平成25年9月13日（金）
経済統計学会　第57回全国研究大会

経済調査における売上高の欠測値補定 ～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

本研究の分析結果は、総務省・経済産業省『平成24年経済センサス - 活動調査』の速報結果の調査票情報を基に著者が独自集計したものである。また、本発表の内容は、発表者の個人的見解を示すものであり、機関の見解を示すものではない。

独立行政法人 統計センター NSTAC

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
- 経済センサス - 活動調査の速報データを用いた分析結果
- 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
  - 効率性が低下
- 観測値と欠測値に体系的な差
  - 統計分析の結果に偏りが発生するおそれ
- 妥当な統計分析を行うために、何らかの形で欠測値に対処することが必要

欠測値の対処法：単一代入法

- 直感的な方法として頻繁に使用されている
- 欠測値を、何らかの予測値で置き換える方法
  - 平均値補定
  - コールドデック補定
  - ホットデック補定
  - 回帰補定（確定的補定と確率的補定）
  - その他

単一代入法の例：回帰補定（確定的補定）

ID	身長	年齢	性別	体重	補定
1	欠測	40	男	63	170.3
2	174	31	男	62	174
3	161	45	女	48	161
4	158	24	女	42	158
5	欠測	40	男	63	170.3
6	163	52	女	58	163
7	172	29	男	70	172
8	153	38	女	46	153
9	178	28	男	70	178

多重代入法と単一代入法の比較

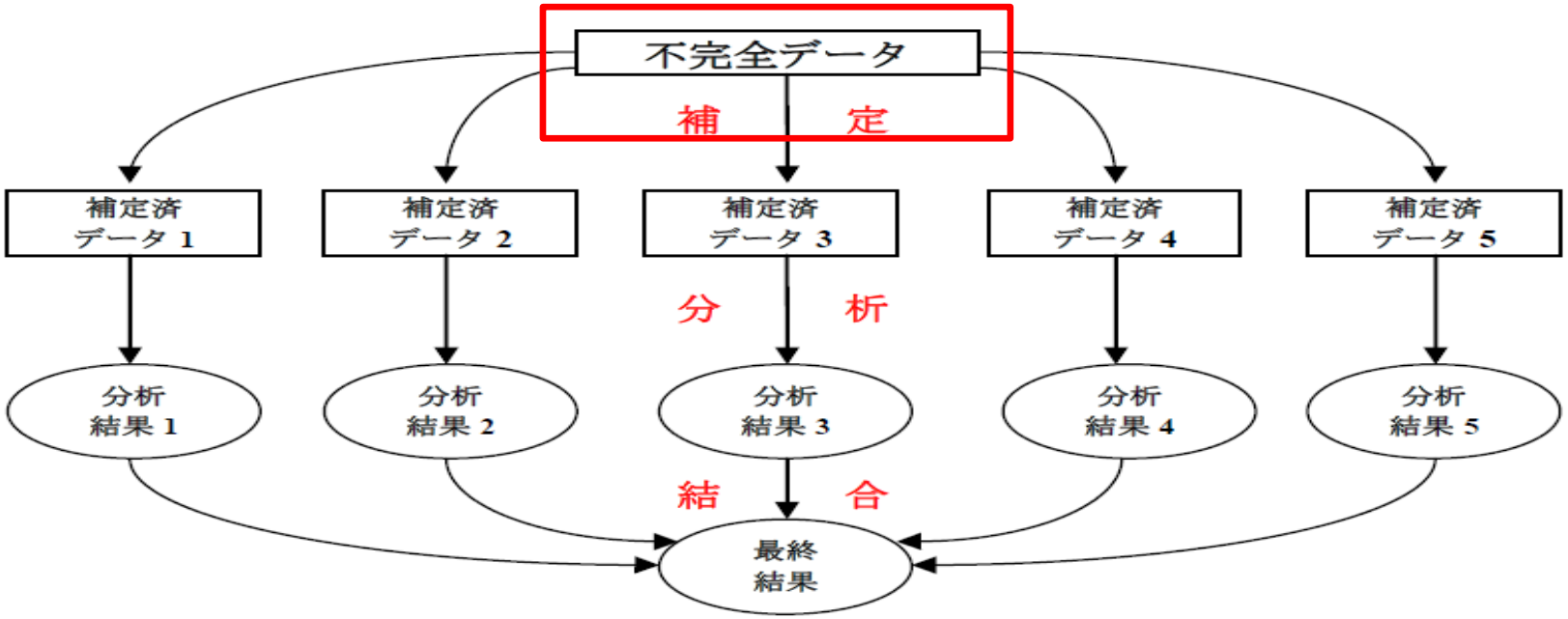
- 「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」
  - 『統計研究彙報』第70号（2013年3月）
  - EDINETデータを使用
  - EMBアルゴリズムを用いた多重代入法と単一代入法（確定的補定と確率的補定）の精度比較

