R による多重補完法の比較に関する研究

2014SS028 金武芽実
指導教員：松田眞一

1　はじめに

欠測値が存在するデータに対して用いられる統計的手法の 1 つに多重補完法がある．この多重補完法は，R において複数パッケージが存在する．本研究では，どのパッケージがデータの補完に優れているかを比較，検証する．

2　多重補完法

多重補完法 (Multiple Imputation) とは 1987 年に Donald Rubin 氏が提唱した統計的方法である．多重補完法は欠測のあるデータについて複数回補完を行い，完全なデータセットを作成すること，得られたデータセットに対してそれぞれ任意の統計的方法を適用すること，そしてその結果を統合することの 3 のステップで行う．補完回数については，欠測率に応じて設定する．(松山 [2]，高橋・伊藤 [3] 参照)

3　多重補完法のパッケージ

研究に用いた R のパッケージについて紹介する．

Amelia Amelia パッケージは EMB アルゴリズムを用いてデータの補完を行う．(高橋・伊藤 [3] 参照)

mice mice パッケージは FCS（Fully Conditional Specification）を用いてデータの補完を行う．(高橋・伊藤 [3] 参照)

mi mi パッケージは条件付き分布を用いて，欠測値の補完を行う．また加えて事前分布のベイズモデルを用いて補完を行う．(Su et al. [4] 参照)

4　使用したデータ

今回，第 I 相臨床試験に参加した健康な成人男性 100 名のデータを使用した．変数は実験前における最高血圧，最低血圧，脈拍数，体温，年齢，身長，体重の 7 変数である．(小林・志村 [1] 参照) データは標準化し，標準化したデータに対して欠測を発生させた．

5　欠測メカニズム

今回シミュレーションで用いた欠測メカニズムについて説明する．(松山 [2] 参照)

MCAR MCAR(Missing Complete at Random) とは，完全にランダムな欠測である状態を指す．

MAR MAR(Missing at Random) とは，欠測が起きる確率は他の変数に依存する状態であることを指す．

6　プログラミング

今回，乱数を用いて欠測を発生させるプログラム，各パッケージにおけるシミュレーションプログラム，そして結果を比較するための最大値探索プログラムを作成した．

7　シミュレーション

計算時間がかかるため，シミュレーション回数を 100 回，補完回数を 5 回としてシミュレーションを行った．欠測率は，MCAR，MAR 共に 5%，10%，20%，30% とした．欠測率，及び変数ごとに，3 つのパッケージの補完値の平均値の二乗誤差の最大値を比較し，3 つの中で最大であった回数を集計し，考察を行った．なお，以下では紙面の都合上，一部の結果のみを示す．

7.1　MCAR の結果

MCAR では，mi が一番性能がよくないことが分かった．欠測率が上がると mi が最大となる回数が増えたので，欠測率が高いほど mi は真値に近い値を補完できないと思われる．合計数だけで見ると，欠測率が低いとき Amelia の方がより欠測値の補完に優れていると考えられる．しかし，変数によっては，mice の方が優れているものもあった．

7.2　MAR の結果

MAR では，欠測を起こすかどうかの決める基準の変数を年齢とし，基準値は第 3 四分位数 0.3599(35 歳相当) とした．年齢を 2 区分に分け，それぞれの欠測率のオッズを設定して欠測を起こした．オッズは 2 と 3 と設定した。年齢を除いた 6 変数の補完値を求め，比較した．欠測率が 20%，オッズが 2 のときの結果を表 1 に示す．

表 1　欠測率 20%，オッズ 2 の結果

	Amelia	mice	mi
最高血圧	21	28	51
最低血圧	25	31	44
脈拍数	30	26	44
体温	22	38	40
身長	19	31	50
体重	31	26	43
合計	148	180	272

すべての欠測率，オッズにおいて mi が一番真値に近い値を補完できないことが分かった．MAR では，ほとんど最大にならなかった mice が一番補完に優れていると考えられる．Amelia は欠測率，オッズ比ごとで優れているものもあれば，そうでないものもあった．しかし，欠測率が高い方がいい結果を得ているものが多いので，欠測率が高くても優れているのではないかと考えられる．

