原 著

欠測を伴う経時測定データにおける MMRM
(Mixed-Effects Model for Repeated Measures)
の並べ替え法に基づく推測手法

Permutation Inference Methods for the MMRM
(Mixed-Effects Model for Repeated Measures)
in Incomplete Longitudinal Data Analysis

右京芳文^{*1,*2}・野間久史^{*3}

Yoshifumi Ukyo^{*1,*2} and Hisashi Noma^{*3}

^{*1} ヤンセンファーマ株式会社生物統計部

^{*2} 総合研究大学院大学複合科学研究科統計科学専攻

^{*3} 統計数理研究所データ科学研究系

^{*1} Department of Biostatistics, Janssen Pharmaceutical K.K.

^{*2} Department of Statistical Science, School of Multidisciplinary Sciences,
The Graduate University of Advanced Studies

^{*3} Department of Data Science, The Institute of Statistical Mathematics

e-mail : noma@ism.ac.jp

Missing data is a common problem in longitudinal clinical trials, and mixed-effects models for repeated measures (MMRM) have been widely applied to circumvent the resulting bias effectively. However, many standard inference methods of MMRM lead to the inflation of type I error rates for the tests of regression coefficient parameters when the longitudinal dataset is small and incomplete. Permutation inference methods have been established as accurate inference methods under small sample settings. In this article, we propose two effective permutation-based inference methods for the analyses using MMRM. One is the permutation of the treatment assignment variable and the other is the permutation of weighted residuals estimated by the reduced model under null hypothesis. We conducted numerical evaluations via simulation studies under realistic situations to evaluate performances of the proposed methods. The two methods generally provided valid inference results and performed relatively well compared with the current standard methods, even for small and incomplete datasets. Applications to a postnatal depression clinical trial are also presented.

Key words: MMRM, permutation test, small sample, missing data, longitudinal data.

1. はじめに

新医薬品及び新治療の開発に関する臨床試験では、経時的に測定されたアウトカムによって、治療効果の評価が行われることが多い。こうした経時測定データに対する解析では、同一個人に繰り返し測定されるアウトカムの独立性が一般的に成り立たないため、相関を適切に考慮した解法を用いる必要がある。加えて、経時測定データには必ずと言ってよいほど欠測が発生する。臨床試験における欠測の取り扱いについては、欧米の規制当局において、ガイドラインの作成が進められており (European Medicines Agency, 2010; National Research Council, 2010)、近年では、臨床試験・臨床研究の統計解析の実践においても、欠測データの適切な取り扱いが求められるようになってきている (Newgard and Lewis, 2015)。

線形混合モデル (linear mixed model) は、このような欠測を伴う経時測定データの解析において、ひとつの標準的な方法として広く用いられてきた (Diggle et al., 2002; Fitzmaurice et al., 2012; Laird and Ware, 1982)。特に、近年では、時点間の結果変数間の相関構造を周知的にモデル化した Mixed-effect Models for Repeated Measures (MMRM) が、その実践的有用性により急速に普及している (Mallinckrodt et al., 2001; 2004; 2008; 2003)。

しかしながら、これらの方法において、標準的な推測手法として用いられている制限付き最尤法 (Restricted Maximum Likelihood: REML, Patterson and Thompson, 1971) などの方法は、大標本近似に基づく方法であるため、サンプルサイズが十分に大きくない条件下では、その推測の妥当性が損なわれる可能性があることが古くから知られている (Kenward and Roger, 1997; Schluchter and Elashoff, 1990)。Gosho et al. (2017) によるシミュレーション実験では、MMRM を用いた解析においても、欠測を伴う小標本の条件下では、従来の標準的な推測手法の多くにおいて、固定効果パラメータの推測における妥当性が失われることが示されている。これらの問題を解決するために、これまでに、高次漸近理論に基づくいくつかの方法が提案されている (Kenward and Roger, 1997; Kenward and Roger, 2009)。これらの手法の有用性を示す数値的なエビデンスも多く得られているが (例えば, Gosho et al., 2017)、いずれも大標本近似を用いた手法であり、その推測の妥当性が絶対的に保証されることはない。

一方で、小標本下における統計的推測の有効なアプローチとして、古くから並べ替え法に基づく推測手法も広く用いられてきた (Good, 2005; Mielke and Berry, 2007)。MMRM においても、長らく蓄積された経験的な知見から、並び替え法に基づく推測手法は、臨床試験の実践において有用なアプローチになることが期待できるが、これまでに、MMRM において、並び替え法について検討された研究は報告されていない。

本論文では、これらの欠測を伴う小標本下での MMRM の解析における推測の問題を解決するためのひとつのアプローチとして、並び替え法に基づく推測手法を提案する。特に、複数の回帰パラメータや分散共分散パラメータを含む回帰モデルを対象とした並び替え法として、2つの方法を提案するものとする。1つは、R. A. Fisher (Fisher, 1935) による、古典的な比較群を表す指示変数の並び替えを行う方法である。ランダム化臨床試験において、試験治療がランダムに割り付けられていること及びデータの交換可能性が成立していることを前提にすれば、経時測定データにおける多変量アウトカムに対しても、この古典的な並び替え検定により、妥当な P 値を得る