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(Rubin, 1978)
  - 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

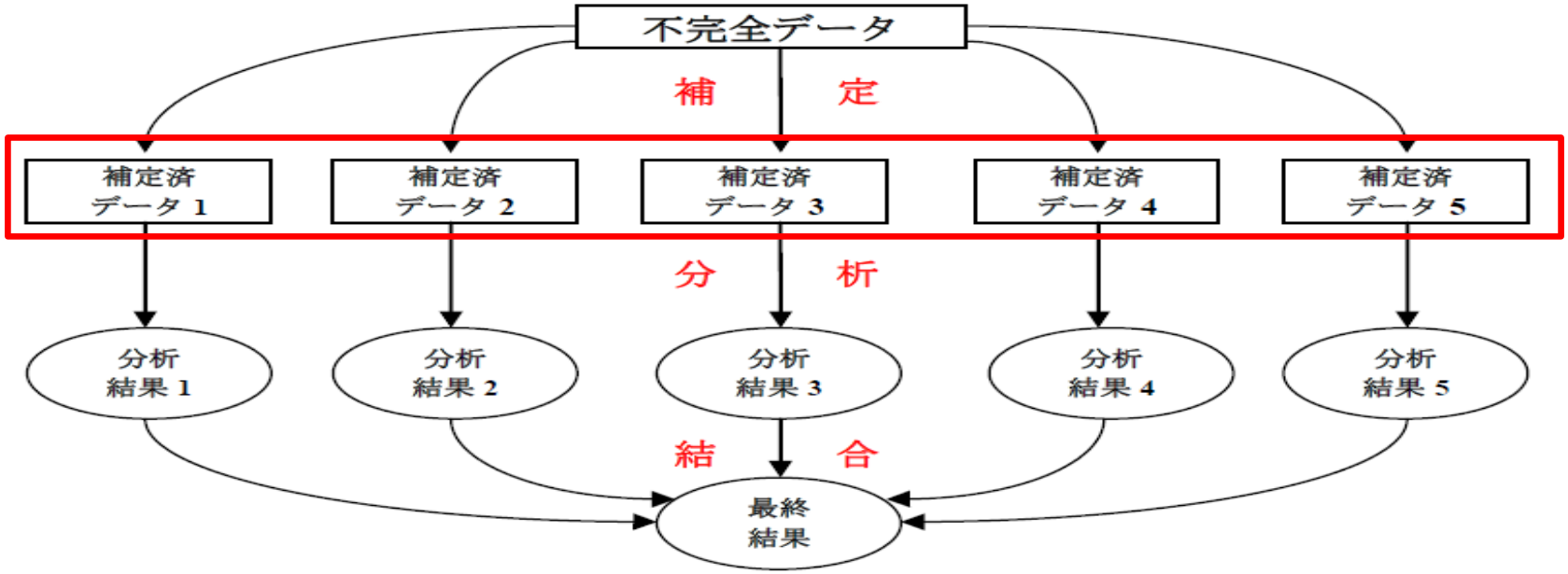
多重代入法の概略1

- 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、パラメータを無作為抽出し、補定



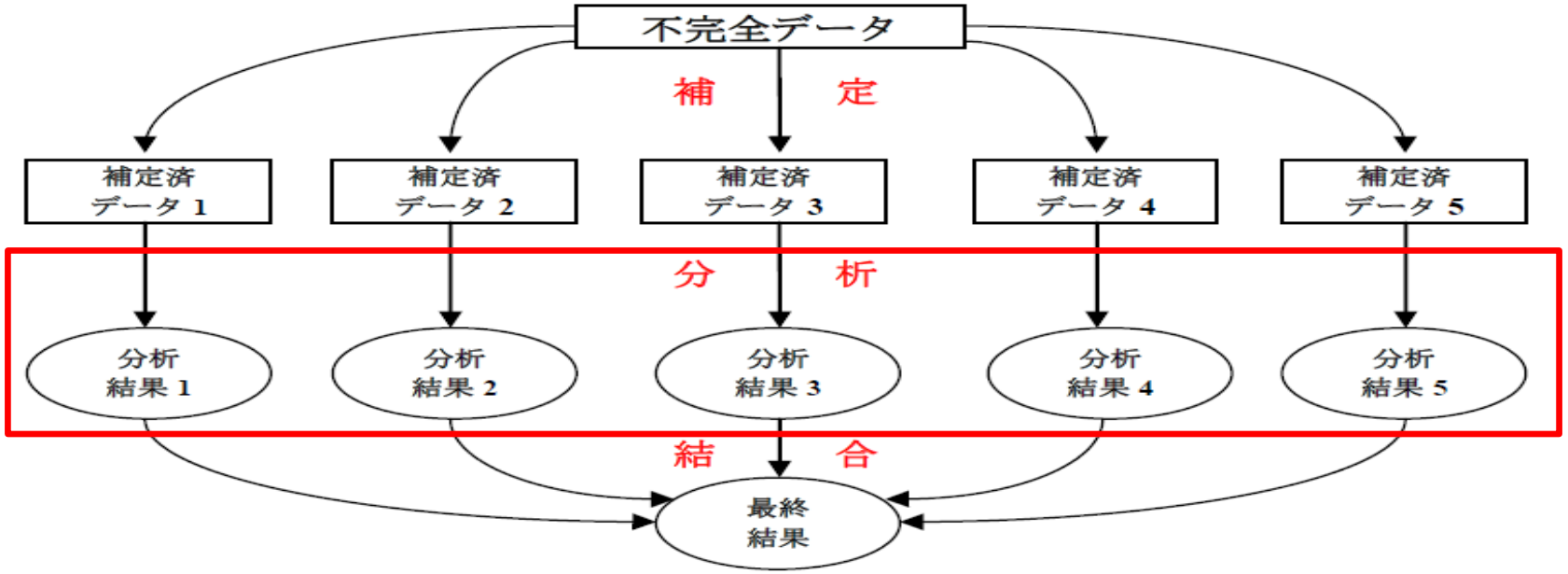
多重代入法の概略2

- *M*個(*M* ＞ 1)の補定済データセットを作成し、補定にまつわる不確実性を反映



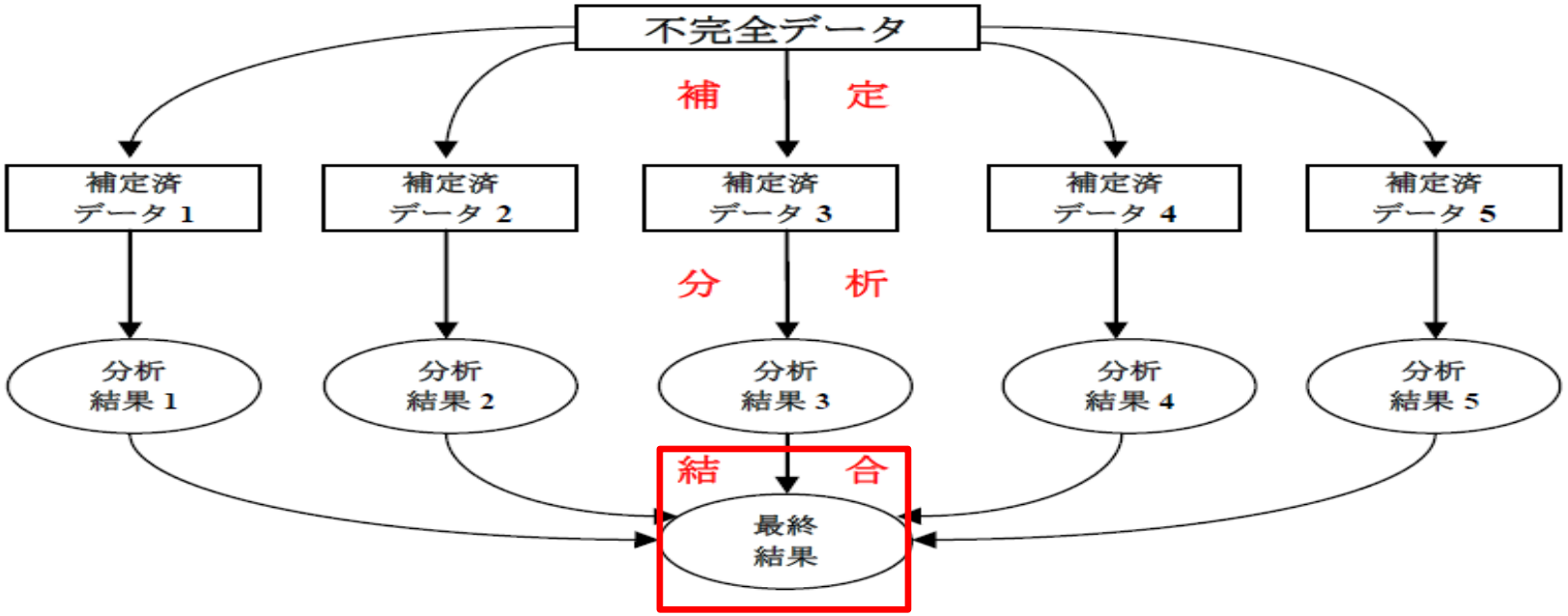
多重代入法の概略3

- *M*個の補定済データセットを別々に使用して統計分析



多重代入法の概略4

- 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：具体例

補定間の不確実性

ID	身長	年齢	性別	体重	補定1	補定2	補定3
1	欠測	40	男	63	171.5	165.3	173.2
2	174	31	男	62	174	174	174
3	161	45	女	48	161	161	161
4	158	24	女	42	158	158	158
5	欠測	40	男	63	168.6	172.2	168.1
6	163	52	女	58	163	163	163
7	172	29	男	70	172	172	172
8	153	38	女	46	153	153	153
9	178	28	男	70	178	178	178

補定内の不確実性

多重代入法モデル1

$$D \sim N_p(\mu, \Sigma)$$

D＝n×pデータセット

nは観測数、pは変数の数

- 真のデータは、もしデータが欠測していないならば、平均値ベクトルμと分散共分散行列Σで多変量正規分布

多重代入法モデル2

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

