経済統計学会関東支部7月例会

公的統計のデータエディティング:
混淆正規分布モデル
及び
多重代入法の適用可能性

平成25年7月6日

独立行政法人　統計センター
統計技術研究課
高橋　将宜

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。

内容（目次）

- 公的統計調査
- データエディティング
- UNECEワークセッション
- 選択的エディティング(Selective Editing)
- 混淆正規分布モデルとSeleMix
- 多重代入法(Multiple Imputation)
- 結語

統計センター

- 国勢調査や消費者物価指数などの国の基本となる統計の作成
- 各府省や地方公共団体の委託による各種統計の作成
- これらに必要な統計技術の研究
- 我が国の中央統計機関の一翼を担う独立行政法人

統計センターの発足

- 1871年：太政官正院に政表課
- 1947年：総理庁統計局
- 1949年：総理府統計局製表部
- 1984年：総務庁統計センター
- 2001年：総務省統計センター
- 2003年：総務省所管の独立行政法人
  統計センター

統計調査の企画から公表まで

調査の企画・設計

調査事項・調査方法の企画、調査票の設計

総務省統計局など

実施調査

統計調査員の任命、調査票の配布・収集

地方公共団体など

製表

調査票の情報を基に行う統計の作成

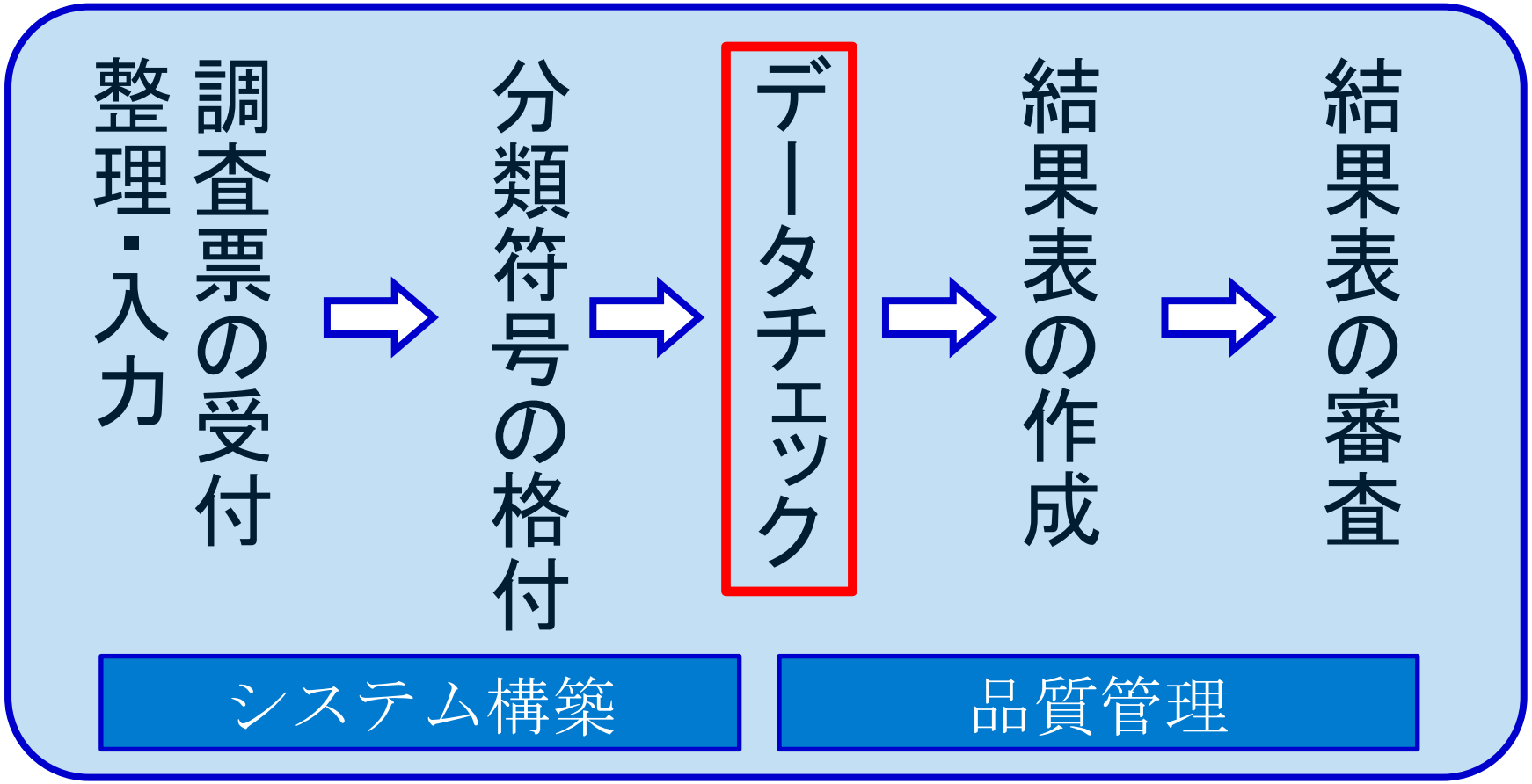
統計センター

分析・公表

結果の分析
公表・報告書の刊行

総務省統計局など

製表の企画・設計



約580万事業所

データエディティングの研究

【経済センサス－活動調査】

- すべての産業分野における事業所及び企業が対象
- 売上高や費用総額といった経理項目などを調査
- 全国的及び地域別に経済の構造を明らかに

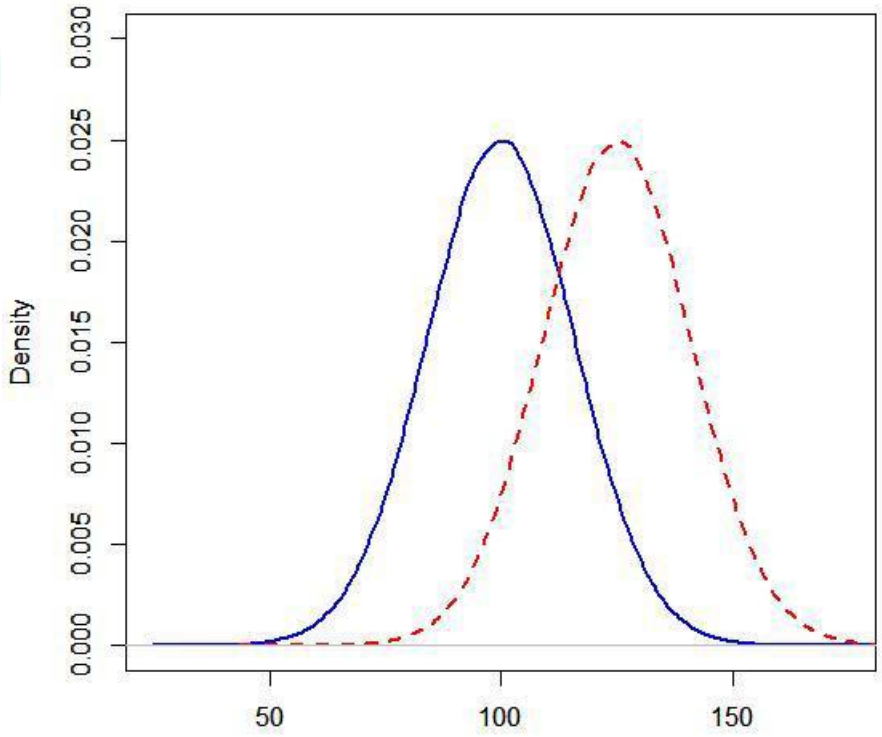
経理項目の記入漏れ、記入誤りが発生

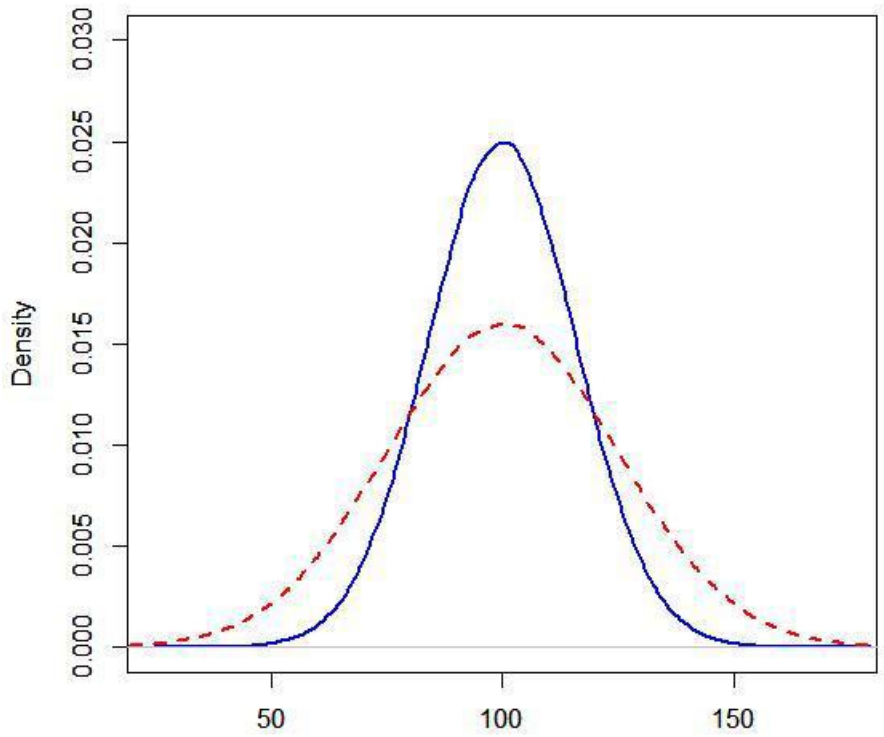
データエディティング及び補定方法の研究が必須

データエディティングとは

- 統計データエディティング(statistical data editing)
  - 単にデータエディティングとも呼ぶ
  - エラーを検出し訂正するプロセス
  - データ収集→エラーの審査→エラーの訂正