ことができる。2つ目は、帰無仮説のもとでの制約付きパラメータ空間において推定された回帰モデルにおいて、得られた残差の重み付き統計量の並べ替えを行う方法である。この重み付き残差に基づく並び替え法は、線形混合モデルの分散成分の検定のために、Lee and Braun (2012) によって提案されており、シミュレーション実験などでも、広範な条件下での妥当性が示されているが、固定効果パラメータの推測にこれを用いる検討はこれまでにない。本論文では、また、シミュレーション実験によって、これらの提案法の性能を評価し、産後うつ病の臨床試験 (Gregoire et al., 1996) の事例解析を通して、その実践的な有用性についての検討を行った。

2. MMRM による経時測定データ解析

連続量の結果変数が経時的に複数回繰り返し測定される 2 群比較のランダム化臨床試験を想定する。最終測定時点を n 時点目とする。2 群の合計参加者数を N とし (それぞれの群の参加者数を N_0, N_1 とする)、参加者 i の結果変数の測定時点を j 、測定回数を n_i とおく ($j=1, 2, \dots, n_i, n_i \leq n$)。主要な統計解析は、最終測定時点である n 時点目の 2 群の結果変数の平均値の差とする。本研究では、単調な欠測を想定するものとするが、非単調な欠測においても同様に、3 節における提案法を適用することは可能である。

線形混合モデル (Laird and Ware, 1982; Verbeke and Molenberghs, 2000) は、同一個人における複数時点での結果変数の相関を変量効果を用いてモデル化した回帰モデルであり、以下の式で表現される。

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

$\mathbf{Y}_i$ は、参加者 i における $n_i \times 1$ の結果変数ベクトルを表す。 $\boldsymbol{\beta}$ は、 $p \times 1$ の固定効果パラメータベクトルである。 $\mathbf{b}_i$ は、変量効果を表す $q \times 1$ のベクトルで、多変量正規分布 $MVN(\mathbf{0}, \mathbf{D})$ に従って分布するものとする。ここで、 $\mathbf{D}$ は (i, j) 成分 d_{ij} に対して $d_{ij} = d_{ji}$, ($i \neq j$) となる $q \times q$ の共分散行列である。 $\mathbf{X}_i$ と $\mathbf{Z}_i$ は、それぞれ固定効果・変量効果に対応する $n_i \times p$ 及び $n_i \times q$ のデザイン行列である。 $\boldsymbol{\varepsilon}_i$ は、 $n_i \times 1$ の誤差ベクトルであり、それぞれ独立に多変量正規分布 $MVN(\mathbf{0}, \boldsymbol{\Sigma}_i)$ に従う。 $\boldsymbol{\Sigma}_i$ は、対角成分に各時点での誤差分散を持つ $n_i \times n_i$ の対角行列にすることが多い。変量効果 $\mathbf{b}_i$ と誤差 $\boldsymbol{\varepsilon}_i$ は独立であり、異なる個人間のデータも全て独立と仮定する。

MMRM は、本質的に、上記の慣習的な線形混合モデル (1) と同じモデルを想定したモデルであるが、結果変数 $\mathbf{Y}_i$ の周辺分布を直接モデル化した、異なるモデル化に基づくモデルである。線形混合モデル (1) において、変量効果 $\mathbf{b}_i$ に対して周辺化した結果変数 $\mathbf{Y}_i$ の周辺分布は、以下のように書くことができるが、

$$\mathbf{Y}_i \sim MVN(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{V}_i), \quad \mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i \quad (2)$$

MMRM は、上式 (2) における変量効果の分布を明示的にモデル化することはせず、周辺分布の共分散行列 $\mathbf{V}_i$ そのものをモデル化したモデルとなる (Mallinckrodt et al., 2008)。臨床試験における応用では、関心のある治療効果の指標を表すパラメータは、一般的に固定効果パラメータ $\boldsymbol{\beta}$ の成分によって表されるため、変量効果の分布は、実質的には局外要因となることが多い。した

がって、周辺共分散行列 V_i に、無構造などのモデルを用いることによって、MMRM では、局外要因となる変量効果の分布についてのモデルの誤特定を可能な限り回避した固定効果パラメータ β の推測を行うことができる。パラメータの推測は、最尤法、REML 法などの直接尤度法の枠組みでの推測を行うことができるため、仮に結果変数に部分的な欠測があったとしても、その欠測データメカニズムが MAR (missing at random) であれば、妥当な推測を行うことができる (Molenberghs and Kenward, 2007; 五所, 丸尾, 2017)。

3. 並べ替え法による推測手法

3.1 検定問題の設定

いま、一般性を失うことなく、関心のある最終時点の平均値の差の検定は、固定効果パラメータ $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ の第一成分 β_1 を治療群の主効果と扱い、以下の検定問題と設定することができる。

$$\text{帰無仮説 } H_0: \beta_1 = \beta_1^{null}$$

$$\text{対立仮説 } H_1: \beta_1 \neq \beta_1^{null}$$

ここで、 β のうち、 β_1 を除いた部分ベクトルを $\beta_c = (\beta_2, \beta_3, \dots, \beta_p)^T$ とし、結果変数ベクトルの周辺共分散行列 V_i の成分で構成されるベクトルを u とおく。

パラメータベクトル $\theta = (\beta_1, \beta_c, u)^T$ の最尤推定量は、結果変数の周辺分布モデル (2) による多変量正規分布に基づく対数尤度関数 $l_{ML}(\beta_1, \beta_c, u)$ を最大化することによって得られる推定量となる。一方、分散成分 u についての固定効果パラメータの自由度を調整した REML 推定量は、以下の対数制限付き尤度関数を最大化することによって得られる推定量となる。

$$l_{REML}(\beta_1, \beta_c, u) = l_{ML}(\beta_1, \beta_c, u) - \frac{1}{2} \log \left| \sum_{i=1}^N X_i^T V_i^{-1} X_i \right| \quad (3)$$