7.3 まとめ

欠測メカニズムが MCAR であっても，MAR であっても mi は補完の性能が他の 2 つに比べてよくないことが分かった．特に，欠測率が高くなればなるほど，真値に近い補完をすることが難しいと考えられる．MCAR では Amelia の方が mice より優れていると思われるが，MAR では欠測率によっては mice の方が優れているときもあり，また変数によって大きく差があったものがあるので，これらの欠測値に対する補完は変数の分布にもよると考えられる．

8 対数変換後のシミュレーション

上記の結果は，分布のゆがみが影響していると考えられる．そこで，元のデータに対して対数変換を行った場合，二乗誤差の最大値が変化するのか比較，検証を行った．今回は脈拍数に対して対数変換を行い，標準化してシミュレーションを行った．なお，欠測率は 5% と 20% のみ取り扱い，それ以外のシミュレーションの条件は変更していない．欠測メカニズムが MAR，欠測率が 20%，オッズが 2 のときの結果を表 2 に示す．

表 2 MAR 欠測率 20%，オッズ 2 対数変換後の結果

	Amelia	mice	mi
最高血圧	24	32	44
最低血圧	21	24	55
脈拍数	35	29	36
体温	29	23	48
身長	22	45	33
体重	27	28	45
合計	158	181	261

対数変換を行った脈拍数において一番改善が見られたのは mi であった．しかし，最大値がよくなっても，Amelia と mice にはかなわないことが分かった．mice は対数変換をする前に比べてそこまで変化がなかった．Amelia は合計数が減ったり，3 つの中で最大となった変数が対数変換をすることによってなくなったので，一番効果があったと考えられる．

8.1 最大値の変化個数

最大値を比較し，個数を調べることだけでは，どれだけの効果があったのかが分かりにくい．そこで，対数変換後，補完値の二乗誤差の最大値が変化があったのかをパッケージごとに比較を行った．比較の対象は対数変換を行った脈拍数のみ取り扱い，最大値が小さくなったときの回数を調べた．一番改善したものは mi であった．したがって，mi はデータの分布をみて対数変換をするなど，適切な処理をしてから多重補完法を適用した方がより真値に近い補完ができると考えられる．しかし，3 つのパッケージを比較すると，対数変換を行ったとしても，Amelia や mice に比べると劣ってしまっている．

8.2 改善率

対数変換を行った脈拍数に対して，改善率を求めた．改善率は，対数変換後に二乗誤差の最大値が小さくなった個数を各回における発生した欠測個数で割ることで計算した．MCAR においては mi は効果があったと言いにくい．逆に MAR においては mi に効果があったと思われる．しかし，改善していても Amelia や mice と同程度の補完ができないと考えられる．また，Amelia と mice については，条件によっては効果があると思われる．

8.3 対数変換後のまとめ

対数変換を行ったことにより，効果が見られたのは mi であったが，それでも Amelia と mice と同程度の補完はできないことが分かった．mice は対数変換を行っても結果が大きく変化しなかった．Amelia は，最大となった個数が減ったので，一番効果があったと思われる．

9 考察

シミュレーションの結果を比較して，mi が一番真値に近い補完を行うことが他の 2 のパッケージに比べてできないことが分かった．Amelia と mice に関しては，対数変換を行っても結果が大きく変化しなかった．今回の結果とそれぞれの特徴から，真値により近い補完値を求めるときは Amelia を，さまざまな変数が混在するデータの補完を行いたいときは mice を用いるべきだと考える．しかし，今回はデータ数が少ないために明確な結果とはならなかった．したがって，シミュレーション回数を増やすなどをして，比較，検証すべきであったと思われる．

10 おわりに

本研究を通して，多重補完法の R 内におけるパッケージの欠測値に対する補完の性能，利点について知ることができた．今後，大学院での研究においてシミュレーションを行うときは，条件，データ数に気を付けて研究していきたい．

参考文献

[1] 小林宏行・志村政文：BAY o 9867 (Ciprofloxacin) の臨床第一相試験，『CHEMOTHERAPY』，**33**(S-7)，140-170，1985.

[2] 松山裕：経時観察研究における欠測データの解析，『計量生物学』，**25**(2)，89-116，2004.

[3] 高橋将宜・伊藤孝之：様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～，『統計研究彙報』，**71**，39-82，2014.

[4] Su, Y. , Gelman, A. , Hill, J. and Yajima, M. : Multiple Imputation with Diagnostics(mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*, 45(2), DOI:10.18637/jss.v045.i02, 2011.