体系的エラーとランダムエラー

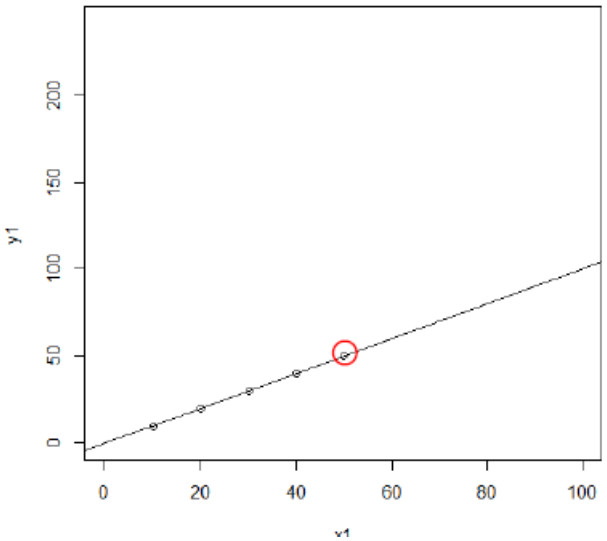


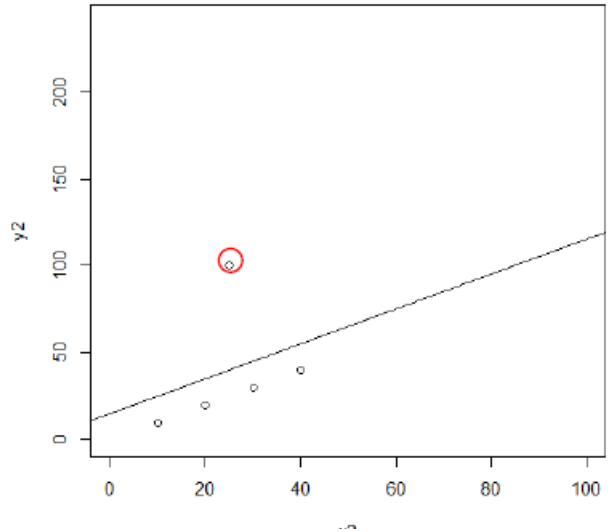


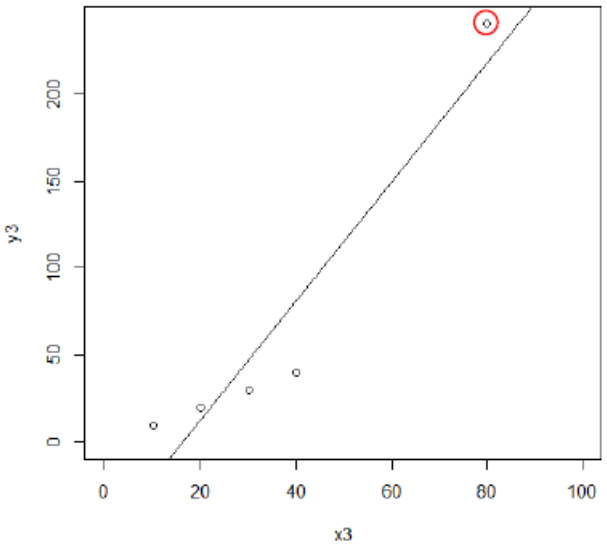
青線＝正データの分布

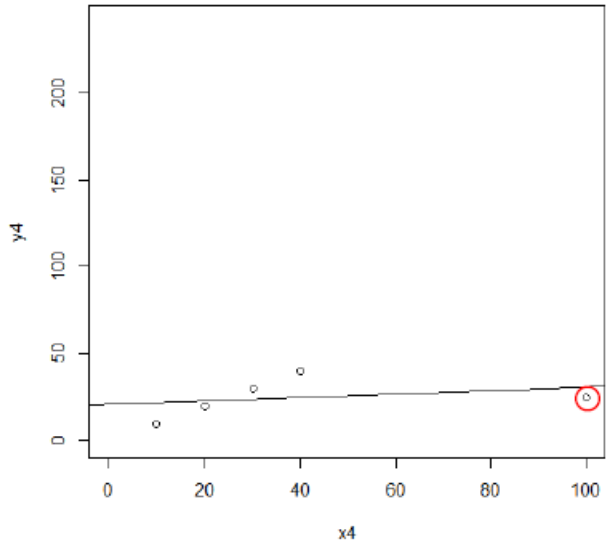
赤線＝エラーデータの分布

外れ値と影響力


y1
0
50
100
150
200
0
20
40
60
80
100


y2
0
50
100
150
200
0
20
40
60
80
100


y3
0
50
100
150
200
0
20
40
60
80
100
x3


y4
0
50
100
150
200
0
20
40
60
80
100
x4

Editing and Imputation (de Waal et. al, 2011)

- 体系的エラーの訂正
- ミクロ分析（スコア関数）
  - ランダムエラーの特定
  - 影響力のあるエラー→人手審査
- 欠測値及びエラーの補定
- 補定値の調整
- マクロ分析
  - 影響力のあるエラーが残っていなければ、ミクロデータとして完成

諸外国のデータエディティングとその起源

- 日本だけではなく、各国公的統計機関において行われている
- 諸外国におけるデータエディティングによるエラー訂正の研究
  - 1950年代から行われている
  - (Nordbotten, 1955)