ここでは、次節以降における並び替え検定を議論する上での基本統計量を、上記の検定問題における、REML 推定量に基づく以下の Wald 型の検定統計量とする。

$$T(\beta_1^{null}) = \frac{\hat{\beta}_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} \quad (4)$$

$\hat{\beta}_1$ は、 β_1 の REML 推定量であり、 $\widehat{\text{Var}}(\hat{\beta}_1)$ はそのモデル分散の推定量である。 $\hat{\beta}_1$ は、単純な最尤推定量に置き換えても、以降の議論は同様に成立する。自由度の調整方法も任意である。また、この基本統計量には、尤度比統計量やスコア統計量などを用いてもよい。

3.2 方法 1: 比較群指示変数の並べ替えによる方法

治療のランダム割り付けが成立していれば、帰無仮説のもとで結果変数と比較群を表す指示変数は独立とみなすことができる。Fisher (1935) は、この考え方から、比較群の指示変数を並び替える並び替え検定を考案した。すなわち、一定の確率でいずれかの治療がランダムに割り付けら

れるというルールをもとに、帰無仮説のもとでの検定統計量の分布を推定し、これに基づいて検定を構成するという方法である。原理的には、すべての割り付けの組み合わせに対して、検定統計量の値を計算し、その分布を P 値を計算するための参照分布に用いるという方法になる。ただし、サンプルサイズが大きくなると、割り付けの組み合わせの数が膨大なものとなってしまう、すべての組み合わせを取り尽すための計算コストが非現実的に大きくなってしまうため、その中から十分な数の組み合わせをランダムに選択し、並び替え分布をモンテカルロ推定値で代用することが一般的である (Dwass, 1957)。MMRM に対しての並び替え検定は、以下のようなアルゴリズムによって構成することができる。

Algorithm 1 (比較群指示変数の並び替え法)

Step 1. (観測されたデータに対する基本統計量 $T(\beta_1^{null})$ の計算)

観測されたデータに対して、3.1 節の手順にしたがって、基本統計量 $T(\beta_1^{null})$ を計算する。

Step 2. (比較群指示変数の並び替え)

N 人の比較群を示す指示変数を $\{N_0, N_1\}$ の組み合わせに並び替えを行う。ここでは、その数を B 通りとする。すべての組み合わせを取り尽すことが可能である場合には、 $B = N C_{N_0}$ とし、そうでない場合には、この中から十分大きな数の組み合わせ（例えば、1000 組）をランダムに選択する。

並び替え後の B 組のデータセットに対して、Step 1 と同じ手順で、基本統計量 $T^{(b)}(\beta_1^{null})$ を計算する ($b=1, 2, \dots, B$)。

Step 3. (P 値の計算)

Step 2 で得られた統計量 $T^{(1)}(\beta_1^{null}), T^{(2)}(\beta_1^{null}), \dots, T^{(B)}(\beta_1^{null})$ の分布を参照分布に用いて、P 値を計算する。

$$p = \frac{1}{B+1} \left\{ 1 + \sum_{b=1}^B I[|T^{(b)}(\beta_1^{null})| > |T(\beta_1^{null})|] \right\} \quad (5)$$

ここで、 $I(x)$ は指示関数であり、引数の命題 x が真のとき 1、偽のとき 0 を返す関数である。

ここで得られる並び替え P 値は、 β_1 以外の局外パラメータを含むモデルに基づくものであることに注意が必要である。局外パラメータを含まない、モデルによらない単純なランダム化に基づく検定では、すべての並び替えの組み合わせを取り尽すことによって、正確な P 値を得ることができるが、回帰モデルなどの局外パラメータを含むモデルにおける検定では、これは一般的に正確な P 値とはならない。その厳密な正当性は、大標本理論によって説明することが可能であるが、このモデルに基づく近似的な P 値を用いた検定も、経験的に小標本下で良好な性能を有することが知られている (Anderson and Robinson, 2001)。なお、上記のアルゴリズムから得られる並び替え P 値を用いて、これに対応する $100 \times (1 - \alpha)\%$ 信頼区間は、有意水準 α の検定で有意にならない帰無仮説に対応する β_1^{null} の集合によって構成することができる。

3.3 方法2: 重み付き残差の並び替えによる方法

回帰モデルの残差に基づく並び替え検定は, モデルに基づく妥当な検定方法として古くから知られており, Freedman and Lane (1983) によって提案された. Freedman and Lane (1983) は, 線形回帰モデルにおける並び替え法の議論を行っており, 線形回帰モデルでは, 残差の分布が独立かつ同一の分布に従っていると見なすことができるため, その並び替えを行うことによって検定統計量の帰無仮説のもとでの分布を求めることができる. これによって, モデルに基づく妥当な P 値を計算することができる (Anderson and Robinson, 2001).

変量効果を含む, 線形混合モデルにおいても, Lee and Braun (2012) が, 分散成分の検定において, この残差に基づく並び替え検定を用いた方法を提案している. 線形混合モデルでは, 変量効果がモデルに含まれるため, 単純な残差は, 独立でも同一分布に従うわけでもないため, 単純に残差の並び替えするだけでは妥当な P 値を計算することはできない. このため, Lee and Braun (2012) は, 帰無仮説のもとでの周辺共分散行列 $\mathbf{V}_i$ の制約付き推定値をもとにした重み付き残差統計量を用いることによって, 分散成分の検定を構成することを提案した.

