利である. γ BHC と DDT と γ BHC の併用薬の場合の効力比について計算結果を示す. SAS の出力から, 対応する回帰係数および分散共分散マトリックスを Excel に張付けると自動的に計算される.

表 2.4 Excel による γ BHC と併用薬の効力比の計算

ロジットの回帰係数から効力比の信頼区間の計算

x	ln(dose)	Var [^] Cov [^]	0	1	2
薬剤 S 切片 beta0 [^] =	-3.8425	0	0.11067	0.06762	-0.06548
薬剤 N 切片 beta1 [^] =	-1.4248	1	0.06762	0.08126	-0.04804
共通の傾き beta2 [^] =	2.6958	2	-0.06548	-0.04804	0.04652
y% =	50	z=	1.96		
ln(ED50(薬剤 S)) [^] =	1.4254				
ln(ED50(薬剤 S)) [^] =	0.5285	ln_効力比=	0.8968	効力比=	2.4518
Q_a=	7.0886	ln_L95% [^] =	0.7250	L95% [^] =	2.0648
Q_b=	-12.9013	ln_U95% [^] =	1.0950	U95% [^] =	2.9891
Q_c=	5.6275				

γ BHC の併用薬の効力比は 2.45, その 95%信頼区間は (2.06~2.99) である. Excel の表には, Mathematica で計算した DDT と併用薬の効力比の 95%信頼区間の計算結果を示したので, 参考にしてもらいたい.

3. 自然反応

3.1. 殺虫剤実験での自然死

多くの実験の中で、実験ユニットでの反応が刺激物によって曝露させているために単一でないことが、見出される。例えば、殺虫剤の試験で、ある害虫が、殺虫剤に曝露されていない時にでも実験期間中に死ぬかもしれないし、あるいは、曝露されている害虫が、毒素とは関係のない理由で死ぬかもしれない。これらの害虫は、自然反応、自然死があると言われていて、これらは、一般的に死亡のデータの解析の中で考慮されなければならない必要がある。殺虫剤の曝露の結果として死んだと決めるような試験で使われる個々の害虫の死後の処置にで、まれに起きる可能性がある。また、自然反応は、対照群から、または、実際の用量反応データから推定するべきであらう。

殺虫剤の試験で、比率 π が、殺虫剤なしの反応としよう。また、与えられた用量 d_i 、で殺虫剤単独の期待される比率が、 p_i 、 $i=1,2,\dots,n$ であるとしよう。比率 p_i は、自然反応 π のもとで直接観察されたものではない。そこで、用量 d_i での反応に対して期待される害虫全体の比率が、 p_i^* であるとしよう。この量は、自然に反応する害虫の比率 π 、および、殺虫剤に対する残りの比率 $(1-\pi)$ 比率 p_i から構成される。それゆえに、 p_i^* 、これは得られた反応の確率であり、次式が与えられる。

$$p_i^* = \pi + (1-\pi)p_i \quad \text{Colett (4.8)}$$

これは、アボット (Abbott) の式として知られている。 π の値は、観察された反応の最小確率であり、害虫のコントロールのバッチから推定されるであらう。コントロール群は、化合物の効果について情報がないので、他の用量反応データからわけて扱うべきである。ところで、この方法の選択は、説明変数として用量の対数を用いることによる起因する困難さでなない。実験期間中に y_0 が死ぬような n_0 からなるコントロール群の場合に、 π は、 $\pi_0 = y_0/n_0$ によって推定されなければならない。コントロール群で使用される害虫の数は、 π_0 が十分な精度をもつように多くなければならない。このような事情なので、本来の変動が、 π_0 よりも低くなるような毒素の低い用量での反応確率が、導き出されることもあり、 π_0 は、自然の死亡率の過大推定であるかもしれないし、また、それ相応に調整される必要があるかもしれない。

自然反応のゼロでない推定値に気づくための簡単な調整のための解析は、サイズが n_i のバッチの害虫の数の効用を議論することであり、それは、 i 番目の殺虫剤の用量

d_i に影響を受ける数

$$n_i^* = n_i(1 - \pi_0)$$

として、この用量での反応 y_i 、殺虫剤にたいして反応している数は、実際には

$$y_i^* = y_i - n_i \pi_0$$

を議論することである。調整された比率 y_i^*/n_i^* は、 $\text{logit}(p_i^*)$ と d_i の関係のモデル化での元として用いられる。残念なことに、 y_i^* が、もはや 2 項分布に従うことができないので、この単純な手順は、理論的には正当化されるえない。しかしながら、 π_0 が小さく、かなり正確に推定されるならば、その手順は重大な欠陥にはならない。

モデル化に際し、 y_i が $B(n_i, p_i^*)$ 分布に従うと仮定することができるような、観察された比率 y_i/n_i を考えよう。これは、

$$p_i = (p_i^* - \pi)/(1 - \pi)$$

のロジット変換が用量 d_i について直線的であると考えるのである。すなわち、

$$\text{logit}(p_i) = \beta_0 + \beta_1 d_i$$

である。 p_i^* についてのモデルは、次ぎのように変換である。

$$p_i^* = \pi + (1 - \pi) \left\{ \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)} \right\} \quad \text{Colett (4.9)}$$

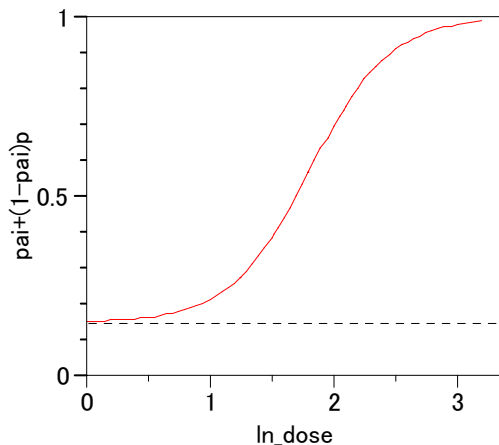


図 3.1 自然反応 $\pi = 0.15$ がある場合の用量反応

このモデルは、元のデータに対して次の尤度関数の最大化によって得られたデータにあてはめることができる。

$$L(\beta) = \prod_i \binom{n_i}{y_i} (p_i^*)^{y_i} (1 - p_i^*)^{n_i - y_i} \quad \text{Colett (4.10)}$$

ここで、 p_i^* は、 β_0 と β_1 に関する式(4.9)で与えられている。代わりに、式(4.9)でのモデルは、次の式で表すことができる。

$$\log \left\{ \frac{p_i^* - \pi_0}{1 - p_i^*} \right\} = \beta_0 + \beta_1 d_i \quad \text{Colett (4.11)}$$

これは、 p_i^* について線形ロジスティックモデルではないが、このモデルは、専門のソフトウェア SAS の PROBIT プロシジャを用いることによりあてはめられる。または、SAS の GENMOD で独自に定義のリンク関数を用いてもあてはめられる。

自然反応についての情報が、ごくわずか、あるいは存在しなければ、最も妥当な方法は、用量反応データから β_0 と β_1 に付け加える π を推定することである。その結果、 p_i^* についてのモデルは次のようになる。

$$\log \left\{ \frac{p_i^* - \pi}{1 - p_i^*} \right\} = \beta_0 + \beta_1 d_i \quad \text{Colett (4.12)}$$

ここで、 π 、 β_0 および β_1 が同時に推定される必要がある。式(4.12)で記述されたモデルは、どのようなパッケージを用いても前のモデル(式(4.11))であてはめることができるれば、あてはめることができる。モデルをあてはめる手順は、 π についての範囲と π の値を選ぶことであり、かつ、逸脱度を最小化するような β_0 および β_1 に対応している。この方法は、逸脱度を最小化するような π の値のサーチ・アルゴリズムに埋め込まれるであろう。詳細は、Colett 9.3.2節で見出すことができる。

この議論は、単一の説明変数 d_i を考えることによって動機づけられるべきである。これらのモデルは、線型の系統だったモデルの一つの説明変数、要因、および要因の組み合わせ、探索的な変数よりも多くの要素にたいし一般化する理由となる。特に、同様な手順は、平行性検定法において自然死について調整することを可能にするために使うことができる。これは、後の事例4.4で例示される。

3.2. 最大反応がある場合

生物検定法で、いくつかの個体がある刺激物、それ自体によって免疫があるような曝露された。そのような個体は、刺激物に反応することはないであろうし、たとえ、どのような高レベルの刺激物であるかにかかわらない。このような特質を自然免疫と称する。

式 4.8 によって与えられた p_i^* のモデルは、最大反応確率がある値 $\omega (<1)$ との仮定によって実験単位の中で自然免疫を一般化できる。観測された反応確率は、 $p_i^* = \pi + (1 - \pi - \omega) p_i$ となるであろう。そのパラメータ π と ω は、前に示したように、データから両方を直接推定することができ、パラメータ β と同時に推定できる。

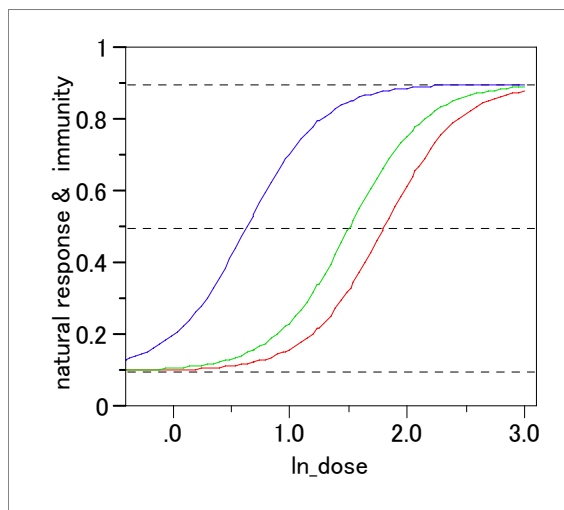


図 3.2 自然反応 10%と最大反応 90%がある場合の用量反応

3.3. 小麦粉の害虫に対する殺虫剤の毒性 (Collett 例 4.4)

Collett (例 4.3) で述べた平行性検定法は、実際に、3 種の殺虫剤のいずれも散布されなかった 200 匹の害虫がコントロール群含んでいた。6 日後にこれらの 20 匹が死んでしまいうことから、10%の自然死の率が示唆されている。

式 (4.11) で観測できる反応確率に対するモデル、10%の死亡率の仮定であるが、SAS の PROBIT でロジスティック回帰を行うか、GENMOD でユーザ指定のリンク関数もちいて平行線をあてはめたときに、このモデル下での逸脱度は、自由度 14 で 26.39 であり、結果を次に示す。

DDT :	$\text{logit}(p) = -5.643 + 3.113 \log(\text{dose})$	(自然死)
γ -BHC :	$\text{logit}(p) = -4.724 + 3.113 \log(\text{dose})$	(自然死)
DDT + γ -BHC :	$\text{logit}(p) = -1.996 + 3.113 \log(\text{dose})$	(自然死)

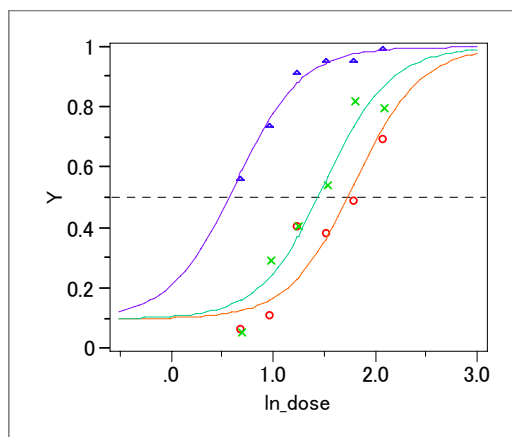


図 3.3 自然死を 10%とした場合のロジスティック曲線
DDT(—○—)、 γ BHC (—×—)、併用 (—△—)

これらの式から、 ED_{50} 値の推定が得られ、相対力価も推定される。これらの結果は、

DDTでは $6.109 \text{ mg}/10\text{cm}^2$ 、

γ -BHCでは $4.561 \text{ mg}/10\text{cm}^2$ 、

DDTと γ -BHCの混合物では $1.899 \text{ mg}/10\text{cm}^2$

であった。これらの値は、式 4.3 での報告よりも幾分大きい。

注) コレットが求めた ED_{50} は、追試できない。再計算するとDDTの ED_{50} は、 $6.109 \text{ mg}/10\text{cm}^2$ ではなくcolett式 4.11 より、

$$\log \left\{ \frac{0.5 - 0.1}{1 - 0.5} \right\} = \hat{\beta}_0 + \hat{\beta}_1 \cdot ED_{50}$$

$$\log \left\{ \frac{0.4}{0.5} \right\} = -5.634 + 3.113 \cdot ED_{50}$$

$$ED_{50} = \frac{-0.223 + 5.634}{3.113} = 1.738$$

$$\exp(ED_{50}) = \exp(1.738) = 5.686$$

$5.686 \text{ mg}/10\text{cm}^2$ となる。注、終わり

用量反応データから自然死 π を推定しようとしたときに、Collet の 9.3.2 節で記述される方法を用いてモデルの線型成分の 2 つのパラメータのについて、 $\hat{\pi}$ がゼロのとき逸脱度が見出される。これは、ゼロより大きいコントロールの死亡率をオリジナルデータの情報で十分正当化されないことが示唆している。 $\hat{\pi} = 0.0$ のときの逸脱度は、21.28 で

あり、あてはめたモデルは Collet 事例 4.3 のごとくである。

3.4. SAS による解析結果

PROBIT プロシジャによる解析結果

マニュアルのオーバービュー

The PROBIT procedure calculates maximum likelihood estimates of regression parameters and the natural (or threshold) response rate for quantal response data from biological assays or other discrete event data. This includes probit, logit, ordinal logistic, and extreme value (or gompit) regression models. Probit analysis developed from the need to analyze qualitative (dichotomous or polytomous) dependent variables within the regression framework. Many response variables are binary by nature (yes/no), while others are measured ordinally rather than continuously (degree of severity). Ordinary least squares (OLS) regression has been shown to be inadequate when the dependent variable is discrete (Collett, 1991 and Agresti, 1990). Probit or logit analyses are more appropriate in this case.

