平方和の分解では解けない
繰り返し不揃いの 2 元配置データの解析入門

高橋 行雄
BioStat 研究所（株）

An Introduction to Analysis of Unbalanced Two-way ANOVA
that cannot be solved by sum of squares decomposition

Yukio Takahashi
BioStat Research Co.,Ltd.

要旨： 各種のフリーソフトが興隆すると共にSASが埋没し，統計解析の質も低下し続けている．歯止めを掛けるために，伝統的で根強いファンがいる平方和の分解による解析ができないポピューラな事例を取り上げ，GLM プロシジャが実用化した線形モデルによる解析方法を復活させることにより，SAS の掘り起こしをする．フリーソフトに対して無償で継続的に使用できるようになったSASの優位性を際立たせるめには，平方和の分解で凝り固まっている統計の教科書が目の上のたんこぶである．他方，GLM プロシジャの理論は，難解で近寄り難い側面もある．今回取り上げる事例は，繰り返しが不揃いの 2 元配置データであり，平方和の分解による分散分析が行えないが，線形モデルによる解析が可能であり，GLM プロシジャにより 30 年以上前から実用化されている．GLMプロシジャの偉大かつブラック・ボックス的な解析方法をExcelの行列関数により再現し,線形モデルの有用性にスポットライトをあてたい.また,他のフリーソフトにないGLMプロシジャの最小 2 乗平均（lsmeans）に焦点をあて，その有用性について論ずるとともに，Excel で再現する．また，GLMプロシジャのHTML出力をExcelに取り込み，見栄えの良いグラフの作成方法も示す.

キーワード：平方和の分解，最小 2 乗平均，線形モデル，SAS/GLM，ダミー変数，デザイン行列，田口の式

1. はじめに

各種のフリーソフトが興隆すると共にSASが埋没し，統計解析の質も低下し続けていると実感しているのは，私だけなのだろうか.「何々による統計解析」と題する書籍，Web上で公開されているPDF版の“書籍”を目にする．それらの多くは,「何々の使い方」であって，統計解析そのものの学習には向いていない.「実験計画法」と題する書籍も数多く見かけるが，シグマによる定式化が極度に標準化されているために，さらなる学習を妨げるようなガラスの天井が張り巡らされているかの如くである.

私自身は，SASのGLM，GENMODなどの数多くのプロシジャを使いつつ，行列計算を主体にした線形モデル，一般化線形モデルなどについてマニュアルに掲載されている多くに事例により統計解析のスキルの向上ができ，その理論的背景についての簡潔な記述によって理解の助けになった．もちろんSAS以外の多くの統計ソフトを使用した経験もあるが，統計解析の理論の学習に役に立ったのは，先進的なSASユーザー達による数多くの書籍でもあった．なお，GLM プロシジャの詳細は，自由にダウンロードできる SAS Institute（2013)，「SAS/STAT13.1 User’s Guide, The GLM Procedure」を参照のこと．また，Little ら（2020）,「SAS for Linear Models, 4th ed.」には，多彩な事例が示されており，必読の書である.

ある時，前値と後値がある群間比較データに「共分散分析」など難しい統計手法など使わなくとも，前後差による対応のない t 検定を使えばいいのだ，と断言する雑誌の査読者が存在することを知った．Excel の回帰分析で簡単にできる「共分散分析」を難しいとは，どういうことだと疑問をもった．また，ある時，直交表の解析で欠測値があると解析できないので，何らかのデータを代入して解析するとの Web 上での記事に遭遇し愕然としたこともあった．詳細は，高橋行雄（2021b），「線形モデルによる欠測値がある直交表の解析」を参照してもらいたい．

SAS の GLM プロシジャを使った経験者ならば，繰り返しが等しくとも，不揃いでも，まったく違和感なく解析するであろう．そもそも，SAS の GLM プロシジャには，繰り返しが（揃っている，不揃いである）を区別する用語が見当たらない．昨年 2021 年の SAS ユーザー総会では，高橋行雄（2021a），「各種のダミー変数を用いた最小 2 乗平均と 95%信頼区間の実際」と題して，守屋ら（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」で例示されている「共変量を含む繰り返し不揃いの 2 元配置データ」について，各種のダミー変数を使い，Excel の行列関数による最小 2 乗平均の算出法を詳しく示した．

さらにショックなのは，「繰り返し不揃いの 2 元配置データ」は，解析できないなどの Web 上の記事が数多く見出され，中には，高橋・大橋・芳賀（1989）の 15 章を参照せよなどの Q&A にも遭遇し，統計解析の質の低下を実感した．なぜ，このような嘆かわしい状況になっているのだろうか．SAS の地盤低下もその原因の一つであり，無償で継続的に使える OnDemand SAS が安定的に使えるようになったこともあり，この嘆かわしい状況を少しでも変える努力をしたい．そのためには，SAS の使用法だけでは，フリーソフトでの使い方の説明と同じレベルであり, 万人が使える Excel を用いた SAS の解析法の詳しい説明が不可欠と考えた．

2. 4 種の平方和と LSMEANS

高橋ら（1989）の第 15 章では，「統計ソフトウェアの充実に伴い，解析過程がブラックボックスとなり，出力結果を適切に解釈し利用できない場合が増加しつつある．GLM プロシジャで出力される 4 種の平方和と最小二乗平均（LSMEANS）について，利用に必要な最小限度の解説を加えるのがこの章の目的である．GLM プロシジャは，特性値 y の総平方和 S_{TOTAL} を MODEL ステートメントで指定した要因で説明できる部分 S_{MODEL} と説明できない部分 S_{ERROR} （残差平方和）に分解する．MODEL ステートメントに含まれる要因が二つ以上あるとき，MODEL ステートメントの SS1，SS2，SS3 および SS4 オプションにより各要因の寄与部分を表す平方和が出力される．この平方和は， それぞれ Type I，Type II，Type III，および Type IV の平方和と呼ばれている．」と指摘し，「15-1 繰り返し数が等しい場合」，「15-2 繰り返し数が等しくないが因子が直交する場合」に引き続き，「15-3 因子が直交しない場合」について例示されている．第 15-3 節に示されているデータを表 1 に示す．