MMRM でも, 同様に, 残差が独立・同一分布に従うことは一般的にないため, 本研究では, 固定効果パラメータの検定のために, 以下の重み付き残差に基づく並び替え法を提案する. いま, モデル (2) において, 帰無仮説のもとで 2 群の治療効果に関する固定効果パラメータが等しいと仮定したモデルを制約付きのモデルと考える. ここで, モデル (2) における帰無仮説のもとでの固定効果パラメータ・周辺共分散行列の制約付きパラメータ推定値を, $(\tilde{\beta}, \tilde{\mathbf{V}}_i)$ とする ($i=1, 2, \dots, N$). また, $\tilde{\mathbf{U}}_i$ を $\tilde{\mathbf{V}}_i$ をコレスキー分解した上三角行列とする. このとき, この $\tilde{\mathbf{U}}_i$ を用いた, 以下の重み付き残差統計量は, 多変量標準正規分布に従う.

$$(\tilde{\mathbf{U}}_i^T)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\tilde{\beta}) \sim MVN(0, \mathbf{E}_{n_i}) \quad (6)$$

$\mathbf{E}_{n_i}$ は, n_i 次の単位行列である. 前節の Fisher 流の並び替え検定とは異なり, 比較群指示変数以外の説明変数もモデルに残っているため, これら説明変数と結果変数 $\mathbf{Y}_i$ の関係が保存されている点が特徴となる. 検定の構成手順は以下の通りである.

Algorithm 2 (重み付き残差の並び替えによる方法)

Step 1. (観測されたデータに対する基本統計量 $T(\beta^{null})$ の計算)

観測されたデータに対して, 3.1 節の手順にしたがって, 基本統計量 $T(\beta^{null})$ を計算する.

Step 2. (重み付き残差の並び替え)

観測されたデータに対して, 帰無仮説のもとでの制約付きモデルのもとでの結果変数の予測値 $\tilde{\mathbf{Y}}_i = \mathbf{X}_i\tilde{\beta}$, 及び, 重み付き残差統計量 $\mathbf{e}_i = (\tilde{\mathbf{U}}_i^T)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\tilde{\beta})$ を計算する.

重み付き残差 $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N$ をランダムに並び替え, B 通りの並び替え重み付き残差統計量 $\mathbf{e}_i^{(b)}, \mathbf{e}_2^{(b)}, \dots, \mathbf{e}_N^{(b)}$ ($b=1, 2, \dots, B$) を得る.

並び替えを行った重み付き残差統計量を, 結果変数の予測値 $\tilde{\mathbf{Y}}_i$ に加えて, 並び替え残差を付与した予測値 $\mathbf{Z}_i^{(b)} = \tilde{\mathbf{Y}}_i + \mathbf{e}_i^{(b)}$ を得る. こうして得られる $\mathbf{Z}_1^{(b)}, \mathbf{Z}_2^{(b)}, \dots, \mathbf{Z}_N^{(b)}$ を結果変数とし, Step 1

と同じ回帰関数モデルと共分散構造を仮定した MMRM に対して、基本統計量 $T^{(b)}(\beta_1^{null})$ を計算する ($b=1, 2, \dots, B$).

Step 3. (P 値の計算)

Step 2 で得られた統計量 $T^{(1)}(\beta_1^{null}), T^{(2)}(\beta_1^{null}), \dots, T^{(B)}(\beta_1^{null})$ の分布を参照分布に用いて、P 値を計算する.

$$p = \frac{1}{B+1} \left\{ 1 + \sum_{b=1}^B I[|T^{(b)}(\beta_1^{null})| > |T(\beta_1^{null})|] \right\} \quad (7)$$

信頼区間については、前節と同様、Algorithm 2 から得られる並び替え P 値を用いて、有意水準 α の検定で有意にならない帰無仮説に対応する β_1^{null} の集合から、対応する $100 \times (1-\alpha)\%$ 信頼区間を構成することができる.

4. シミュレーション実験

4.1 シナリオ

提案した並び替え法の性能と実践的有用性を評価するため、シミュレーション実験を行った. シミュレーションの設定は、Gosho et al. (2017) のシミュレーション実験で用いられたものをもとに設定した.

プラセボ群及び実薬群の 2 群の比較試験を想定する. 各群ともにベースラインを含めた 7 時点の測定時点があるものとし、最終測定時点の群間差に主たる関心があるものとする. 参加者 i , 時点 j の測定値 Y_{ij} は以下の式により多変量正規分布を利用して生成した. ここで、 b_{ij} は参加者 i , 時点 j の群効果とする.

$$Y_{ij} = \text{mean}_j + b_{ij} + \text{subject}_i + \varepsilon_{ij}, \quad i=1, 2, \dots, N, \quad j=1, 2, \dots, 7 \quad (8)$$

ただし、変量効果 subject_i は $N(0, 3^2)$ に従うと仮定した.

今回は、特に、小標本下での性能評価が目的であるため、1 群あたりのサンプルサイズは 10 例 ($N=20$) とした. 結果変数 Y_{ij} の平均構造 mean_{ij} については、図 1 に示す 2 種類のシナリオを仮定した. シナリオ 1 は、2 群間で最終時点での平均値が同じであるとする帰無仮説に対応している. 一方、シナリオ 2 は 2 群間で最終時点での平均値は異なっているという想定で、対立仮説に対応する. シナリオ 1 を用いて第 1 種の過誤確率の評価、シナリオ 2 を用いて検出力の評価を行った.