Y_{ij} ：観測値i及び変数jの欠測値

$\tilde{Y}_{ij}$ ：観測値i及び変数jのシミュレーション値

$Y_{i,-j}$ ：変数jを除くi行のすべての観測値

～ ：適切な事後分布からの無作為抽出

多重代入法モデル3

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

- 回帰係数βを算出するために必要な情報
  - 平均値及び分散・共分散の情報
  - この情報はμとΣにすべて含まれる

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{\sum (X_i - \bar{X})^2/(n-1)} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

多重代入法モデル4

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

- データが欠測していない場合
  - μ と Σ が完全に既知
  - Y_j に基づく真の回帰係数を決定的に算出可能
  - 欠測値を決定的に補定可能
  - そもそも補定をする必要はない！！

多重代入法モデル5

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

- データが欠測している場合
  - μ と Σ を決定的に知ることができない
  - β を確実に知ることは不可能

ID	身長
1	欠測
2	174
3	161
4	158
5	欠測
6	163
7	172
8	153
9	178

$$\mu = \frac{身長_1 + 174 + 161 + 158 + 身長_5 + 163 + 172 + 153 + 178}{9} = \frac{身長_1 + 身長_5 + 1159}{9}$$

$$\mu_{obs} = \frac{174 + 161 + 158 + 163 + 172 + 153 + 178}{7} = 165.5714$$

$$\mu \neq \mu_{obs}$$

ただし、$身長_1 + 身長_5 = 2\mu_{obs}$ の場合を除く

多重代入法モデル6