表 1　繰り返し不揃いで因子が直交しない場合

	B_1		B_2				B_3		平均 y
A_1	10	13	14	12	**15**	**11**	22	19	14.5
A_2	15	14	16	18			21	18	17.0
平均 y	13.0		14.33				20.0		15.57

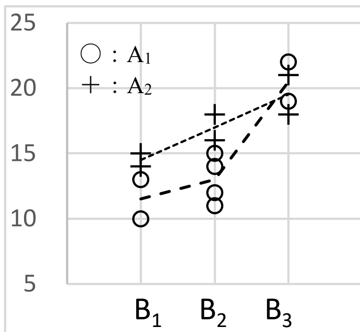


図 1 散布図に平均値を重ね書き

SAS の GLM プロシジャでは，繰り返し数が等しい場合でも，等しくないが因子が直交する場合でも，繰り返し数が不揃いで因子が直交しない場合でも，SAS の GLM プロシジャの使い方は同じだが，結果が微妙に異なるこ

とが強調されている．特に結果に微妙な差が起きるタイプ I，II，III，IV の各平方和の使い分けを主体にしている．当時は，IBM のメインフレーム版の SAS を用いて統計解析業務を行っていたが，現在は，無償で継続的に使える OnDemand SAS を使い，結果を再現する．結果の出力は，HTML 形式が標準設定となっているので，出力結果を全て Excel シート上にペーストし，整形した結果を用い，Excel で作成した 95%信頼区間のひげ付き線グラフを示す．

表 2 左に示したのは，SAS のプログラムでデータを取り込む DATA ステップ，解析のための PROC ステップで，GLM プロシジャを起動し，model ステートメントで解析モデルの設定，（ss1，ss2，ss3，ss4）のオプションで 4 種の平方和の出力，さらに最小 2 乗平均を出力するための lsmeans ステートメントが設定されている．表 2 右は，GLM プロシジャの出力で，分散分析表および 4 種の平方和が示されている．

表 2　SAS の GLM プロシジャによる分散分析表と 4 種の平方和

```
Title1 'TwoWay_unB_ANOVA.sas  2022-5-
22 Y.Takahashi' ;

data  D1 ;
    input A$ B$ @@ ;
    do k = 1 to 4 ;
      input Y @@; output ;
    end  ;
datalines ;
  A1 B1  10  13   .   .
  A1 B2  14  12  15  11
  A1 B3  22  19   .   .
  A2 B1  15  14   .   .
  A2 B2  16  18   .   .
  A2 B3  21  18   .   .
  ;
proc print data=D1 ;  run ;
proc glm data=D1 ;
    class A B ;
    model   y =  A B A*B
        / ss1 ss2 ss3 ss4 ;
    lsmeans     A B A*B / cl ;
    run ;
```

要因	自由度	平方和	平均平方	F 値	Pr > F
Model	5	145.4286	29.0857	8.95	0.0039
Error	8	26.0000	3.2500		
Corrected Total	13	171.4286			

R2 乗	変動係数	Root MSE	Y の平均
0.8483	11.5775	1.8028	15.5714

要因	自由度	Type I	平均平方	F 値	Pr > F
A	1	21.4286	21.4286	6.59	0.0332
B	2	108.8000	54.4000	16.74	0.0014
A*B	2	15.2000	7.6000	2.34	0.1586

要因	自由度	Type II	平均平方	F 値	Pr > F
A	1	16.1333	16.1333	4.96	0.0565
B	2	108.8000	54.4000	16.74	0.0014
A*B	2	15.2000	7.6000	2.34	0.1586

要因	自由度	Type III	平均平方	F 値	Pr > F
A	1	13.0909	13.0909	4.03	0.0797
B	2	105.2000	52.6000	16.18	0.0015
A*B	2	15.2000	7.6000	2.34	0.1586

要因	自由度	Type IV	平均平方	F 値	Pr > F
A	1	13.0909	13.0909	4.03	0.0797
B	2	105.2000	52.6000	16.18	0.0015
A*B	2	15.2000	7.6000	2.34	0.1586

表 3 に示したのは，4 種の平方和を横に並べ，比較した結果である．タイプ III とタイプ IV が完全に一致しているのは，欠測となっている組合せセルが無いためである．これらの違いについて，説明を尽くしてもブラック・ボックスのままである．

表 3　4 種の平方和の比較

要因	自由度	Type I		Type II		Type III		Type IV
A	1	21.4286	≠	16.1333	≠	13.0909	=	13.0909
B	2	108.8000	=	108.8000	≠	105.2000	=	105.2000
A*B	2	15.2000	=	15.2000	=	15.2000	=	15.2000

さて，さらにブラック・ボックス的なのは，表 4 左に示す最小 2 乗平均（lsmeans）である．あえて説明をすれば，表 1 に示す A×B の 2 元表の各セルの平均を計算し，さらにセル平均の平均が A と B の最小 2 乗平

均となる．なぜ，セル平均なのか，内部でこのような計算をしているのか，95%信頼区間を出すための標準誤差 SE は，どのような計算をしているのか，他の各種のフリーソフトで再現するために，計算方法が知りたい，との要望に答えるために，第 16 章の「GLM プロシジャの計算方式」で，SAS の行列計算言語 IML を使って，説明をしたのである．だが，その当時 IML は有償であり，誰でも自由に使えるわけではなく，GLM プロシジャでは，このような行列計算をしているとの言い訳的な説明に過ぎなかった．表 4 右に示すのは，Excel の線グラフによる最小 2 乗平均に 95%信頼区間を重ね書きした結果である．無償で継続的に使える OnDemand SAS の HTML 出力をそっくり Excel シート取り込むことができるので，体裁を整えることも容易である．