また、これとは別に、途中で 2 群間の平均値が異なり薬効が存在しても最終時点で平均値が同じとなるシナリオ (シナリオ 3)、及び、プラセボ群で若干の群間差が存在し 2 群間で最終時点での平均値が異なるシナリオ (シナリオ 4) についての実験も行った. 結果は付録に掲載している.

4.2 相関構造

誤差 ε_{ij} の相関構造には、1 次の不均一自己回帰 (heterogeneous first-order autoregressive; ARH (1)) を仮定した. ARH (1) は時点ごとに分散が異なり、相関は時点が遠くなるほど減少

図 1. シミュレーション実験で用いた平均構造に関する 2 種類のシナリオ

する相関構造を模した分散共分散行列である。本シミュレーションでは、対角成分 σ_j^2 を $\sigma_j^2 = 9 \times \{1 + 3(j-1)/6\}$ の式で定義する。隣接時点間の相関係数は 0.7 と設定する。

4.3 欠測メカニズム

本シミュレーションでは、MCAR (missing completely at random) 及び MAR の 2 種類の欠測メカニズムのもとでの評価を行った (Rubin, 1976)。結果変数 Y_{ij} の欠測確率を λ_{ij} として、その値が、1 時点前の結果変数 $Y_{i,j-1}$ で条件付けたロジスティック回帰モデルによって規定されるものと仮定した。

$$\text{logit}(\lambda_{ij}) = \gamma_0 + \gamma_1 Y_{i,j-1}, j = 2, 3, \dots, 7 \quad (9)$$

このモデルをもとに、MCAR では $\gamma_1 = 0$ 、MAR では、 $\gamma_1 = -1$ とした設定を考えた。欠測メカニズムと比較群ごとの最終的な脱落割合、及び、その脱落割合を達成するロジスティック回帰モデルの係数の設定を表 1 にまとめた。2 群を併せた全体での最終的な脱落割合は、20% 及び 40% の 2 種類を設定した。本研究におけるシミュレーションでは単調な欠測を仮定する。すなわち、一度欠測が発生したら、欠測発生時点以降の測定値はすべて欠測値として取り扱う。

4.4 評価方法

本研究で提案した 2 つの並び替え法と 3 つの標準的な推測手法の比較を行った。以降では、比較群指示変数を並べ替える方法 (Algorithm 1) 及び重み付き残差の並び替えによる方法 (Algorithm 2) を、それぞれ PermTrt, PermWR と略記する。今回のシミュレーション実験では、基本統計量の計算の際の自由度調整法には、一貫して between-within 法 (Schluchter and Elashoff, 1990) を用いた。実際には、多くのソフトウェアで Satterthwaite 法が実装されており

表 1. シナリオ別の最終時点における累積脱落割合とパラメータ設定

シナ リオ	欠測 メカニズム	累積 脱落 割合 (%)	累積脱落割合 (%)		γ_0	
			プラセボ群	実薬群	プラセボ群	実薬群
1	MCAR	20	20	20	3.2	3.2
		40	40	40	2.4	2.4
	MAR	20	20	20	7.1	7.1
		40	40	40	4.2	4.2
2	MCAR	20	26	17	3.0	3.5
		40	47	36	2.2	2.6
	MAR	20	29	11	7.4	7.4
		40	51	27	4.4	4.4

Satterthwaite 法の適用も考えられる。提案法の直接的な比較対象として、REML 法および between-within 法を用いた検定 (MMRM_{BW-REML}) と、最尤法および between-within 法を用いた検定 (MMRM_{BW-ML}) を設定した。また、高次漸近近似を用いた、小標本下での改良法として、広く実践でもスタンダードな方法として用いられている Kenward-Roger 法 (Kenward and Roger, 1997) も、比較対象に設定した (MMRM_{KR})。

回帰関数のモデルには、比較群及び時点の変数をダミー変数としてモデル化し、また、これらの交互作用項を含めたものを用いた。共分散構造には、一貫して無構造を用いた。それぞれの方法において、最終測定時点における群間差に対応する固定効果パラメータの検定を行い、P 値を計算した。MMRM_{KR}、MMRM_{BW-REML} 及び MMRM_{BW-ML} については、固定効果の標準誤差が 50 を超えた場合には推定が不適切であると判断し、評価から除外する処理を行った。また、それぞれの方法について収束割合を算出した。

本研究におけるシミュレーションプログラムは SAS Ver 9.4 (SAS Institute, Cary, USA) で作成しており、MMRM の実装には PROC GLIMMIX を用いた。同様の解析は、PROC MIXED でも可能である。収束判定の基準には SAS GLIMMIX プロシジャのデフォルト設定を適用した。反復回数は 2000 回とした。また、並べ替えを行ったデータセットを MMRM で解析した結果、基本統計量が算出できないケースもあった。今回のシミュレーション実験では、これらのケースを除いて合計 3000 回のランダムな並び替えがなされるように設定した。すべての検定において、有意水準は両側 5% に設定した。

4.5 結 果

4.5.1 第 1 種の過誤確率

シナリオ 1 のもとでの第 1 種の過誤確率の結果を図 2 に示す。有意水準 5% の検定を行っているため、第 1 種の過誤確率は 5% に制御されるはずである。MMRM_{BW-REML} 及び MMRM_{BW-ML}

