REML法入門

ースタッガード型枝分かれ実験を例にしてー

高橋 行雄

BioStat研究所（株）

制限付き最尤法（REML）

Q1. 制限付き最尤法ってどんな方法ですか.

A1. え！　ぼくもよくわからないんだよ.

Q2. 平均と分散に対して別々の尤度関数を立てて推定する方法，とかWebで見たんだ.
http://www012.upp.so-net.ne.jp/doi/math/anova/REML.pdf

A2. わかったような気がするのだが？

経時測定データ

Q3. SASのMIXEDプロシジャを経時測定データの解析で使っているのだけれども，REML法が標準的に使われているとマニュアルに書いてあるが，釈然としないのですよ.

A3. それなら簡単さ. 線形モデル

$$y = X\beta + \varepsilon$$

を拡張して変量効果を入れたモデル

$$y = X\beta + Z\gamma + \varepsilon$$

について，REML法で解いているのだよ.

最尤法 vs REML法

Q4. ところで，最尤法は尤度関数を偏微分して，尤度が最大になるようなパラメータを推定する方法だと，いろいろな教科書に書いてあるのだけれども，REML法はどうなんでしょうか．

A4. 実は，ぼくもよくわからないのだ．

Q5. なんだ，そうか．SASを使っていると言えば，突っ込まれることはないから，まー，いいか．

基本から積み上げる

- 一組のデータ(1，2，3，4)の最小2乗平均

母平均μを未知とした場合に，Excelのソルバーで最小2乗法により$\hat{\mu}$を推定してみよう．最小2乗法を用いた場合に，偏差$(y_i - \hat{\mu})$の平方和S_eを最小にするような推定値$\hat{\mu}$を求めればよい．

$$S_e = \sum_{i=1}^{4} (y_i - \hat{\mu})^2$$

最小2乗法では，推定したパラメータは「平均値」のみ

一組のデータ(1，2，3，4)の最尤法

最尤法は，母平均μの周りでデータが，ある分布に従うと仮定する．データy_iが平均$\hat{\mu}$，分散$\tilde{\sigma}^2$の正規分布に従うとしたとき，y_iの正規分布の確率密度$L_i = \mathrm{N}(y_i; \hat{\mu}, \tilde{\sigma}^2)$を考え，それらの積として尤度$L$を定義する．

$$L = \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{1}{2\tilde{\sigma}^2} (y_i - \hat{\mu})^2 \right)$$

尤度Lが最大となるように$\hat{\mu}$と$\tilde{\sigma}^2$の値を変化させ，尤度Lがこれ以上大きくならないと判定したときに，ストップする．このとき，推定されたパラメータを最尤解という．

一組のデータの制限付き最尤法

制限付き最尤法は，データ y_i が正規分布に従うとした場合の尤度を

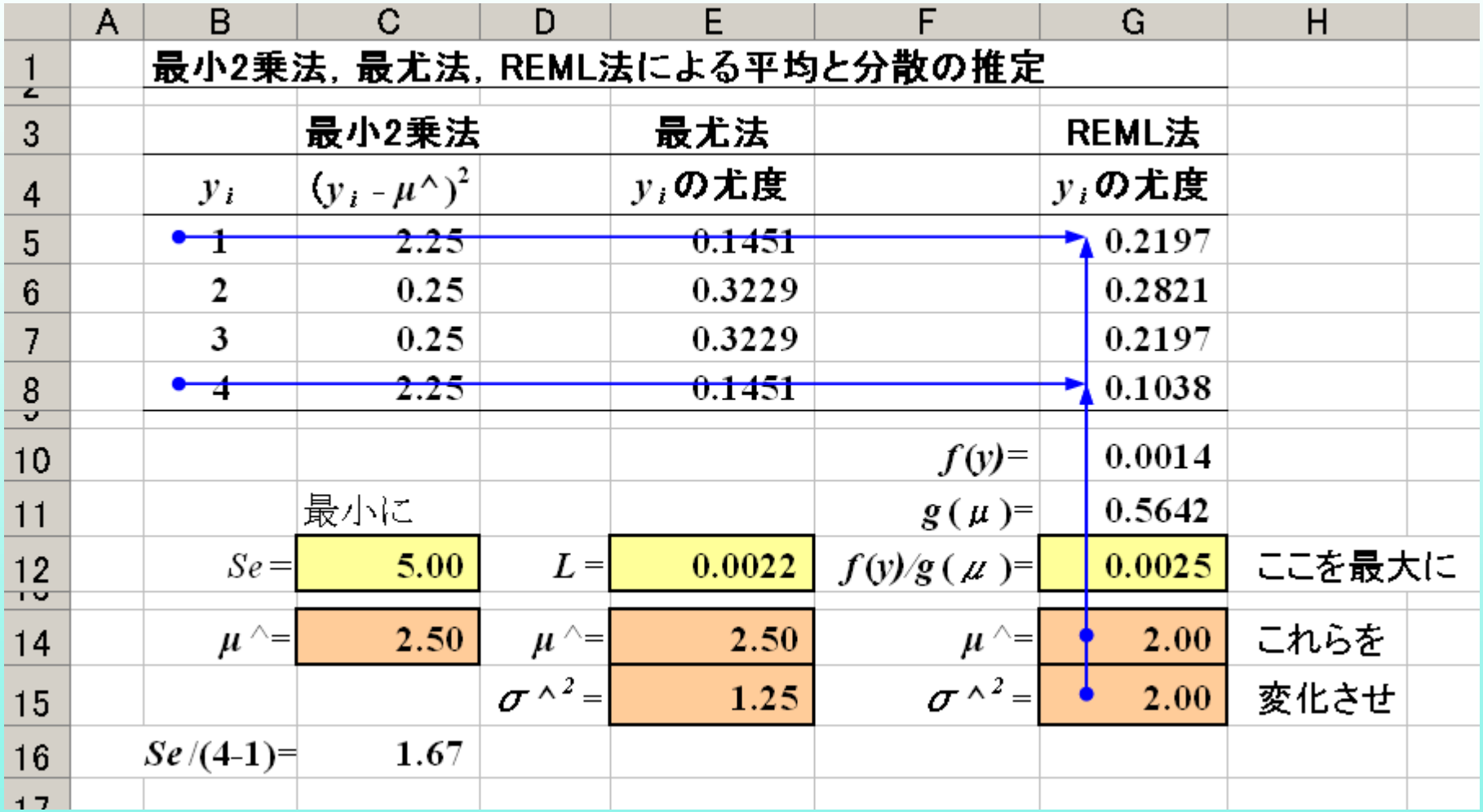
$$f(y) = L = \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}(y_i - \hat{\mu})^2 \right)$$

とし，さらに平均 $\tilde{\mu}$ が正規分布に従うとした場合の尤度を

$$g(\tilde{\mu}) = \frac{1}{\sqrt{2\pi(\hat{\sigma}_e^2/4)}} \exp\left[-\frac{(\tilde{\mu} - \hat{\mu})^2}{2\sqrt{(\hat{\sigma}_e^2/4)}} \right]$$