- 欧米諸国では、統計的手法に基づいたデータエディティングが行われている

統計データエディティングに関するワークセッション

- 国連欧州経済委員会(UNECE)主催
- １年半周期で開催
- 参加者
  - 欧州各国の統計機関
  - 米国、カナダ、オーストラリア、アジアなどの統計機関
  - 欧州統計局(Eurostat)や国連工業開発機関(UNIDO)などの国際機関

統計データエディティングに関するワークセッション

- 主な討議内容
  - データエディティングの革新的な手法や技術開発
  - 統計の加工処理におけるデータエディティングの工程
  - センサスや行政情報源などから得られたデータのエディティングや補定
  - 社会経済的な様々な分野

直近のワークセッション

- 2012年9月24日から26日
- ノルウェーの首都オスロにて開催

http://www.unece.org/stats/documents/2012.09.sde.html

全44論文（英語）をダウンロードして閲覧することができる

直近のワークセッションにおける議題

- 選択的及びマクロエディティング
- エディティングのグローバルな解決策
- 複数情報源及び混合モードからのデータ統合の文脈におけるエディティングと補定
- エディティングプロセスの効率性を分析するためのメタデータ及びパラデータの使用法
- データエディティング及び補定のためのソフトウェアとツール
- 新たな手法
- センサスデータのエディティング及び補定

