様々な多重代入法アルゴリズムの比較

統計センター　高橋　将宜　　統計センター　伊藤　孝之

1. はじめに

データが欠測している場合、利用可能なデータサイズが縮小し、偏りが発生する恐れがある。理想的な欠測値対処法は、欠測値を含む不完全データが、欠測値のない完全データと同一になる方法だが、このような目標は、いかなる補定法を用いても達成できない。多重代入法(Multiple Imputation)は、不完全データを用いた統計分析が、完全データによる統計分析と同様に、統計的に妥当になる欠測値対処法である。多重代入法の理論的概念が発案されて数十年の時が経過したが、事後分布からの無作為抽出の実装は難しく、ソフトウェアに実装されているアルゴリズムには様々なものが存在し、いずれのアルゴリズムがどのような状況において優れているのかは不明である。本研究では、公的経済統計における欠測値の補定に関して、様々な多重代入法アルゴリズム間の相対的優位性を比較検証した。

2. 多重代入法の理論

多重代入法では、観測データを条件として、欠測データの事後分布を構築し、この事後分布からの無作為抽出を行うことで、補定にまつわる不確実性を反映させた M 個($M > 1$)のシミュレーション値を生成する。M 個の補定済データセットを別々に使用して統計分析を行い、しかるべき手法により結果を統合し、点推定値を算出する。

3. アルゴリズムとソフトウェア

伝統的な手法により観測データの尤度関数を算出して事後分布から平均値ベクトルと分散・共分散行列の無作為抽出を行うことは難しい。こういった問題を解決するために、様々な計算アルゴリズムが提唱されている。1980年代に提唱された多重代入法の理論は、ベイズ統計学の枠組みで構築され、マルコフ連鎖モンテカルロ法 (MCMC: Markov chain Monte Carlo)に基づいていた。データ拡大法 (DA: Data Augmentation)は、MCMC の計算アルゴリズムであり、繰り返し手法を用いて推定値を改善していく方法である。このアルゴリズムを使用しているソフトウェアは、R パッケージ Norm 3.0.0 及び SAS PROC MI 9.3 である。MCMC の代替法として、完全条件付指定 (FCS: Fully Conditional Specification)が提唱されており、各々の不完全な変数に対して補定モデルを構築し、それぞれの変数に対して補定値を繰り返し作成する。このアルゴリズムを使用しているソフトウェアは、R パッケージ MICE 2.13、PASW Missing Values 18、SOLAS 4.01 である。また、近年では、伝統的な期待値最大化法 (EM: Expectation-Maximization)にブートストラップ法を応用した EMB アルゴリズムも提唱されている。このアルゴリズムを使用しているソフトウェアは、R パッケージ Amelia II (version 1.6.1)である。

4. データセット及び評価方法

2012 年 2 月に我が国で初めて実施された経済センサス - 活動調査の速報データ及びシミュレーションデータを用いて、補定値と真値との差や計算効率など、様々な多重代入法アルゴリズムの優劣を比較検討した。

参考文献

[1] Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.

[2] Schafer, Joseph L. (2008). *NORM: Analysis of Incomplete Multivariate Data under a Normal Model, Version 3*. Software Package for R. University Park, PA: The Methodology Center, the Pennsylvania State University.

[3] Takahashi, Masayoshi and Takayuki Ito. (2012). "Multiple Imputation of Turnover in EDINET Data: Toward the Improvement of Imputation for the Economic Census," *Work Session on Statistical Data Editing, UNECE*, Oslo, Norway, September 24-26, 2012.

[4] 高橋将宜, 伊藤孝之. (2013). 「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」, 『統計研究彙報』第 70 号 no.2, 総務省統計研修所, pp.19-86.

[5] van Buuren, Stef. (2012). *Flexible Imputation of Missing Data*. London: Chapman & Hall/CRC.

平成25年9月9日（月）
2013年度統計関連学会連合大会

様々な多重代入法アルゴリズムの比較

独立行政法人統計センター
○高橋　将宜
伊藤　孝之

独立行政法人 統計センター NSTAC

本報告の内容は、報告者個人の見解であり、必ずしも機関の見解ではない。

目次

- 研究の目的
- 多重代入法(multiple imputation)の理論
- 多重代入法アルゴリズムとコンピュータソフトウェア
- 分析結果
- 結語と将来の課題

用語について

- **補定**(imputation)
- **多重代入法**(multiple imputation)
- **単一代入法**(single imputation)

欠測データの影響

- 利用可能なデータサイズが縮小し、効率性が低下
- 観測値と欠測値との間に体系的な差異が存在する場合、統計分析の結果に偏りが発生するおそれ
- 統計分析においては、何らかの形で欠測値に対処することが必須
- 欠測データの対処法として多重代入法 (Rubin, 1987)

