PROC MIXED 入門

岸本　淳司
(ＳＡＳ／慶應義塾大学／東京大学)

An Introdunction to PROC MIXED

Junji Kishimoto
SAS Institute Japan / Keio Univ. SFC / Univ. of Tokyo
e-mail address: jpnjak@jpn.sas.com

概 要

PROC MIXED は、固定効果とランダム効果とを同時に持つモデルである「混合モデル」を扱うプロシジャである。その狙いは、相関をもった測定値について自然なモデリングを行うことである。従来からの線形モデルのためのプロシジャ PROC GLM と比較しながら、指定方法の入門的紹介を行う。

キーワード： PROC MIXED, 混合モデル, 変量効果, 分散成分, 反復測定

1　混合モデルとは何か？

1.1　固定効果とランダム効果

固定効果（母数効果）と**ランダム効果**（変量効果）とを同時に持つモデルを総称して混合モデルと呼ぶ。PROC MIXED は混合モデルの推定を行うプロシジャである。固定効果とランダム効果とは、次のような意味である。

固定効果：実際その実験でとられた水準のみに推論の興味があるような要因。

ランダム効果：実験で設定された水準が、無限母集団からランダムにとられたものと想定できる要因。個々の水準値よりも全体的な散らばりの情報（分散成分）に興味がある。統計的推論は、実際に設定された各水準ではなく、想定された母集団について行われる。

固定効果とランダム効果とは厳密に区別できるものではなく、解析者が推論するときの関心によって仮定されることがある。

1.2 例

ある種のガンについて、化学療法と放射線療法とを比較したいとする。特に子供について興味があった。5 つの施設が実験に協力してくれることになった。各施設ごとに大人 10 人と子供 10 人とが選択された。処置 2ヶ月後、腫瘍縮小の程度が測定された。この実験中の各要因の中で：

- 処理（化学療法か放射線療法か）は固定効果である。なぜなら、この 2 つの処理を比較したいのであって、その他の治療法一般について推論したいわけではないからである。
- 年齢（子供か大人か）も固定効果である。
- 施設は、固定効果とも考えられるし、ランダム効果であるとも考えられる。もし研究者の関心が実験に参加した 5 つの病院のみの変動にあるなら、固定効果が想定される。もし施設の母集団を想定して、5 つの病院はそこからサンプリングしたと考え、 結論を母集団について下したいのなら、ランダム効果として扱うべきである。
- 患者はランダム効果である。実験に参加した患者は、同種のガン患者の母集団からのサンプルであると想定される。

1.3 GLM モデル

PROC GLM は固定効果のみの線形モデルを当てはめる。モデルは次のように表される。

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

ここで、

$\mathbf{y}$ は、観測されたデータベクトルである。

$\mathbf{X}$ は、既知の計画行列である。

β は、未知の固定効果パラメタのベクトルである。

ε は、観測されないランダム誤差のベクトルである。

このモデルでは ε の各要素は無相関で、平均ゼロ・分散 σ^2 の正規分布をすると想定される。すなわち、

$$\begin{aligned} \mathrm{E}[\varepsilon] &= \mathbf{0} \\ var[\varepsilon] &= \sigma^2\mathbf{I} \\ &= \begin{pmatrix} \sigma^2 & & & 0 \\ & \sigma^2 & & \\ & & \ddots & \\ 0 & & & \sigma^2 \end{pmatrix} \end{aligned}$$

$\sigma^2\mathbf{I}$ とは、ε の各要素の分散が等しい（同じ記号 σ^2 で表されている）ことと、無相関（対角要素がゼロ）を表現している。y の期待値は $\mathrm{E}[y] = \mathbf{X}\beta$ だが、分散は

$$var[y] = \sigma^2\mathbf{I}$$

となり、測定値間の独立が仮定される。

PROC GLM ではランダム効果を含む検定も実行できるが、当てはめているのはすべて固定効果のモデルである。

1.4 混合 (MIXED) モデル

PROC MIXED が扱う混合モデルでは、PROC GLM が扱うモデルを次のように拡張している。

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$$

ただし、

$\mathbf{y}$ は、観測されたデータベクトルである。

$\mathbf{X}$ は、既知の計画行列である。

β は、固定効果に対する未知パラメタである。

$\mathbf{Z}$ は、既知の計画行列である。

γ は、ランダム効果についての未知パラメタである。

ε は、観測されないランダム誤差のベクトルである。

第一の拡張点は、ランダム効果を表す項 $\mathbf{Z}\gamma$ ををを導入したことである。γ は、固定効果のパラメタ β とは異なり、ある分布をすると想定される。その分布の平均はゼロで、ある分散共分散行列 $\mathbf{G}$ を持つと特定する。分散共分散行列 $\mathbf{G}$ には、構造のバリエーションを何種類か指定することができる。

$$\begin{aligned} \mathrm{E}[\gamma] &= \mathbf{0} \\ var[\gamma] &= \mathbf{G} \\ &= \begin{pmatrix} ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix} \end{aligned}$$