人手審査と選択的エディティング

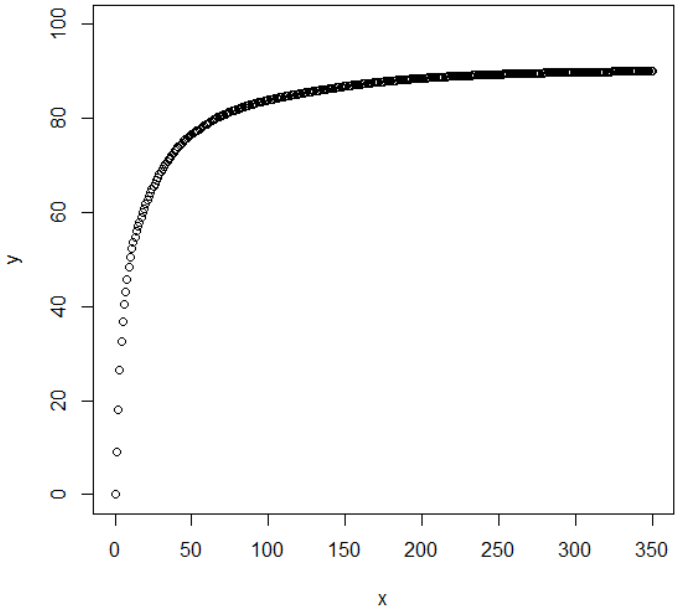
- 人手によるエディティング
  - 審査、処理、照会など
  - 非常に時間と費用がかかる
- 選択的エディティングの目的
  - 前提条件：出力品質の維持
  - 人手によるエディティングの費用をできる限り取り除くこと

選択的エディティングとは

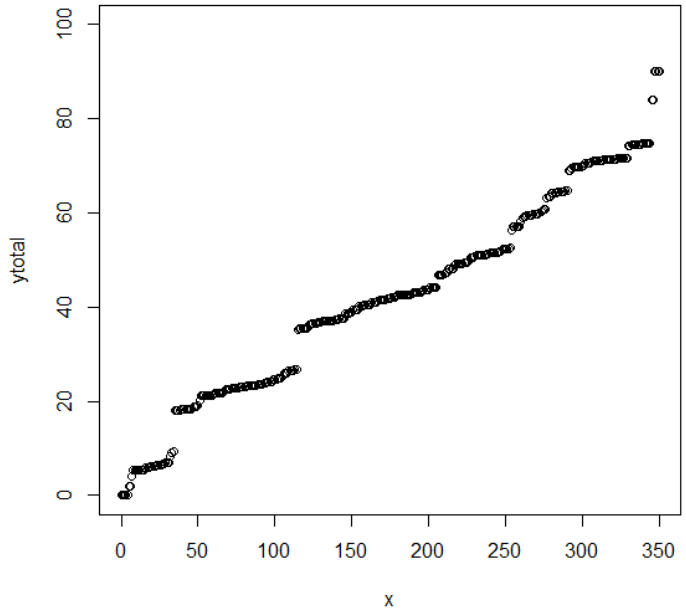
- エラーの可能性があり、調査結果に影響を及ぼし得る回答の修正及び補定を優先化する手法
- 潜在的に影響力のあるエラーを持つユニットを人手審査のために選ぶ
- つまり、重要なものから順番に訂正を行うという趣旨

選択的エディティングの利点

選択的エディティング



人手によるエディティング



訂正	0	1	10	20	50	100	150	200	250	300	350
選択	0%	9.06%	50.50%	61.72%	76.55%	83.91%	86.83%	88.55%	89.38%	89.83%	89.99%
人手	0%	0.03%	5.37%	6.13%	19.24%	24.68%	38.92%	43.68%	52.42%	69.95%	89.99%

選択的エディティングの起源

- Latouche and Berthelot (1990, 1992)
- スコア関数を用いてエディティングに優先付けをした最初期の研究
- 4つの指針
  - 回答ユニットの重要性
  - 疑わしい回答の与える影響の度合い
  - 疑わしい回答の数
  - 項目（変数）の相対的な重要性