図 2. シミュレーション実験の結果 (シナリオ 1, 第 1 種の過誤確率)

では、欠測割合が増加すると MCAR 及び MAR の両方で第 1 種の過誤確率が大きくなった。MMRM_{BW-REML} では、欠測割合 40% のとき、第 1 種の過誤確率は、MCAR 及び MAR でそれぞれ 9.8% 及び 8.6% であった。一方、MMRM_{KR} では、MAR 下で欠測割合が 20% のときに第 1 種の過誤確率が 5% を下回り、保守的な傾向を示した。

シミュレーションで得られた第 1 種の過誤確率のモンテカルロ推定値は、一定の誤差を伴うため、図 2 では、モンテカルロ標準誤差の範囲を青線で表示している。PermTrt は、欠測メカニズムや欠測割合に関わらず、概ね 5% 前後を維持していた。PermWR では、MAR 下で欠測割合が 20% のとき、わずかにモンテカルロ誤差の範囲を超えて保守的な傾向が認められた。

4.5.2 検出力

シナリオ 2 のもとでの検出力の結果を図 3 に示す。シナリオ 2 の最終測定時点の群間差は -9.6 であり、欠測がない条件下で MMRM_{KR}、MMRM_{BW-REML} 及び MMRM_{BW-ML} では検出力はそれぞれ 86.3%, 86.3% 及び 88.5% であった。

図 3 において、MMRM_{BW-ML} の検出力が最も高く、欠測割合に応じて 88.5% から 70.1% の範囲に値をとった。しかしながら、MMRM_{BW-REML} 及び MMRM_{BW-ML} は、シナリオ 1 の下で、第 1 種の過誤確率が 5% に制御されていなかったため、それを反映した結果であるものと考えられる。一方、MMRM_{KR} は、シナリオ 1 のもとで、第 1 種の過誤確率が MAR の下で保守的な傾向を示していたため、それとは逆の傾向を持つものと考えられる。PermTrt の検出力は、欠測割合が 40% の MCAR 及び MAR 下でそれぞれ 59.5% 及び 59.4% であった。また、PermWR の検出

図 3. シミュレーション実験の結果 (シナリオ 2, 検出力)

力は、欠測割合が 40% の MCAR 及び MAR 下でそれぞれ 56.1% 及び 53.8% であった。

4.5.3 収束割合

収束割合の結果を図 4 及び図 5 に示した。比較したすべての方法で欠測割合が 40% となった場合でも約 90% 前後の収束割合が得られた。また、方法間での違いもほとんど認められなかった。

5. 産後うつ病の臨床試験データ解析への応用

Gregoire et al. (1996) は、61 人の産後 3 か月以内の女性を対象に、産後うつ病に対するエストロゲンの経皮投与の有効性を評価するための二重盲検プラセボ対照比較試験を実施した。試験の計画時点では 100 人の登録を見込んでいたが、実際には、想定よりも登録が進捗せずに 61 人の女性がプラセボ群 (27 人) もしくはエストロゲン群 (34 人) にランダム割付された (Everitt, 2001)。

本試験は 1 か月ごとにうつ症状の指標である Edinburgh Postnatal Depression Scale (EPDS) スコア等の測定を行い、8 か月間のフォローアップを行っている。このうち、はじめの 2 回の測定は、ランダム化前に行われており、治療開始後のフォローアップ期間は、6 か月間となる。EPDS スコアは、スコアが高いほどうつ症状の重症度が高いことを表す指標である。EPDS スコアのベースラインは 1 か月目及び 2 か月目の EPDS スコアの平均値と定義されていた。8 か月目の最終評価時点における全体集団では、プラセボ群で 37.0% (10/27)、エストロゲン群で 17.6% (6/34) の参加者に EPDS スコアの欠測が認められた。すべての欠測は単調な欠測となっていた。

図 4. シミュレーション実験の結果 (シナリオ 1, 収束割合)

図 5. シミュレーション実験の結果 (シナリオ 2, 収束割合)

表 2. 産後うつ臨床試験における 8 カ月時点の平均値の差の推定値, 95% 信頼区間と P 値

方法†	全体集団		部分集団	
	群間差 [95%CI]	p-value	群間差 [95%CI]	p-value
PermTrt	4.34 [1.48, 7.08]	0.0037	3.93 [-1.26, 9.16]	0.1436
PermWR	4.34 [1.52, 7.15]	0.0033	3.93 [-1.37, 9.22]	0.1383
MMRM _{KR}	4.34 [1.45, 7.23]	0.0040	3.93 [-1.23, 9.09]	0.1288
MMRM _{BW-REML}	4.34 [1.51, 7.18]	0.0033	3.93 [-1.03, 8.90]	0.1160
MMRM _{BW-ML}	4.34 [1.57, 7.12]	0.0027	3.93 [-0.81, 8.67]	0.1004

図 6. 産後うつ臨床試験における並べ替え統計量の分布 ($B=3000$)

また, 今回は, ベースライン時点での EPDS スコアが 21 を超える集団を, うつ症状が重症な部分集団と定義し, この部分集団に対しての解析も行った. 部分集団における参加者数は 30 人 (プラセボ群及びエストロゲン群それぞれ 15 人) であり, 8ヶ月目には 9 人が脱落し, 21 人 (プラセボ群 9 人及びエストロゲン群 12 人) の EPDS スコアが得られている.

