多重代入法による欠測データの統計解析：Q&A

野間 久史
統計数理研究所
2016年9月10日
Research Memorandum
e-mail: noma(at)ism.ac.jp
URL: http://www.ism.ac.jp/~noma/

1

Question 1

- MAR (Missing At Random)の仮定が成り立つことをどうやって検定？できるのかできないのか、できないとしたら MNAR (Missing Not At Random) として感度分析がbetterなのか？
- MARであるか、MNARであるかは、データから証明することができない、いわば「検証不可能」な仮定であるといわれます。MARでないということは、観測されていない情報で欠測メカニズムが規定されるということであり、そもそもそれを評価する情報が得られていないためです。
- ReasonableかつPlausibleと思われるシナリオのもとで、十分な感度解析を行い、主要な解析に基づく結論の頑健性を評価しておくことが重要です。

2

補足：欠測のメカニズム（Rubin, 1976）

- Missing Completely At Random (MCAR)
  - 欠測メカニズムは、他のいかなる要因とも関係がなく、完全に（純粋に）ランダムに起こる
- Missing At Random (MAR)
  - 欠測メカニズムはランダムではないが、観測されたデータによって説明することができる
- Missing Not At Random (MNAR) = Not MAR
  - 欠測メカニズムは観測されたデータだけでは説明することができず、欠測した（観測されていない）データにも依存する
  - 感度解析をするしかない（選択モデル，パターン混合モデル）

※ MARはその名称から「ランダムな欠測」と誤解されることがありますが、実際には「非ランダムな欠測」です。ご注意ください。

3

Question 2

- 正しくないMIをしたことによって、complete case analyses や欠損値をダミー化した場合よりも真値から遠ざかってしまうことはないのか？（正しくないというのは、収束はしたが、推定に使用したモデルや変数がおかしいということです。）
- 補完値の生成モデルが、真の構造を大きく外してしまった場合、MIの結果にはバイアスが入ることが知られています。検定・信頼区間も妥当性を失います。
- しかしながら、Complete-Case AnalysisはMCARでない場合には、一般的にバイアスが入りますので、優劣の議論以前に、MCARが保証できない場合には推奨できません。

4

Question 3

- 上記でいうところの「正しいMI」「正しくないMI」を判定することはできるのか？
- （私の知る限り）現段階で、スタンダードといわれる評価（判定）方法は確立されておらず、どの統計ソフトウェアにも実装されていないように思います。これは、生物統計の分野で今後解決されるべき研究課題といえるかと思います。
- 一方で、Data-Drivenな盲目的な評価は、しばしば的外れな結論に行き着くリスクを孕みますので（例えば、Stepwise法による変数選択のように）、いずれにしても、ReasonableかつPlausibleと思われるシナリオのもとで、十分な感度解析を行い、主要な解析に基づく結論の頑健性を評価しておくことが重要です。

5

Question 4

- パネルデータ分析に適したMI手法はあるのでしょうか？
- モデルとセッティングによってケースバイケースかと思われます。

6

Question 5

- MIで作成するデータセットはいくつが良いか？（議論がありそうです＆PCのスペック向上に伴い増えているようですが、現状の相場観を。できればその根拠の文献紹介も合わせて）
- 統計家によって意見は違うかと思います。現段階でのコンセンサスはありません。ただ、少なくとも、従来から言われている3～5などの回数では、現在の計算機環境からしても、Scientificな妥当性を担保するための十分な努力がなされたとはいえないかと思われます。最近の文献を見る限り、100～1000回程度あれば、相応の精度は保証されるように思います。
- 詳細は、例えば、Carpenter and Kenward (2013) をご参照ください。