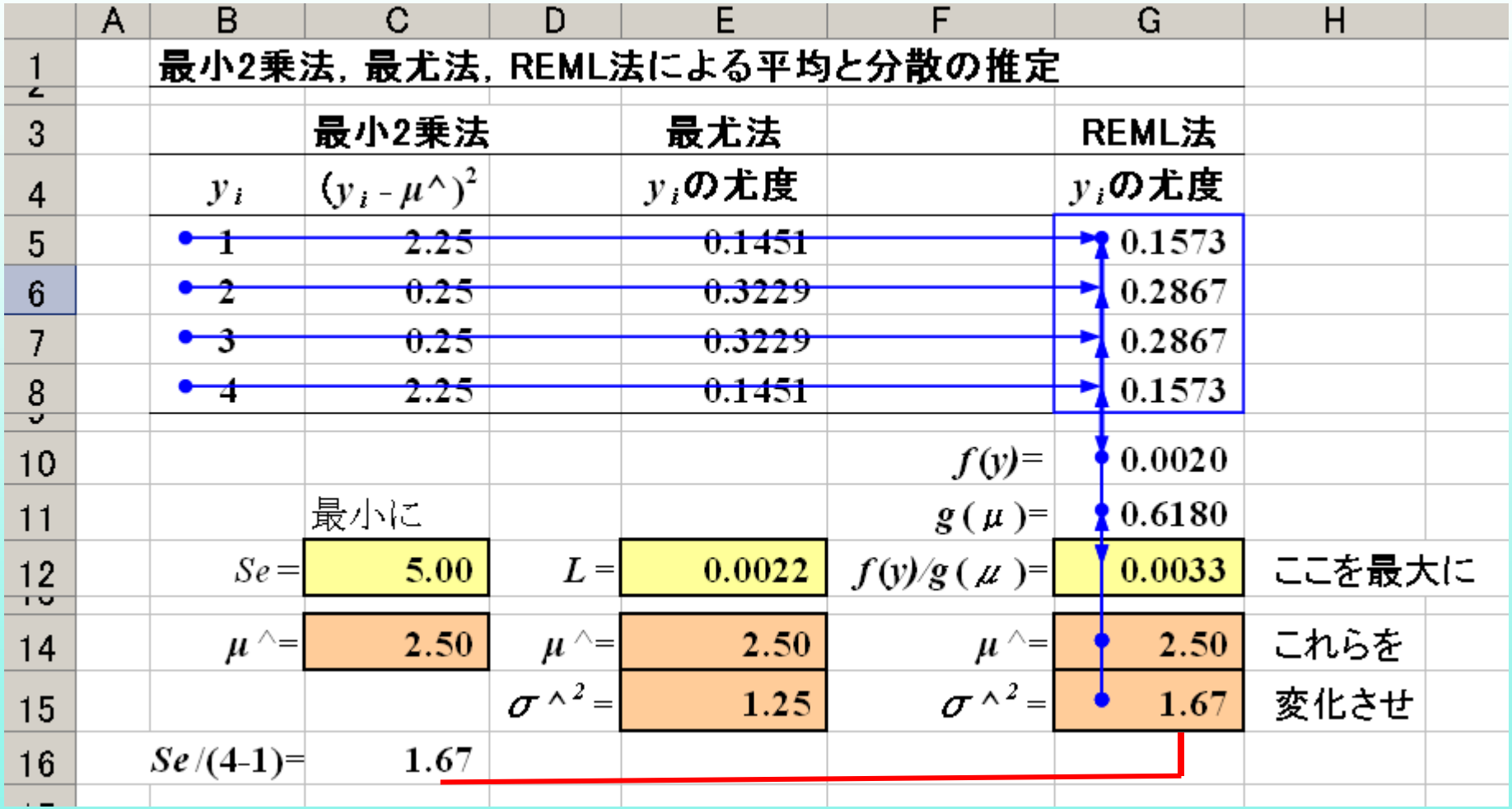
として，その比 $L = f(y)/g(\tilde{\mu})$ を最大にする方法.

Excel ソルバーによる REML法



	A	B	C	D	E	F	G	H
1		最小2乗法，最尤法，REML法による平均と分散の推定						
3			最小2乗法		最尤法		REML法	
4		y_i	$(y_i - \mu\hat{})^2$		y_iの尤度		y_iの尤度	
5		1	2.25		0.1451		0.2197	
6		2	0.25		0.3229		0.2821	
7		3	0.25		0.3229		0.2197	
8		4	2.25		0.1451		0.1038	
10						$f(y)$=	0.0014	
11			最小に			$g(\mu)$=	0.5642	
12		Se=	5.00	L=	0.0022	$f(y)/g(\mu)$=	0.0025	ここを最大に
14		$\mu\hat{}$=	2.50	$\mu\hat{}$=	2.50	$\mu\hat{}$=	2.00	これらを
15				$\sigma\hat{}^2$=	1.25	$\sigma\hat{}^2$=	2.00	変化させ
16		$Se/(4-1)$=	1.67					

REML法による分散の推定



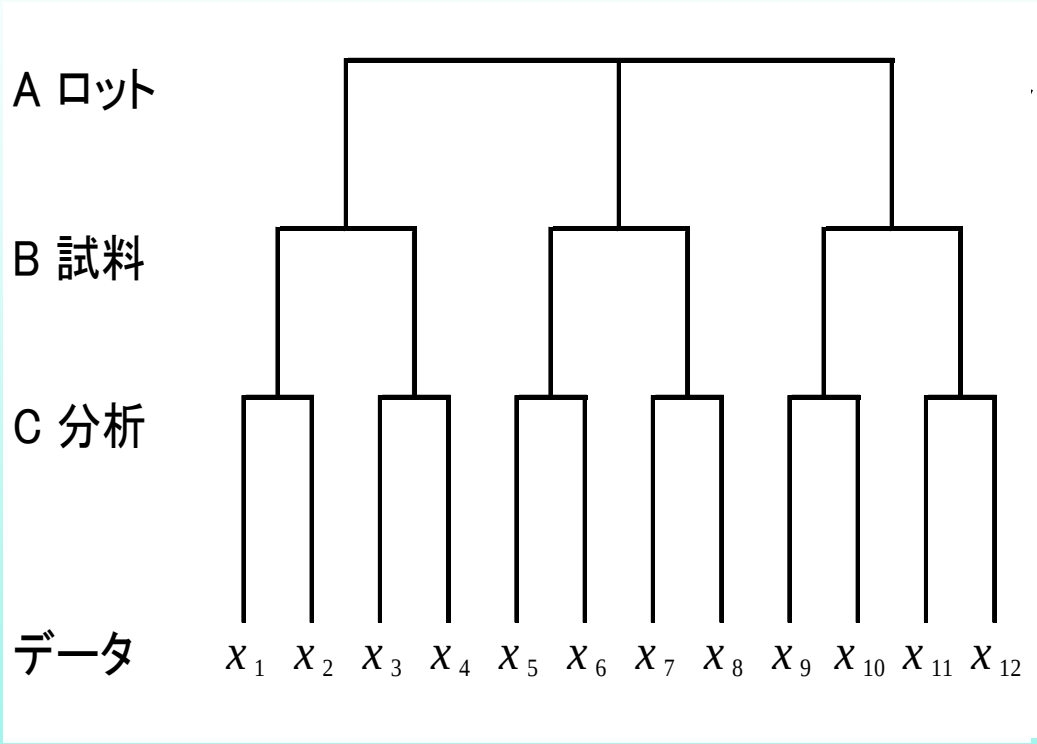
	A	B	C	D	E	F	G	H
1		最小2乗法，最尤法，REML法による平均と分散の推定						
3			最小2乗法		最尤法		REML法	
4		y_i	$(y_i-\mu\hat{})^2$		y_iの尤度		y_iの尤度	
5		1	2.25		0.1451		0.1573	
6		2	0.25		0.3229		0.2867	
7		3	0.25		0.3229		0.2867	
8		4	2.25		0.1451		0.1573	
10						$f(y)=$	0.0020	
11			最小に			$g(\mu)=$	0.6180	
12		$Se=$	5.00	$L=$	0.0022	$f(y)/g(\mu)=$	0.0033	ここを最大に
14		$\mu\hat{}=$	2.50	$\mu\hat{}=$	2.50	$\mu\hat{}=$	2.50	これらを
15				$\sigma\hat{}^2=$	1.25	$\sigma\hat{}^2=$	1.67	変化させ
16		$Se/(4-1)=$	1.67					