第二の拡張点は、誤差ベクトル ε の分散共分散行列 $\mathbf{R}$ にも構造のバリエーションを指定することができるようにしたことである。この拡張により、誤差に相関があるモデル（反復測定）の指定ができる。

$$\begin{aligned} \mathrm{E}[\varepsilon] &= \mathbf{0} \\ var[\varepsilon] &= \mathbf{R} \\ &= \begin{pmatrix} ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix} \end{aligned}$$

ただし γ と ε とは無相関であるとする。このとき $\mathbf{y}$ の分散共分散行列 $\mathbf{V}$ は次のように表される。

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\emptyset} + \mathbf{R}$$

すなわち、$\mathbf{G}$ と $\mathbf{R}$ の構造をいろいろ指定することによって、$\mathbf{y}$ の共分散行列に対するさまざまなモデリングができるようになる。固定効果 $\mathbf{X}\beta$ は、$\mathbf{y}$ の期待値をモデリングするが、混合モデルではそれに併せて誤差の非等分散性や相関についてのモデリングを可能にする。

固定効果モデルでは、パラメタ β と σ^2 だけを推定すればよかったが、混合効果モデルでは、$\gamma, \mathbf{G}, \mathbf{R}$ についても推定しなければならない。そのためには最小二乗法は適切な方法ではなく、γ と ε とは正規分布していると仮定して、尤度に基いた推定法が採用される。その 1 つは、最尤法 (ML; Maximum Likelihood) であり、もう 1 つは制約つき最尤法 (REML; REstricted Maximum Likelihood) である。PROC MIXED では、REML がデフォルトの推定法である。

1.5 PROC MIXED の基本文法

PROC MIXED では、当てはめるモデルの骨格部分を次のように指定する。

```
PROC MIXED <オプション>;
   CLASS 分類変数群;
   MODEL 従属変数 = 固定効果群 </ オプション>;
   RANDOM ランダム効果群 </ オプション>;
   REPEATED 反復効果群 </ オプション>;
RUN;
```

各ステートメントの意味を概説する。

CLASS: 分類変数を指定する。PROC GLM の CLASS ステートメントと同じである。

MODEL: 等号の左辺に従属変数を、右辺に独立変数（固定効果または共変量）のリストを指定する。従属変数は連続変数でなければならない。

RANDOM: ランダム効果 γ についての指定を行う。ランダム効果に対する計画行列 **Z** を指定し、オプションで γ の分散共分散行列 **G** の構造を指定する。実用的に用いられる構造は、TYPE=VC（分散成分；デフォルト）か TYPE=UN（無構成）のどちらかである。

REPEATED: 誤差 ε の分散共分散行列 **R** の構造を指定する。これにより、反復測定の効果が指定される。反復効果の分散構造にはさまざまなパターンが用いられる。REPEATED ステートメントで効果を指定する場合は、分類変数でなければならない。

PROC GLM の RANDOM ステートメントは、平均平方の期待値の構造を出力させるためのものであった。PROC MIXED の RANDOM ステートメントはそれとは意味が全く異なる。

PROC MIXED では、次のステートメントも有用である。

LSMEANS: 固定効果について、一般化最小二乗平均を計算する。リリース 6.10 から最小二乗平均についての多重比較の機能が追加された。

CONTRAST: 固定効果またはランダム効果についてのユーザー指定の仮説を検定する。

ESTIMATE: 固定効果またはランダム効果の線形結合を推定する。

MAKE: PROC MIXED が作る数表を SAS データセットに変換する。PROC GLM にはこのような機能はないため、たとえばパラメタ推定値や多重比較の結果をデータセットに出力することはできない。PROC MIXED に同様の仕事をさせると、MAKE ステートメントでデータセット出力ができるようになる。

2　ランダム効果を含む例

2.1　身長の例題

次に示すデータには、4 つの家族 18 人について性別と身長（インチ単位）が記録されている。

```
data heights;
  input family gender $ height @@;
  cards;
1 F 67  1 F 66  1 F 64  1 M 71  1 M 72
2 F 63  2 F 63  2 F 67  2 M 69  2 M 68  2 M 70
3 F 63  3 M 64
4 F 67  4 F 66  4 M 67  4 M 67  4 M 69
;
```

2.2　固定効果モデル

家族の効果と性別の効果とから身長を説明する固定効果モデル

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

を当てはめてみよう。

PROC GLM では、次のようになる。

```
proc glm data=heights;
  class gender family;
  model height = gender family;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GENDER	1	57.4568182	57.4568182	19.07	0.0008
FAMILY	3	33.9345960	11.3115320	3.75	0.0385

PROC MIXED でも、全く同じように指定できる。