自然反応 π は、PROBIT ステートメントで、`C=rate` で与えるか、あるいは、データより推定する場合は、OPTC オプションを用いる。

controls how the natural response is handled. Specify the OPTC option to request that the natural response rate C be estimated. Specify the `C=rate` option to set the natural response rate or to provide the initial estimate of the natural response rate. The natural response rate value must be a number between 0 and 1.

自然反応として `c=0.1` を与えた場合の解析事例を示す。PROBIT プロシジャは、この名称から類推されるようにデフォルトでは、2 値反応に対してプロビット変換が用いられるが、`model` ステートメントのオプションで `dist=logistic` とすることによりロジット変換にも対応している。

次にプログラムを示す。2 値反応は、`r/n` の形式で与え、`class` ステートメントで `drug` を、`model` ステートメントで切片なしのモデルとするために `noint` オプションを指定している。モデルの当てはまりの良さの統計量、逸脱度は、`lackfit` オプションの指定で出力される。

プログラム

```
Title2 ' <<< Natural Response >>> ' ;
proc probit data=C403 c=0.1 ;
  class drug ;
  model r/n = drug ln_dose / dist=logistic noint lackfit;
run ;
```

実行結果

		Probit Procedure		Model Information		
		Name of Distribution		Logistic		
		Log Likelihood		-418.2341575		
Goodness-of-Fit Tests						
Statistic		Value		DF	Pr > ChiSq	
Pearson	Chi-Square	24.2963		14	0.0422	
L.R.	Chi-Square	26.3892		14	0.0231	
Type III Analysis of Effects						
		Wald				
Effect	DF	Chi-Square	Pr > ChiSq			
drug	3	155.0194	<.0001			
ln_dose	1	126.9122	<.0001			
Analysis of Parameter Estimates						
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000 0.0000	.	.
drug 1:DDT	1	-5.6343	0.5049	-6.6238 -4.6448	124.55	<.0001
drug 2:BHC	1	-4.7244	0.4471	-5.6007 -3.8481	111.66	<.0001
drug 3:mix	1	-1.9968	0.3395	-2.6621 -1.3315	34.60	<.0001
ln_dose	1	3.1135	0.2764	2.5718 3.6552	126.91	<.0001
C	0	0.1000	0.0000	0.1000 0.1000		

実行結果から、それぞれの薬剤の切片が、-5.6343、-4.7244、-1.9968、共通の切片として 3.1135 が推定され、_c_ に与えた自然反応が出力されている。

なお、このデータで、

GENMODプロシジャに解析

SAS の GENMOD プロシジャでは、リンク関数および逆リンク関数をユーザが定義できるようになっている。SAS の GENMOD プロシジャは、リリース 6.09 で初登場の新しいプロシジャである。そのオーバービューを次に示す。

The GENMOD procedure fits generalized linear models, as defined by Nelder and Wedderburn (1972). The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an

exponential family of distributions. Many widely used statistical models are generalized linear models. These include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data. Many other useful statistical models can be formulated as generalized linear models by the selection of an appropriate link function and response probability distribution. Refer to McCullagh and Nelder (1989) for a discussion of statistical modeling using generalized linear models. The books by Aitkin, Anderson, Francis, and Hinde (1989) and Dobson (1990) are also excellent references with many examples of applications of generalized linear models. Firth (1991) provides an overview of generalized linear models.

FWDLINK Statement

FWDLINK variable = expression ;

You can define a link function other than a built-in link function by using the FWDLINK statement. If you use the MODEL statement option LINK= to specify a link function, you do not need to use the FWDLINK statement. The variable identifies the link function to the procedure. The expression can be any arithmetic expression supported by the DATA step language, and it is used to define the functional dependence on the mean.

Alternatively, the link function can be defined by using programming statements (see the "Programming Statements" section) and assigned to a variable, which is then listed as the expression. The second form is convenient for using complex statements such as if-then-else clauses. The GENMOD procedure automatically computes derivatives of the link function required for iterative fitting. You must specify the inverse of the link function in the INVLINK statement when you specify the FWDLINK statement to define the link function. You use the automatic variable `_MEAN_` to represent the mean in the preceding expression.

INVLINK Statement

INVLINK variable = expression ;

If you define a link function in the FWDLINK statement, then you must define the inverse link function using the INVLINK statement. If you use the MODEL statement option LINK= to specify a link function, you do not need to use the INVLINK statement. The variable identifies the inverse link function to the procedure. The expression can be any arithmetic expression supported by the DATA step language, and it is used to define the functional dependence on the linear predictor.

Alternatively, the inverse link function can be defined using programming statements (see the section "Programming Statements") and assigned to a variable, which is then listed as the expression. The second form is convenient for using complex statements such as if-then-else clauses. The automatic variable `_XBETA_` represents the linear predictor in the preceding expression.

これらのマニュアル上の記載に基づき、自然反応を 10%とする GENMOD プロシジャを次に示す。Fwdlink ステートメントで用いる発現率 `p` を用いる。このために `p = _mean_` と前もって定義しておく。Invlink ステートメントで引用する回帰式は、を `xbeta = _xbeta_` として定義する。`_mean_` および `_xbeta_` は SAS の内部変数である。

プログラム

```
Title2 '<<< Parallel lines (Natural Response) >>>';
proc genmod data=C403 ;
  class drug ;
  pai = 0.1 ;
  p = _mean_ ;
  xbeta = _xbeta_ ;
  fwdlink link = log((p-pai)/(1-p)) ;
  invlink ilink = pai+(1-pai)*(exp(xbeta)/(1+exp(xbeta))) ;
  model r/n = drug ln_dose / dist=binomial noint covb ;
run ;
```

実行結果

```

<<< C_ex404a.sas  Parallel lines (Natural Response) >>>
The GENMOD Procedure
Model Information

Criteria For Assessing Goodness Of Fit
Criterion          DF          Value          Value/DF
Deviance           14          26.3892          1.8849
Scaled Deviance    14          26.3892          1.8849
Pearson Chi-Square  14          24.2963          1.7354
Scaled Pearson X2  14          24.2963          1.7354
Log Likelihood                    -418.2342

Estimated Covariance Matrix
Prm2          Prm3          Prm4          Prm5
Prm2    0.25488    0.19666    0.13315    -0.13041
Prm3    0.19666    0.19989    0.11761    -0.11519
Prm4    0.13315    0.11761    0.11523    -0.07799
Prm5   -0.13041   -0.11519   -0.07799    0.07638

```

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
drug 1:DDT	1	-5.6343	0.5049	-6.6238	-4.6448	124.55	<.0001
drug 2:BHC	1	-4.7244	0.4471	-5.6007	-3.8481	111.66	<.0001
drug 3:mix	1	-1.9968	0.3395	-2.6621	-1.3315	34.60	<.0001
ln_dose	1	3.1135	0.2764	2.5718	3.6552	126.91	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

4. 非線型ロジスティック回帰モデル

4.1. ロジスティック直線が当てはまらない場合

観測された 2 値反応の出現確率の経験ロジット変換が、与えられた説明変数に対してプロットされた場合に、2つの関係が、直線でないかもしれない。もっと複雑なモデルが、同定される必要があるだろう。ロジスティック回帰のあてはめの目的が、ED90 あるいは、もっと端の ED99 値のような極端な反応の用量の推定であるならば、端の用量での観測値に近づくようなモデルの選定は特に重要である。

真の反応確率 p と説明変数 x の間の関連のモデル、実用的な見地での同定は、多項式ロジスティック回帰モデルである。一般的な形式は次のようである。

$$\text{logit}(p) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

この $\text{logit}(p)$ についてのモデルは、パラメータ β について依然として線型であり、他のモデルよりもより簡単にあてはめられる。Morgan (1985) は、用量の端について 3 次の多項式回帰 ($k=3$) が多くのデータセットへの良いあてはめとなることを見出した。彼は、また、3つのパラメータモデル

$$\text{logit}(p) = \beta_1 (x - \mu) + \beta_2 (x - \mu)^3 \quad \text{Collett (4.13)}$$

ここで、 μ は ED50 値

を見出し、また、このモデルと 3 次多項式とのいずれかを選ぶことはほとんどないことを見出した。式 (4.13) で与えられたモデルのあてはめの一つの方法は、GLIM のようなパッケージを用いることであり、 μ の値の格子点についてパラメータ β_1 と β_2 を推定し、逸脱度を最小にするような組み合わせを得ることである。この方法の難点の一つは、推定された $\hat{\mu}$ の分散だけでなく、 $\hat{\mu}$ を含む共分散を得ることができない。必要となるのであれば、例えば、推定された ED90 値の標準誤差が必要とされるならばである。

4.2. 小麦粉の害虫 (confused flour beetles、こくぞう虫?) の死亡率 (Collett 事例 4.5)

元々 Strand (1930) に最初に報告され、Bliss (1935) によって引用された実験のねらいは、*Tribolium confusum*, ガス状の二硫化炭素 (CS_2) に対する小麦粉の害虫の反応を評価することであった。実験では、液状の二硫化炭素の規定量が、約 30 匹の害虫のバッチがはいっている平らな布のケージが吊るされたフラスコへ加えられた。同様な害虫のバッチが、 CS_2 各濃度に対して使われた。5 時間後に死亡数、および、フラスコ中の用

量分析によって測定されたガス状の CS₂実濃度、測定値はmg/L、が記録された。死亡率は、表 4.1で与えられている。

表 4.1 二硫化炭素ガスの異なる濃度で曝露された n 匹の死亡した害虫の数 y

CS ₂ の 濃度	反復 1		反復 2	
	y	n	y	n
49.06	2	29	4	30
52.99	7	30	6	30
56.91	9	28	9	34
60.84	14	27	14	29
64.76	23	30	29	33
68.69	29	31	24	28
72.61	29	30	32	32
76.54	29	29	31	31

多くの論文でこのデータを引用しており、最初の 2 つの濃度は、見かけ上の非線型性の理由で落とされている。ブリス自身も次ぎのように注釈をつけている。

残りの観察データでの比較で、2 つの低濃度は、例外的に高死亡を与えた。残りの濃度全体で、プロットした値 ほどよく直線に見えるので、データは、2 つに区別して扱われ、リッターあたり CS₂ の 96.91mg での結果だけが、両方のデータセットに含まれるであろう。

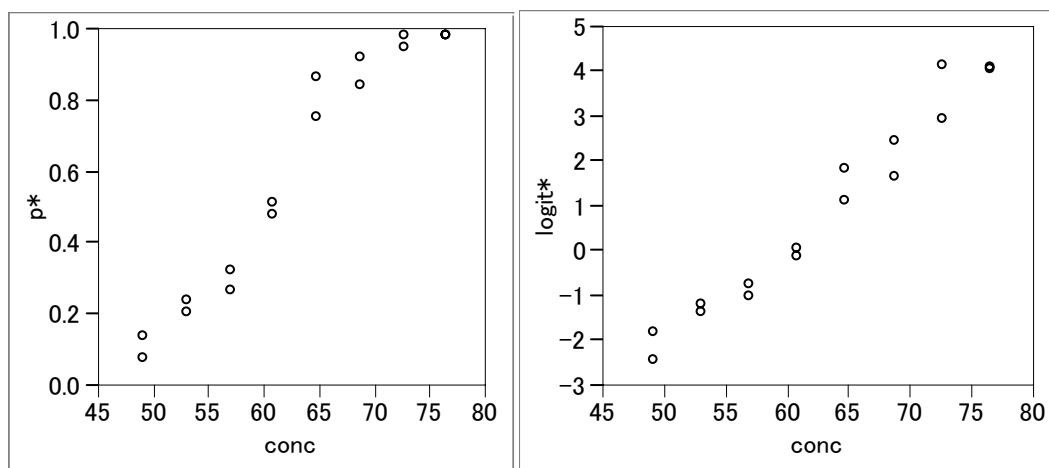


図 4.1 死亡率とロジット変換後のプロット

しかしながら、これに対するいかなる生物学的な動議づけも見出されなく、データセットの中に残されている。

4.3. 2次曲線のあてはめ

2つの反復のデータを合わせて、濃度に対して得られた出現率の経験ロジットのプロットが、図 4.2 に示されている。死んだ害虫の比率のロジット変換と濃度 x の間の非線型のような関連は、濃度の対数をとっても除くことができず、濃度自身が、説明変数として持ちいられる。 p_i をCS₂の x_i mg/Lの濃度で曝露されたときの害虫が死んだ確率としたときに、モデル $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$ のあてはめで、逸脱度は自由度 14 で 12.51 である。このモデルに対して平均逸脱度は、1 よりも小さいので、表面上このモデルは受けいられるようにみえる。あてはめられたモデル の式は、 $\text{logit}(\hat{p}) = -14.81 + 0.249 x$ であり、図 4.4 でデータについてプロットされているが、観測されたデータにあてはまっていないように見える。