不偏分散に一致する

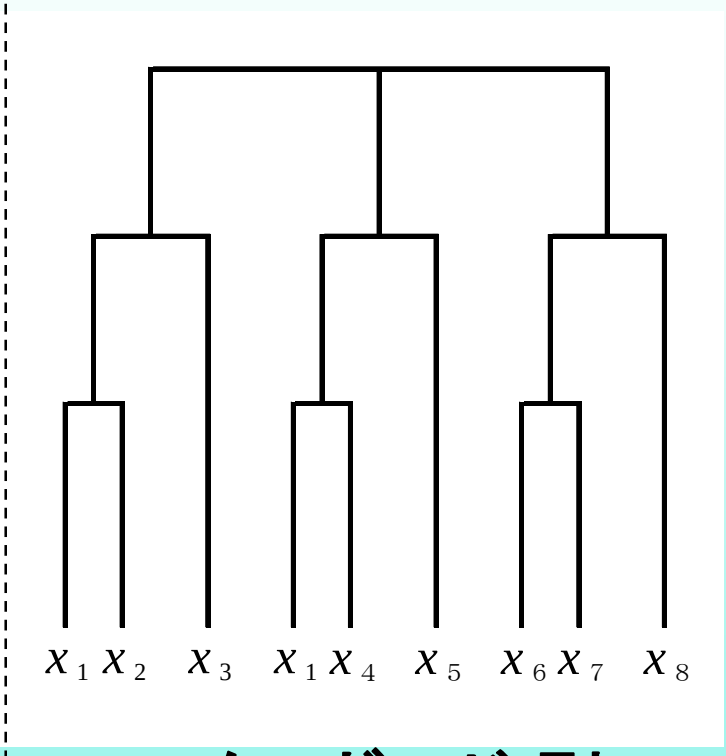
枝分かれ実験はREML法の入門

- 基本的なものから一つ一つ複雑なものへ段階的に進めることが，理解するための原則.
- 枝分かれ実験では，平均が一つの場合で，複数の分散成分の同時推定へと拡張している.
- 枝分かれ実験を，線形混合モデルの最も基本的な問題である．REML法の入門として，枝分かれ実験データの解析が適している.

スタッガード 型枝分かれ実験



完備型



スタッガード 型
データが不揃いの例

スタッカード型とは

http://www.bscycle.co.jp/company/release20090126.html

ホリゾンタル（ダイヤモンド型）

スタッガード 型

スタッガード 型は理想的な実験計画

- スタッガード 型は，手間のかかる最下層の余分な分析の繰返しを排除する.
- 平均平方の期待値を用いた分散成分の推定は統計の教科書には見出せない.
- したがって，適用事例も見出せない.
- SASを使えば，20年以上前から，VARCOPプロシジャで簡単に計算できるのだが，統計の教科書に解説がないので，実務家は使うことができない.

REML法

- REML法は，完備型の枝分かれ実験データでも，繰返しが不揃いな場合でも分散成分を適切に推定する.
- スタッカード型も繰返しが不揃いの場合の枝分かれ実験計画の一例である.
- MIXEDプロシジャ，VARCOMPプロシジャで簡単に分散成分をREML法で求めることができる.

枝分かれ:2段階

枝分れが 2 段データ

B	C	A_1	A_2	A_3
B_1	C_1	94	107	86
	C_2	96	108	87
B_2	C_1	90	117	85
	C_2	(92)	(116)	(83)

( ) 内を欠測値：スタッカード型

SASプログラム

Program 3　VARCOMプロシジャのREMLオプション

```
proc varcomp data=d02 method=reml ;
    class          A  B  C ;
    model  y  =  A  B(A) ;
    run ;
```

Program 4　MIXEDプロシジャによる分散成分の推定

```
proc mixed data=d02 cl=wald covtest ;
    class          A  B  C ;
    model  y  =      ;
    random         A  B(A) ;
    run ;
```

MIXEDプロシジャでの結果

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
A	178.71	188.59	0.95	0.1717	0.05	46.2398	10118
B(A)	18.7485	15.9378	1.18	0.1197	0.05	5.8236	310.35
Residual	1.0042	0.8234	1.22	0.1113	0.05	0.3212	14.2087

拡大

Cov Parm	Estimate	Lower	Upper
A	178.71	46.2398	10118
B(A)	18.7485	5.8236	310.35
Residual	1.0042	0.3212	14.2087

分散成分の95%信頼区間

MIXED プロシジャの分散成分の 95%信頼区間は，χ^2 分布のパーセント点を用いた次式が用いられている．

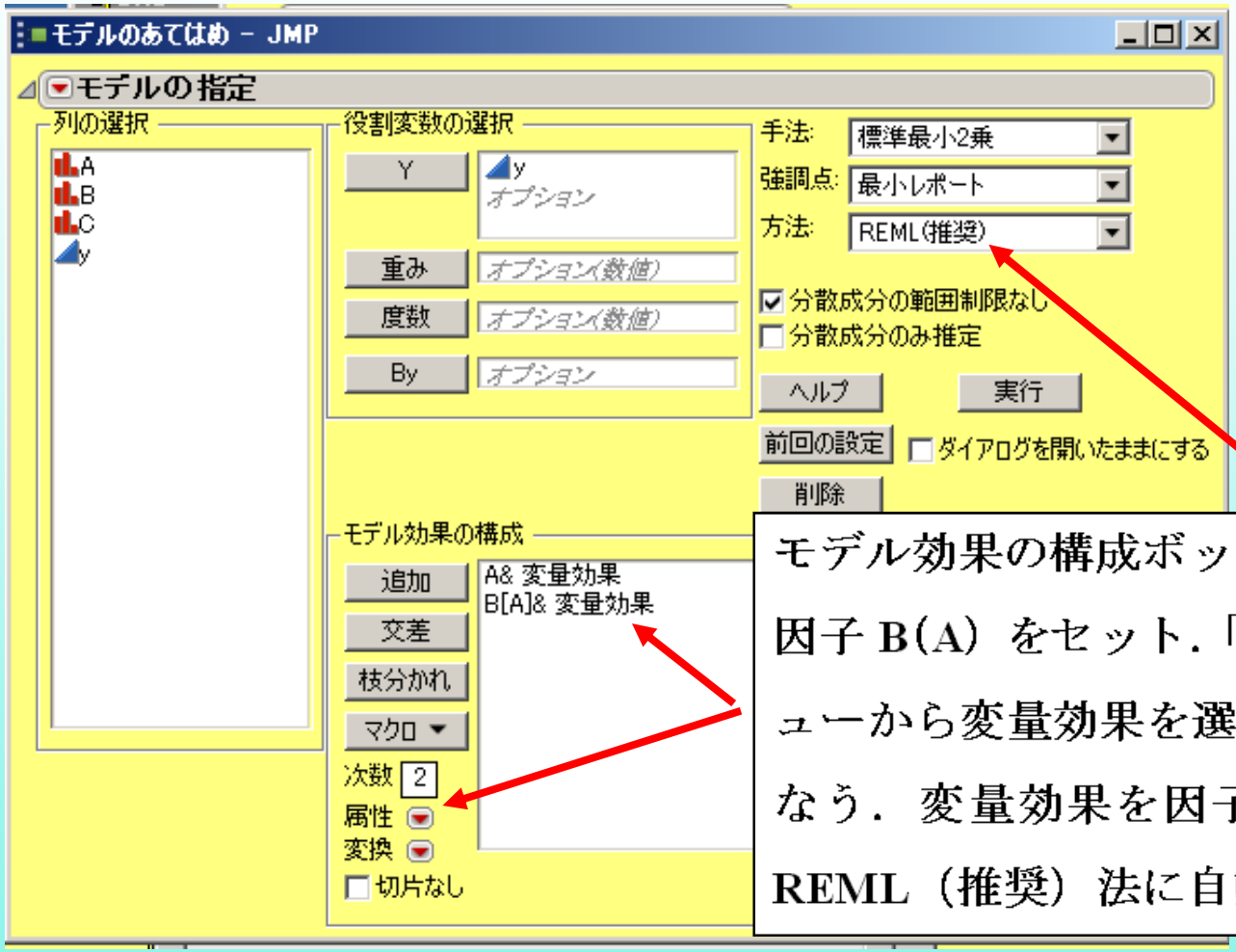
$$\frac{\nu\hat{\sigma}^2}{\chi^2_{\nu,1-\alpha/2}} \le \sigma^2 \le \frac{\nu\hat{\sigma}^2}{\chi^2_{\nu,\alpha/2}}, \quad \text{ただし，} \quad \nu = 2\left(\hat{\sigma}^2 / SE(\hat{\sigma}^2)\right)^2$$