```
proc mixed data=heights;
  class gender family;
  model height = gender family;
run;
```

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
GENDER	1	13	19.07	0.0008
FAMILY	3	13	3.75	0.0385

固定効果の検定出力の部分を比較すると、全く同じ結論であることがわかる。PROC MIXED では尤度に基づいて検定を行うため、平方和は表示されない。

2.3 混合効果モデル

固定効果モデルでは、測定値の独立が仮定されている。身長のデータでは、同一家族内の身長には相関があることが想定できる。このような同一クラスター内の相関を表現するために、家族をランダム効果として扱う方法がある。この場合、データ中の 4 家族は大きな母集団からランダムにサンプリングされたと仮定される。

混合モデル

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$$

の各項は、次のように設定される。

$\mathbf{y}$ は、18 人の身長観測値のベクトルである。

$$\mathbf{y} = \begin{pmatrix} 67 & 66 & 64 & 71 & 72 & 63 & 63 & 67 & 69 & 68 & 70 & 63 & 64 & 67 & 66 & 67 & 67 & 69 \end{pmatrix}'$$

$\mathbf{X}$ は、固定効果である性別に対応する計画行列である。切片項も含まれる。

$\mathbf{Z}$ は、ランダム効果である家族に対応する計画行列である。切片項は含まれない。

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

β は、切片と性別とを表す固定パラメタである。

$$\beta = \begin{pmatrix} \mu & \beta_1 & \beta_2 \end{pmatrix}'$$

測定値の期待値は、固定効果で規定される。

$$\mathrm{E}(Y_{ijk}) = \mu + \beta_i$$

γ は家族を表し、$\gamma_j = \mathrm{N}(0, \sigma_\gamma^2)$ の分布をするランダム効果パラメタである。

$$\gamma = \begin{pmatrix} \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \end{pmatrix}'$$

ε は、ランダム誤差を表す。

$$\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_{18} \end{pmatrix}'$$

変量ベクトル γ は多変量正規分布をし、その平均は 0 で、分散は次のようになる。

$$\mathbf{G} = \sigma_\gamma^2 \mathbf{I}_4 = \begin{pmatrix} \sigma_\gamma^2 & 0 & 0 & 0 \\ 0 & \sigma_\gamma^2 & 0 & 0 \\ 0 & 0 & \sigma_\gamma^2 & 0 \\ 0 & 0 & 0 & \sigma_\gamma^2 \end{pmatrix}$$

誤差ベクトル ε も多変量正規分布し、その分散構造は固定効果モデルとかわらない。

$$\mathbf{R} = \sigma_\varepsilon^2 \mathbf{I}_{18} = \begin{pmatrix} \sigma_\varepsilon^2 & & & 0 \\ & \sigma_\varepsilon^2 & & \\ & & \ddots & \\ 0 & & & \sigma_\varepsilon^2 \end{pmatrix}$$

測定値の分散は、

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\emptyset} + \mathbf{R}$$

であるが、対角成分に注目すれば、測定値の分散成分はランダム効果の分散と誤差分散の和で表される。

$$var(Y) = \sigma_\gamma^2 + \sigma_\varepsilon^2$$

2.4 PROC MIXED の指定と出力

PROC MIXED では、MODEL ステートメントには固定効果のみを指定し、ランダム効果は RANDOM ステートメントで指定する。