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

- 事後分布からμとΣの無作為抽出を行い、推定不確実性を反映させるために、複数の計算アルゴリズムが共存

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

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

- 初期値 θ_0
- I-Step: Imputation Step
  - $P\left(Y_{mis} \mid Y_{obs}, \theta_t\right)$ に基づき、$Y_{mis}^{(t+1)}$ を生成
- P-Step: Posterior Step
  - $P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right)$ に基づき、θ_{t+1}を生成
- 注
  - θ ：未知のパラメータ
  - t: 繰り返し回数を意味する。

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

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

- データセット内の観測値と回答指示行列Rに基づいて、各々の変数Y_jの補定モデルを構築

$$P\left(Y_{j,mis} \mid Y_{j,obs}, Y_{-j}, R\right)$$

- 各々の変数に対し、観測値 $Y_{j,obs}$ からの無作為抽出により補定の初期値 $\tilde{Y}_{j,0}$ を設定
- このプロセスを、t = 1,...,Tまで繰り返す。また、j = 1,...,pまで繰り返す。

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

- Y_jを除くt番目の繰り返し時の完全データ

$$\tilde{Y}_{-j,t} = \left(\tilde{Y}_{1,t}, \ldots, \tilde{Y}_{j-1,t}, \tilde{Y}_{j+1,t-1}, \ldots, \tilde{Y}_{p,t-1}\right)$$