表 4 左に示すのは，lsmeans ステートメントで出力された各水準の最小 2 乗平均と 95%信頼区間を Excel に取り込んで整形し，最小 2 乗平均と 95%信頼区間の差から「幅」を別途計算し，Excel の線グラフで（A1 から A2 B3）まで連続した折れ線を描き，上下の幅を重ね書きし，線種を整え，切れ目を入れた結果が示されている．さらに，標準誤差 SE = 幅 / T.inv.2T(0.05, 8) を Excel シート上で計算した結果である．このような連係プレーが，容易にできるようになったことは，画期的なことである．

表 4　GLM プロシジャの最小 2 乗平均と Excel による 95%信頼区間のひげ付き線グラフ

A	B	最小 2 乗平均	95% 信頼限界		幅	SE
A1		15.00	13.45	16.55	1.55	0.6719
A2		17.00	15.30	18.70	1.70	0.7360
	B1	13.00	10.92	15.08	2.08	0.9014
	B2	15.00	13.20	16.80	1.80	0.7806
	B3	20.00	17.92	22.08	2.08	0.9014
A1	B1	11.50	8.56	14.44	2.94	1.2748
A1	B2	13.00	10.92	15.08	2.08	0.9014
A1	B3	20.50	17.56	23.44	2.94	1.2748
A2	B1	14.50	11.56	17.44	2.94	1.2748
A2	B2	17.00	14.06	19.94	2.94	1.2748
A2	B3	19.50	16.56	22.44	2.94	1.2748

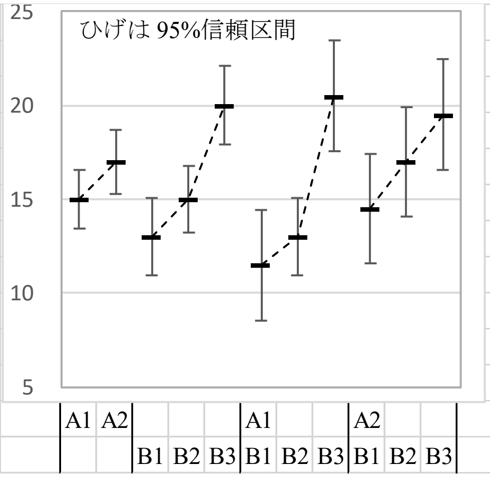
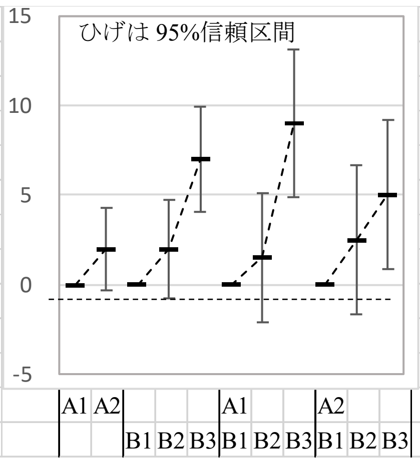


表 5 に estimate ステートメントを使って，A1，B1 を基準とした水準間の差，（A1 B1），（A2 B1）を基準とした組合せ水準の差と SE を求めた結果を示す．水準間の差について lsmeans ステートメントのオプション pdiff，tdiff などで出力することができる．ただし，p 値と t 値がマトリックス状に出力されるが，95%信頼区

表 5　Estimate ステートメントによる最小 2 乗平均の差に関する 95%信頼区間の線グラフ

パラメータ	推定値	標準誤差	t 値	Pr > |t|	$t_{0.05}SE$	L95%	U95%
	0.00						
A2-A1	2.00	0.9965	2.01	0.0797	2.2980	-0.30	4.30
	0.00						
B2-B1	2.00	1.1924	1.68	0.1320	2.7497	-0.75	4.75
B3-B1	7.00	1.2748	5.49	0.0006	2.9396	4.06	9.94
	0.00						
A1*B2-A1*B1	1.50	1.5612	0.96	0.3648	3.6002	-2.10	5.10
A1*B3-A1*B1	9.00	1.8028	4.99	0.0011	4.1572	4.84	13.16
	0.00						
A2*B2-A2*B1	2.50	1.8028	1.39	0.2029	4.1572	-1.66	6.66
A2*B3-A2*B1	5.00	1.8028	2.77	0.0242	4.1572	0.84	9.16



間，あるいは，*SE* の出力が無いために，ひげ付き線グラフが書けない．そこで，estimate ステートメントを用いて標準誤差 *SE* を得ることにした．Estimate ステートメントは，lsmeans ステートメントで推定できる最小 2 乗平均も含め，あらゆる推定ができる．ただし，GLM プロシジャが用いているダミー変数についての知識が必要であり，ここでは，estimate ステートメントの使用法に立ち入らない．

3. 平方和の分解によるチャレンジ

繰り返し不揃いの 2 元配置では，平方和の分解による分散分析が，どのような理由によりできないのかを明らかにする．表 6 に示すのは，繰り返しが等しい場合の 2 元配置の平方和の分解の手順を，不揃いの場合に適用した結果である．総平方和 $S_{\mathrm{T}} = 171.43$，残差平方和 $S_e = 26.00$ は，表 2 の分散分析表 Error の平方和に一致する．モデル全体の平方和は，$S_{\mathrm{Model}} = S_{\mathrm{T}} - S_e = 145.43$ となり，平方和の分解は成り立っている．さて，$S_{\mathrm{A}} = 21.43$，$S_{\mathrm{B}} = 114.10$，$S_{\mathrm{A \times B}} = 16.10$ と計算されているので，それらを合計すると $S'_{\mathrm{Model}} = 151.62$ となり，$S_{\mathrm{Model}} = 145.43$ に一致しない．それぞれの平方和を表 2 に示した GLM プロシジャのタイプ I，II，III，IV の平方和と比較しても一致する結果がない（ただし，$S_{\mathrm{A}} = 21.43$ のみは，タイプ I の平方和に一致している）．したがって，平方和の分解による分散分析ができないことは，明らかである．