```
proc mixed data=heights;
  class gender family;
  model height = gender;
  random family;
run;
```

REML 推定法による分散パラメタの推定値は、次のように出力される。

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
FAMILY	0.81863290	2.53458987	3.12568569	0.81	0.4174
Residual	1.00000000	3.09612512	1.25232525	2.47	0.0134

“Estimate” の列には、家族による分散成分の推定値と誤差分散の推定値とが示されている。各成分はそれぞれ次のように推定される。

$$\widehat{\sigma}_\gamma^2 = 2.53458987$$
$$\widehat{\sigma}_\varepsilon^2 = 3.09612512$$

“Ratio” の列には、各成分の誤差分散 σ_ε^2 との比が表示されている。“Std Error” の列には、2 次微分行列の逆行列によって求められた近似標準誤差が示されている。分散成分推定値を対応する近似標準誤差で割ることにより “Z” 統計量が得られ、標準正規分布から計算した両側 p 値が “Pr $>$ |Z|” の列に表示される。この検定は標本数が少ないときは正確ではない。

ランダム効果 γ の共分散行列 $\mathbf{G}$ は、

$$\widehat{\mathbf{G}} = \widehat{\sigma}_{\gamma}^2 \mathbf{I}_4 = 2.534 \mathbf{I}_4 = \begin{pmatrix} 2.534 & 0 & 0 & 0 \\ 0 & 2.534 & 0 & 0 \\ 0 & 0 & 2.534 & 0 \\ 0 & 0 & 0 & 2.534 \end{pmatrix}$$

誤差ベクトル ε の共分散行列 $\mathbf{R}$ は、

$$\widehat{\mathbf{R}} = \widehat{\sigma}_{\varepsilon}^2 \mathbf{I}_{18} = 3.096 \mathbf{I}_{18} = \begin{pmatrix} 3.096 & & & 0 \\ & 3.096 & & \\ & & \ddots & \\ 0 & & & 3.096 \end{pmatrix}$$

と推定される。

測定値 $\mathbf{y}$ の分散共分散行列 $\mathbf{V}$ は、

$$\widehat{\mathbf{V}} = \mathbf{Z}\widehat{\mathbf{G}}\mathbf{Z}^{\prime} + \widehat{\mathbf{R}} =$$

$$\begin{pmatrix}
5.63 & 2.53 & 2.53 & 2.53 & 2.53 & & & & & & & & & & & & & & & & & \\
2.53 & 5.63 & 2.53 & 2.53 & 2.53 & & & & & & & & & & & & & & & & & \\
2.53 & 2.53 & 5.63 & 2.53 & 2.53 & & & & & & & & & & & & & & & & & \\
2.53 & 2.53 & 2.53 & 5.63 & 2.53 & & & & & & & & & & & & & & & & & \\
2.53 & 2.53 & 2.53 & 2.53 & 5.63 & & & & & & & & & & & & & & & & & \\
 & & & & & 5.63 & 2.53 & 2.53 & 2.53 & 2.53 & 2.53 & & & & & & & & & & & \\
 & & & & & 2.53 & 5.63 & 2.53 & 2.53 & 2.53 & 2.53 & & & & & & & & & & & \\
 & & & & & 2.53 & 2.53 & 5.63 & 2.53 & 2.53 & 2.53 & & & & & & & & & & & \\
 & & & & & 2.53 & 2.53 & 2.53 & 5.63 & 2.53 & 2.53 & & & & & & & & & & & \\
 & & & & & 2.53 & 2.53 & 2.53 & 2.53 & 5.63 & 2.53 & & & & & & & & & & & \\
 & & & & & 2.53 & 2.53 & 2.53 & 2.53 & 2.53 & 5.63 & & & & & & & & & & & \\
 & & & & & & & & & & & 5.63 & 2.53 & & & & & & & & & \\
 & & & & & & & & & & & 2.53 & 5.63 & & & & & & & & & \\
 & & & & & & & & & & & & & 5.63 & 2.53 & 2.53 & 2.53 & 2.53 & & & & \\
 & & & & & & & & & & & & & 2.53 & 5.63 & 2.53 & 2.53 & 2.53 & & & & \\
 & & & & & & & & & & & & & 2.53 & 2.53 & 5.63 & 2.53 & 2.53 & & & & \\
 & & & & & & & & & & & & & 2.53 & 2.53 & 2.53 & 5.63 & 2.53 & & & & \\
 & & & & & & & & & & & & & 2.53 & 2.53 & 2.53 & 2.53 & 5.63 & & & &
\end{pmatrix}$$

と推定される。

すなわち、測定値 Y の分散成分は、

$$\begin{aligned} \widehat{var}(Y) &= \widehat{\sigma}_{\gamma}^2 + \widehat{\sigma}_{\varepsilon}^2 \\ 5.63 &= 2.53 + 3.10 \end{aligned}$$

と分解される。

また、同一家族内については共分散 2.53 の相関があるが、家族が異なると相関はゼロというモデルになっている。

固定効果である性別についての検定は、次のように出力される。

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
GENDER	1	13	18.29	0.0009

固定効果モデルによる F 値 = 19.07 とは僅かに異なっている。

3　反復測定の解析

3.1　降圧剤の比較のデータ

次に示すのは、2 種類の降圧剤を投与したときの血圧変化パターンを比較した仮想データである。全部で 8 人の被験者が、A, B 2 種類の降圧剤にランダムに割り当てられた。また 1 人の被験者について投与前・1 時間後・2 時間後・3 時間後の 4 時点で血圧が測定された。

血圧の経時的変化

降圧剤	被験者	投与前	1h	2h	3h
A	1	119	113	106	116
	2	120	115	110	112
	3	123	126	116	115
	4	121	118	123	124
B	5	125	118	107	105
	6	133	120	114	107
	7	125	110	107	103
	8	126	118	115	110

次の 2 種類の効果を区別する。

被験者間要因 (Between-Subject effects):一人の被験者に対しては同じ値をとる要因。今の例では、2 種類の降圧剤 (DRUG)。

被験者内要因 (Within-Subject effects):同一被験者に対して複数の水準をとる要因。今の例では、4 時点の時間 (TIME)。

1 変量アプローチ用 (forUNI) と多変量アプローチ用 (forMUL) の 2 種類のデータセットを作成する。

```
data forUNI(keep=drug subj time bp)
     forMUL(keep=drug subj t1-t4);
  array bpt{*} t1-t4;
  input drug $ subj @;
  do time=1 to 4;
     input bp @;
     bpt{time} = bp;
     output forUNI;
  end;
  output forMUL;
  cards;
A 1 119 113 106 116
A 2 120 115 110 112
A 3 123 126 116 115
A 4 121 118 123 124
B 5 125 118 107 105
B 6 133 120 114 107
B 7 125 110 107 103
B 8 126 118 115 110
;
```

1 変量アプローチ用データセット (forUNI)

OBS	DRUG	SUBJ	TIME	BP
1	A	1	1	119
2	A	1	2	113
3	A	1	3	106
4	A	1	4	116
5	A	2	1	120
6	A	2	2	115
7	A	2	3	110
8	A	2	4	112
9	A	3	1	123
10	A	3	2	126
11	A	3	3	116
12	A	3	4	115
13	A	4	1	121
14	A	4	2	118
15	A	4	3	123
16	A	4	4	124
17	B	5	1	125
18	B	5	2	118
19	B	5	3	107
20	B	5	4	105
21	B	6	1	133
22	B	6	2	120
23	B	6	3	114
24	B	6	4	107
25	B	7	1	125
26	B	7	2	110
27	B	7	3	107
28	B	7	4	103
29	B	8	1	126
30	B	8	2	118
31	B	8	3	115
32	B	8	4	110

多変量アプローチ用データセット (forMUL)

OBS	T1	T2	T3	T4	DRUG	SUBJ
1	119	113	106	116	A	1
2	120	115	110	112	A	2
3	123	126	116	115	A	3
4	121	118	123	124	A	4
5	125	118	107	105	B	5
6	133	120	114	107	B	6
7	125	110	107	103	B	7
8	126	118	115	110	B	8

3.2 分割実験モデルによる１変量分散分析

経時的データの分析は、古典的には、被験者を 1 次単位、ある時点での被験者を 2 次単位とした分割実験型分散分析で解析される。分割実験型の解析を PROC GLM で行うには、1 変量アプローチ用に整形されたデータセット (forUNI) を用いる。また、被験者間要因 (DRUG) の検定には、TEST ステートメントで誤差項を正しく指定しなければならない。

分割実験型の 1 変量分散分析は、次のプログラムで実施される。