x の 2 次項含んだ場合に、逸脱度は、自由度 13 で 7.93 に減少する。この減少は、自由度 1 で、4.58 であり、高度に有意であり、このロジスティック回帰の 2 次式は、直線回帰のものよりもあてはまっている。3 次の項 x^3 が、モデルに加えられたときに、逸脱度は、さらに自由度 1 で 0.045 だけ減少し、あてはめ特性は、そんなに改善されない。あてはめられた 2 次式を次に示す。

$$\text{logit}(\hat{p}) = 7.968 - 0.517 x + 0.00637 x^2$$

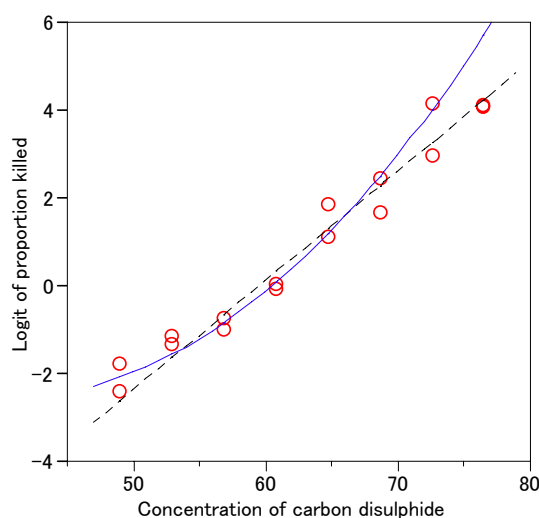


図 4.2 ロジット変換値での直線および 2 次曲線のあてはめ

----- : 直線, — : 2 次曲線

この図は、ロジスティック 2 次曲線が、CS₂の 76.54mg/Lの濃度に対して得られたデー

タにあてはまっていないことを示唆しているが、ミスリーディングである。2つのバッチの害虫の得られた出現率は、この濃度で両方とも一致した作用であるが、しかし、これらの値の経験ロジット変換の使用は、グラフ上でそれらを含むようにしている。??? この濃度あてはめられた反応の確率は、正確には 0.997 であり、この値は得られた値に非常に近い。次ぎの図は、異なるガス濃度での作用にたいして害虫の得られた出現率が、2次のロジスティック回帰モデルによってよくあてはまっていることを示している。

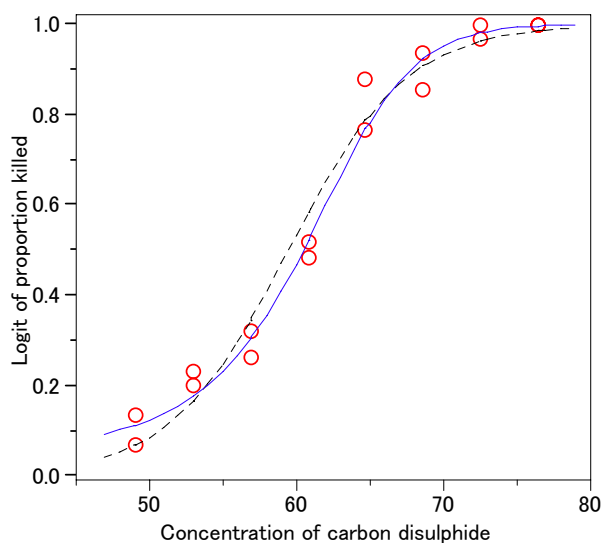


図 4.3 死亡率での2次曲線のあてはめ

----- : 直線シグモイド, — : 2次曲線シグモイド

2次のロジスティック回帰モデルに基づいて ED_{90} 値を推定するために、次ぎの式を用いる。

$$\text{logit}(0.9) = \log(9) = 7.968 - 0.517 ED_{90} + 0.00637 ED_{90}^2$$

この $\log(9)$ を移行して、次ぎの2次式を得る。

$$0.00637 ED_{90}^2 - 0.517 ED_{90} + 5.771 = 0$$

この式の2つの根は、67.80 および 13.36 であり、67.80mg/Lが、 CS_2 ガスに曝露された害虫の90%反応を起こすような推定値である。

$$\text{logit}(0.9) = \log(9) = 7.968 - 0.517 ED_{90} + 0.00637 ED_{90}^2$$

JMPによる2次式のロジスティック回帰

2次式をJMPで当てはめることは容易である。モデルの当てはめでモデル式を構成するときに、conc*concのように解析時に交互作用を生成する。データに交互作用項を作る必要はない。

名義ロジスティックのあてはめ

重み: n

モデル全体の検定

モデル	(-1)*対数尤度	自由度	カイ2乗	p値(Prob>ChiSq)
差	140.60698	2	281.214	<.0001
完全	182.11353			
縮小	322.72051			

R2乗(U) 0.4357
 オブザベーション(または重みの合計) 481

勾配による収束

あてはまりの悪さ(LOF)

要因	自由度	(-1)*対数尤度	カイ2乗	p値(Prob>ChiSq)
あてはまりの悪さ(LOF)	5	1.49424	2.988483	
飽和モデル	7	180.61929		
あてはめたモデル	2	182.11353		0.7018

パラメータ推定値

項	推定値	標準誤差	カイ2乗	p値(Prob>ChiSq)
切片	-17.211187	1.9737161	76.04	<.0001
conc	0.28452486	0.0314599	81.79	<.0001
(conc-62.8611)*(conc-62.8611)	0.00637211	0.0031429	4.11	0.0426

推定値は次の対数オッズに対するものです: 0/1

効果に対するWald検定

要因	パラメータ数	自由度	Waldカイ2乗	p値(Prob>ChiSq)
conc	1	1	81.7946695	0.0000
conc*conc	1	1	4.11053155	0.0426

効果の尤度比検定

要因	パラメータ数	自由度	尤度比カイ2乗	p値(Prob>ChiSq)
conc	1	1	278.086496	0.0000
conc*conc	1	1	4.57793404	0.0324

薬物濃度のパラメータの推定値は、薬物濃度データの平均値を引いて算出されているので、元の濃度で2次式になることを次に示した。このような2乗の項に対する変換は、多項式回帰係数に意味付けしやすくなる工夫である。

$$\begin{aligned}
 \text{logit}(p) &= -17.21119 + 0.28452x + 0.00637211(x - 62.8611)^2 \\
 &= -17.21119 + 0.28452x + 0.00637211(x^2 - 2 \times 62.8611x + 62.8611^2) \\
 &= 7.968 - 0.517x + 0.00637x^2
 \end{aligned}$$

$\text{logit}(p) = b_0 - b_1x + b_2x^2$ の場合、1 次項 b_1 は負になっていて、用量が増えると反応が減るかのような錯覚が生じる。JMP のパラメータ推定値の標準誤差の使用は注意が必要である。また、回帰係数の検定においても、 b_1 は、次の GENMOD プロシジャで示すが、 $p = 0.1668$ と有意でなく、単回帰分析の結果とまったく異なる。これは、多項式回帰において、1 次式の解析の後に 2 次の項を追加し、さらに 3 次の項を追加するという逐次的解析が必要なのであるが、このために、GENMOD プロシジャでは model ステートメントで TYPE1 オプションの追加が必要となる。JMP では、このような煩雑さを除くために、 $\text{logit}(p) = b_0 - b_1x + b_2(x - \bar{x})^2$ のような変換をしている。

JMP では、同じ用量での反復を無視して、データをプールした上でロジスティック回帰を行っている。そのために、逸脱度、JMP では、あてはめの悪さ (LOF) の自由度が 5 となっている。同じ濃度で 2 回の実験が繰り返し行なわれ、それを考慮した解析にはなっていないことに注意が必要である。

SASのGENMODプロシジャによるあてはめ

$\text{logit}(p) = b_0 - b_1x + b_2x^2$ の解析となっている。過大発散度の自由度は、13 となっている。適切な解析となっている。

元の薬物濃度のままでの解析

プログラム

```
Title1 'C_ex405a.sas 2002-7-19 Y.Takahashi' ;

data C405 ;
  input conc @@ ;
  conc2 = conc - 62.8611 ;
  do rep = 'R1', 'R2' ;
    input r n @@;
    R1 = (rep='R1') ;
    R2 = (rep='R2') ;
    output ;
  end ;
datalines ;
49.06 2 29 4 30
52.99 7 30 6 30
56.91 9 28 9 34
60.84 14 27 14 29
64.76 23 30 29 33
68.69 29 31 24 28
72.61 29 30 32 32
76.54 29 29 31 31
;
```

```
proc genmod data=C405 ;
  model r/n = conc conc*conc / link=logit dist=binomial type1 ;
run ;
```

実行結果

Criteria For Assessing Goodness Of Fit							
Criterion	DF	Value	Value/DF				
Deviance	13	7.9273	0.6098				
Scaled Deviance	13	7.9273	0.6098				
Pearson Chi-Square	13	7.3590	0.5661				
Scaled Pearson X2	13	7.3590	0.5661				
Log Likelihood		-182.1135					

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	7.9684	11.0363	-13.6624	29.5992	0.52	0.4703
conc	1	-0.5166	0.3736	-1.2489	0.2157	1.91	0.1668
conc*conc	1	0.0064	0.0031	0.0002	0.0125	4.11	0.0426
Scale	0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	289.1413			
conc	12.5053	1	276.64	<.0001
conc*conc	7.9273	1	4.58	0.0324

薬物濃度の平均値を引いた2次の項を用いたGENMODプロシジャ

JMP 流に $\text{logit}(p) = b_0 - b_1x + b_2(x - \bar{x})^2$ のモデルを当てはめた結果を示す。ただし、元のデータセットの作成時に、薬物濃度の平均値を引いておくことが必要で、データにより、プログラムの変更が必要で、実用性には欠ける。

プログラム

```
proc genmod data=C405 ;
  model r/n = conc conc2*conc2 / link=logit dist=binomial type1 ;
run ;
```

実行結果

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13	7.9273	0.6098
Scaled Deviance	13	7.9273	0.6098
Pearson Chi-Square	13	7.3590	0.5661
Scaled Pearson X2	13	7.3590	0.5661
Log Likelihood		-182.1135	

Analysis of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-17.2112	1.9737	-21.0796	-13.3428	76.04	<.0001
conc	1	0.2845	0.0315	0.2229	0.3462	81.80	<.0001
conc2*conc2	1	0.0064	0.0031	0.0002	0.0125	4.11	0.0426
Scale	0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	289.1413			
conc	12.5053	1	276.64	<.0001
conc2*conc2	7.9273	1	4.58	0.0324

LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	289.1413			
conc	12.5053	1	276.64	<.0001
conc*conc	7.9273	1	4.58	0.0324

SASのLOGISTICプロシジャによるあてはめ

単純な使い方では、問題が多い。過大発散度の章で詳しく示すので、ここでは割愛する。

4.4. 非線形モデル

多項式回帰モデルは、しばしば課題の領域の文脈の言葉の中で正当化が困難であり、そのために相当の注意が、それらを使うためにとられるべきである。それらは、実験中に使用された用量水準の外側での奇妙な振舞いよりむしろ著しく影響を受けやすいことである。それらをもちいるための一つの代替案は、未知のパラメータに対して非線形の関数によって反応確率のロジスティック変換モデル化することである。例えば、

$$\text{logit}(p) = \beta_0 e^{\beta_1 x}$$

このモデルは、反応確率のロジスティック変換において指数的に増加するか、あるいは、減少し、そして一方、

$$\text{logit}(p) = \beta_0 + \beta_1 e^{\beta_2 x}$$

このモデルは、ロジットスケール上で漸近的な回帰である。原則として、これらのモデル最尤法によってあてはめられることができる、しかし、通常は多項式モデルよりもっと困難であり、それらをあてはめることができる一般的なコンピュータソフトウェアはない。

2値データに対するモデルのあてはめを改善するための他の方法は、ロジスティック回帰よりもフレキシブルであるような反応確率の変換を用いることである。Colett、

Section 5.3 を参照.