多重代入法の理論と様々な多重代入法アルゴリズム

- 多重代入法の理論的概念
  - 発案されてから数十年の時が経過
  - 事後分布からの無作為抽出の実装は難しい
- 計算アルゴリズム
  - いずれのアルゴリズムがどのような状況において優れているのかは不明

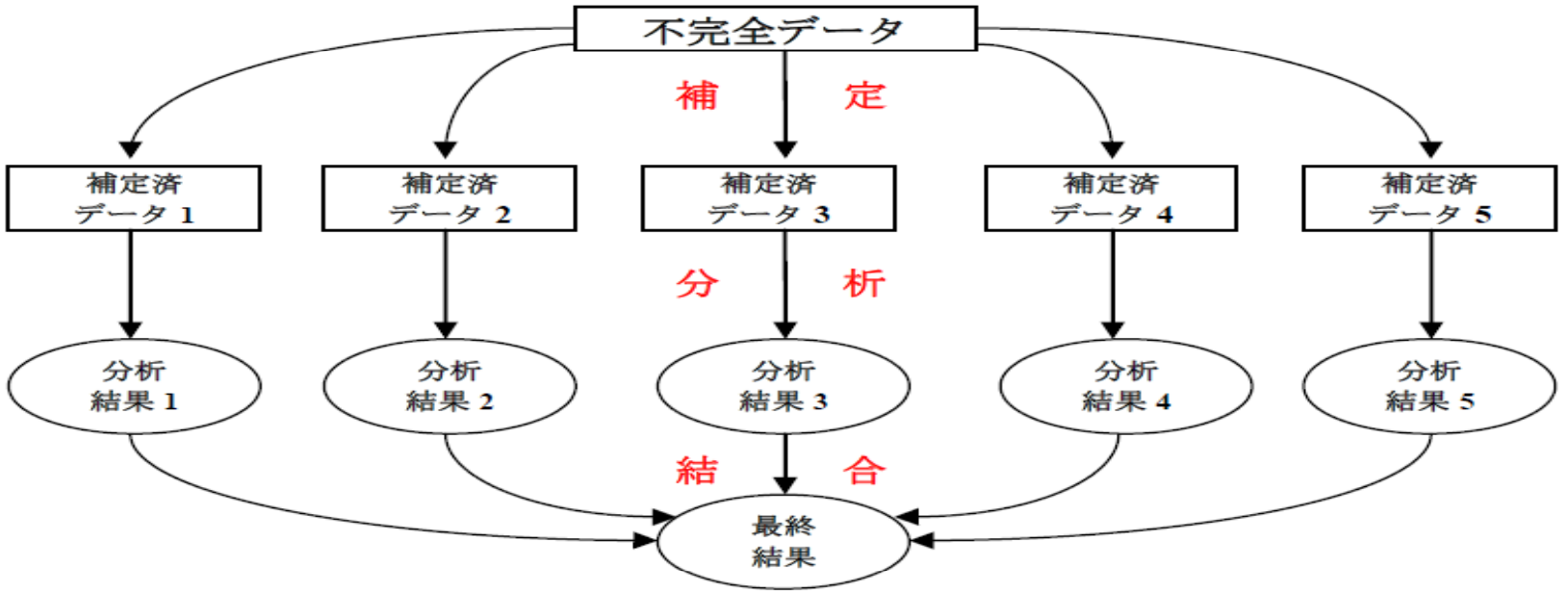
研究内容

- 公的経済統計における欠測値補定に関して、いずれの多重代入法アルゴリズムが優れているかを検証

多重代入法

- 観測データを条件として、欠測データの事後分布を構築し、無作為抽出を行う
- 補定にまつわる不確実性を反映させたM個($M > 1$)の補定済データセットを生成
- これらM個の補定済データセットを別々に使用して統計分析を行い、しかるべき手法により結果を統合し、点推定値を算出

多重代入法の概念図


不完全データ
補
定
補定済
データ 1
補定済
データ 2
補定済
データ 3
補定済
データ 4
補定済
データ 5
分
析
分析
結果 1
分析
結果 2
分析
結果 3
分析
結果 4
分析
結果 5
結
合
最終
結果

補定モデル

$$\tilde{Y}_{ij} = Y_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i$$

完全データの尤度

$$L(\mu, \Sigma \mid D) \propto \prod_{i=1}^{n} N(Y_i \mid \mu, \Sigma)$$

観測データの尤度

$$L(\mu, \Sigma \mid Y_{obs}) \propto \prod_{i=1}^{n} N(Y_{i,obs} \mid \mu_{i,obs}, \Sigma_{i,obs})$$

- 伝統的な手法により、上式を算出して、事後分布から平均値ベクトルμと分散・共分散行列Σの無作為抽出を行うことは難しい
- 様々な計算アルゴリズム

マルコフ連鎖モンテカルロ法 (MCMC): データ拡大法 (Data Augmentation)