因子 A について 95%信頼区間の下限 46.24 は，

$$\nu = 2\left(\hat{\sigma}_A^2 / SE(\hat{\sigma}_A^2)\right)^2 = 2\times(178.71/188.59)^2 = 1.7959$$

$$\frac{\nu\hat{\sigma}^2}{\chi^2_{\nu,1-\alpha/2}} = \frac{1.7959\times 178.71}{6.9410} = 46.24$$

JMPによる解析



モデル効果の構成ボックスに，因子 A，枝分かれ因子 B(A) をセット.「属性▼」のプルダウンメニューから変量効果を選択して因子の属性変換を行なう．変量効果を因子に含めると，計算方法は REML（推奨）法に自動的に切り替わる.

JMPでの分散成分の推定

REML法による分散成分推定値

変量効果	分散比	分散成分	標準誤差	95%下側	95%上側	全体に対する百分率
A	177.9607	178.7096	188.5907	-190.92	548.34	90.0
B[A]	18.6699	18.7485	15.9378	-12.49	49.99	9.4
残差		1.0042	0.8234	0.32	14.21	0.5
合計		198.4623				100.0

-2対数尤度= 47.362362396

分散成分推定値の共分散行列

変量効果	A	B[A]	残差
A	35566.459	-110.7515	-0.7730
B[A]	-110.7515	254.0137	-0.5638
残差	-0.7730	-0.5638	0.6780

因子 A の分散成分 178.71 となり，MIXED プロシジャの結果と一致する．95%信頼区間は因子 A の場合に（-190.92～548.34）となり $SE(\hat{\sigma}_A^2)$ と t 分布のパーセント点から便宜的に計算しているので使うべきでない．Var(B(A)) も同様．Var(Error) は，χ^2 分布を用いている．分散共分散行列が出力されているので，その対角要素から別途 χ^2 分布のパーセント点を用いた計算を行なうこと．

枝分かれ：１段階

3台の貨車Aから2個の硫化鉱の試料Bを分析

試料番号	貨車 A_1	A_2	A_3
1	42.0	41.4	41.1
2	41.8	41.5	40.8

要因	*df*	平方和	平均平方	平均平方の期待値
貨車間 A	2	0.9033	0.4517	$\sigma_e^2 + 2\sigma_A^2$
貨車内 e	3	0.0700	0.0233	σ_e^2
全体 T	5	0.9733		

$$\hat{\sigma}_e^2 = V_e = 0.0233\,, \quad \hat{\sigma}_A^2 = \frac{0.4517 - 0.0233}{2} = 0.2142$$

平均平方の期待値から分散成分を算出する古典的な方法

枝分かれ実験の分散共分散構造

貨車間の誤差を $\varepsilon_i^{(1)}$，貨車内の誤差を $\varepsilon_{ij}^{(2)}$

$$y_{ij} = \mu + \varepsilon_i^{(1)} + \varepsilon_{ij}^{(2)}$$

			分散共分散構造 V_i	
A_i	r_{ij}	y_{ij}	$r_{ij}=1$	$r_{ij}=2$
A1	**1**	**42.0**	$\sigma_A^2+\sigma_e^2$	σ_A^2
	2	**41.8**	σ_A^2	$\sigma_A^2+\sigma_e^2$
A2	**1**	**41.4**	$\sigma_A^2+\sigma_e^2$	σ_A^2
	2	**41.5**	σ_A^2	$\sigma_A^2+\sigma_e^2$
A3	**1**	**41.1**	$\sigma_A^2+\sigma_e^2$	σ_A^2
	2	**40.8**	σ_A^2	$\sigma_A^2+\sigma_e^2$

最尤法による分散成分の推定

同一貨車内のデータ（y_{i1}，y_{i2}）から 2 次元正規分布の確率密度を求める．

$$L_i = \frac{1}{(2\pi)^{p/2}\left|V_i\right|^{1/2}} \exp\left[-\frac{1}{2}(y_i-\hat{\mu})^T V_i^{-1}(y_i-\hat{\mu})\right]$$

$p=2$ として Excel の行列式 Mdeterm 関数，行列の転置 Transpose 関数，行列の積 Mmult 関数，逆行列 Minverse 関数で計算し，ソルバーで $L=L_1\cdot L_2\cdot L_3$ を最大化する．

多(p)次元正規分布の確率密度

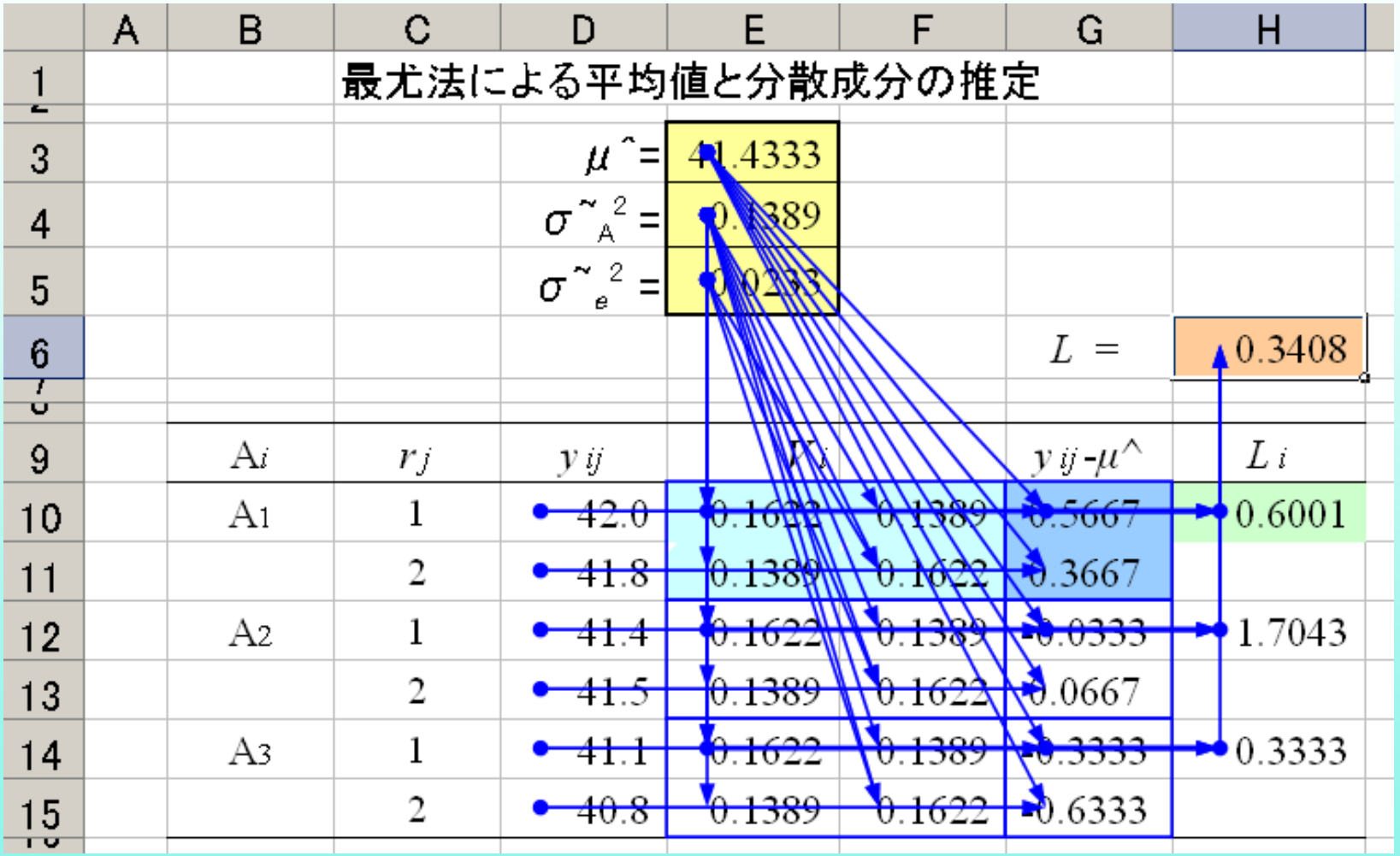
- Excel, SAS, JMPにも 多次元の関数はない.
- 自前で計算式を定義する必要がある.
- 幸いExcel の行列関数で計算できる.