最尤法ではないが、経験ロジットを反応変数とした非線型最小 2 乗法によるモデルによるあてはめを行なってみる。

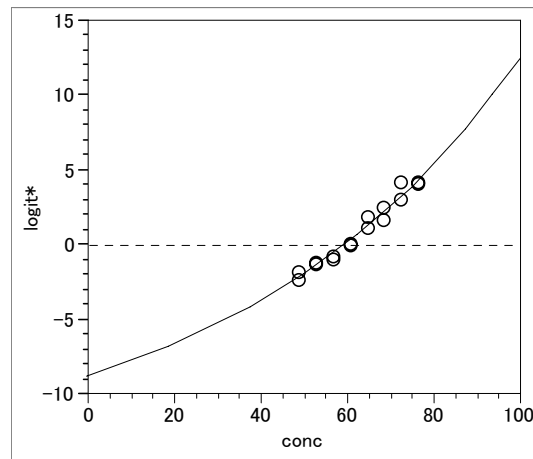


図 4.4 指数関数のあてはめ

JMPの非線形モデル

解				
	SSE	DFE	MSE	RMSE
	77.13672235	478	0.1613739	0.4017137
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
b0	-14.63278704	2.23779943	-20.609234	-11.247507
b1	5.8874712127	1.6798748	3.50149337	10.6987418
b2	0.0152966673	0.0022096	0.01103867	0.0195676

$$\text{logit}(p) = -14.63 + 5.887 e^{0.0153 x}$$

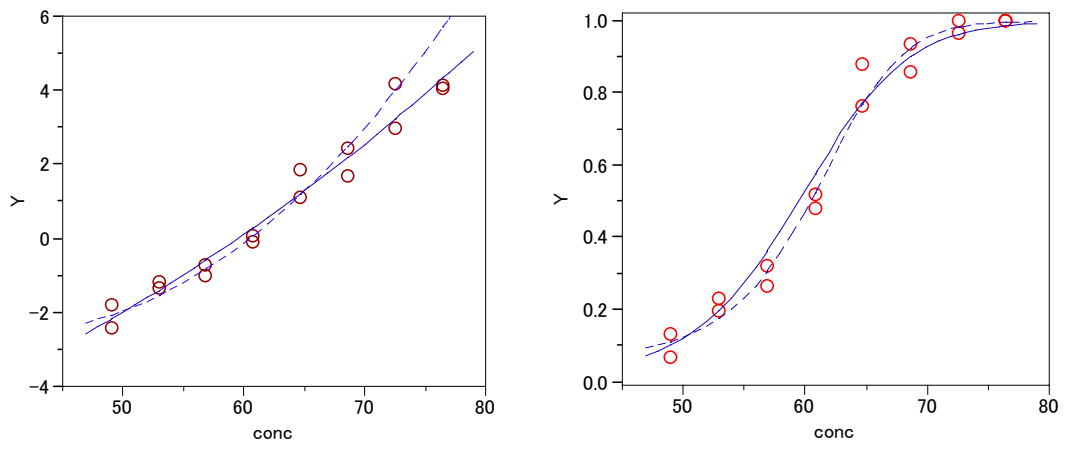


図 4.5 2次シグモイドおよび指数シグモイド曲線のあてはめ

----- : 2次曲線シグモイド, —— : 指数シグモイド

5. 補 2 重対数の適用例

2 値反応データのモデリングにおいて、ときには有益となる確率のスケールの変換は、 $\log[-\log(1-p)]$ が説明変数の線型関数としてのモデル化されるような補 2 重対数である。この変換は、3.5.3 節で記述された。生物検定法の文脈の中で、補 2 重対数モデルは、個体の認容性が、確率密度関数について Gumbel 分布として知られている極値の分布を仮定することによって推論される。

$$f(u) = \frac{1}{\kappa} e^{(u-\alpha)/\kappa} \exp[-e^{(u-\alpha)/\kappa}], \quad -\infty < u < \infty$$

ここで、 $-\infty < \alpha < \infty$ 、 $\kappa > 0$ は、未知パラメータである。プロビットおよびロジット変換が基盤となっている正規およびロジスティック分布とは違って、 α を最頻値 (mode)、 α より大きな平均値を持つような近似的な認容性の分布である。分布の分散は、 1.645κ である。用量 d_i で曝露された時の反応の確率は、

$$p_i = \int f(u) du = 1 - \exp[-e^{(d_i-\alpha)/\kappa}]$$

であり、そこで、 $\log[-\log(1-p_i)] = \beta_0 + \beta_1 d_i$ 、ここで、 $\beta_0 = -\alpha/\kappa$ および $\beta_1 = 1/\kappa$ が未知のパラメータである。補 2 重対数関数は、再び $(-\infty, \infty)$ の値に対して $(0, 1)$ 範囲の確率に変換するが、この関数は $p = 0.5$ で対称でない：実際に、 $p = 0.5$ のとき変換値は -0.3665 である。

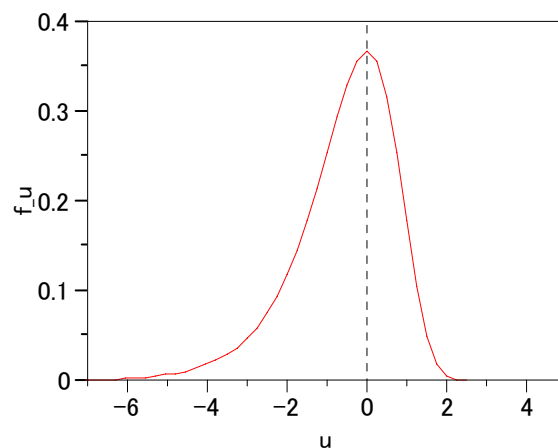


図 5.1 $\alpha=0$ 、 $\kappa=1$ の場合のガンベル確率密度関数

ガンベル耐性分布は、同様に物質の破壊応力に関する問題で遭遇する。最も弱いリンク理論、それは、適用される力が、最も弱い部分またはリンクの耐性より長い時に成分

破壊に関してであり、??? 極小の分布、ガンベル分布であるように示唆する ???
 これは、失敗の確率に関して補2重対数モデルを導く。補2重対数モデルに直接導くよ
 うな幾つかの他の文脈は、次ぎの小節でさらに詳しく記述する。

5.1. 希釈法

溶液中の感染有機物の多くは、直接測定されえなく、希釈法がたぶん存在する有機物
 の数を推定するために使われる。このタイプの分析法は、有機物を含んでいる溶液が、
 逐次的に希釈され、それぞれの希釈の検体が、成長培地を含む多くのプレートにたいし
 て適用される。培養期間後に、成長したプレートの割合が記録される。元の溶液が、2
 倍ごとに希釈されるとすると、半分の強さ、4分の1の強さ、などのような溶液となる。
 N を元の溶液中の一単位容積あたりの有機物の数としたときに、 $1/2^i$ ごとに希釈された溶
 液中の対応する数は、 $N/2^i$ 、 $i=0, 1, 2, \dots, m$ であると期待される。ランダムサンプリング
 の元で、プレート上に堆積しているであろう希釈剤の一単位容積あたりの有機物の数は、
 ポアソン分布になると仮定することができる。ポアソン分布は、平均値が μ_i で、 μ_i は、
 i 番目の希釈の一単位容積あたりの有機物の数と期待される。従って、 $\mu_i = N/2^i$ となり、
 $\log \mu_i = \log N - i \log 2$ である。ポアソン分布の仮定の基で、 i 番目の希釈物でのプレート
 が、何も有機物含まないチャンスは、 $e^{-\mu_i}$ であり、同様に、 p_i がこの希釈物上での
 成熟が起きる確率であれば、 $p_i = 1 - e^{-\mu_i}$ である。それゆえに、 $\mu_i = -\log(1 - p_i)$ 、
 $\log \mu_i = \log[-\log(1 - p_i)]$ となり、

$$\log[-\log(1 - p_i)] = \log N - i \log 2$$

となる。このモデルが成り立つデータは、プレートの数、 y_i から成り立つ。ここで、 y_i
 は、元の溶液の i 番目の希釈液に接種された n_i プレート中の成長が起きるプレート数で
 ある。 $\beta_0 = \log N$ とし、 $\beta_1 = -\log 2$ とすると、

$$\log[-\log(1 - p_i)] = \beta_0 + \beta_1 i$$

となる。ここで、 i は説明変数であり、 β_1 は、回帰係数として知られている。ここでの
 直接的な興味は、 $\hat{N} = \exp(\hat{\beta}_0)$ によって推定される N である。ここで、 $\hat{\beta}_0$ は、上記のモ
 デルの β_0 の最尤推定量である。 N についての $100(1-\alpha)\%$ 信頼区間は、
 $\hat{\beta}_0 \pm z_{\alpha/2} \text{s.e.}(\hat{\beta}_0)$ から $\log N$ についての最初に計算された信頼区間によって得ること
 ができる。ここで、 $z_{\alpha/2}$ は、標準正規分布の上側 $100\alpha/2$ 点である。信頼限界の結果は、
 従って、 N 自身に対する信頼区間の要求について、指数を取ることである。

表 5.1 希釈法のデータ例

i	x	myu	$\log(\text{myu})$	1-p	p	$-\log(1-p)$	$\log(-\log(1-p))$
0	1	10	2.3	0.0001	1.0000	10	2.30
1	0.5	5	1.61	0.0067	0.9933	5	1.61
2	0.25	2.5	0.92	0.0821	0.9179	2.500	0.92
3	0.125	1.25	0.22	0.2865	0.7135	1.250	0.22
4	0.0625	0.625	-0.47	0.5353	0.4647	0.625	-0.47
5	0.0313	0.3125	-1.16	0.7316	0.2684	0.3125	-1.16
6	0.0156	0.1563	-1.86	0.8554	0.1447	0.1563	-1.86
7	0.0078	0.0781	-2.55	0.9249	0.0752	0.0781	-2.55
8	0.0039	0.0391	-3.24	0.9617	0.0383	0.0391	-3.24
9	0.0020	0.0195	-3.94	0.9807	0.0193	0.0195	-3.94
10	0.0010	0.0098	-4.63	0.9903	0.0097	0.0098	-4.63
11	0.0005	0.0049	-5.32	0.9951	0.0049	0.0049	-5.32
12	0.0002	0.0024	-6.02	0.9976	0.0024	0.0024	-6.02

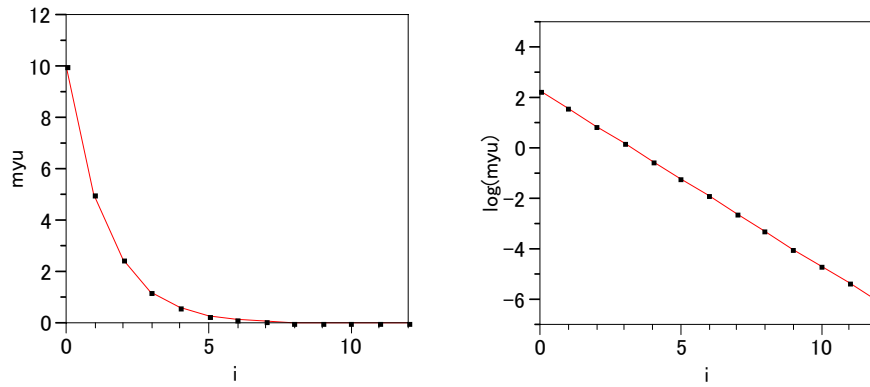


図 5.2 希釈法での μ および $\log(\mu)$

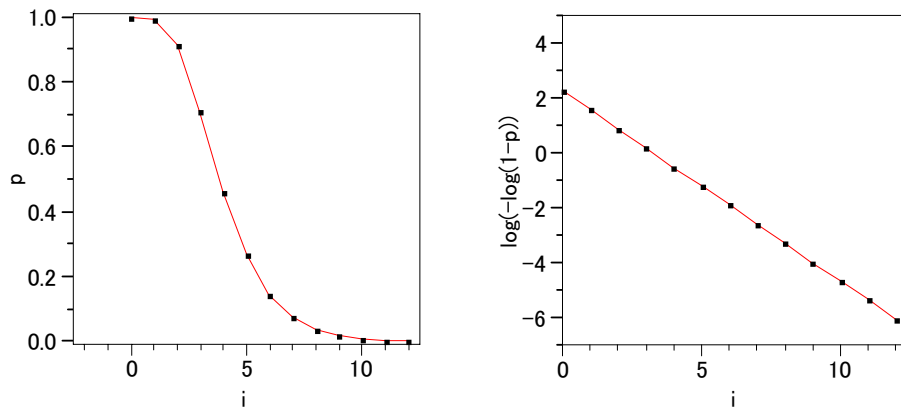


図 5.3 出現率と補二重対数変換のグラフ

5.2. 4 グループ化された生存時間解析

生存時間解析で、一般的に故障という用語であるベントまでの測定されたいくらかの時間の元の長さについて興味のある打ち切り???