表 6　繰り返し不揃いで因子が直交しない場合の平方和の計算

			$\hat{\mu}$			$\hat{\alpha}$		$\hat{\beta}$			$\widehat{(\alpha\beta)}$	$\hat{\varepsilon}$
		y_{ijk}	$\bar{y}_{\cdots}$	$y - \bar{y}_{\cdots}$	$\bar{y}_{\mathrm{A}}$	$\bar{y}_{\mathrm{A}} - \bar{y}_{\cdots}$	$\bar{y}_{\mathrm{B}}$	$\bar{y}_{\mathrm{B}} - \bar{y}_{\cdots}$	$\bar{y}_{\mathrm{AB}}$	$\hat{y}_{\mathrm{AB}}$	$\bar{y}_{\mathrm{AB}} - \hat{y}_{\mathrm{AB}}$	$y - \bar{y}_{\mathrm{AB}}$
$\mathrm{A_1}$	$\mathrm{B_1}$	10	15.57	-5.57	14.50	1.07	13.00	2.57	11.50	11.93	-0.43	-1.50
		13	15.57	-2.57	14.50	1.07	13.00	2.57	11.50	11.93	-0.43	1.50
	$\mathrm{B_2}$	14	15.57	-1.57	14.50	1.07	14.33	1.24	13.00	13.26	-0.26	1.00
		12	15.57	-3.57	14.50	1.07	14.33	1.24	13.00	13.26	-0.26	-1.00
		15	15.57	-0.57	14.50	1.07	14.33	1.24	13.00	13.26	-0.26	2.00
		11	15.57	-4.57	14.50	1.07	14.33	1.24	13.00	13.26	-0.26	-2.00
	$\mathrm{B_3}$	22	15.57	6.43	14.50	1.07	20.00	-4.43	20.50	18.93	1.57	1.50
		19	15.57	3.43	14.50	1.07	20.00	-4.43	20.50	18.93	1.57	-1.50
$\mathrm{A_2}$	$\mathrm{B_1}$	15	15.57	-0.57	17.00	-1.43	13.00	2.57	14.50	14.43	0.07	0.50
		14	15.57	-1.57	17.00	-1.43	13.00	2.57	14.50	14.43	0.07	-0.50
	$\mathrm{B_2}$	16	15.57	0.43	17.00	-1.43	14.33	1.24	17.00	15.76	1.24	-1.00
		18	15.57	2.43	17.00	-1.43	14.33	1.24	17.00	15.76	1.24	1.00
	$\mathrm{B_3}$	21	15.57	5.43	17.00	-1.43	20.00	-4.43	19.50	21.43	-1.93	1.50
		18	15.57	2.43	17.00	-1.43	20.00	-4.43	19.50	21.43	-1.93	-1.50
		平方和		**171.43**		21.43		114.10			16.10	26.00
				S_{T}		S_{A}		S_{B}			$S_{\mathrm{A \times B}}$	S_e
						$S_{\mathrm{T}}' = S_{\mathrm{A}} + S_{\mathrm{B}} + S_{\mathrm{A \times B}} + S_{\mathrm{e}} =$				**177.62**		

4. SAS の GLM プロシジャを使うしかないのか

GLM プロシジャにより，繰り返し不揃いの 2 元配置の解析が行えることは分かったが，正しいのだろうか．SAS なので，使い方さえ正しければ，信頼できる結果として受け入れていいのだろうか．守屋ら（2018）にフリーソフト R で lm() 関数と drop1() 関数を組み合わせ，共変量を含む繰り返し不揃いの 2 元配置のデータに対する分散分析の結果が示されていて，GLM プロシジャのタイプ III の平方和が一致することが確認された．最小 2 乗平均については，別途 R の lsmeans パッケージを使った結果も示されている．もちろん，

GLM プロシジャの結果に一致する．高橋（2021a）では，Excel のみを用いて，守屋ら（2018）が示した最小 2 乗平均が再現できることを示した．これらの一連の結果から，GLM プロシジャの正しさが，再確認された．高橋ら（1989）の第 16 章で，SAS の行列計算言語 IML によって，繰り返し不揃いの 2 元配置の解析が再現できることを示した．これらの経験を元に，Excel のみを用い統計ソフトに依存しない解析法を示すことにより，繰り返し不揃いの 2 元配置データの解析が誰にでも追試ができ，また，フリーソフトを用いた解析にチャレンジされることを願っている．

5. Excel の回帰分析が救いの神

実務に明け暮れていた時代には，統計解析に Excel を使うことなど論外であった．啓蒙活動に軸足を移した時に，統計ソフトの解析結果を解釈し説明するためには，Excel で再現できることを示す重要性を痛感するようになった．平方和の分解で解決できない繰り返し不揃いの 2 元配置の解析が Excel で出来ることを示すことにより，平方和の分解に代わる新たな方法が多くの悩める人達への救いになってほしいと願っている．

GLM プロシジャは，平方和の分解ではなく，ダミー変数をベースにした線形モデル（回帰モデル）による解析を行っている．線形モデルでは，すべての変数が連続変数でなければならないので，質的変数の場合はダミー変数に置き換える必要がある．「ダミー変数とは（なし・あり）を数値化して（0，1）とすること」など，断定的な表現を見かけるたびにうんざりする．（0，1）ではなく線形モデルでは，（1，-1）のように足して 0 となるような対比型ダミー変数とすることが，繰り返し不揃いの場合の最小 2 乗平均および分散の推定を容易にする．

表 7 左に繰り返し不揃いの 2 元配置データの解析に対し，線形モデルを適用するために必要なデザイン行列を示す．因子 A の A_1 に対し $a_1=1$，A_2 に対し $a_1=-1$，因子 B の B_1 に対し $(b_1=1,\ b_2=0)$，B_2 に対し $(b_1=0,\ b_2=1)$，B_3 に対し $(b_1=-1,\ b_2=-1)$ を与え，交互作用 A×B は，因子 A と因子 B のダミー変数の積で与える．さらに切片の推定のための変数とし，$x_0=1$ を加えている．表 7 右に Excel の回帰分析を用いた結果を示す．回帰全体に対する分散分析表は，表 2 に示した GLM プロシジャのモデル全体に対する分散分析表に一致するが，4 種の平方和の出力はなく，回帰パラメータの推定値が出力される．