$p=2$ の場合

$$L_i = \frac{1}{(2\pi)^{p/2}\left|V_i\right|^{1/2}} \exp\left[-\frac{1}{2}(y-\hat{\mu})^T V_i^{-1}(y-\hat{\mu})\right]$$

```
=(1/((2*pi())^(2/2)*Mdeterm(E10:F11)^(1/2)))
 * Exp(-(1/2) * Mmult(Transpose(G10:G11),
                     Mmult(Minverse(E10:F11),(G10:G11)) ))
```

Excelのソルバーによる最尤解



	A	B	C	D	E	F	G	H
1			最尤法による平均値と分散成分の推定					
3				$\hat{\mu}$=	41.4333			
4				$\tilde{\sigma}_A^2$ =	0.1389			
5				$\tilde{\sigma}_e^2$ =	0.0233			
6							L =	0.3408
9		Ai	rj	yij		Vi	yij -μ^	Li
10		A1	1	42.0	0.1622	0.1389	0.5667	0.6001
11			2	41.8	0.1389	0.1622	0.3667	
12		A2	1	41.4	0.1622	0.1389	-0.0333	1.7043
13			2	41.5	0.1389	0.1622	0.0667	
14		A3	1	41.1	0.1622	0.1389	-0.3333	0.3333
15			2	40.8	0.1389	0.1622	-0.6333	

固定効果（パラメータ）の尤度

- 4個の平均値の分散（共分散を考慮する）

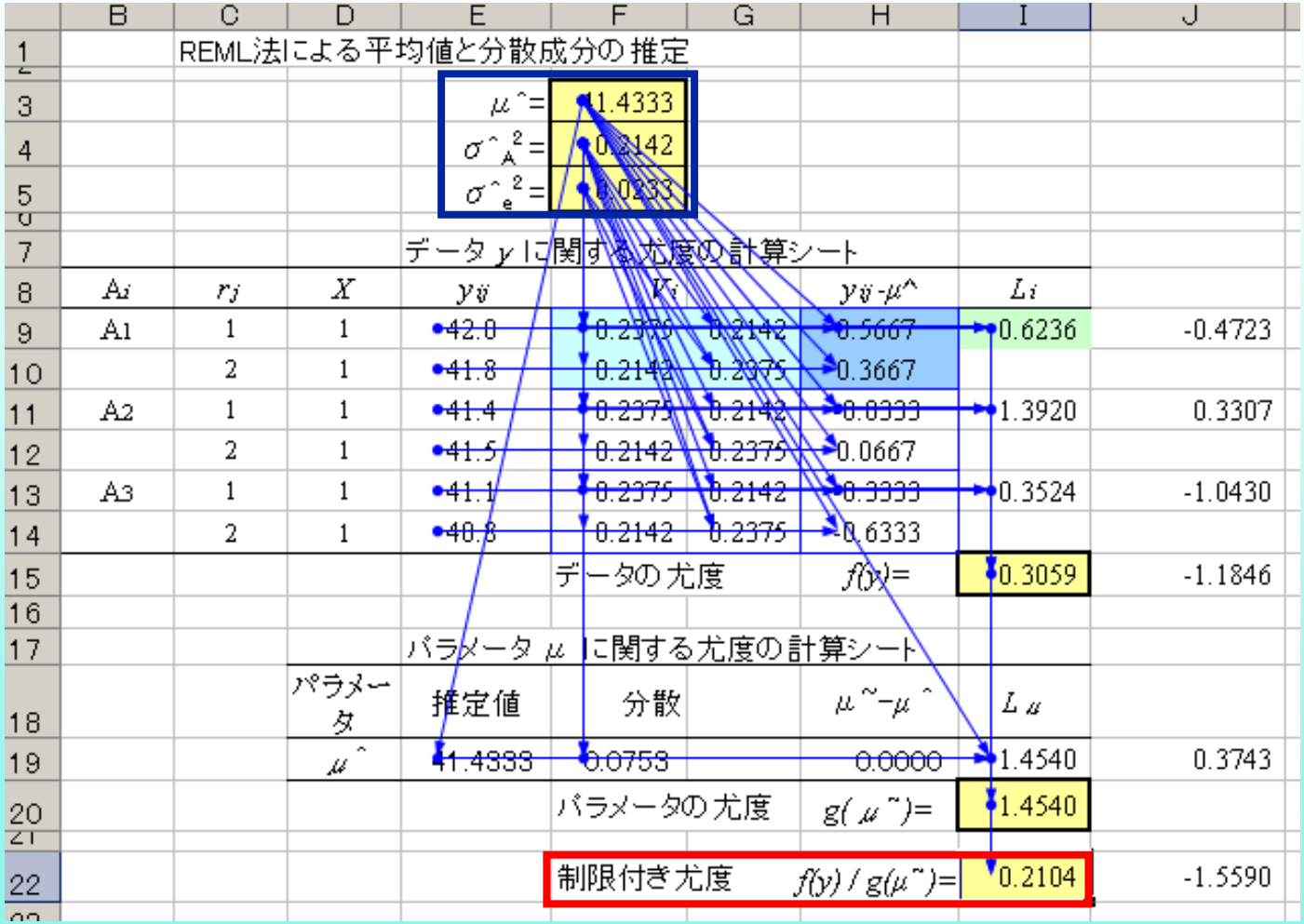
$$\tilde{\mu} = \frac{y_{11} + y_{12} + y_{21} + y_{22} + y_{31} + y_{32}}{6}$$

$$V(\tilde{\mu}) = \frac{1}{6^2}\{V(y_{11} + y_{12}) + V(y_{21} + y_{22}) + V(y_{31} + y_{32})\}$$

$$= \frac{1}{6^2}\sum_{i=1}^{3}\{(V(y_{i1}) + V(y_{i2}) + 2Cov(y_{i1}, y_{i2})\}$$

$$= \frac{3}{6^2}\{2(\hat{\sigma}_A^2 + \hat{\sigma}_e^2) + 2\hat{\sigma}_A^2\} = \frac{(2\hat{\sigma}_A^2 + \hat{\sigma}_e^2)}{6}$$