未完

5.3. 血清学的テスト

個体が曝露されたとき,... 次の事例を含めて, 未完

5.4. マラリアについての血清学的テスト (事例 4.6)

年間の感染率

Draper, Voller and Carpenter (1972) は、アフリカと南アメリカの幾つかの地域から集められたマラリアについて血清学データの解釈について議論している。、ブラジルのアマゾンの2つの地域で1971年に実施された血清学的な調査で、村のなかでマラリア寄生虫 *Plasmodium Vivax* 抗体の存在を検出するために間接的な蛍光抗体検査が用いられた。データは、血清学的陽性が見出された7つの年齢グループでの個体の出現率を参照して表4.3に再現されている。同様に表の中にそれぞれの年齢の範囲の midpoint が与えられているが、最後の年齢グループは、元の論文の図の一つから推定された。モデル $\log[-\log(1-p_i)] = \beta_0 + \beta_1 A_i$ をあてはめたGLIMの出力は、表5.2の後に与えられている。ここで p_i は、血清学的陽性である年齢 A_i の個人ごとの期待出現率であり、 A_i は、年齢の範囲の midpoint が取られている。

表 5.2

Age group	mid_age	y	n	p	log[-log(1-p)]
0-11 months	0.5	3	10	0.300	-1.03
1-2 years	1.5	1	10	0.100	-2.25
2-4 years	3	5	29	0.172	-1.66
5-9 years	7	39	69	0.565	-0.18
10-14 years	12	31	51	0.608	-0.07
15-19 years	17	8	15	0.533	-0.27
>=20 years	30	91	108	0.843	0.61

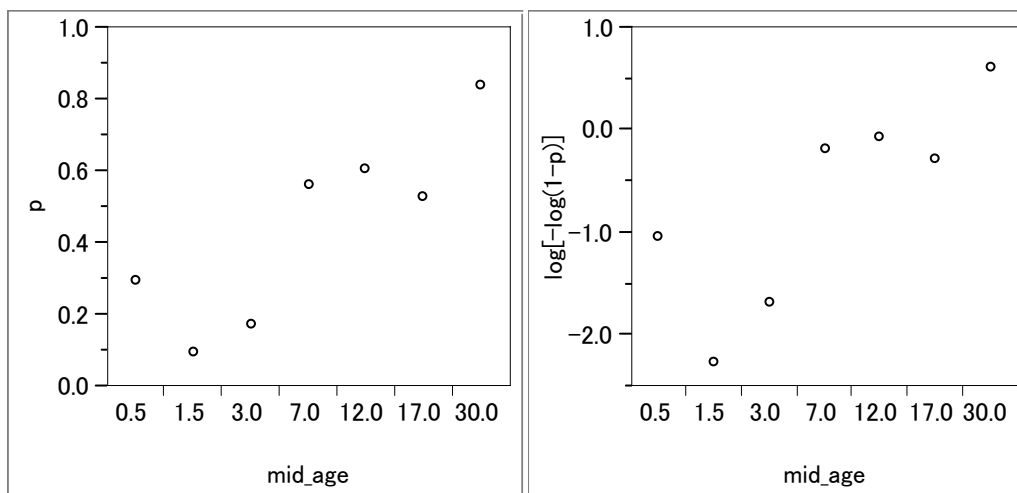


図 5.4

一般化線型モデルでの用語で、GLIM パッケージの暗黙の了解で、回帰係数についての用語で、オフセット `offset` として知られている???。この例で、用語 $\log A_i$ は、オフセットとして定義される。この出力の中で直接的な興味のある量は、 $\hat{\beta}_0 = -2.607$ であるような $\beta_0 = \log \mu$ の推定値である。これから、 $\hat{\mu} = e^{-2.607} = 0.074$ となり、感染率 h は、 $\hat{h} = 1 - e^{-\mu} = -0.071$ によって推定される。アマゾンでのこのタイプのマラリアについて年間の感染率は、それゆえに 1000 人中、約 71 人の感染者と推定される。

推定された年間の感染率の標準誤差は、GLIM の出力にある $\beta_0 = \log \hat{\mu}$ の標準誤差から決定することができる。あるパラメータ $\hat{\theta}$ の推定値関数 $g(\hat{\theta})$ の分散の近似は、次式で得られる。

$$\text{Var}[g(\hat{\theta})] \approx \left\{ \frac{\partial g(\hat{\theta})}{\partial \hat{\theta}} \right\}^2 \text{Var}[\hat{\theta}]$$

この結果の特別な場合が敷き式 (4.1) で与えられている。この場合には、 \hat{h} の分散は、次式で要求されている。

$$\hat{h} = 1 - \exp[-e^{\hat{\beta}_0}]$$

式 (4.16) を用いて \hat{h} の近似的な分散が得られ、 \hat{h} の標準誤差は次式で与えられる。

$$\text{s.e.}(\hat{h}) \approx e^{\hat{\beta}_0} \exp[-e^{\hat{\beta}_0}] \text{s.e.}(\hat{\beta}_0)$$

$\hat{\beta}_0 = 2.507$ 、 $\text{s.e.}(\hat{\beta}_0) = 0.085$ なので、 $\text{s.e.}(\hat{h}) \approx 0.0059$ を見出せる。

SASのGENMODプロシジャによる解析

プログラム

```
Title 'malaria.sas 2002-10-08 Y.Takahashi' ;

data d01 ;
  input Age_group$15. mid_age y n p c_log2 ;
  log_age = log(mid_age) ;
datalines ;
0-11_months 0.5 3 10 0.300 -1.03
1-2_years 1.5 1 10 0.100 -2.25
2-4_years 3.0 5 29 0.172 -1.66
5-9_years 7.0 39 69 0.565 -0.18
10-14_years 12.0 31 51 0.608 -0.07
15-19_years 17.0 8 15 0.533 -0.27
>=20_years 30.0 91 108 0.843 0.61
;

proc genmod data=d01 ;
  model y/n = / dist=binomial link=c11 offset=log_age ;
run ;
```

実行結果

Criteria For Assessing Goodness Of Fit						
Criterion	DF	Value	Value/DF			
Deviance	6	20.2533	3.3755			
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-2.6068	0.0831	-2.7698 -2.4438	982.98	<.0001
Scale	0	1.0000	0.0000	1.0000 1.0000		

年齢階層ごとのあてはめ

プログラム

```
proc genmod data=d01 ;
  class age_group ;
  model y/n = age_group / dist=binomial link=c11 offset=log_age noint ;
run ;
```

実行結果

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood		-161.4651	

Analysis Of Parameter Estimates							
	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
	0	0.0000	0.0000	0.0000	0.0000	.	.
1:0-11_months	1	-0.3378	0.5804	-1.4754	0.7998	0.34	0.5606
2:1-2_years	1	-2.6558	1.0005	-4.6167	-0.6950	7.05	0.0079
3:2-4_years	1	-2.7633	0.4479	-3.6412	-1.8855	38.07	<.0001
4:5-9_years	1	-2.1287	0.1648	-2.4517	-1.8057	166.86	<.0001
5:10-14_years	1	-2.5509	0.1862	-2.9160	-2.1859	187.62	<.0001
6:15-19_years	1	-3.1048	0.3622	-3.8147	-2.3950	73.49	<.0001
7:>=20_years	1	-2.7866	0.1204	-3.0226	-2.5506	535.57	<.0001
	0	1.0000	0.0000	1.0000	1.0000		

6. 過大発散 (Overdispersion)

線形ロジスティックモデルが、2 値データの解析、あるいは、比率の形のデータが、用いられているときに、反応確率のロジスティック変換は、測定されている説明変数について線形独立であると仮定されており、かつ、観測された成功の数は、一般的に 2 項分布になると仮定されている。もし、線形ロジスティック回帰モデルが、満足されているのであれば、そのモデルは、得られた反応確率に適切にあてはまるはずであり、付け加えれば、データ内のランダムな変動は、適切にモデル化されなければならない。 n 個の 2 項比率にあてはめた線形ロジスティックモデルが満足するとき、残差逸脱度は、近似的に p をモデルのあてはめでの未知パラメータ数であるとしたときに、自由度 $(n-p)$ の χ^2 分布に従う。自由度 $(n-p)$ の確率変数 χ^2 の期待値は、 $(n-p)$ であるので、良くあてはまったモデルに対する残差逸脱度は、近似的にその自由度の数になるべきであり、平均逸脱度は、1 に近いであろう。もしも、当てはめたモデルが、得られた割合を適切に記述していないと、残差平均逸脱度 (the residual mean deviance) が得られたものより大きくなりやすい。同時に、もしも、データの変動が、二項サンプリングのものよりもの大きくなれば、残差平均逸脱度は、得られたものより大きくなりやすい。

線形ロジスティックモデルが正しいと考えられる場合であっても、残差平均逸脱度は、それにもかかわらず均一性 (unity) を超えることもあり、二項分布の仮定が、確かでないかもしれない。このようなデータは、過大発散 (overdispersion)、二項変動に割増 (extra binomial variation) として知られている現象、または、簡単に異質性があるという。多くはない共通な出来事 (a much less common occurrence) は、割合についての線形ロジスティックモデルでの逸脱度が、自由度の数よりもかなり少ない場合に、見出される。これは、過小発散 (underdispersion) として参照される。この現象は、実際にたまに起きるので、この章の力点は、過大発散にある。さらに、この章では、線形ロジスティックモデル作成に関連した過大発散のみを考えるのであるが、反応確率のプロビット、および対数-対数補完の変換したモデルも同様に適用される。

6.1. 過大発散の潜在的な原因 (Potential Causes of Overdispersion)

これまでのものより実質的に大きい残差平均逸脱度を導くような、多くの異なる状況が割合として説明したグループ化された二項データのモデル化に存在する。考えねばな

らない最初の説明は、モデルの系統的な要素が、ある方法によって不十分であることである。例えば、要因配置実験の割合のモデル化において、大きな残差逸脱度 (residual deviance) が、その誤差の自由度の数に関連し、モデルの中に十分な数の交互作用項が含まれていない結果として起きる。同様に、用量反応のモデル化において、成功確率のロジスティック変換と用量の線形モデルが、実際の関連が2次的であるとき、あるいは、投与量の対数が説明変数として取られるべきであるとき、に仮定されるべきであるかもしれない。

例題 6.1 抗肺炎 (pneumococcus) 血清, Collett, Ex 6.1

例題 1.6 として与えられた保護血清 (protective serum) の異なる用量での接種した後にマウスの肺炎による死亡の割合について考えよう。

表 6.1 肺炎による死亡数

dose	ln_dose	n	n_total
0.0028	-5.88	35	40
0.0056	-5.18	21	40
0.0112	-4.49	9	40
0.0225	-3.79	6	40
0.0450	-3.10	1	40

例題 4.1 で、用量を説明変数とした線形ロジスティックモデルの当てはめの逸脱度は、自由度 3 で 15.90 である。図 6.1に示すように、ロジスティック直線のあてはめは、高用量、および低用量で当てはまりが悪い。

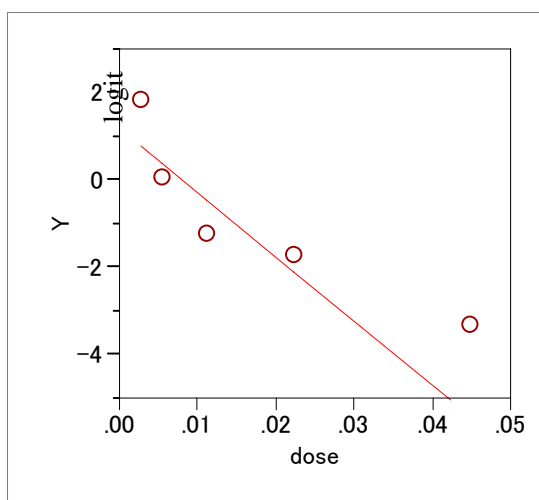


図 6.1 ロジスティック直線のあてはめ

対数用量を説明変数にした場合の残差逸脱度は、自由度 3 で 2.81 となり、当てはまりは十分である。

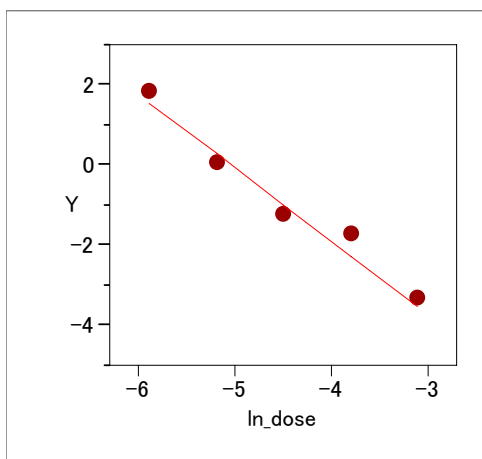


図 6.2 対数用量に対するロジスティック回帰直線

JMPによるLOFの解析

元の用量での LOF

あてはまりの悪さ(LOF)			
要因	自由度	(-1)*対数尤度	カイ2乗
あてはまりの悪さ(LOF)	3	7.950009	15.90002
飽和モデル	4	85.657864	p値(Prob>ChiSq)
あてはめたモデル	1	93.607872	0.0012

対数用量での LOF

あてはまりの悪さ(LOF)			
要因	自由度	(-1)*対数尤度	カイ2乗
あてはまりの悪さ(LOF)	3	1.404449	2.808898
飽和モデル	4	85.657864	p値(Prob>ChiSq)
あてはめたモデル	1	87.062313	0.4220

- ・ 明らかな過大発散は、データの外れ値によっても起こる。
 - ・ ロジスティック・リンク関数が不適當な場合にも起きる。この場合に補 2 重対数変換に基づけばよりよいあてはめとなる場合もある。
- ・ モデルの体系的な成分 (the systematic component) の定義での欠点は、アウトロイヤーの出現、および、リンク関数の選択の適切さは、標準化逸脱度、あるいは、残差尤度、あるいは、他の診断統計量のインデックスプロットによって見抜くことができる。

Colett 3.8.2 節では、実験ユニット小さい場合、逸脱度のカイ 2 乗近似が破綻する

(breaks down), ことを指摘した. グループ化されていない二項データでは, 逸脱度はカイ 2 乗分布しない.

逸脱度の大きさは, 過大発散の指標となり得ない.

極めて重要である

明らかな過大発散の説明

大きな残差平均逸脱度は,

反応確率間の変動

2 値反応間の相関

によって説明される

反応確率間の変動Variation between the response probabilities

実験ユニットでバッチの数が, 同じ条件で得られたとき, 反応確率がバッチとバッチの間で, それにもかかわらず異なるかもしれない.

胎児の例, 催奇形性試験として知られている.

母体の遺伝的影響により同じ実験条件でも奇形出現率は異なる.

奇形出現率の分散は, それが定数の場合よりも大きい.

この変動を説明する変数を実験者は知ることができない.

実験条件が十分にコントロールされていない場にも起きる.

インゲン豆ときゅうりの引き抜いた根の中で, 2つのハマウツボの種の発芽の割合を考える.