表 7　繰り返し数が不揃いの 2 元配置データに対する回帰分析

			—— デザイン行列 $\boldsymbol{X}$ ——					
		y	x_0	a_1	b_1	b_2	a_1b_1	a_1b_2
A_1	B_1	10	1	1	1	0	1	0
		13	1	1	1	0	1	0
	B_2	14	1	1	0	1	0	1
		12	1	1	0	1	0	1
		15	1	1	0	1	0	1
		11	1	1	0	1	0	1
	B_3	22	1	1	-1	-1	-1	-1
		19	1	1	-1	-1	-1	-1
A_2	B_1	15	1	-1	1	0	-1	0
		14	1	-1	1	0	-1	0
	B_2	16	1	-1	0	1	0	-1
		18	1	-1	0	1	0	-1
	B_3	21	1	-1	-1	-1	1	1
		18	1	-1	-1	-1	1	1

Excel の回帰分析（x_0 を除く）
分散分析表

		自由度	変動	分散	分散比
	回帰	5	145.4286	29.0857	8.9495
	残差	8	26.0000	3.2500	
	合計	13	171.4286		
		係数	標準誤差	t	P-値
$\hat{\theta}_0$	切片	16.00	0.4983	32.1117	0.0000
$\hat{\theta}_1$	a_1	-1.00	0.4983	-2.0070	0.0797
$\hat{\theta}_2$	b_1	-3.00	0.7205	-4.1639	0.0031
$\hat{\theta}_3$	b_2	-1.00	0.6719	-1.4884	0.1750
$\hat{\theta}_4$	a_1b_1	-0.50	0.7205	-0.6940	0.5073
$\hat{\theta}_5$	a_1b_2	-1.00	0.6719	-1.4884	0.1750

表 2 に示した GLM プロシジャのタイプ I，II，III の平方和は，表 7 左に示したデザイン行列 $\boldsymbol{X}$ の変数の一部の変数を用いた回帰分析を繰り返し行ない，得られた分散分析表のモデルのそれぞれの平方和 S_{MODEL} の差分から得られることを示す．表 7 右に示した分散分析表のモデル（回帰）の平方和をフルモデル $S_{(\mathrm{A+B+A\times B})}=145.4286$ と表す．表 8 に示したのは，Excel の回帰分析で選択する変数を○印で示し，6 種の回帰モデルが示され，Excel の回帰分析を実行した場合のモデルの平方和を抜き出した結果が示されている．因子 A のみのモデルの平方和は，$S_{(\mathrm{A})}=21.4286$ で，因子 B のみのモデルの平方和は，$S_{(\mathrm{B})}=114.0952$ で，因子 A と因子 B を合わせたモデルの平方和は，$S_{(\mathrm{A+B})}=130.2286$ である．フルモデル $S_{(\mathrm{A+B+A\times B})}$ から因子 A を除いた場合は，$S_{(\mathrm{B+A\times B})}=132.3377$ であり，因子 B を除いた場合は，$S_{(\mathrm{A+A\times B})}=40.2286$ である．

表 8　繰り返し数が不揃いの 2 元配置データに対する回帰分析

		デザイン行列 $\boldsymbol{X}$						モデルの
		切片	A	—B—		—A×B—		平方和
番号	6 種の回帰モデル	x_0	a_1	b_1	b_2	a_1b_1	a_1b_2	S_{MODEL}
1	因子A	△	○					21.4286
2	因子B	△		○	○			114.0952
3	因子A＋因子B	△	○	○	○			130.2286
4	因子B＋交互作用A×B	△		○	○	○	○	132.3377
5	因子A　＋交互作用A×B	△	○			○	○	40.2286
6	因子A＋因子B＋交互作用A×B	△	○	○	○	○	○	145.4286

表 9 に示す「タイプ $\mathrm{I_A}$」の平方和は，$S^{(1)}_{(\mathrm{A})}=S_{(\mathrm{A})}=21.4286$，$S^{(1)}_{(\mathrm{B})}=S_{(\mathrm{A+B})}-S_{(\mathrm{A})}=108.8000$，$S^{(1)}_{(\mathrm{A\times B})}=S_{(\mathrm{A+B+A\times B})}-S_{(\mathrm{A+B})}=15.2000$ として求められており，表 2 のタイプ I の平方和に一致する．タイプ I の平方和は，モデルに含める順番に依存していて，因子 B，因子 A の順にした場合は，「タイプ $\mathrm{I_A}$」の場合は，$S^{(1)\prime}_{(\mathrm{B})}=S_{(\mathrm{B})}$，$S^{(1)\prime}_{(\mathrm{A})}=S_{(\mathrm{A+B})}-S_{(\mathrm{B})}$ となる．タイプ II の平方和は，$S^{(2)}_{(\mathrm{A})}=S_{(\mathrm{A+B})}-S_{(\mathrm{B})}=16.1333$，$S^{(2)}_{(\mathrm{B})}=S_{(\mathrm{A+B})}-S_{(\mathrm{A})}=108.8000$ となり，$S^{(2)}_{(\mathrm{A})}$ と $S^{(1)}_{(\mathrm{A})}$ は異なる．タイプ III の平方和は，$S^{(3)}_{(\mathrm{A})}=S_{(\mathrm{A+B+A\times B})}-S_{(\mathrm{B+A\times B})}=13.0909$，$S^{(3)}_{(\mathrm{B})}=S_{(\mathrm{A+B+A\times B})}$