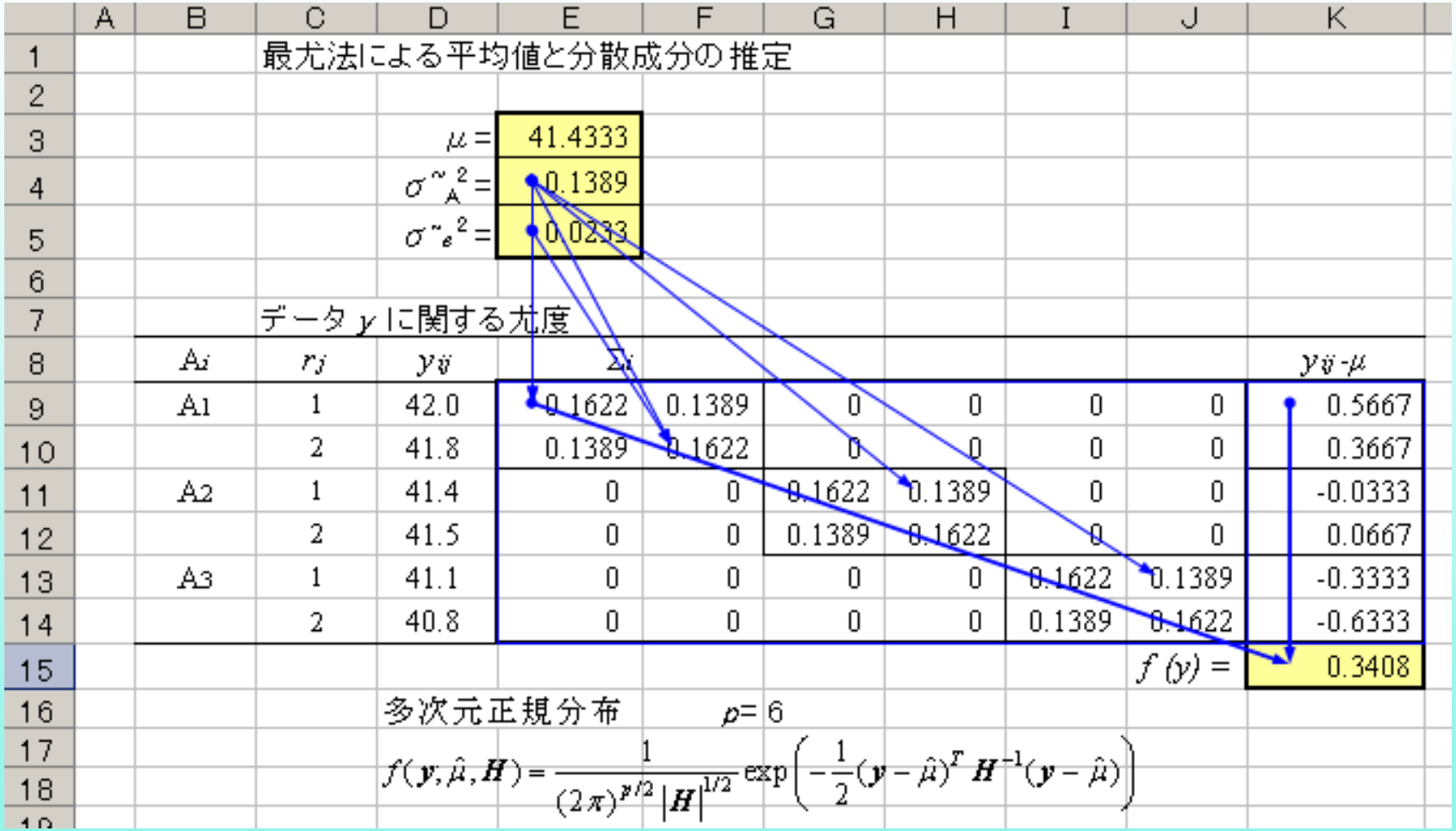
尤度の比をソルバーで最大化



	B	C	D	E	F	G	H	I	J
1		REML法による平均値と分散成分の推定							
3				$\hat{\mu}$=	41.4333				
4				$\hat{\sigma}_A^2$=	0.2142				
5				$\hat{\sigma}_e^2$=	0.2375				
7				データ y に関する尤度の計算シート					
8	A_i	r_j	X	y_{ij}		V_i	$y_{ij}-\hat{\mu}$	L_i	
9	A1	1	1	42.0	0.2375	0.2142	0.5667	0.6236	-0.4723
10		2	1	41.8	0.2142	0.2375	0.3667		
11	A2	1	1	41.4	0.2375	0.2142	0.0333	1.3920	0.3307
12		2	1	41.5	0.2142	0.2375	0.0667		
13	A3	1	1	41.1	0.2375	0.2142	-0.3333	0.3524	-1.0430
14		2	1	40.8	0.2142	0.2375	-0.6333		
15					データの尤度		$f(y)$=	0.3059	-1.1846
16									
17				パラメータ μ に関する尤度の計算シート					
18			パラメータ	推定値	分散		$\tilde{\mu}-\hat{\mu}$	L_μ	
19			$\hat{\mu}$	41.4333	0.0753		0.0000	1.4540	0.3743
20					パラメータの尤度		$g(\tilde{\mu})$=	1.4540	
22					制限付き尤度		$f(y)/g(\tilde{\mu})$=	0.2104	-1.5590

REML法により分散成分を直接推定することができる.