ハマウツボの種類 (species) : 1:ageyptica75, 2:ageyptica75

引抜いた根の種類 (extrast) : 1:Bean, 2:Cucumber

5 ないし 6 のバッチからの種, 最低 7 個から最大 81 個の種

反復測定が行われていると同様の実験データ

表 6.2 ハマウツボの発芽

1:ageyptica75				2:ageyptica75			
1:Bean		2:Cucumber		1:Bean		2:Cucumber	
<i>y</i>	<i>n</i>	<i>y</i>	<i>n</i>	<i>y</i>	<i>n</i>	<i>y</i>	<i>n</i>
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

Example 6.2 Germination of Orobanche (岩波生物学辞典：シソ花類；ハマウツボ)

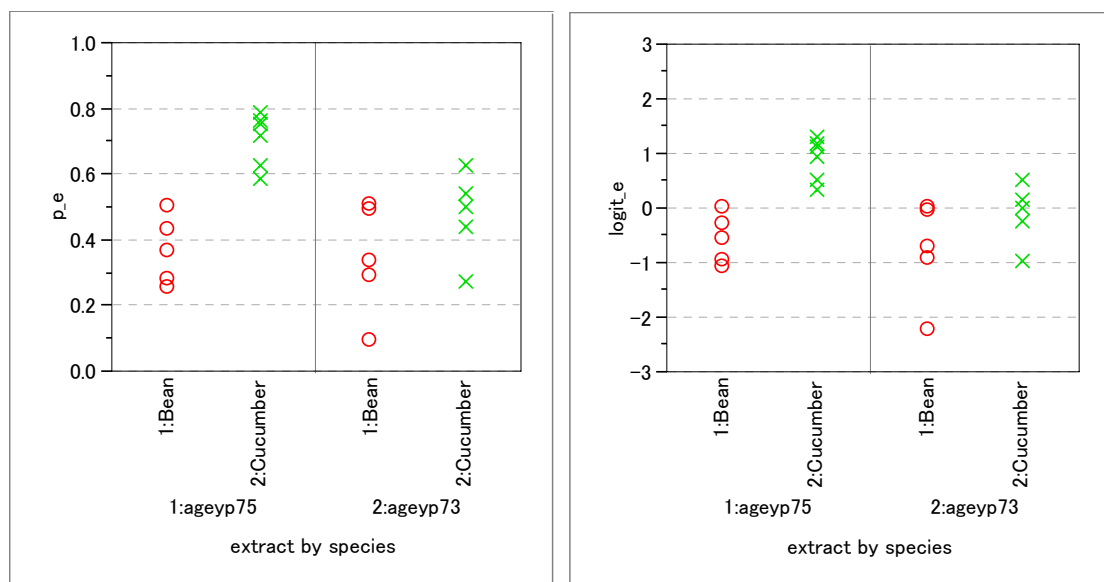


図 6.3 ハマウツボの各バッチの発芽割合と経験ロジット

表 6.3 ハマウツボの発芽データにロジスティックモデルのあてはめ後の逸脱度

Terms fitted in model	Deviance	d.f.
β_0	98.72	20
$\beta_0 + species_j$	96.18	19
$\beta_0 + extract_k$	42.75	19
$\beta_0 + species_j + extract_k$	39.69	18
$\beta_0 + species_j + extract_k + (species \times extract)_{jk}$	33.28	17

交互作用が有意, $39.28 - 33.28 = 6.41$. しかし, 17 の自由度の残差逸脱度が 33.28 と有意に大きい. モデルがあてはまっていると言い難いが, 何が原因かは分からない. どのようにモデルに体系的な成分 (the systematic component) 追加できるか明らかでない.

外れ値もない。したがって、個のデータは、過大発散であると結論づけられる。

この発散の説明は、

- 種のバッチの違い、
- 引き抜いた根の不均一性、
- 均一に発芽しない異なるバッチ、
- などであろう。

経験的に、実験者の知識は、これらの説明がよりもっともらしい光明を投じることを期待するものだ。

SASのGENMODプロシジャによる解析

逐次型 Type1 のモデル

プログラム

```
proc genmod data=ex103grm ;
  class species extract ;
  model y/n = extract species species*extract
        / dist=binomial link=logit type1 ;
run ;
```

実行結果

Criteria For Assessing Goodness Of Fit				
Criterion	DF	Value	Value/DF	
Deviance	17	33.2778	1.9575	
Scaled Deviance	17	33.2778	1.9575	
Pearson Chi-Square	17	31.6511	1.8618	
Scaled Pearson X2	17	31.6511	1.8618	
Log Likelihood		-543.1106		

LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	98.7195			
species	96.1752	1	2.54	0.1107
extract	39.6859	1	56.49	<.0001
species*extract	<u>33.2778</u>	1	6.41	0.0114

Type3 のモデル

プログラム

```
proc genmod data=ex103grm ;
  class species extract ;
  model y/n = species extract species*extract
        / dist=binomial link=logit type3 ;
  lsmeans species*extract ;
run ;
```

実行結果

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
species	1	2.53	0.1121
extract	1	37.48	<.0001
species*extract	1	6.41	0.0114

		Least Squares Means	
species	extract	Estimate	SE
1:agevp75	1:Bean	-0.5582	0.1260
1:agevp75	2:Cucumber	0.7600	0.1250
2:agevp73	1:Bean	-0.4122	0.1842
2:ageyp73	2:Cucumber	0.1278	0.1688

2値反応間の相関

説明が役立つように、催奇形成試験を再度考える。もしも、ある特定のラビットからの胎児の1例が、ある異常を持つとしたとき、同じ異常は、同じラビットからの他の胎児に見出されやすいだろう。同様に、他のラビットの胎児に異常のない場合は、残りの胎児が奇形であることは少ないであろう。1腹のそれぞれの胎児の反応の間に、妊娠ラビットからの異常な胎児の割合は、結果として正の相関は、これらの2値反応が独立であった場合よりもさらに変動させる傾向になるであろう。なぜなら、独立した2値反応は、2項変動を導き、異常な胎児の割合は、胎児の奇形数は、2項分布であると仮定のもとであるよりも、より変動するであろう。従って、その割合に適切なモデルを当てはめたあとの平均逸脱度は、単一であるよりも大きくなると期待される。

ほかにも、いろいろ説明があるが、疑わしい。

6.2. 反応確率のモデル化変動 (Modeling Variability in Response Probabilities)

反応確率が、同じ実験の条件で曝露された実験ユニットの群間で多様であるとき、ある仮定が、この変動の形式についてなされるべきである。この節では、我々は、Williams (1982a) によって論述された一般モデルを考えよう。

n 個の観測された割合からなるデータ, $y_i/n_i, i=1,2,\dots,n$, および, 第 i 番目の対応する反応確率が, 線形ロジスティック・モデルを通じて k 個の説明変数 x_1, x_2, \dots, x_k に依存するとしよう. 第 i 番目の観測値に実際の反応確率, g_i (θ_i の別表現) と呼ぶ, が, p_i の平均に関して変動すると仮定しよう. 実際の反応確率, g_i の分散は, p_i がゼロまたは同一であるときにゼロとなり, これが正しいとしたときの最も簡単な関数として,

$\phi(0 \geq)$ が未知のスケール・パラメータであるとする、

$$\text{Var}(\mathcal{G}_i) = \phi p_i (1 - p_i)$$

である。 \mathcal{G}_i の量は未観測のランダム変数、または、潜在変数である。しかしながら、 \mathcal{G}_i の特定の値が与えられた場合に、第 i 番目の観測された成功の数、 y_i は、平均が $n_i \mathcal{G}_i$ 、分散が $n_i \mathcal{G}_i (1 - \mathcal{G}_i)$ の 2 項分布に従うであろう。 y_i の平均は、 \mathcal{G}_i の条件付きで、

$$E(y_i | \mathcal{G}_i) = n_i \mathcal{G}_i$$

よって与えられ、 y_i の条件付き分散は、

$$\text{Var}(y_i | \mathcal{G}_i) = n_i \mathcal{G}_i (1 - \mathcal{G}_i)$$

である。記号 ‘|’ は、条件付きで、と読む。ただし、 \mathcal{G}_i は、推定できないので、観測された割合 y_i / n_i は、 p_i の推定値であるように得られるべきであり、 \mathcal{G}_i の期待値、および y_i の条件付きでない平均と分散が、要求されるであろう。これは、 $E(y_i)$ と $\text{Var}(y_i)$ 式 (6.1) での反応確率におけるランダム変数についての仮定の効果を探るように我々を導く。

条件付き確率の理論の標準的な結果によって、確率変数 Y の条件付きでない期待値は、

$$E(Y) = E\{E(Y | X)\}$$

を用いて、与えられた X についての Y の条件付き確率から得ることができる。そして、条件付きでない Y の分散は、

$$\text{Var}(Y) = E\{\text{Var}(Y | X)\} + \text{Var}\{E(Y | X)\}$$

で与えられる。

これらの 2 つの結果の利用して、

$$E(y_i) = E\{E(y_i | \mathcal{G}_i)\} = E(n_i \mathcal{G}_i) = n_i E(\mathcal{G}_i) = n_i p_i$$

と

$$\text{Var}(y_i) = E\{\text{Var}(y_i | \mathcal{G}_i)\} + \text{Var}\{E(y_i | \mathcal{G}_i)\}$$

ここで、

$$\begin{aligned}
 \text{Var}(y_i | \mathcal{G}_i) &= E\{n_i \mathcal{G}_i (1 - \mathcal{G}_i)\} \\
 &= n_i \{E(\mathcal{G}_i) - E(\mathcal{G}_i^2)\} \\
 &= n_i \{E(\mathcal{G}_i) - \text{Var}(E(\mathcal{G}_i)) - [E(\mathcal{G}_i)]^2\} \\
 &= n_i \{p_i - \phi p_i (1 - p_i) - p_i^2\} \\
 &= n_i p_i (1 - p_i) [1 + (n_i - 1)\phi]
 \end{aligned}$$

同様に,

$$\text{Var}(E(y_i | \mathcal{G}_i)) = \text{Var}(n_i \mathcal{G}_i) = n_i^2 \text{Var}(\mathcal{G}_i) = n_i^2 \phi p_i (1 - p_i)$$

そこで,

$$\text{Var}(E(y_i)) = n_i p_i (1 - p_i) [1 + (n_i - 1)\phi]$$

となる.

ランダム変動が y_i になれば, $\phi = 0$, これは通常の 2 項分布が仮定される.

ランダム変動が y_i にあれば, $[1 + (n_i - 1)\phi]$ だけ分散は大きくなる.

表 6.4 ウィリアムの分散の補正, $\phi = 0.0249$, $\pi = 0.2$ の場合

n のサイズ		平均と分散		SD と CV		補正後の SD と CV		
n	$\log_{10} n$	y	Var	SD	CV	$1+(n-1)\phi$	SD'	CV'
1	0.00	0.2	0.2	0.4	200.0	1.000	0.4	200.0
2	0.30	0.4	0.3	0.6	141.4	1.025	0.6	143.2
5	0.70	1	0.8	0.9	89.4	1.100	0.9	93.8
10	1.00	2	1.6	1.3	63.2	1.224	1.4	70.0
20	1.30	4	3.2	1.8	44.7	1.473	2.2	54.3
30	1.48	6	4.8	2.2	36.5	1.722	2.9	47.9
50	1.70	10	8	2.8	28.3	2.220	4.2	42.1
100	2.00	20	16	4.0	20.0	3.465	7.5	37.2
1000	3.00	200	160	12.7	6.3	25.875	64.3	32.2
10000	4.00	2000	1600	40.0	2.0	249.975	632.4	31.6

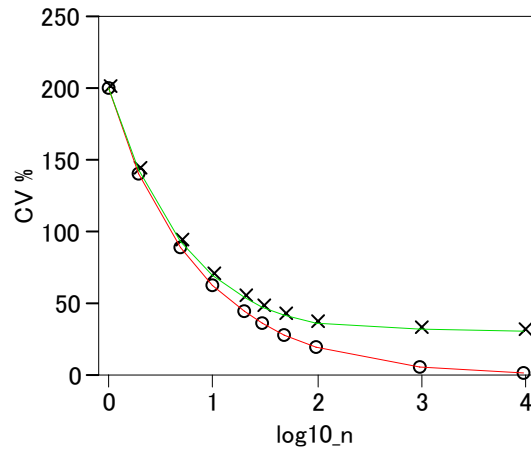


図 6.4 ウィリアムの補正後の CV

—○—：補正無し，—×—：補正後

6.3. 2 値反応間の相関

中略

6.4. 過大発散データのモデル化

中略

事例 6.4 ハマウツボの発芽

ウィリアムの方法によって過大発散の調整をしよう．Example 6.2 のハマウツボの発芽の取り上げる．因子は，種（species）と根（extract）である．

交互作用を含むモデルのロジスティックモデルの逸脱度は，df=3, chi2=33.28 である．

y_i の分散は，

$$\text{Var}(y_{jkl}) = n_{jkl} p_{jk} (1 - p_{jk}) [1 + (n_{jkl} - 1)\phi]$$

である．

SAS の LOGISTIC プロシジのにより， $\phi = 0.0249$ を得る．分散の過大評価分を

$$\sigma_{jkl}^2 = 1 + (n_{jkl} - 1)\phi$$

とにおいて、その逆数を重みとしたロジスティック回帰を行う。

表 6.5 重み付きロジスティック回帰モデルの逸脱度

Terms fitted in model	Deviance	d.f.
β_0	47.26	20
$\beta_0 + species_j$	44.94	19
$\beta_0 + extract_k$	24.63	19
$\beta_0 + species_j + extract_k$	21.99	18
$\beta_0 + species_j + extract_k + (species \times extract)_{jk}$	18.45	17