```
/* Univariate ANOVA Using a Split-Plot Model */
proc glm data=forUNI;
  class drug subj time;
  model bp = drug subj(drug) time time*drug;
  test h=drug e=subj(drug);
run;
```

```
 Source                   DF      Type I SS   Mean Square   F Value     Pr > F

 DRUG                      1      36.125000     36.125000      3.07     0.0966
 SUBJ(DRUG)                6     325.875000     54.312500      4.62     0.0052
 TIME                      3     797.000000    265.666667     22.60     0.0001
 DRUG*TIME                 3     291.375000     97.125000      8.26     0.0011

 Source                   DF    Type III SS   Mean Square   F Value     Pr > F

 DRUG                      1      36.125000     36.125000      3.07     0.0966
 SUBJ(DRUG)                6     325.875000     54.312500      4.62     0.0052
 TIME                      3     797.000000    265.666667     22.60     0.0001
 DRUG*TIME                 3     291.375000     97.125000      8.26     0.0011

 Tests of Hypotheses using the Type III MS for SUBJ(DRUG) as an error term

 Source                   DF    Type III SS   Mean Square   F Value     Pr > F

 DRUG                      1     36.1250000    36.1250000      0.67     0.4459
```

分割実験型 1 変量アプローチでは、被験者内要因とそれにかかわる効果（ここでは TIME と TIME*DRUG の交互作用）の検定には、時点間での誤差の相関が一定であるという仮定が必要であり、それを満たさない場合はリベラルな（p 値が過小な）検定になる。この仮定を検証する方法は次に紹介する。

3.3 REPEATED ステートメントを使った反復測定の解析

PROC GLM では、REPEATED ステートメントを使ってで反復測定の解析を実行することもできる。これにより多変量型のアプローチが可能になる。多変量アプローチとは、今の例でいうと 1 人の被験者につき血圧を 4 回測定しているので、その 4 回の測定値をそれぞれ別の従属変数とみなす解析法である。4 回の測定誤差の間には自然な相関関係が想定されていて、特別な相関構造は仮定しない。

REPEATED ステートメントを使った PROC GLM のプログラムは次のようになる。

```
/* Multivariate Repeated Measures Analysis */
proc glm data=forMUL;
  class drug;
  model t1-t4 = drug / nouni;
```

```
  repeated time / printe;
run;
```

多変量アプローチの指定には、次のような特徴がある。

- 入力データには、多変量アプローチ用に整形されたもの (forMUL) を用いる。
- MODEL ステートメントの左辺に複数回の測定値を指定する。
- 個々の測定値についての検定に興味がないときには、MODEL ステートメントに NOUNI オプションを指定する。
- REPEATED ステートメントで反復測定であることを指定し、また複数回の測定 (T1–T4) をまとめた被験者内要因の名称を指定する。
- 個々の応答 T1–T7 間の偏相関係数を計算し、また 1 変量アプローチが有効であるための球面性の仮説を検定するため、PRINTE オプションを指定する。

PROC GLM で多変量アプローチを扱うときには、次の事項に注意する。

- ある被験者の複数回の測定について 1 個でも欠測があった場合、その被験者についてのデータ全体が解析から削除されてしまう。
- 被験者内反復測定効果 (TIME) は、分類変数として扱われる。
- 被験者間要因と被験者内要因との間の交互作用は、自動的にモデルに追加される。

PRINTE オプションの効果によって、4 つの応答間の偏相関係数と球面性 (Sphericity) の検定の結果が表示される。

```
                    General Linear Models Procedure
                Repeated Measures Analysis of Variance

 Partial Correlation Coefficients from the Error SS&CP Matrix / Prob > |r|

            DF = 6            T1        T2        T3        T4

            T1          1.000000  0.638306  0.496055  0.166382
                          0.0001    0.1229    0.2575    0.7214

            T2          0.638306  1.000000  0.560376  0.244757
                          0.1229    0.0001    0.1907    0.5968

            T3          0.496055  0.560376  1.000000  0.798081
                          0.2575    0.1907    0.0001    0.0315

            T4          0.166382  0.244757  0.798081  1.000000
                          0.7214    0.5968    0.0315    0.0001
```

測定時間の間隔が大きくなるほど相関係数が小さくなっているので、TIME の主効果と TIME*DRUG の交互作用についての 1 変量分散分析の F 検定は有効でないことが示唆される。