更なる一般化 $n \times n$ の行列計算



最尤法による平均値と分散成分の推定

$\mu =$	41.4333
$\tilde{\sigma}_A^2 =$	0.1389
$\tilde{\sigma}_e^2 =$	0.0233

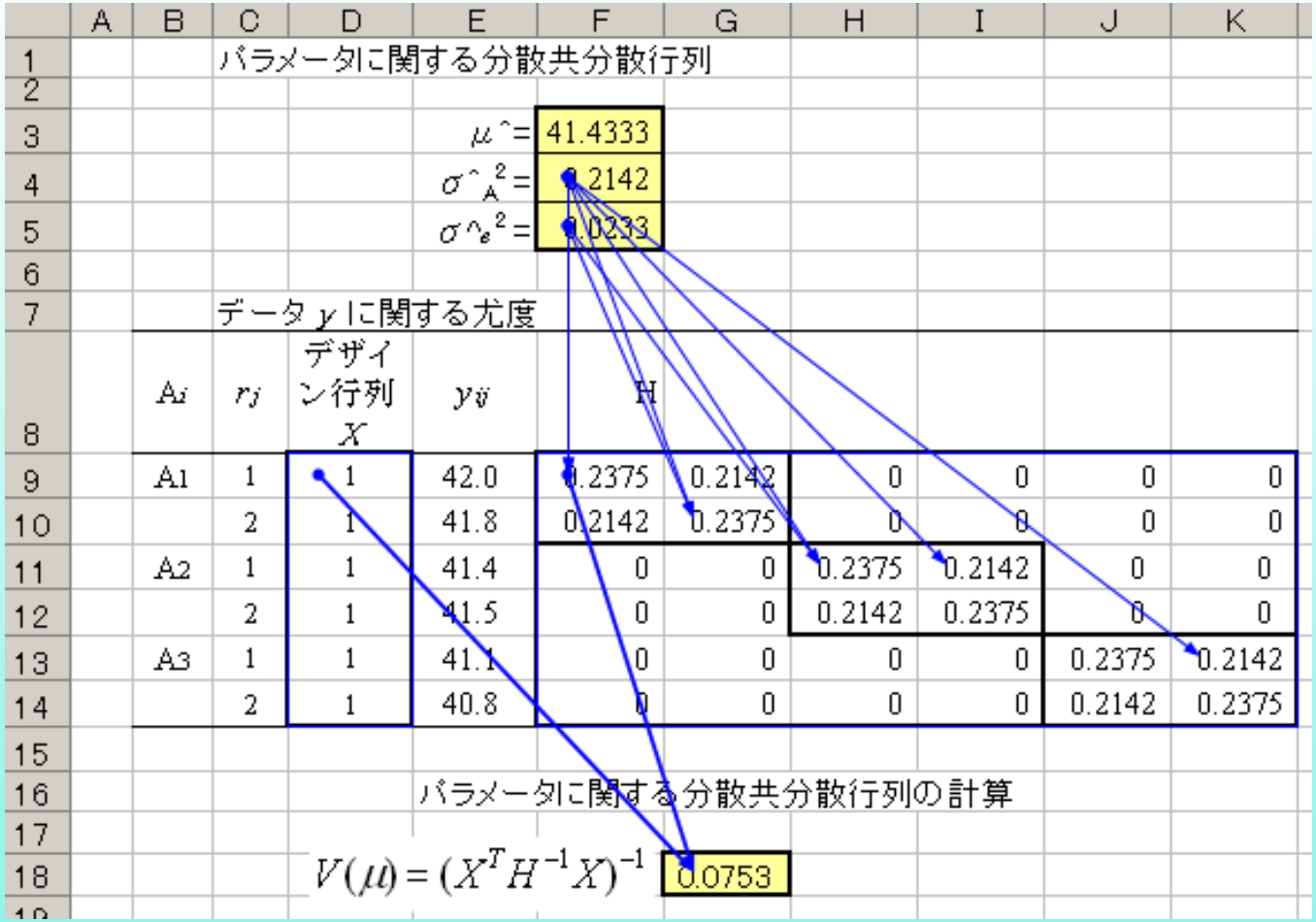
データ y に関する尤度

Ai	rj	yij	Σi						yij-μ
A1	1	42.0	0.1622	0.1389	0	0	0	0	0.5667
	2	41.8	0.1389	0.1622	0	0	0	0	0.3667
A2	1	41.4	0	0	0.1622	0.1389	0	0	-0.0333
	2	41.5	0	0	0.1389	0.1622	0	0	0.0667
A3	1	41.1	0	0	0	0	0.1622	0.1389	-0.3333
	2	40.8	0	0	0	0	0.1389	0.1622	-0.6333
								$f(y) =$	0.3408

多次元正規分布　　$p = 6$

$$f(\boldsymbol{y}, \hat{\mu}, \boldsymbol{H}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{H}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \hat{\mu})^T \boldsymbol{H}^{-1} (\boldsymbol{y} - \hat{\mu}) \right)$$

パラメータ(平均値)の分散



パラメータに関する分散共分散行列

$\hat{\mu}$ =	41.4333
$\hat{\sigma}_A^2$ =	0.2142
$\hat{\sigma}_e^2$ =	0.0233

データ y に関する尤度

A_i	r_j	デザイン行列 X	y_{ij}	H					
A1	1	1	42.0	0.2375	0.2142	0	0	0	0
	2	1	41.8	0.2142	0.2375	0	0	0	0
A2	1	1	41.4	0	0	0.2375	0.2142	0	0
	2	1	41.5	0	0	0.2142	0.2375	0	0
A3	1	1	41.1	0	0	0	0	0.2375	0.2142
	2	1	40.8	0	0	0	0	0.2142	0.2375

パラメータに関する分散共分散行列の計算

$V(\hat{\mu}) = (X^T H^{-1} X)^{-1}$ 0.0753

- $n \times n$ の行列計算を用いると平均値の分散が簡単に求めることができる.
- 固定効果が増えても同じ計算式が適用できる.
- どのような分散共分散行列に対しても同じ計算式が適用できる.