表 9　Excel の回帰分析の分散分析表のモデルの平方和の差分から得られる平方和

要因	*df*	回帰の平方和			差し引く平方和			タイプ $\mathrm{I_A}$	*df*	平方和
A	1	$S_{(\mathrm{A})}$	21.4286				=	$S_{(\mathrm{A})}^{(1)}$	1	21.4286
B	3	$S_{(\mathrm{A+B})}$	130.2286	-	$S_{(\mathrm{A})}$	21.4286	=	$S_{(\mathrm{B})}^{(1)}$	2	108.8000
A×B	5	$S_{(\mathrm{A+B+A\times B})}$	145.4286	-	$S_{(\mathrm{A+B})}$	130.2286	=	$S_{(\mathrm{A\times B})}^{(1)}$	2	15.2000
e	13	$S_{(\mathrm{T})}$	171.4286	-	$S_{(\mathrm{A+B+A\times B})}$	145.4286	=	$S_{(e)}^{(1)}$	8	26.0000
T									13	171.4286
要因	*df*	回帰の平方和			差し引く平方和			タイプ II	*df*	平方和
A	3	$S_{(\mathrm{A+B})}$	130.2286	-	$S_{(\mathrm{B})}$	114.0952	=	$S_{(\mathrm{A})}^{(2)}$	1	16.1333
B	3	$S_{(\mathrm{A+B})}$	130.2286	-	$S_{(\mathrm{A})}$	21.4286	=	$S_{(\mathrm{B})}^{(2)}$	2	108.8000
A×B	5	$S_{(\mathrm{A+B+A\times B})}$	145.4286	-	$S_{(\mathrm{A+B})}$	130.2286	=	$S_{(\mathrm{A\times B})}^{(2)}$	2	15.2000
e	13	$S_{(\mathrm{T})}$	171.4286	-	$S_{(\mathrm{A+B+A\times B})}$	145.4286	=	$S_{(e)}^{(2)}$	8	26.0000
T			各平方和の合計は，S_{T}=171.4286に一致しない						13	166.1333
要因	*df*	回帰の平方和			差し引く平方和			タイプ III	*df*	平方和
A	3	$S_{(\mathrm{A+B+A\times B})}$	145.4286	-	$S_{(\mathrm{B+A\times B})}$	132.3377	=	$S_{(\mathrm{A})}^{(3)}$	1	13.0909
B	3	$S_{(\mathrm{A+B+A\times B})}$	145.4286	-	$S_{(\mathrm{A+A\times B})}$	40.2286	=	$S_{(\mathrm{B})}^{(3)}$	2	105.2000
A×B	5	$S_{(\mathrm{A+B+A\times B})}$	145.4286	-	$S_{(\mathrm{A+B})}$	130.2286	=	$S_{(\mathrm{A\times B})}^{(3)}$	2	15.2000
e	13	$S_{(\mathrm{T})}$	171.4286	-	$S_{(\mathrm{A+B+A\times B})}$	145.4286	=	$S_{(e)}^{(3)}$	8	26.0000
T			各平方和の合計は，S_{T}=171.4286に一致しない						13	159.4909

$-S_{(\mathrm{A+A \times B})} = 105.2000$ となり，タイプ II の平方和と異なる．なお，交互作用は，$S^{(1)}_{(\mathrm{A \times B})} = S^{(2)}_{(\mathrm{A \times B})} = S^{(3)}_{(\mathrm{A \times B})} = 15.2000$ と共通である．

6. 最小 2 乗平均の 95%信頼区間

繰り返しが不揃いの 2 元配置に対し，表 7 で推定されたパラメータ $\hat{\boldsymbol{\theta}}$ を用い，因子 A，因子 B，交互作用 A×B の最小 2 乗平均を推定するための線形和 $L^{(i)} = \boldsymbol{l}^{(i)}\hat{\boldsymbol{\theta}}$ を表 10 に示す．この結果は，表 4 に示した GLM プロシジャの lsmeans ステートメントによって推定された最小 2 乗平均に一致する．このようにパラメータの推定値に関する線形和で計算された推定値を，SAS では最小 2 乗平均と言っている．「最小 2 乗平均」は，竹内ら（1989），「統計学辞典」の索引でも見いだせない SAS の方言であることを認識する必要があるが，R のパーケージに lsmeans が登場したことにより，SAS の方言からの脱却しつつあることは嬉しいことである．

表 10　線形モデルの推定値を用いた線形和による最小 2 乗平均

			l_0	l_1	l_2	l_3	l_4	l_5				最小
A	B	L	x_0	a_1	b_1	b_2	a_1b_1	a_1b_2	θ^		lθ^	2 乗平均
A_1		$L^{(1)}$	1	1	0	0	0	0	16.00	=	15.00	15.00
A_2		$L^{(2)}$	1	-1	0	0	0	0	-1.00		17.00	17.00
	B_1	$L^{(3)}$	1	0	1	0	0	0	-3.00		13.00	13.00
	B_2	$L^{(4)}$	1	0	0	1	0	0	-1.00		15.00	15.00
	B_3	$L^{(5)}$	1	0	-1	-1	0	0	-0.50		20.00	20.00
A_1	B_1	$L^{(6)}$	1	1	1	0	1	0	-1.00		11.50	11.50
	B_2	$L^{(7)}$	1	1	0	1	0	1			13.00	13.00
	B_3	$L^{(8)}$	1	1	-1	-1	-1	-1			20.50	20.50
A_2	B_1	$L^{(9)}$	1	-1	1	0	-1	0			14.50	14.50
	B_2	$L^{(10)}$	1	-1	0	1	0	-1			17.00	17.00
	B_3	$L^{(11)}$	1	-1	-1	-1	1	1			19.50	19.50

さて，最小 2 乗平均の 95%信頼区間を求めるためには，パラメータ $\hat{\boldsymbol{\theta}}$ に関する共分散行列を必要とする．パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ は，デザイン行列 $\boldsymbol{X}$，分散分析表の残差の平均平方（誤差分散の推定値）$\hat{\sigma}^2 = 3.2500$ を用いて