- Augmentation＝「拡大」
- 欠測値に適当な値（初期値）を付置することで擬似的にデータを「拡大」
- 一時的な完全データを作成
- 繰り返し手法を用いて推定値を徐々に改善
- Rパッケージ Norm 3.0.0
- SAS PROC MI 9.3

完全条件付指定 (FCS): 連鎖方程式 (Chained Equations)

- 各々の不完全な変数に対して補定モデルを構築
- 周辺分布を利用して、単純無作為抽出を行う
- 条件付で指定した補定モデルを使用して、補定を繰り返す
- Rパッケージ MICE 2.13
- PASW (SPSS) Missing Values 18
- SOLAS 4.01

EMBアルゴリズム

- Expectation-Maximization with Bootstrapping
- **伝統的な期待値最大化法**(EM: Expectation-Maximization)
- **ノンパラメトリック・ブートストラップ**
- R**パッケージ**Amelia II

データセット

- 自然対数に変換したEDINETデータの情報（平均値、分散・共分散など）をもとに、多変量正規分布によって観測数100万、5変量のシミュレーションデータセットを生成した。

	最小値	第1四分位	中央値	平均値	第3四分位	最大値	標準偏差
売上高	2.201	8.998	10.110	10.110	11.230	18.480	1.656
資産	2.584	9.210	10.300	10.300	11.390	18.370	1.617
資本金	0.691	7.097	8.127	8.126	9.156	15.780	1.529
売上原価	1.367	8.533	9.746	9.747	10.960	18.800	1.800
従事者数	0.000	4.221	5.053	5.054	5.888	11.080	1.237

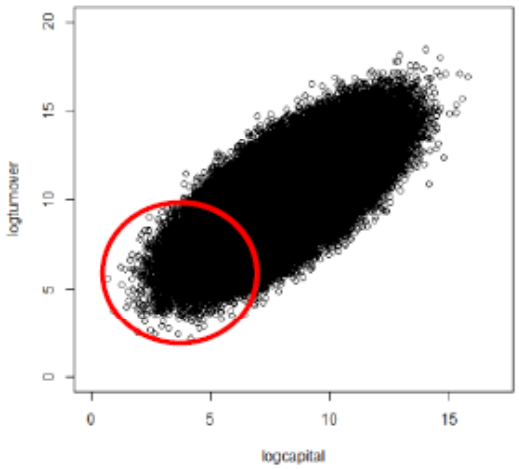
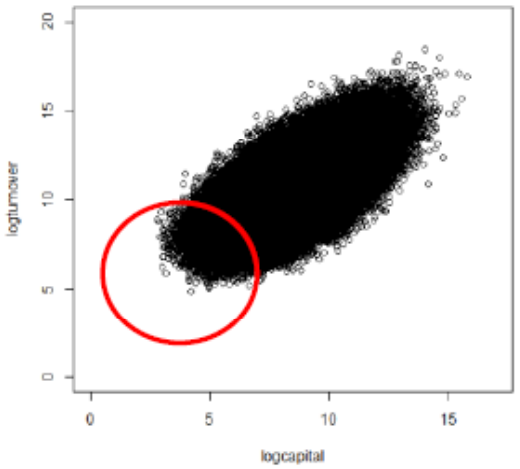
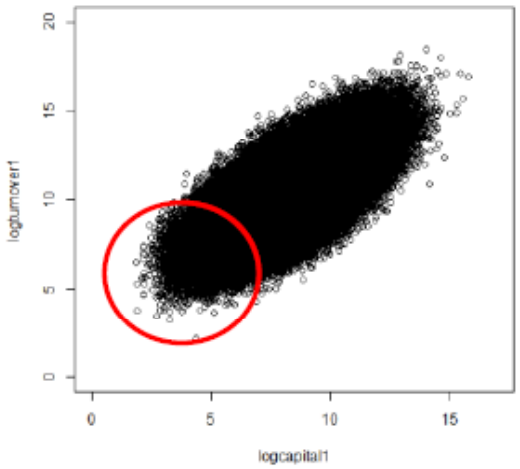
欠測発生メカニズム

- MAR
- **売上高**
  - 10%＝10万個
- **資産、資本金、売上原価**
  - 5%＝5万個
- **事業従事者数**
  - 1%＝1万個
- 500万レコードのうち、26万レコードを欠測
- 100万ユニットのうち、12万7453ユニットに欠測値が含まれている（12.7%）

結果1

	真値	List-Wise	Norm	SAS	MICE	SOLAS	SPSS	Amelia
傾き	0.7862	0.6973	NA	0.7598	0.7568	NA	0.7530	0.7613
t値	1054.7180	839.0746	NA	927.7656	960.7229	NA	938.9850	930.9872
n	1000000	872547	NA	998848	1000000	NA	998514	998848
欠測率	0.0000	12.7453	NA	0.1152	0.0000	NA	0.1486	0.1152