```
          Test for Sphericity: Mauchly's Criterion = 0.302765
Chisquare Approximation = 5.6421038 with 5 df   Prob > Chisquare = 0.3426
```

球面性 (Sphericity) の検定とは、被験者内分散共分散行列がタイプ H 共分散構造（Huynh - Feldt 条件とも呼ばれる）をしていることを調べるものである。この条件は、どのような直交対比を選んでも独立で等分散であるということと等価である。この検定が有意であれば、1 変量アプローチによる TIME と TIME*DRUG の効果の検定は適切でないことが示唆される。今の例では、有意にはなっていない。

3.4 多変量検定

多変量検定として、次の 4 つが有名である。

- Wilks のラムダ
- Pillai のトレース
- Hotelling-Lawley のトレース
- Roy の最大根

TIME の効果について、これらの多変量検定の結果が次のように出力される。

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no TIME Effect
H = Type III SS&CP Matrix for TIME E = Error SS&CP Matrix

S=1 M=0.5 N=1

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.07735391	15.9035	3	4	0.0109
Pillai's Trace	0.92264609	15.9035	3	4	0.0109
Hotelling-Lawley Trace	11.92759538	15.9035	3	4	0.0109
Roy's Greatest Root	11.92759538	15.9035	3	4	0.0109

TIME*DRUG の効果についても同様に多変量検定が実施される。

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no TIME*DRUG Effect
H = Type III SS&CP Matrix for TIME*DRUG E = Error SS&CP Matrix

S=1 M=0.5 N=1

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.13341306	8.6607	3	4	0.0318
Pillai's Trace	0.86658694	8.6607	3	4	0.0318
Hotelling-Lawley Trace	6.49551834	8.6607	3	4	0.0318
Roy's Greatest Root	6.49551834	8.6607	3	4	0.0318

これらの多変量検定には次の特徴がある：

- Huynh-Feldt 条件を必要としない。
- 被験者内分散共分散行列に構造制約はない（無構成である）。
- 分散共分散行列はモーメント法により推定される。
- 4 つの各検定法で p 値が異なることがある。通常、Wilks のラムダが最も広く用いられている。
- Huynh-Feldt 条件が成立しているときの被験者内効果の検定については、1 変量検定の方が検出力が高い。

3.5 自由度を調整した１変量検定

REPEATED ステートメントを使った指定形式でも、分割実験型の 1 変量検定の結果が報告される。被験者間要因である DRUG については、次のように出力される。

```
                    General Linear Models Procedure
                Repeated Measures Analysis of Variance
           Tests of Hypotheses for Between Subjects Effects

Source                   DF    Type III SS   Mean Square   F Value     Pr > F

DRUG                      1        36.1250       36.1250      0.67     0.4459

Error                     6       325.8750       54.3125
```

TEST ステートメントで誤差項を指定しなくても正しい検定が実施される。被験者間要因の検定には、球面性の仮定は必要ない。

いままでに説明したように、分割実験型の 1 変量検定は H–F 条件が満たされないときリベラルな（p 値が過小な）検定にる。一方多変量検定は保守的な（p 値が過大になる）検定である。1 変量検定と多変量検定との妥協的方法として、1 変量 F 検定の自由度を調整する方法がある。TIME と TIME*DRUG についての 1 変量 F 統計量は、球面性の仮定が成立しないときであっても近似的に F 分布する。そのような F 値を得るため、F 統計量の分母子の自由度を小さめに調整する方法が 2 種類ある。1 つは Greenhouse–Geisser(G–G) の調整であり、もう 1 つが Huynh–Feldt(H–F) の調整である。

PROC GLM で REPEATED ステートメントを使って反復測定の解析を行うと、被験者内効果の TIME と TIME*DRUG とについて、この 2 種類の自由度を調整した 1 変量 F 検定も出力される。

```
                    General Linear Models Procedure
                Repeated Measures Analysis of Variance
         Univariate Tests of Hypotheses for Within Subject Effects

Source: TIME

                                                                 Adj  Pr > F
     DF     Type III SS     Mean Square   F Value   Pr > F    G - G    H - F
      3    797.00000000    265.66666667     22.60   0.0001   0.0001   0.0001

Source: TIME*DRUG

                                                                 Adj  Pr > F
     DF     Type III SS     Mean Square   F Value   Pr > F    G - G    H - F
      3    291.37500000     97.12500000      8.26   0.0011   0.0037   0.0011

Source: Error(TIME)

     DF     Type III SS     Mean Square
     18    211.62500000     11.75694444

                  Greenhouse-Geisser Epsilon = 0.7496
                         Huynh-Feldt Epsilon = 1.4209
```

このデータでは H–F 修正のための ε が 1 を超えているので、実際には H–F 修正は実施されない。

3.6 PROC MIXED による反復測定の指定

PROC MIXED を用いると、反復測定の解析がより柔軟に行える。特に、

- 欠測値があっても解析が行える。
- 被験者内分散共分散行列について、さまざまな構造を指定することができる。

という点に特徴がある。

PROC MIXED による反復測定の解析は、次のように行う。

- 入力データは、1 変量アプローチ型で用意する。
- すべての固定効果（被験者間効果も被験者内効果も含む）を MODEL ステートメント中に指定する。
- 被験者内分散共分散行列を REPEATED ステートメントで指定する。

被験者内分散共分散行列に複合対称型 (Compound Symmetry) 構造を指定した例を次に示す。