MMRM による解析では, 前節のシミュレーション実験で用いたモデルと同様, 比較群・時点のダミー変数をモデル化した回帰関数と, 無構造の共分散行列を仮定したモデルを用いた. 比較手法も, 同様に, PermTrt, PermWR, MMRM_{BW-REML}, MMRM_{BW-ML}, MMRM_{KR} の 5 つとした.

表 2 に, 最終測定時点である 8 か月目の EPDS スコア群間差及びその 95% 信頼区間について解析結果を示している. 全体集団及び部分集団のいずれにおいても, PermTrt 及び PermWR の P 値は, MMRM_{BW-REML} よりも大きな値をとる傾向が確認できた. 部分集団解析結果においては, MMRM_{KR} よりも PermTrt 及び PermWR の P 値は大きい傾向があった. また, MMRM_{BW-ML} は比較対象とした方法の中で最も小さい P 値を示した.

図 6 に、最終測定時点である 8 か月目において、並べ替え検定を MMRM に適用した手法によって得られる基本統計量の並び替え分布のヒストグラムを示した。並び替えの回数は 3000 回であり、上側が全体集団、下側が部分集団解析でのヒストグラムとなっている。それぞれのヒストグラムの赤点線は観測データの基本統計量の値を示しており、全体集団及び部分集団でそれぞれ 3.06 及び 1.62 であった。

6. 考 察

本研究では、近年、臨床試験・臨床研究の実践において、広く普及している MMRM における並び替え法に基づく推測手法を提案した。並べ替え検定の手法自体は、これまでも臨床試験のデータ解析で長らく用いられてきている。一方、並べ替え検定を MMRM に適用して詳細に検討・評価した研究は、著者らの知る限り、本研究が初めてとなる。シミュレーション実験の結果、提案した 2 つの並び替え法は、いずれも適切に第 1 種の過誤確率を制御することができることが示された。従来のスタンダードな推測手法であった Kenward-Roger 法などと比較すると、その性能は著しく異なるわけではなかったが、現状、利用することができる手法は、いずれも大標本近似に基づく方法であり、その推測の妥当性が絶対的に保証されることはない。このような条件下で、正確性が期待できる方法の選択肢を多く用意しておくことは、実践的にも重要であると考えられる。理論上は、これらの手法の優劣関係を議論することはできないが、シミュレーション実験の結果、少なくとも一定の条件下で、提案する方法は、既存の手法よりも優れた性能を有する可能性があることが示唆された。

提案法では、並べ替えたデータセットに対して MMRM で解析した結果、収束しなかったものの結果を除外している。これは、多くの場合、並べ替えを行った結果、最終時点までに脱落を起こした参加者が、偶然に片方の群に偏り、基本統計量が計算できないことが原因である。しかし、シミュレーション実験の結果を精査したところ、該当するケースが起こることは稀であり、帰無分布の妥当性に影響を与えることはほとんどないと考えられた。

シミュレーション実験の結果、シナリオ 1 及び 3 の MAR 下で欠測割合が大きくなると、MMRM_{KR} は過度に保守的な第 1 種の過誤確率を示した。これは固定効果の標準誤差が過大に推定されているためである可能性が高い（付録、図 12 を参照）。

本研究のシミュレーションでは、データ生成モデル ARH (1) を仮定したが、実際のデータではこれ以外の相関構造を持つ場合も多いと考えられる。この点については、今回のシミュレーションでは、解析に、最も自由度の高い無構造モデルを用いているため、仮に相関構造が ARH (1) 以外の構造であったとしても異なる傾向を持つ結果が得られる見込みは低いと考えられた。また、より複雑な相関構造を仮定した場合、無構造モデルに基づく推測が不安定化する可能性はある。

MMRM は、医薬品開発や臨床研究及・疫学研究において、欠測を伴う経時測定データ解析の標準的な解析方法として、今後、ますます普及していくものと思われる。評価項目によっては、MMRM のモデルの平均構造や共分散構造の不現実性が大きく、必ずしも第一選択手法にはならないが、希少疾患を対象とした臨床試験も、近年、ますます増加傾向にある。本研究で議論した

小標本下での妥当な推測方法は、今後も実践的な有用性を増していくであろう。

謝 辞

本研究は、科学研究費補助金（課題番号：JP15K15954, JP15H03390）の助成を受けた。

参考文献

- Anderson, M.J., and Robinson J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics* **43**, 75–88.
- Diggle, P., Diggle P.J., Heagerty P., Heagerty P.J., Liang K.-Y., and Zeger S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.
- Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* **28**, 181–187.
- European Medicines Agency (2010). *Guideline on Missing Data in Confirmatory Clinical Trials*. https://www.ema.europa.eu/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf (accessed on 11 Jan, 2019).
- Everitt, B.S. (2001). *A Handbook of Statistical Analyses Using S-Plus*. New York: CRC Press
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- Fitzmaurice, G. M., Laird N. M., and Ware J. H. (2012). *Applied Longitudinal Analysis*, vol 998. New York: John Wiley & Sons.
- Freedman, D., and Lane D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics* **1**, 292–298.
- Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York: Springer-Verlag.
- Gosho, M., Hirakawa A., Noma H., Maruo K., and Sato Y. (2017). Comparison of bias-corrected covariance estimators for MMRM analysis in longitudinal data with dropouts. *Statistical Methods in Medical Research* **26**, 2389–2406.
- Gregoire, A. J., Kumar R., Everitt B., Henderson A. F., and Studd J. W. (1996). Transdermal oestrogen for treatment of severe postnatal depression. *Lancet* **347**, 930–933.
- Kenward, M. G., and Roger J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.
- Kenward, M. G., and Roger J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis* **53**, 2583–2595.
- Laird, N. M., and Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.
- Lee, O. E., and Braun T. M. (2012). Permutation tests for random effects in linear mixed models. *Biometrics* **68**, 486–493.