プログラム, ウィリアムの補正

```
Title 'C_ex604_Williams.sas 2001-07-02 Y.Takahashi ' ;
proc logistic data=ex103grm ;
  class species extract / param=ref ref=first ;
  model y / n = species extract species*extract
        / scale=williams lackfit ;
run ;
```

結果の出力

Data Set	WORK. EX103GRM
Response Variable (Events)	y
Response Variable (Trials)	n
Number of Observations	21
Weight Variable	1 / (1 + 0.024938 * (n - 1))
Sum of Weights	378.57262374
Model	binary logit
Optimization Technique	Fisher's scoring

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	17	18.4418	1.0848	0.3615
Pearson	17	16.9996	1.0000	0.4544

Number of events/trials observations: 21

NOTE: Since the Williams method was used to accommodate overdispersion, the Pearson chi-squared statistic and the deviance can no longer be used to assess the goodness of fit of the model

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error
Intercept	1	-0.5354	0.1937
species 2:ageyp73	1	0.0701	0.3115
extract 2:Cucumber	1	1.3295	0.2782
species*extract 2:ageyp73 2:Cucumber	1	-0.8193	0.4352

6.5. n_i が同数の特別な場合

同一の 2 値の観察数の場合，式 (6.2) $\text{Var}(y_i) = n_i p_i (1 - p_i) [1 + (n_i - 1)\phi]$ は，

$$\text{Var}(y_i) = n_0 p_i (1 - p_i) [1 + (n_0 - 1)\phi]$$

となる． $\sigma^2 = 1 + (n_0 - 1)\phi$ とおけば，この分散は， $n_0 p_i (1 - p_i) \sigma^2$ と書ける．

モデル全体のピアソンの X^2 統計量の期待値は，近似的に， p を未知のパラメータとしたときに

$$(n - p) [1 + (n_0 - 1)\phi]$$

となり， $(n - p)\sigma^2$ となる．パラメータ σ^2 は，それゆえに， $X^2 / (n - p)$ によって推定される． σ^2 の推定は，不均一因子 (heterogeneity factor) としられている．逸脱度は， X^2 統計量に非常に近いので， σ^2 は，モデル全体の平均逸脱度として推定される．

$E(X^2) \approx (n - p)\sigma^2$ であるという結果は， X^2 が， $\sigma^2 \chi_{n-p}^2$ 分布に従うことを連想する．個結果は，より専門的な論法で，マックラフとネルダー，McCuiagh and Nelder (1998) により裏付けられている．

モデル (1) の逸脱度 D_1 ，自由度 ν_1

モデル (2) の逸脱度 D_2 ，自由度 ν_2

モデル全体の逸脱度 D_0 ，自由度 ν_0

$$\frac{(D_1 - D_2) / (\nu_1 - \nu_2)}{D_0 / \nu_0}$$

が，自由度 $(\nu_1 - \nu_2)$ ， ν_0 の F 分布に従う．

σ^2 は，スケール・パラメータとよばれている．フィラーの定理による DE50 では，信頼限界の計算で標準誤差として， $\sqrt{\hat{\sigma}^2}$ とし，自由度 $(n - p)$ の t 分布が用いられている．

事例 6.5 粉虫の例

表 6.6 エチレンオキシドの曝露に作用した粉虫の数

subj	conc	ln_conc	y	n	p*	logit*
1	24.8	3.21	23	30	0.76	1.14
2	24.6	3.20	30	30	0.98	4.11
3	23.0	3.14	29	31	0.92	2.47
4	21.0	3.04	22	30	0.73	0.97
5	20.6	3.03	23	26	0.87	1.90
6	18.2	2.9	7	27	0.27	-1.01
7	16.8	2.82	12	31	0.39	-0.44
8	15.8	2.76	17	30	0.56	0.26
9	14.7	2.69	10	31	0.33	-0.72
10	10.8	2.38	0	24	0.02	-3.89

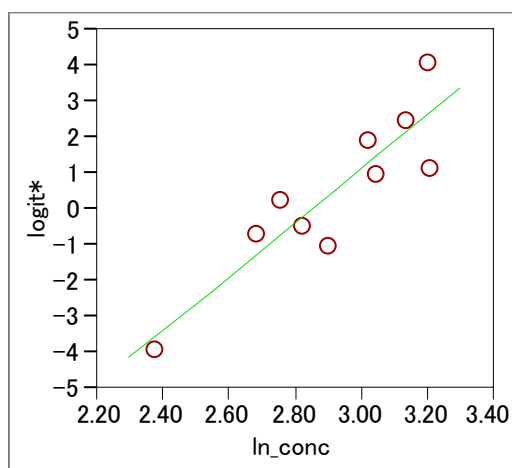


図 6.5 ロジスティック直線のあてはめ

図 6.5から、あてはまりは良くないと思われる。JMPでのLOFの結果から、自由度 8 で、逸脱度は 36.44 と明らかに大きい。

あてはまりの悪さ(LOF)				
要因	自由度	(-1)*対数尤度	カイ2乗	
あてはまりの悪さ(LOF)	8	18.22182	36.44365	
飽和モデル	9	126.57137	p値(Prob>ChiSq)	
あてはめたモデル	1	144.79320		<.0001

パラメータ推定値				
項	推定値	標準誤差	カイ2乗	p値(Prob>ChiSq)
切片	-17.867021	2.2690148	62.01	<.0001
ln_conc	6.26542849	0.7781254	64.83	<.0001

推定値は次の対数オッズに対するものです: 0/1

効果の尤度比検定				
要因	パラメータ数	自由度	尤度比カイ2乗	p値(Prob>ChiSq)
ln_conc	1	1	101.556952	0.0000

JMPの結果から、分散分析風に尤度をまとめないし、過大発散を逸脱度を用いて調整する。

表 6.7 有意な逸脱度がある場合の F 検定による調整

要因		自由度	逸脱度	平均逸脱度	F 比	p 値
ln_conc	回帰	1	101.56	101.56	22.29	0.0015
あてはまりの悪さ(LOF)	残差	8	36.44	4.56		
切片のみのモデル	全体	9	138.00			

SASのGENMODプロシジャでの逸脱度の調整

プログラム

```
proc genmod data=ex105bt1 ;
  model y / n = ln_conc
    / dist=binomial link=logit
    type1 scale=deviance covb ;
run ;
```

出力結果

Criterion	DF	Value	Value/DF
Deviance	8	36.4436	4.5555
Scaled Deviance	8	8.0000	1.0000
Pearson Chi-Square	8	33.4340	4.1793
Scaled Pearson X2	8	7.3393	0.9174
Log Likelihood		-31.7846	

	Prm1	Prm2
Prm1	23.45345	-8.02661
Prm2	-8.02661	2.75824

Parameter	DF	Estimate	Error
Intercept	1	-17.8670	4.8429
ln_conc	1	6.2654	1.6608
Scale	0	2.1344	0.0000

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF

Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	138.0006						
ln_conc	36.4436	1	8	22.29	0.0015	22.29	<.0001

SASのLOGISTICプロシジャによる過大発散の調整

プログラム

```
proc logistic data=ex105btl ;
  model y / n = ln_conc / itprint scale=deviance covb ;
run ;
```

出力結果

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	8	36.4436	4.5555	<.0001
Pearson	8	33.4340	4.1793	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-17.8670	4.8429	13.6112	0.0002
ln_conc	1	6.2654	1.6608	14.2322	0.0002

SASのPROBITプロシジャによる、ロジスティック回帰およびED50の推定

```
proc probit data=ex105btl ;
  model y / n = ln_conc / distribution=logistic
  lackfit covb inversecl ;
run ;
```

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ln_conc	1	15.5133	<.0001

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-17.8670	4.6386	-26.9585 -8.7755	14.84	0.0001
ln_conc	1	6.2654	1.5907	3.1476 9.3832	15.51	<.0001

Estimated Covariance Matrix /* 結果が異なる, 再チェック */

	Intercept	ln_conc
Intercept	21.516578	-7.363744
ln_conc	-7.363744	2.530450

Probit Model in Terms of Tolerance Distribution

	MU	SIGMA
MU	2.85168376	0.159606
SIGMA		

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	0.002453	-0.000601
SIGMA	-0.000601	0.001642

Probability	ln_conc	95% Fiducial Limits	
0.10	2.50099	1.89922	2.67606
0.50	2.85168	2.67709	2.96539
0.90	3.20237	3.06177	3.64791

事例 6.6 ハマウツボの発芽

プログラム

```
proc genmod data=ex103grm ;
  class species extract ;
  model y / n = species extract species*extract
        / scale=deviance type3 ;
run ;
```

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	17	33.2778	1.9575
Scaled Deviance	17	17.0000	1.0000
Pearson Chi-Square	17	31.6511	1.8618
Scaled Pearson X2	17	16.1690	0.9511
Log Likelihood		-277.4487	

LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
species	1	17	1.29	0.2718	1.29	0.2561
extract	1	17	19.15	0.0004	19.15	<.0001
species*extract	1	17	3.27	0.0881	3.27	0.0704

6.6. ベータ 2 項分布モデル

中略

6.7. ロジスティック回帰でのランダム係数モデル

事例 6.8 毒性実験での例

表 6.8 生存胎児数

コントロール群			投与群		
subj.	y	n	subj.	y	n
1	13	13	17	12	12
2	12	12	18	11	11
3	9	9	19	10	10
4	9	9	20	9	9
5	8	8	21	10	11
6	8	8	22	9	10
7	12	13	23	9	10
8	11	12	24	8	9
9	9	10	25	8	9
10	9	10	26	4	5
11	8	9	27	7	9
12	11	13	28	4	7
13	4	5	29	5	10
14	5	7	30	3	6
15	7	10	31	3	10
16	7	10	32	0	7

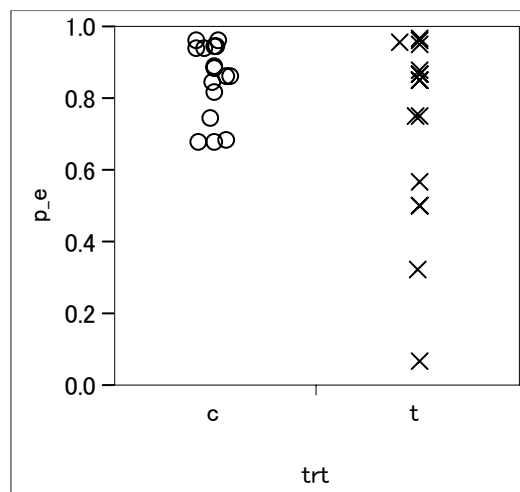


図 6.6 群ごとの生存胎児の割合

SASのNL MIXEDを用いた線形ロジスティックモデルでのランダム効果

我々は、過大発散が、同一に処理された実験ユニットの異なるバッチについての反応確率の変動で説明できることを見てきた。6.1 節で説明したように、この変動は、記録されていない関連する説明変数に起因すると考えることができるか、あるいは、適切に測定されるか制御されないような確かな説明変数に含まれる。

Example 46.2: Probit-Normal Model with Binomial Data

For this example, consider the data from Weil (1970), also studied by Williams (1975), Ochi and Prentice (1984), and McCulloch (1994). In this experiment 16 pregnant rats receive a control diet and 16 receive a chemically treated diet, and the litter size for each rat is recorded after 4 and 21 days. The SAS data set is as follows.

```
data rats;
  input trt$ m x ;
  if (trt='c') then do;
    x1 = 1;
    x2 = 0;
  end;
  else do;
    x1 = 0;
    x2 = 1;
  end;
  litter = _n_;
  datalines;
c 13 13 | t 10 9
c 12 12 | t 9 8
c 9 9 | t 9 8
c 9 9 | t 5 4
c 8 8 | t 9 7
c 8 8 | t 7 4
c 13 12 | t 10 5
c 12 11 | t 6 3
c 10 9 | t 10 3
c 10 9 | t 7 0
c 9 8 | run;
c 13 11
c 5 4
c 7 5
c 10 7
c 10 7
t 12 12
t 11 11
t 10 10
t 9 9
t 11 10
t 10 9
```

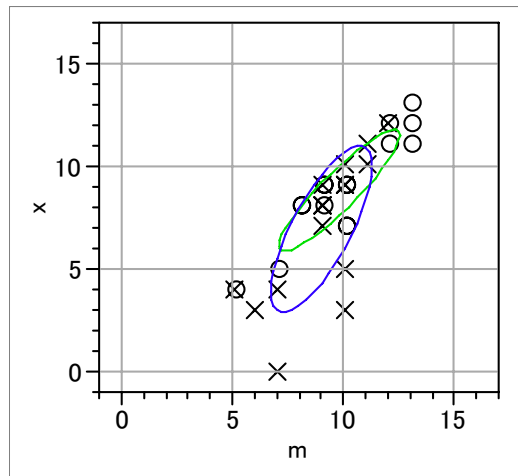


図 6.7 リッターサイズ m と生存胎児数の確率楕円

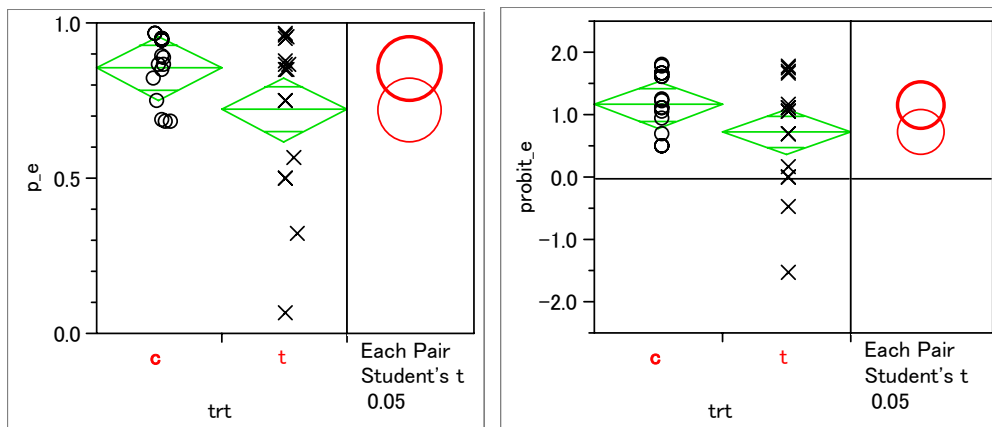


図 6.8 生存胎児数の割合と経験ロジットでの平均値の比較

Here, M represents the size of the litter after 4 days, and X represents the size of the litter after 21 days. Also, indicator variables X_1 and X_2 are constructed for the two treatment levels.