- 観測値、補定値（t時点）、回答メカニズムを条件として、補定モデルの未知のパラメータを抽出

$$\tilde{\lambda}_{j,t} \sim P\left(\lambda_{j,t} \mid Y_{j,obs}, \tilde{Y}_{-j,t}, R\right)$$

- 補定値を抽出

$$\tilde{Y}_{j,t} \sim P\left(Y_{j,mis} \mid Y_{j,obs}, \tilde{Y}_{-j,t}, R, \tilde{\lambda}_{j,t}\right)$$

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

EMBアルゴリズム

- 不完全データ（標本サイズ＝ n）
  - q個の値が観測
  - $n - q$個の値が欠測
- ブートストラップ法
  - 不完全データから標本サイズnのブートストラップ副標本の抽出をM回（復元抽出）
- EMアルゴリズム
  - μとΣの点推定値をM個算出
  - 欠測値の補定を行う

3アルゴリズムのまとめ

MCMC
- 理論的に正統
- 多変量正規分布を仮定
- 全変数の補定を一括

FCS
- 理論的根拠が希薄
- 多変量正規分布を仮定しない
- 変数ごとに補定

EMB
- 理論的根拠が希薄
- 多変量正規分布を仮定
- ブートストラップによりコレスキー分解を回避

データセット

- 経済センサス - 活動調査の速報データ
- 2012年2月に実施
- 全データ数は、580万企業・事業所
- 産業大分類I（卸売業、小売業）の単独事業所（個人経営以外）
- 観測数277,263

データセット：経済センサスとは(1)

- 調査対象
  - 日本全国の事業所及び企業
- 目的
  - 我が国における包括的な産業構造を明らかにし、各種経済統計のための母集団情報を整備すること

データセット：経済センサスとは(2)

- 基礎調査（平成21年）
  - 事業所・企業の基本的構造を明らかにするもの
- 活動調査（平成24年）
  - 事業所・企業の経済活動の状況を明らかにするもの
  - 経営組織、従業者数、売上金額といった様々な情報を収集

欠測のメカニズム

- MCAR: *Missing Completely At Random*
  - 欠測は完全にランダム
  - $P(R|D) = P(R)$
- MAR: *Missing At Random*
  - 欠測はランダム
  - $P(R|D) = P(R|D_{obs})$
- NI: *NonIgnorable*
  - 欠測は無視できない
  - $P(R|D)$は単純化することができない

R ＝(0, 1)
D ＝データ
obs ＝観測

欠測発生メカニズム

- 全変数を自然対数に変換（M ＝ 5に設定）
- 売上高
  - MAR、欠測率20%（55,500個）を人工的に欠測
- 資本金
  - MCAR、欠測率5%（13,600個）を人工的に欠測
- 事業従事者数
  - 欠測させていない（欠測率0%）

結果1：補定値の精度

	平均値	標準偏差	傾きの係数	傾きのt値
真値	8.7636	1.5099	1.2075	534.2876
リストワイズ	9.1326	1.3330	1.1431	408.1007
AMELIA	8.7820	1.4597	1.1818	428.9757
MICE	8.7819	1.4598	1.1820	420.2365
NORM	NA	NA	NA	NA
SAS	8.7819	1.4598	1.1819	421.9047
SOLAS	8.7810	1.4605	1.1830	443.6289
SPSS	8.7818	1.4599	1.1820	414.1378

・傾きの係数：log(turnover)=a+b*log(worker)のbの値
・傾きの*t*値：b の*t*値
・100個のシードの結果について、Welchの二標本の平均に関する*t*検定により、SOLASと他のソフトウェアの間で、95%水準で有意差を検証

結果2：計算効率(M = 5)

	AMELIA	MICE	NORM	SAS	SOLAS	SPSS
PC1	1分24秒	10分35秒	動作せず	NA	22分15秒	NA
PC2	55秒	7分18秒	NA	NA	NA	4分2秒
PC3	1分14秒	9分17秒	NA	1分15秒	NA	NA

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
- 計算効率
  - アルゴリズム間に大きな差
  - SASとAmeliaは、計算効率に関して、十分な性能を発揮

将来の課題

- 大規模データセットの多重代入
  - 観測数(n)
  - 変数(p)
- 様々な状況における多重代入
  - カテゴリカルな変数の補定
  - 外れ値の影響

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。