- Mallinckrodt, C.H., Clark W.S., and David S.R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics* **11**, 9-21.
- Mallinckrodt, C.H., Kaiser C.J., Watkin J.G., Detke M.J., Molenberghs G., and Carroll R.J. (2004). Type I error rates from likelihood-based repeated measures analyses of incomplete longitudinal data. *Pharmaceutical Statistics* **3**, 171-186.
- Mallinckrodt, C.H., Lane P.W., Schnell D., Peng Y., and Mancuso J.P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal* **42**, 303-319.
- Mallinckrodt, C.H., Sanger T.M., Dubé S., DeBrota D.J., Molenberghs G., Carroll R.J., Potter W.Z., and Tollefson G.D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry* **53**, 754-760.
- Mielke, P.W., and Berry K.J. (2007). *Permutation Methods: a Distance Function Approach*. New York: Springer-Verlag.
- Molenberghs, G., and Kenward M. (2007). *Missing Data in Clinical Studies*, New York: John Wiley & Sons.
- Newgard, C.D., and Lewis R.J. (2015). Missing data: how to best account for what is not known. *JAMA* **314**, 940-941.
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: National Academies Press.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.
- Schluchter, M.D., and Elashoff J.T. (1990). Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computation and Simulation* **37**, 69-87.
- Verbeke, G., and Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- 五所正彦, 丸尾和司. (2017). 経時測定データ解析における mixed-effects models for repeated measures (MMRM) の利用. *応用統計学* **46**, 53-65.

付 録

4 節に記したシミュレーション実験（シナリオ 1, 2）に加えて，補足的に，2つのシナリオのもとでの実験も行った．それぞれのシナリオをシナリオ 3, 4 とし，その平均構造については，図 7 に示している．シナリオ 3 は，途中で 2 群間の平均値が異なり薬効が存在しても最終時点で平均値が同じとなるシナリオであり，シナリオ 4 は，プラセボ群で若干の薬効が存在し，2 群間で最終時点での平均値が異なるシナリオである．それぞれシミュレーションの繰り返し回数は 1000 回とした．

シナリオ 3 のもとでの第 1 種の過誤確率の結果を図 8 に示す．PermTrt は，欠測メカニズムや欠測割合に関わらず，概ね第 1 種の過誤確率は 5% 前後に制御されていた．PermWR 及び

図 7. シミュレーション実験で用いた平均構造に関するシナリオ (シナリオ 3, 4)

図 8. シミュレーション実験の結果 (シナリオ 3, 第 1 種の過誤確率)

図 9. シミュレーション実験の結果 (シナリオ 4, 検出力)

MMRM_{KR} では, MAR 下で欠測割合が 20% 及び 40% のとき, わずかにモンテカルロ誤差の範囲を超えて保守的な傾向が認められた.

シナリオ 4 のもとでの検出力の結果を図 9 に示す. 欠測がない条件下で, MMRM_{KR}, MMRM_{BW-REML} 及び MMRM_{BW-ML} では検出力がそれぞれ 87.1%, 87.1% 及び 89.2% であった. 図 9 において, MMRM_{BW-ML} の検出力が最も高く, 欠測割合に応じて 89.2% から 68.5% の範囲に値をとった.

シナリオ 3 のもとでの収束割合の結果を図 10 に示す. 比較したすべての方法で欠測割合が 40% となった場合でも MCAR 及び MAR 下でそれぞれ約 95% 前後及び約 85% 前後の収束割合が認められた. 方法間での違いはほとんど認められなかった.

シナリオ 4 のもとでの収束割合の結果を図 11 に示す. 比較したすべての方法で欠測割合が 40% となった場合でも約 90% 以上の収束割合が認められた.

また, 補足的に, シナリオ 1 のもとで, MMRM_{SAT}, MMRM_{KR} 及び MMRM_{BW-REML} の 3 つの方法について固定効果の標準誤差を評価した結果を図 12 に示している. 反復回数は 10000 回であり, MMRM_{SAT} が固定効果の自由度調整を Satterthwaite 法とした REML 推定を用いた Wald 型検定, MMRM_{KR} が固定効果の標準誤差の推定及び自由度調整を Kenward-Roger 法とした Wald 型検定, MMRM_{BW-REML} が固定効果の自由度調整を between-within 法とした REML 推定を用いた Wald 型検定に対応している. 図 12 より, MCAR 及び MAR の欠測割合が 40% の下では, MMRM_{KR} の標準誤差は他の 2 つの方法よりも大きくなっていった. 本研究のシミュレーシ

図 10. シミュレーション実験の結果 (シナリオ 3, 収束割合)

ン設定では, 2 群の比較試験を想定しており, この想定の下では MMRM_{KR} の自由度調整と Satterthwaite 法の自由度調整は等価であった. したがって, シナリオ 1 の MAR 下で欠測割合が大きくなると, MMRM_{KR} が過度に保守的な第 1 種の過誤確率を示した原因として, 固定効果の標準誤差が過大に推定されている可能性が考えられた.

図 11. シミュレーション実験の結果 (シナリオ 4, 収束割合)

図 12. シミュレーション実験の結果 (シナリオ 1, 第 1 種の過誤確率)