Following McCulloch (1994), assume a latent survival model of the form

$$y_{ijk} = t_i + \alpha_{ij} + \epsilon_{ijk}$$

where i indexes treatment, j indexes litter, and k indexes newborn rats within a litter. The t_i represent treatment means, the α_{ij} represent random litter effects assumed to be iid $N(0, s_i^2)$, and the e_{ijk} represent iid residual errors, all on the latent scale.

Instead of observing the survival times y_{ijk} , assume that only the binary variable indicating whether y_{ijk} exceeds 0 is observed. If x_{ij} denotes the sum of these binary variables for the i th treatment and the j th litter, then the preceding assumptions lead to the following generalized

linear mixed model:

$$x_{ij} | \alpha_{ij} \sim \text{Binomial}(m_{ij}, p_{ij})$$

where m_{ij} is the size of each litter after 4 days and

$$p_{ij} = \Phi(t_i + \alpha_{ij})$$

The PROC NLMIXED statements to fit this model are as follows.

```
proc nlmixed data=rats;
  parms t1=1 t2=1 s1=.05 s2=1;
  eta = x1*t1 + x2*t2 + alpha;
  p = probnorm(eta);
  model x ~ binomial(m, p);
  random alpha ~ normal(0, x1*s1*s1+x2*s2*s2) subject=litter;
  estimate 'gamma2' t2/sqrt(1+s2*s2);
  predict p out=p;
run;
```

As in the previous example, the PROC NLMIXED statement invokes the procedure and the PARMs statement defines the parameters. The parameters for this example are the two treatment means, T1 and T2, and the two random-effect standard deviations, S1 and S2.

The indicator variables X1 and X2 are used in the program to assign the proper mean to each observation in the input data set as well as the proper variance to the random effects. Note that programming expressions are permitted inside the distributional specifications, as illustrated by the random-effects variance specified here.

The ESTIMATE statement requests an estimate of $\gamma_2 = t_2 / \sqrt{1 + s_2^2}$, which is a location-scale parameter from Ochi and Prentice (1984).

The PREDICT statement constructs predictions for each observation in the input data set. For this example, predictions of P and approximate standard errors of prediction are output to a SAS data set named P. These predictions are functions of the parameter estimates and the empirical Bayes estimates of the random effects α_i .

The output for this model is as follows.

	Specifications	
Data Set		DD. RATS
Dependent Variable		x

Distribution for Dependent Variable	Binomial
Random Effects	alpha
Distribution for Random Effects	Normal
Subject Variable	litter
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

The "Specifications" table provides basic information about this nonlinear mixed model.

Dimensions	
Observations Used	32
Observations Not Used	0
Total Observations	32
Subjects	32
Max Obs Per Subject	1
Parameters	4
Quadrature Points	7

The "Dimensions" table provides counts of various variables.

Parameters				
t1	t2	s1	s2	NegLogLike
1	1	0.05	1	54.9362323

The "Parameters" table lists the starting point of the optimization.

Iteration History					
Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	2	53.9933934	0.942839	11.03261	-81.9428
2	3	52.875353	1.11804	2.148952	-2.86277
3	5	52.6350386	0.240314	0.329957	-1.05049
4	6	52.6319939	0.003045	0.122926	-0.00672
5	8	52.6313583	0.000636	0.028246	-0.00352
6	11	52.6313174	0.000041	0.013551	-0.00023
7	13	52.6313115	5.839E-6	0.000603	-0.00001
8	15	52.6313115	9.45E-9	0.000022	-1.68E-8

NOTE: GCONV convergence criterion satisfied.

The "Iterations" table indicates successful convergence in 8 iterations.

Fit Statistics	
-2 Log Likelihood	105.3
AIC (smaller is better)	113.3
AICC (smaller is better)	114.7
BIC (smaller is better)	119.1

The "Fitting Information" table lists some useful statistics based on the maximized value of the log likelihood.

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
t1	1.3063	0.1685	31	7.75	<.0001	0.05	0.9626	1.6499	-0.00002
t2	0.9475	0.3055	31	3.10	0.0041	0.05	0.3244	1.5705	9.283E-6
s1	0.2403	0.3015	31	0.80	0.4315	0.05	-0.3746	0.8552	0.000014
s2	1.0292	0.2988	31	3.44	0.0017	0.05	0.4198	1.6385	-3.16E-6

The "Parameter Estimates" table indicates significance of all of the parameters except S1.

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
gamma2	0.6603	0.2165	31	3.05	0.0047	0.05	0.2186	1.1019
t2 - t1	-0.3588	0.3489	31	-1.03	0.3117	0.05	-1.0703	0.3528

The "Additional Estimates" table displays results from the ESTIMATE statement. The estimate of γ_2 equals 0.6602, agreeing with that obtained by McCulloch (1994). The standard error 0.2166 is computed using the delta method (Billingsley 1986).

Not shown is the P data set, which contains the original 32 observations and predictions of the p_{ij} .

表 6.9 ランダム係数の推定

コントロール								投与群									
trt	litter	m	x	x1	x2	alpha^	eta^	p^	trt	litter	m	x	x1	x2	alpha^	eta^	p^
c	1	13	13	1	0	0.1177	1.424	0.9228	t	17	12	12	0	1	0.9182	1.8657	0.969
c	2	12	12	1	0	0.11	1.4163	0.9217	t	18	11	11	0	1	0.8895	1.837	0.9669
c	3	9	9	1	0	0.0857	1.392	0.918	t	19	10	10	0	1	0.8582	1.8057	0.9645
c	4	9	9	1	0	0.0857	1.392	0.918	t	20	9	9	0	1	0.8238	1.7713	0.9617
c	5	8	8	1	0	0.0772	1.3835	0.9167	t	21	11	10	0	1	0.3084	1.2559	0.8954
c	6	8	8	1	0	0.0772	1.3835	0.9167	t	22	10	9	0	1	0.2633	1.2108	0.887
c	7	13	12	1	0	0.0223	1.3286	0.908	t	23	10	9	0	1	0.2633	1.2108	0.887
c	8	12	11	1	0	0.0138	1.3201	0.9066	t	24	9	8	0	1	0.2129	1.1604	0.8771
c	9	10	9	1	0	-0.0041	1.3022	0.9036	t	25	9	8	0	1	0.2129	1.1604	0.8771
c	10	10	9	1	0	-0.0041	1.3022	0.9036	t	26	5	4	0	1	-0.0763	0.8712	0.8082
c	11	9	8	1	0	-0.0133	1.293	0.902	t	27	9	7	0	1	-0.1516	0.7959	0.787
c	12	13	11	1	0	-0.0669	1.2394	0.8924	t	28	7	4	0	1	-0.6313	0.3162	0.6241
c	13	5	4	1	0	-0.053	1.2533	0.8949	t	29	10	5	0	1	-0.8251	0.1224	0.5487
c	14	7	5	1	0	-0.1276	1.1787	0.8807	t	30	6	3	0	1	-0.7596	0.1879	0.5745
c	15	10	7	1	0	-0.1828	1.1235	0.8694	t	31	10	3	0	1	-1.2687	-0.3212	0.374
c	16	10	7	1	0	-0.1828	1.1235	0.8694	t	32	7	0	0	1	-1.9976	-1.0501	0.1468

Parameter({t1 = 1.3063, t2 = 0.9475}, :x1 * t1 + :x2 * t2 + :Name("alpha^"))

Normal Distribution(:Name("eta^"))

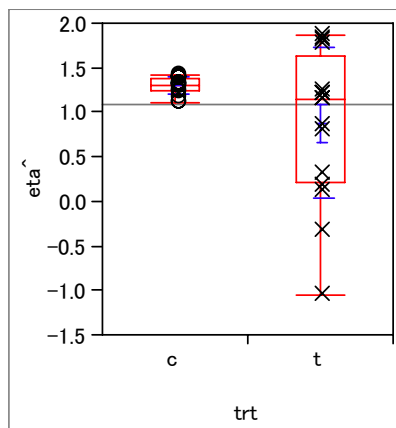


図 6.9 ランダム効果を調整した推定値 $\hat{\eta}_i$

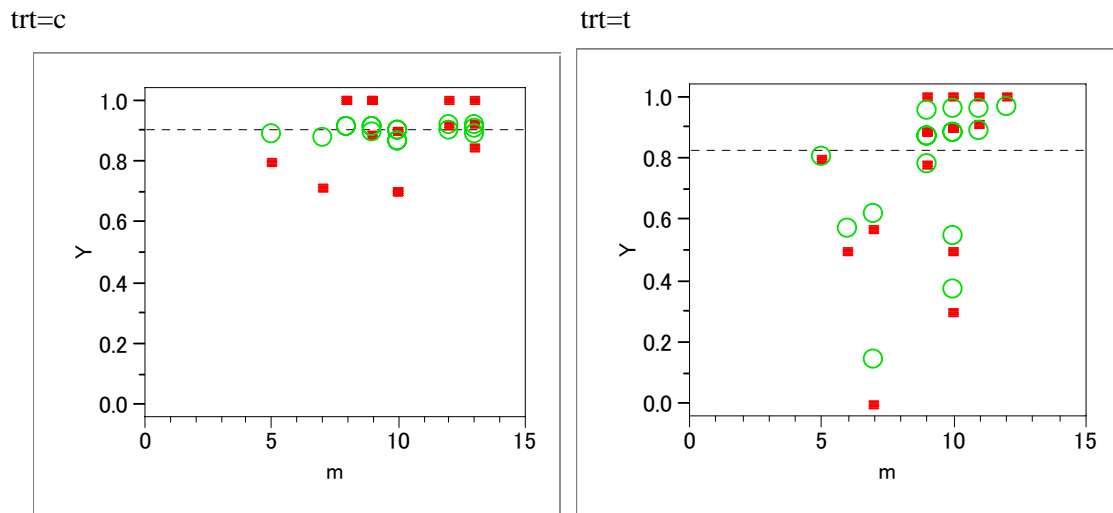


図 6.10 リッターサイズを含めたランダム効果を調整した胎児の生存率

■ : 生データによる生存率, ○ : ランダム効果を調整した胎児の生存率

6.8 手法の比較

この節では、我々は、過大発散のモデルに使うことができる多くの手法を検討してきた。便宜的に、次に一覧されている。

- 手法 1： ウィリアム法 (6.4 節)
- 手法 2： 不均一要素の使用 (6.5 節)
- 手法 3： ベータ 2 項でのモデリング (6.6 節)
- 手法 4： ランダム効果の包含 (6.7 節)
- 手法 5： モデル 4 の近似 (6.7.2 節)

我々は、手法 3 が、手法 1 に基づく形の手順に比べて柔軟性に乏しいこと見てきた。そして、この手法 3 の手順は、さらに検討される必要はない。手法 5 は、しばしば手法 1 に非常に似た結果を与え、近似が、ランダム効果の分散が小さいときにのみ妥当であり、手法 5 は、一般的に勧めることができない。なおまた、ランダム効果を含むモデルの当てはめで使用できるコンピュータ ソフトウェアは、より広く利用可能となっており、もはや正確なアプローチより簡単に実践できるような近似的な方法にたよることではない。類似した議論は、手法 2 で適用することができ、それは、さらにまた簡単に適用できるのであるが、2 項分母 ?? (the binomial denominators) がすべて等しいときにのみ妥当である。

手法 4 は、関係ある説明変数の値が記録されていないときや、モデルの中の変数が適切に測定されていないかコントロールされていないときに、過大発散のモデル化の自然な方法である。これは、また、2 値応答データでの過大発散のモデルに用いることが考慮されているただ一つの手法である。あいにく、ランダム効果を含むモデルのあてはめた後では、過大発散が実際に適用されているかどうかは明快ではない。手法 1 は、反応確率間の変動があるときのグループ化された 2 値データの過大発散のモデリングの一般的な手順であり、手法 2 の特殊な場合を含んでいる。手法 1 と手法 4 は、それるえに一般的に使用にたいして勧められる手順である。