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2 \tag{1}$$

として定義される．線形和 $L = \boldsymbol{l}\hat{\boldsymbol{\theta}}$ の分散 $Var(\boldsymbol{l}\hat{\boldsymbol{\theta}})$ は，$\boldsymbol{l}$ に関する 2 次形式

$$Var(\boldsymbol{l}\hat{\boldsymbol{\theta}}) = \boldsymbol{l}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}^2]\boldsymbol{l}^T = \boldsymbol{l}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\boldsymbol{l}^T \tag{2}$$

によって推定される．パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ は，行列計算を必要とするので，伝統的な分散分析では使われてこなかったが, Excel の行列計算を用いれば表 11 に示すように容易に計算できる. 表 11 右の *SE* は，表 7 に示した Excel の回帰分析のパラメータの「標準誤差」に一致する．これは，$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ の対角要素が，回帰パラメータの分散 $Var(\hat{\boldsymbol{\theta}})$ となるので，その平方根を取ったものである．

表 11　パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$

	パラメータの共分散行列　Σ (θ^)=(X^T X)^-1 σ^2						Var(θ^)	SE
x_0	**0.2483**	-0.0226	0.0226	-0.0451	0.0226	-0.0451	0.2483	0.4983
a_1	-0.0226	**0.2483**	0.0226	-0.0451	0.0226	-0.0451	0.2483	0.4983
b_1	0.0226	0.0226	**0.5191**	-0.2257	-0.0226	0.0451	0.5191	0.7205
b_2	-0.0451	-0.0451	-0.2257	**0.4514**	0.0451	-0.0903	0.4514	0.6719
a_1b_1	0.0226	0.0226	-0.0226	0.0451	**0.5191**	-0.2257	0.5191	0.7205
a_1b_2	-0.0451	-0.0451	0.0451	-0.0903	-0.2257	**0.4514**	0.4514	0.6719
	=Minverse(Mmult(Transpose (Xの範囲), Xの範囲))*σ^2						対角要素	=Sqrt(Var)

表 12 に示すのは，表 10 に示した因子 A，因子 B，交互作用 A×B の最小 2 乗平均に，その分散 $Var(\boldsymbol{l}\hat{\boldsymbol{\theta}})$ が追加されている．分散の右隣の「幅 $t_{0.05}\times SE$」は，Excel の線グラフで 95%信頼区間の幅を重ね書きするために必要であり，95%信頼区間を計算にも使われている．この結果は，表 4 に示した GLM プロシジャの lsmeans ステートメントで求められた最小 2 乗平均の 95%信頼区間に一致する．

繰り返しが等しい場合には，パラメータの推定値に関する線形和による計算ではなく，計算された各種の平均を元の y_{ijk} を用いた式に展開し，分散の加法性を活用した分散の計算が定式化されている．この計算が面倒なので，田口の式，あるいは，伊奈の式により有効反復数 n_e を計算し，分散を $Var(\hat{\sigma}^2 / n_e)$ で計算する方法が知られている．このような繰り返しが等しい場合の定式化が，繰り返しが不揃いの場合に解析できないとの迷信が流布している原因でもある．

表 12　パラメータの共分散行列 $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ を用いた最小 2 乗平均に対する 95%信頼区間

			l_0	l_1	l_2	l_3	l_4	l_5					幅		
A	B	L	x_0	a_1	b_1	b_2	a_1b_1	a_1b_2	θ^		lθ^	$Var(l\theta\hat{})$	$t_{0.05}\times SE$	L95%	U95%
A_1		$L^{(1)}$	1	1	0	0	0	0	16.00	=	15.00	0.4514	1.5493	13.45	16.55
A_2		$L^{(2)}$	1	-1	0	0	0	0	-1.00		17.00	0.5417	1.6972	15.30	18.70
	B_1	$L^{(3)}$	1	0	1	0	0	0	-3.00		13.00	0.8125	2.0786	10.92	15.08
	B_2	$L^{(4)}$	1	0	0	1	0	0	-1.00		15.00	0.6094	1.8001	13.20	16.80
	B_3	$L^{(5)}$	1	0	-1	-1	0	0	-0.50		20.00	0.8125	2.0786	17.92	22.08
A_1	B_1	$L^{(6)}$	1	1	1	0	1	0	-1.00		11.50	1.6250	2.9396	8.56	14.44
	B_2	$L^{(7)}$	1	1	0	1	0	1			13.00	0.8125	2.0786	10.92	15.08
	B_3	$L^{(8)}$	1	1	-1	-1	-1	-1			20.50	1.6250	2.9396	17.56	23.44
A_2	B_1	$L^{(9)}$	1	-1	1	0	-1	0			14.50	1.6250	2.9396	11.56	17.44
	B_2	$L^{(10)}$	1	-1	0	1	0	-1			17.00	1.6250	2.9396	14.06	19.94
	B_3	$L^{(11)}$	1	-1	-1	-1	1	1			19.50	1.6250	2.9396	16.56	22.44
	lθ^=Mmult(lの範囲, θ^の範囲)								SE=sqrt(Var(lθ^))			$t_{0.05}$= T.Inv.2T(0.05, 8)=2.3060			
	Var(lθ^)=Mmult(Mmult(lの範囲, Σ (θ^) の範囲),Transpose(lの範囲))														

表 13 に示すのは，因子 A 内の A_1 と A_2 の最小 2 乗平均の差，因子 B 内の B_1 と B_2，B_1 と B_3 の最小 2 乗平均の差，交互作用 A×B の A_1B_1 と A_2B_1 を基準とした差についての 95%信頼区間の計算結果である．この結果は，表 5 に示した GLM プロシジャの estimate ステートメントを用いて推定した結果に一致する．因子 A 内の（A_2-A_1）は，A_1 のベクトル $\boldsymbol{l}^{(1)}$ と A_2 のベクトル $\boldsymbol{l}^{(2)}$ の差 $\boldsymbol{l}^{(12)} = \boldsymbol{l}^{(2)} - \boldsymbol{l}^{(1)}$ として設定されている．他の場合も同様に表 12 に示したそれぞれのベクトル間の差として求めたものである．