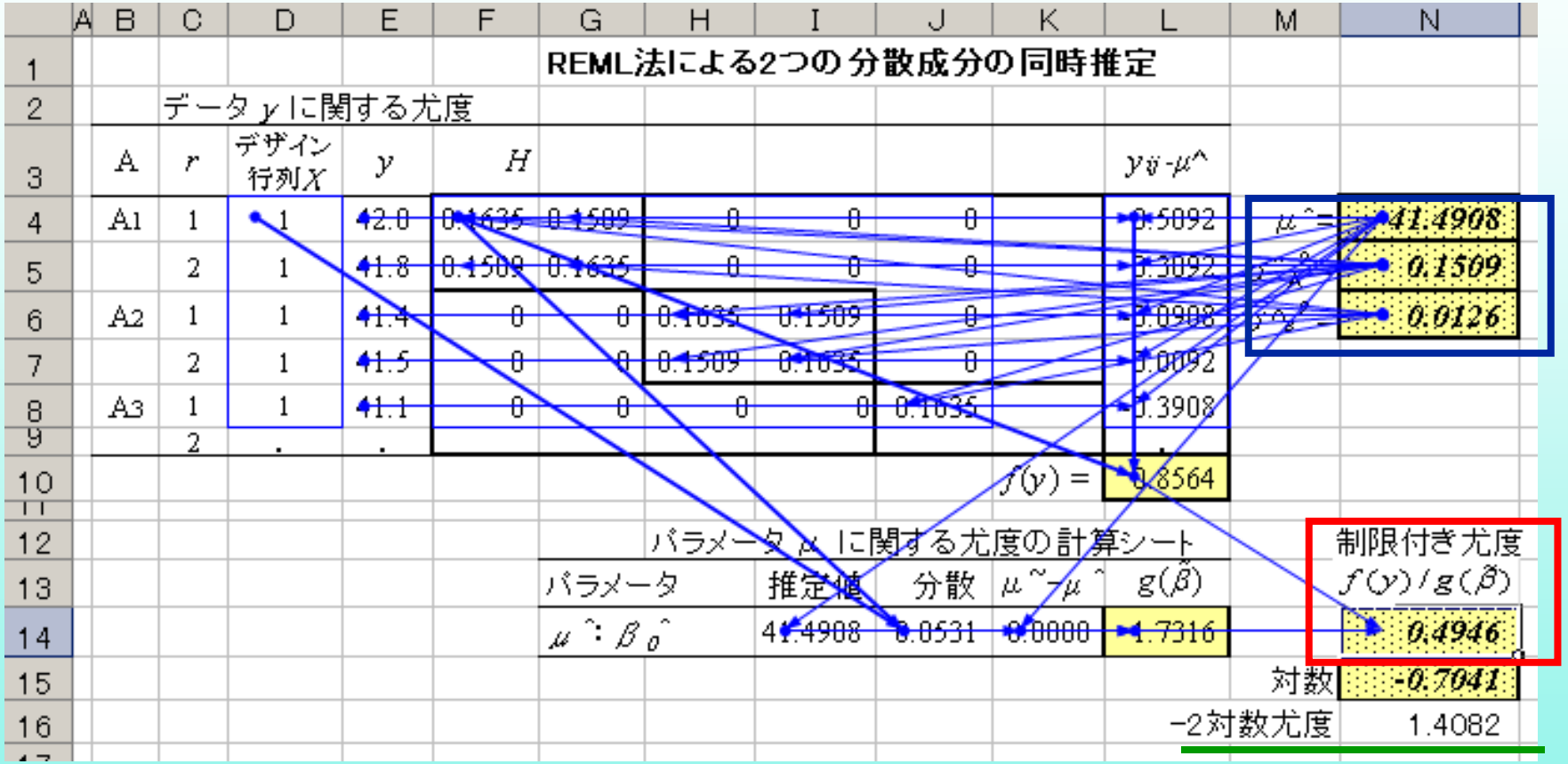
枝分かれ：１段階 繰返しが不揃い

表 5.1　繰返しが不揃いの場合の分散共分散分行列

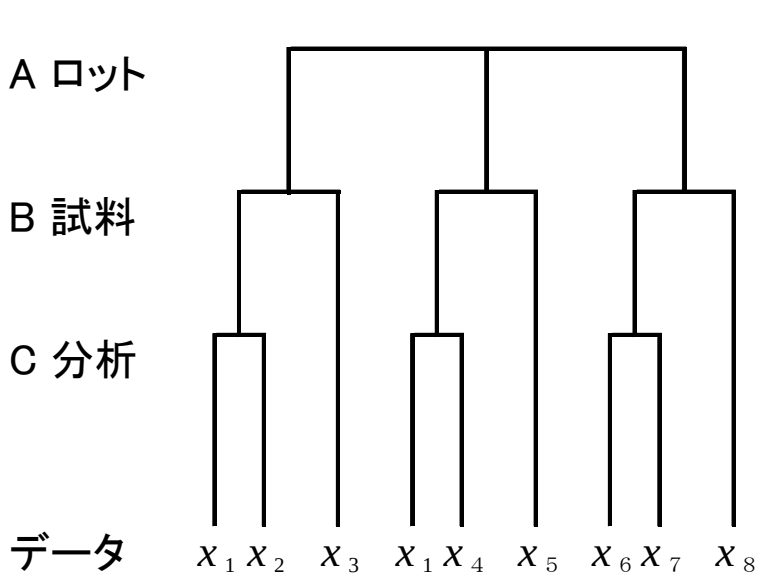
			$A1$		$A2$		$A3$
A_i	r_j	y_{ij}	$r_j=1$	$r_i=2$	$r_j=1$	$r_i=2$	$r_j=1$
$A1$	1	42.0	$\sigma_A^2+\sigma_e^2$	σ_A^2			
	2	41.8	σ_A^2	$\sigma_A^2+\sigma_e^2$			
$A2$	1	41.4			$\sigma_A^2+\sigma_e^2$	σ_A^2	
	2	41.5			σ_A^2	$\sigma_A^2+\sigma_e^2$	
$A3$	1	41.1					$\sigma_A^2+\sigma_e^2$

- 貨車A3は繰返しが１であり，分散 $\sigma_A^2+\sigma_e^2$ のみとなる．

繰返しが不揃いでも計算方法は同じ



スタッカード型２段枝分かれ実験

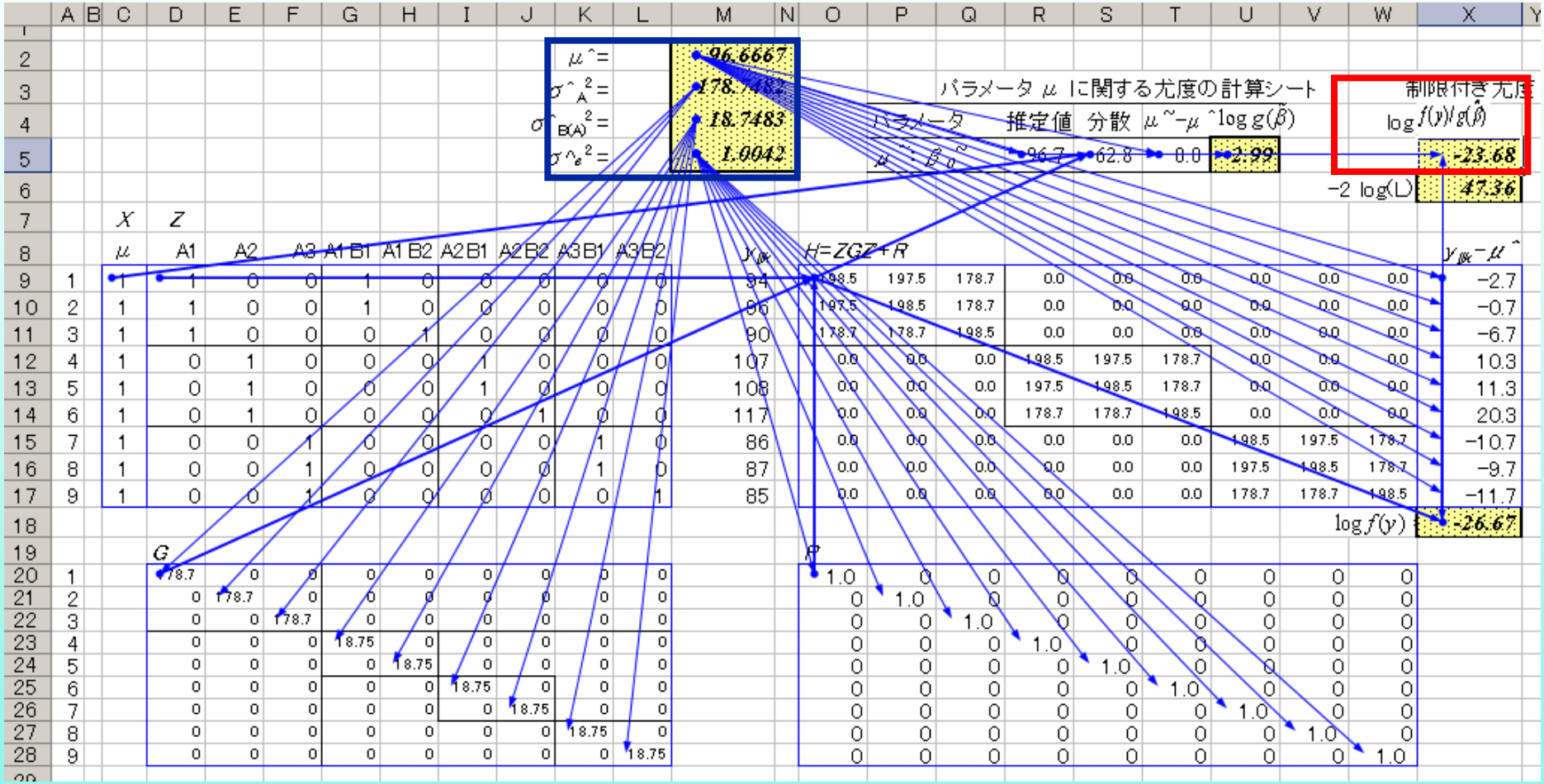


試料Bの第2水準目について分析Cの繰返し数が１である.スタッカード型の枝分かれ実験といわれている.

表 6.3　じぐざぐ型の枝分かれデータ

B	C	A1	A2	A3
B1	C1	94	107	86
	C2	96	108	87
B2	C1	90	117	85
	C2	.	.	.

スタッカード型をREML法で解析



まとめ

- SAS・JMP で簡単に解決できるデータ解析でも，それが一般的に知られていないために，古典的な手計算レベルの手法が現在でも応用分野ではメジャーな方法として使われ続けている.
- 応用分野での統計のオピニオンリーダー達に対する啓蒙活動の必要性を痛感している.
- その第一ステップとして，繰返しが不揃いな枝分かれ実験データを取り上げ，でもREML法による複数の分散成分の同時推定が，Excelでも行えることを示した，