散布図

真の散布図	リストワイズ除去	Amelia II (m = 1)[12]
		

結果2

	NORM	SAS	MICE	SOLAS	SPSS	AMELIA
PC1	動作せず	NA	48分16秒	動作せず	NA	5分30秒
PC2	NA	NA	28分21秒	NA	21分35秒	3分41秒
PC3	NA	4分33秒	40分56秒	NA	NA	4分38秒

PC1：Windows Vista、**プロセッサ**：Intel Core 2 Duo CPU T9400、**メモリ**(RAM)：2.00 GB、32**ビットオペレーティングシステム**
PC2：Windows Vista、**プロセッサ**：Intel Core 2 Duo CPU E8400、**メモリ**(RAM)：2.00 GB、32**ビットオペレーティングシステム**
PC3：Windows 7、**プロセッサ**：Intel Core i5 CPU 670、**メモリ**(RAM)：4.00 GB、32**ビットオペレーティングシステム**。

結語

- 補定の精度
  - いずれのアルゴリズムにも決定的な差はなかった
  - わずかながらMICEが優位
- 計算効率
  - アルゴリズム間に大きな差
  - SASとAmeliaは、シミュレーションデータにおいて、十分な性能を発揮

将来の課題

- 今回の結果は、1つのシード値にのみ基づくもの
- ランダムな影響を排除するために複数のシード値を用いて比較検証を行っている

参考文献1

- Allison, Paul D. (2000). "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research* vol.28, no.3: 301-309.
- Allison, Paul D. (2002). *Missing Data*. CA: Sage Publications.
- Drechsler, Jörg. (2009). "Far From Normal - Multiple Imputation of Missing Values in a German Establishment Survey," *Work Session on Statistical Data Editing, UNECE*, Neuchâtel, Switzerland, October 5-7, 2009.
- Gill, Jeff. (2008). *Bayesian Methods—A Social Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.
- Honaker, James and Gary King. (2010). "What to do About Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* vol.54, no.2: 561–581.
- Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.
- Horton, Nicholas J. and Ken P. Kleinman. (2007). "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician* vol.61, no.1: 79-90.
- Horton, Nicholas J. and Stuart R. Lipsitz. (2001). "Multiple Imputation in Practice: Comparison of Sotfware Packages for Regression Models with Missing Variables," *The American Statistician* vol.55, no.3: 244-254.
- **岩崎学**. (2002). **『不完全データの統計解析』. 東京：エコノミスト社.**

参考文献2

- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* vol.95, no.1: 49-69.
- Leon, Steven J. (2006). *Linear Algebra with Applications*, Seventh Edition. Upper Saddle River, NJ: Pearson/Prentice Hall.
- Lin, Ting Hsiang. (2010). "A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data," *Quality & Quantity* vol.44, no.2: 277-287.
- Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons.
- Rubin, Donald B. (1978). "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section, American Statistical Association*: 20–34.
- Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute Inc.
- Schafer, Joseph L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.
- Schafer, Joseph L. (1999). "Multiple Imputation: A Primer," *Statistical Methods in Medical Research* vol.8: 3-15.

参考文献3

- Schafer, Joseph L. (2008). *NORM: Analysis of Incomplete Multivariate Data under a Normal Model, Version 3*. Software Package for R. University Park, PA: The Methodology Center, the Pennsylvania State University.
- SPSS Inc. (2009). *PASW Missing Values 18*. Chicago, IL: SPSS Inc.
- Statistical Solutions. (2011). *SOLAS Version 4.0 Imputation User Manual*. http://www.solasmissingdata.com/wp-content/uploads/2011/05/Solas-4-Manual.pdf. (Accessed on July 9, 2013).
- Takahashi, Masayoshi and Takayuki Ito. (2012). "Multiple Imputation of Turnover in EDINET Data: Toward the Improvement of Imputation for the Economic Census," *Work Session on Statistical Data Editing, UNECE*, Oslo, Norway, September 24-26, 2012.
- **高橋将宜, 伊藤孝之**. (2013). **「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」**, **『統計研究彙報』第70号** no.2, **総務省統計研修所**, pp.19-86.
- van Buuren, Stef and Karin Groothuis-Oudshoorn. (2011). "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software* vol.45, no.3.
- van Buuren, Stef. (2012). *Flexible Imputation of Missing Data*. London: Chapman & Hall/CRC.
- **渡辺美智子, 山口和範　編著**. (2000). **『EMアルゴリズムと不完全データの諸問題』. 東京：多賀出版**.
- Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

ご清聴ありがとうございました。