7

Question 6

- 研究プロジェクト内で同じMIデータセットを使用したい場合、コードをシェアしていれば各自が同じデータセットを容易に作成可能か？
- まったく共通のプログラムを利用して、共通の乱数のSeedを用いれば、同じデータセットを再現することが可能です。
- 乱数のSeedとは？　現在の計算機技術では、実は純粋な意味での「乱数」を生成することは原理的に不可能で、実際に使われているのは、本物の乱数とほとんど同じ性質を持つ「擬似乱数」です。擬似乱数はDeterministicな数列であり（見かけ上は乱数ですが）、それを生成するアルゴリズムは、当然ながらStarting Pointを定める引数（入力値）が必要です。これを乱数のSeedといいます。すべての統計ソフトの擬似乱数生成モジュールには、Seedを定める引数が必ずあります。これを共有すれば、独立な研究グループでも、まったく同じ乱数系列を発生させることができます。ちなみに、現在、世界で最も広く用いられている擬似乱数生成アルゴリズムは、Mersenne-Twister法というもので、広島大学の松本眞教授のグループによって開発されたものです。

8

Question 7

- 欠損値が存在するデータセットを解析する場合には、その対応には、大きく分けて
  a）complete case analysis
  b）dummy categories(variables)
  c）MI含めて各種imputation
  があると理解しています。 1 はMCARを仮定しているのでバイアスを生むのは理解できるが、 2 はなぜあまりよろしくないのか？MIにくらべてどのような点が統計学的に劣る（よりrobustではない？より緩くなる？）のか。あればその根拠文献紹介も合わせてご教示いただければ幸いです。

9

Answer to Question 7

- Complete-Case Analysisは、MCARの仮定のもとのみで妥当な方法なので、推奨されません
- Dummy Categoriesによる方法も、原則としてMCARなどの強い仮定がなくては妥当性を失います。例えば、A, B, Cの3カテゴリに33%ずつの回答が等しく分布するという場合に、それぞれの欠測確率が5%, 50%, 80%である場合、欠測した対象者をDummy Categoryに分類して、A, B, Cのカテゴリ間の比較を行ったらいかがでしょうか？明らかに元の結果からは、大きなバイアスの入った結果が得られます。
- MIは、MARの仮定のもとで妥当な方法となりますので、妥当な補完値を用いれば、上記のような場合でもバイアスのない評価が可能です。

10

Question 8

- モデルのアウトカムもＭＩしてしまうと、"Congeniality"という問題が起きてしまい、複雑性が増すと聞いています。MI modelにはアウトカムを含めるとしても、メインの解析はimputed outcome observationを除いたものをメインにするほうがいいらしいとのことですが、いかがでしょうか。
- 結果変数の欠測について、多重代入法を用いると、単にパラメータの推定誤差が増すだけ（ノイズが加わるだけ）という議論は古くからされています。ご関心のある方は、Little (1992), White et al. (2011) などをご参照ください。

11

Question 9

- マルチレベル構造やパネルデータ構造はMIの第一段階（欠損の補完）の際に考慮すべきでしょうか？
- 補完される欠測変数の分布を考慮する上で必要なものであれば、考慮すべきであるといえます。
- 補完値の生成モデルをマルチレベルモデルにするケースは、私も詳しくは存じ上げませんが、Carpenter and Kenward (2013) に解説があります。

12

Question 10

- 連続変数の欠測データに対する単一代入法の平均代入法や回帰代入法は、分散の推定にバイアスが入ることはわかるのだが、点推定だけを考えた場合、どのような条件下で推定の妥当性は成り立つのですか？
- Robins and Wang (2000) によると、回帰モデルによる代入値の生成では、代入値生成の回帰モデル（1次モーメントのモデル）が正しければ、分散構造は誤特定していても、概ね一致性は成り立つ。平均代入法や回帰代入法は、この回帰モデルによる代入値の生成において、分散を強制的に0に固定したモデルと考えることができる。
- つまり、代入値生成の回帰関数のモデルが正しければ（分散構造は誤っているわけだが）、一致性は概ね成り立つものと考えてよいであろう。

13

参考文献

- Carpenter, J., and Kenward, M. G. (2013). Multiple Imputation and Its Application. Chichester: Wiley.
- Little, R. J. A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association* **87**, 1227-1273.
- Robins, J. M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113-124.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63: 581-592.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 377-399.

14