表 13　最小 2 乗平均の差に対する 95%信頼区間

			l_0	l_1	l_2	l_3	l_4	l_5					幅		
A	B	L	x_0	a_1	b_1	b_2	a_1b_1	a_1b_2	θ^		lθ^	$Var(l\theta\hat{})$	$t_{0.05}\times SE$	L95%	U95%
A_1-A_1		$L^{(12)}$	0	0	0	0	0	0	16.00	=	0.00	0.0000	0.0000	0.00	0.00
A_2-A_1		$L^{(13)}$	0	-2	0	0	0	0	-1.00		2.00	0.9931	2.2980	-0.30	4.30
	B_1-B_1	$L^{(14)}$	0	0	0	0	0	0	-3.00		0.00	0.0000	0.0000	0.00	0.00
	B_2-B_1	$L^{(15)}$	0	0	-1	1	0	0	-1.00		2.00	1.4219	2.7497	-0.75	4.75
	B_3-B_1	$L^{(16)}$	0	0	-2	-1	0	0	-0.50		7.00	1.6250	2.9396	4.06	9.94
A_1B_1	$-A_1B_1$	$L^{(17)}$	0	0	0	0	0	0	-1.00		0.00	0.0000	0.0000	0.00	0.00
A_1B_2	$-A_1B_1$	$L^{(18)}$	0	0	-1	1	-1	1			1.50	2.4375	3.6002	-2.10	5.10
A_1B_3	$-A_1B_1$	$L^{(19)}$	0	0	-2	-1	-2	-1			9.00	3.2500	4.1572	4.84	13.16
A_2B_1	$-A_2B_1$	$L^{(20)}$	0	0	0	0	0	0			0.00	0.0000	0.0000	0.00	0.00
A_2B_2	$-A_2B_1$	$L^{(21)}$	0	0	-1	1	1	-1			2.50	3.2500	4.1572	-1.66	6.66
A_2B_3	$-A_2B_1$	$L^{(22)}$	0	0	-2	-1	2	1			5.00	3.2500	4.1572	0.84	9.16

7. 考察

多くの大学でデータ・サイエンス学科が新設され，実データを主体にした多様な教育がなされるようになったことは，嬉しい限りである．実データから何らかの規則性を的確に見いだすためには，統計解析の基礎知識と実行能力が必要である．統計解析の専門家を目指すならば統計に関する英語の専門書を読むのはあたりまえであるが，それぞれの学問分野でのデータ解析を行う人達には，日本語で書かれた教科書と統計ソフトが必要である．多くの「統計解析」の書籍は，これらの人達を対象として出版されている．データ・サイエンティストを目指す人達は, 幾つかのプログラミング言語をマスターすることが必須である. 最近の Python などのフリーのプログラミング言語には，行列計算のための関数，さらに重回帰分析など統計解析のための関数が含まれていて，身近な存在となっている．その結果として，Excel の行列関数で示した繰り返しが不揃いの 2 元配置データの解析などは，簡単にプログラミングできるであろう．

実データで，2 つの質的変数，量的な変数を反応として解析しようとすると必然的に「繰り返し不揃いの 2 元配置」に帰着するが，この問題を扱っている日本語の教科書を見い出すことができない．そのためか，Web 上では，「繰り返しが等しくないと解析できない」などが蔓延している．その原因は，平方和の分解に頼り切りで，ダミー変数を用いたデザイン行列 $\boldsymbol{X}$ を用いた解析法が，統計の入門書では，忌み嫌われているためである．そのために，多くの人達にとって身近にある Excel で，簡単に解決できることを示すことにより，「平方和の分解」から脱却してもらいたいと願っている．

SAS の GLM プロシジャが偉大であったのは，分散分析表の作成に留まらず，各種の推定値に対し，最小 2 乗平均とその 95%信頼区間をコンパクトに出力したことにある．伝統的な解析法では，分散分析表を完成させることが主目的で，各種の推定の問題に対しては冷淡であり，その結果として他の統計ソフトには，最小 2 乗平均が欠如しているように思われる．

データ・サイエンティストを目指す人達に対し，適切な事例を提示し，信頼できる無償で継続的に使える統計ソフト SAS の使い方と結果の提示，Excel による解析法の解説と，各種の統計グラフの作成法など一式を提供することが，統計解析の質の向上にかかせないと考え，実践活動を続けている．

参 考 文 献

1） 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.

2） 高橋行雄（2013）, 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－, https://scientist-press.com/wp-content/uploads/2019/07/seminar7.pdf

3） 高橋行雄（2020）, 続高橋セミナー第 9 回 最尤法によるポアソン回帰分析入門 ＜第 13 章＞最小 2 乗平均の謎を予測プロファイルで解く, https://www.yukms.com/biostat/takahasi2/rec/009-13.htm

4） 高橋行雄（2021a）, 各種のダミー変数を用いた最小 2 乗平均と 95%信頼区間の実際, SAS ユーザー総会 2021 論文集, 123-132.

5） 高橋行雄（2021b）,線形モデルによる欠測値がある直交表の解析，https://www.yukms.com/biostat/takahasi2/rec/010.htm

6） 竹内啓ら（1989）, 統計学辞典, 東洋経済.

7） 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/

8） Lenth RV.（2016）, Least-Squares Means : The R Package lsmeans, *Journal of Statistical Software* 69, 1-33. https://www.jstatsoft.org/article/view/v069i01/v69i01.pdf

9） Little R.C., Stroup W.W. and Freund R.j.（2020）, SAS for Linear Models, 4th ed.，SAS Institute.

10）SAS Institute（2013）, SAS/STAT13.1 User's Guide, The GLM Procedure, https://support.sas.com/documentation/onlinedoc/stat/131/glm.pdf