```
/* PROC MIXED */
proc mixed data=forUNI;
  class drug subj time;
  model bp = drug time drug*time;
  repeated time / type=cs subject=subj R;
run;
```

REPEATED ステートメントには、反復変数を指定している。PROC GLM の REPEATED ステートメントとは異なり、ここで反復変数を指定しないこともできる。指定する場合には、分類変数（CLASS ステートメントで指定した変数）でなければならない。

なぜ反復効果を指定するかというと、欠測オブザベーションに対処するためである。たとえば被験者 1 と被験者 2 について、次のように欠測パターンがあったとしよう。

欠測パターン

被験者	投与前	1h	2h	3h
1	119	113	.	116
2	120	.	110	112

1 変量アプローチ用のデータセットでは、上のデータは次のように表現されている。

OBS	SUBJ	TIME	BP
1	1	1	119
2	1	2	113
3	1	4	116
4	2	1	120
5	2	3	110
6	2	4	112

同一被験者の反応を上から順に対応させると、被験者 1 の 1 時間後の観測値と被験者 2 の 2 時間後の観測値とが対応してしまう。そこで、反復変数の TIME によって正しい対応を保持するのである。もし反復変数が連続変数であればこのような対応はとれない。そのため REPEATED ステートメントで反復変数を指定する場合は分類変数でなければならないのである。

以下の場合には、REPEATED ステートメントに反復変数を指定しなくてもよい。

- データがバランスしている場合
- 欠測オブザベーションに対して、オブザベーションは削除しないで目的変数を欠測値として表現した場合
- 欠測オブザベーションが、各被験者についての記録の最後の方のみに現れる場合

REPEATED ステートメントでは、オプションで誤差ベクトル ε の共分散行列 $var(\varepsilon) = \mathbf{R}$ の構造を指定する。TYPE=CS の指定により、各被験者の被験者内分散共分散行列は次のような複合対称型 (Compound Symmetry) に制約される。

$$\mathbf{R}_i = \begin{pmatrix} (\sigma^2+\sigma_1^2) & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & (\sigma^2+\sigma_1^2) & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & (\sigma^2+\sigma_1^2) & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & (\sigma^2+\sigma_1^2) \end{pmatrix}$$

R オプションの指定により、最初の被験者に対応する分散共分散行列 $\mathbf{R}_1$ が表示される。
また、SUBJ= オプションの指定により被験者が区別される。この効果により全体の誤差の分散共分散行列 $\mathbf{R}$ は、4x4 の個人ごとの分散共分散行列を 8 人分ブロック対角形式に並べたものになる。

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & & & 0 \\ & \mathbf{R}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{R}_8 \end{pmatrix}$$

3.7 PROC MIXED による反復測定の出力

PROC MIXED で求めた血圧データ反復測定の結果は以下のようになる（一部）。

R Matrix for SUBJ 1

Row	COL1	COL2	COL3	COL4
1	22.39583333	10.63888889	10.63888889	10.63888889
2	10.63888889	22.39583333	10.63888889	10.63888889
3	10.63888889	10.63888889	22.39583333	10.63888889
4	10.63888889	10.63888889	10.63888889	22.39583333

R オプションの効果により、最初の被験者に対応する誤差の分散共分散行列 $\mathbf{R}_1$ が表示される。

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
TIME CS	0.90490254	10.63888889	7.90032028	1.35	0.1781
Residual	1.00000000	11.75694444	3.91898148	3.00	0.0027

続いて分散成分のパラメタが表示される。この例では、

$\widehat{\sigma}_1^2 = 10.63888889$
$\widehat{\sigma}^2 = 11.75694444$

であるから、行列 $\mathbf{R}_i$ の対角要素は、

$\widehat{\sigma}_1^2 + \widehat{\sigma}^2 = 10.639 + 11.757 = 22.396$

と推定される。
各分散パラメタの検定は Wald 検定であるから、小標本のときは信頼できない。

```
                Model Fitting Information for BP

Description                          Value

Observations                       32.0000
Variance Estimate                  11.7569
Standard Deviation Estimate         3.4288
REML Log Likelihood               -73.7640
Akaike's Information Criterion    -75.7640
Schwarz's Bayesian Criterion      -76.9420
-2 REML Log Likelihood            147.5279
Null Model LRT Chi-Square           6.2845
Null Model LRT DF                   1.0000
Null Model LRT P-Value              0.0122
```

次に、モデルの当てはめ全体の情報が表示される。AIC 基準を使えば、さまざまな分散構造の当てはまりを比較することができる。また、対数尤度がレポートされているので、2 つの分散構造が階層型であれば尤度比検定を実行することもできる。たとえば、Huynh–Feldt 構造の検定を行うには、無構成 (TYPE=UN) と HF 構造 (TYPE=HF) とで別々に PROC MIXED を実行し、求められた 2 つの尤度を比較する。

最後の 3 行は、現在仮定されている分散構造と PROC GLM で仮定されている分散構造 $\mathbf{R} = \sigma^2\mathbf{I}$ とを比較した尤度比検定の結果である。今の例では、単純な構造 ($\sigma^2\mathbf{I}$) よりも複合対称型構造の方がよいことが示唆される (p=0.0122)。

```
                  Tests of Fixed Effects

Source       NDF   DDF  Type III F  Pr > F

DRUG           1     6        0.67  0.4459
TIME           3    18       22.60  0.0001
DRUG*TIME      3    18        8.26  0.0011
```

最後に、固定効果の検定が表示される。ここではデータがバランスしていて、かつ複合対称型の分散構造を指定しているので、PROC GLM による反復測定の解析結果と一致している。

3.8 成長曲線モデル

被験者内要因である TIME を連続変数として扱いたいときもある。このようなモデルを成長曲線モデルという。次に示したのは、時間による血圧の変化は直線を仮定しているが降圧剤によって傾きが異なるというモデルで、さらに個人内分散共分散行列には 1 階の自己相関構造 (AR1) を仮定したときの PROC MIXED の例である。

```
/* Growth Curve Model */
proc mixed data=forUNI;
  class drug subj;
  model bp = drug time drug*time;
  repeated / type=ar(1) subject=subj R;
run;
```

反復効果の変数 (TIME) は CLASS ステートメントに指定されていないので、連続変数として扱われる。この場合、REPEATED ステートメントには反復効果の変数を指定してはいけない。

上のコードによる出力の一部を示す。

R Matrix for SUBJ 1

Row	COL1	COL2	COL3	COL4
1	23.03467816	11.73405952	5.97742898	3.04495279
2	11.73405952	23.03467816	11.73405952	5.97742898
3	5.97742898	11.73405952	23.03467816	11.73405952
4	3.04495279	5.97742898	11.73405952	23.03467816

R オプションの効果により、最初の被験者の誤差分散構造が出力される。今回は AR1 の構造を仮定したので、時期の差が大きくなるほど相関が小さくなる様子がモデリングされている。

共分散パラメタの推定結果は、次のように表示される。

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
DIAG AR(1)	0.02211485	0.50940844	0.15985723	3.19	0.0014
Residual	1.00000000	23.03467816	7.60652246	3.03	0.0025

AR(1) 構造では、誤差の構造は次の形式を仮定する。

$u_{\mathrm{t}} = \rho u_{\mathrm{t}-1} + e_{\mathrm{t}}$

ゆえに、

$$\mathbf{R} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

となる。ここでは

$\widehat{\rho} = 0.50940844$

$\widehat{\sigma}^2 = 23.03457816$

と推定された。

固定効果については、次のように出力される。

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
DRUG	1	6	7.04	0.0379
TIME	1	22	31.79	0.0001
TIME*DRUG	1	22	13.86	0.0012

固定効果の検定が有効であるためには、指定した誤差構造が適切である必要がある。AR(1) の誤差構造が正しいとすると、降圧剤による傾きの差は p=0.0012 で有意と判定される。

固定効果はタイプ III 仮説で表されるが、時間に関して多項式をあてはめた場合などでタイプ I 仮説を表示させたい場合には “htype=1” というオプションが用意されている。

4　おわりに代えて（文献紹介）

混合モデルは発展途上の方法論であり、また数学的にも難解であるため、良い入門的教科書は未だに存在しない。一つ言えるのは、混合モデルはコンピュータによる計算を前提として実用化された方法論であるので、古い教科書は役に立たないということである。

この夏、Raman C. Littell, George A. Milliken, Walter W. Stroup, and Russell D. Wolfinger らの共著によりブレークスルーとなる本が出版される予定である。内容は次のようなものになると予告されている。

Discover the latest capabilities available for mixed-model applications featuring the MIXED procedure in SAS/STAT software. This book is a comprehensive, practical guide to the use of mixed models for data analysis. It attempts to integrate the theory underlying the models, the specific forms of the models, for various appropriate SAS code with interpretation of results.

Experienced users and statisticains new to mixed models will find discussions of specific applications, including simple random effects-only models, simple mixed models with a single fixed and random effect, split-plot models, multilocation models, repeated measures, analysis of covariance, random coefficients, and spatial correlation. The book also includes extension to generalized linear mixed models. In addition to the MIXED procedure, the GLIMMIX and NLINMIX macros are discussed as well as some analysis using the GLM, GENMOD, or NLIN procedures for comparison.

混合モデルの歴史については、次の本がおもしろい。

Searle,S.R., Casella,G., and McCulloch,C.E.(1992) Variance Components, New York: John Wiley and Sons, Inc.

本稿の執筆に関しては、次の文献を直接参考にした。

SAS Institute Inc.(1996) SAS/STAT Software: Changes and Enhancements through Release 6.11, Cary, NC: SAS Institute Inc.

Latour,D.(1995) Introduction to the MIXED procedure Course Notes, Cary, NC: SAS Institute Inc.

Latour,D., Latour,K., and Wolfinger,R.D.(1994) Getting Started with PROC MIXED, Cary, NC: SAS Institute Inc.

Wolfinger,R.D.[岸本 訳](1993) An Introduction to Mixed Modeling with the SAS/STAT MIXED Procedure, 日本 SAS ユーザー会論文集.

PROC MIXED 自体も休みなく進歩を続けている。マニュアル中の解説もどんどん良くなってきている。文献を参照するときには、できるだけ新しいものを探すべきである。