選択的エディティングの起源と現状

- 指針は漠然としたもの
- スコア関数を具体的にどのようにして作成するか？
  - 今日においても未解決の問題
  - 各国の統計機関において、各々の調査のデータごとに、多数の手法が存在
  - 汎用的な方法は存在しない

混淆正規分布モデル

$$f(x) = p(2\pi\sigma^2)^{-1/2} exp\left(-\frac{1}{2\sigma^2}[x-\mu]^2\right) + (1-p)g(x)$$

- もし確率密度関数g(x)で汚染している側の分布の分散が大きい場合、あるいは、平均値がμとは大幅に異なる場合、汚染している側の分布から得られた観測値は、外れ値と見なせる

(高橋, 2012, pp.9-19)

混淆正規分布モデルによる選択的エディティング

$$f_{Y|X}(y|x) = pN(y; X\beta, \Sigma) + (1-p)N(y; X\beta, \Sigma_c)$$

- Xを条件としたYの観測値の式
- 同じ切片と同じ傾きを持つが異なる残差分散を持つ2つの回帰モデルを表している
- Σは正しいデータの分散を表す
- Σ_cはエラーデータの分散を表す

(高橋, 2012, pp.9-19)

混淆正規分布モデルによる選択的エディティング

$$SF_i = \left| \frac{w_i(\hat{y}_i - y_i)}{\sum w_i \hat{y}_i} \right|$$

- 上式により算出したSFiの値に応じて、ローカルスコアとグローバルスコアを算出し、観測値を並び替え、影響力の強い順にエディティングを行う

(高橋, 2012, pp.9-19)

SeleMixの検証

- イタリア国家統計局の開発したRパッケージ
  http://cran.r-project.org/web/packages/SeleMix/index.html
- ランダムエラーを影響力のある外れ値として検出するプログラム
- 高橋(2013)にて検証
  - 経済センサス - 活動調査の経理項目のエディティングに向けた研究の一環
  - EDINETデータを模擬試験データとして利用し検証

EDINETとは

- **EDINET**とは、**E**lectronic **D**isclosure for **I**nvestors' **NET**workの略
- 『金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システム』
- 「提出された開示書類について、インターネット上においても閲覧を可能とするもの」（金融庁, 2012）

EDINETを用いた検証

- 4変量における多変量外れ値の検出
- 被説明変数（エラーを含む変数）
  - 売上高
- 説明変数（エラーのない補助変数）
  - 売上原価合計
  - 資本金
  - 事業従事者数

EDINETを用いた検証

- 想定
  - 売上原価への支出↑　⇒　売上↑
  - 資本金↑　⇒　事業規模↑
  - 事業従事者数↑　⇒　事業規模↑
  - 事業規模↑　⇒　売上↑
- 企業データ
  - 分布に偏り
  - 自然対数に変換→正規分布を近似

EDINETデータにおける外れ値とエラー

- EDINETデータ
  - 外れ値は存在
  - エラーは存在しない（非常に稀な虚偽報告の例を除く）
- 選択的エディティングの最終目標
  - 単に外れ値を検出することではない
  - エラーを効率的に抽出し対処すること

人工的エラー

- 報告値（真値）の約15%を人工的にエラー化
  - 2,871観測数のうち、425個のエラー
  - 桁違いによるランダムエラーを模した

検証結果

- 観測数2,871個
  - 外れ値829個
  - 影響力のある外れ値232個
- エラーデータ425個
  - 外れ値として検出できたものは424個
  - 正答率：**99.765%**
- 影響力のある外れ値232個
  - エラーデータであったものは199個
  - 正答率：**85.776%**

課題

- モデルの検出力
  - 設定した閾値の値に応じて変化
  - *SeleMix*プログラムのデフォルト設定
  - 残差エラーが0.01未満のとき、影響力のある外れ値として検出

課題

閾値	0.001	0.005	0.100	0.150	0.200
外れ値	604	295	232	206	167
エラー	367	242	199	178	147
正答率	60.8%	82.0%	85.8%	86.4%	88.0%

- 検出できる外れ値及びエラーの絶対数
  - 閾値の値が大きくなるにつれて減少
- 正答率
  - 閾値の値を大きくすればするほど改善
- 統計的検定の第1種過誤と第2種過誤と同じ

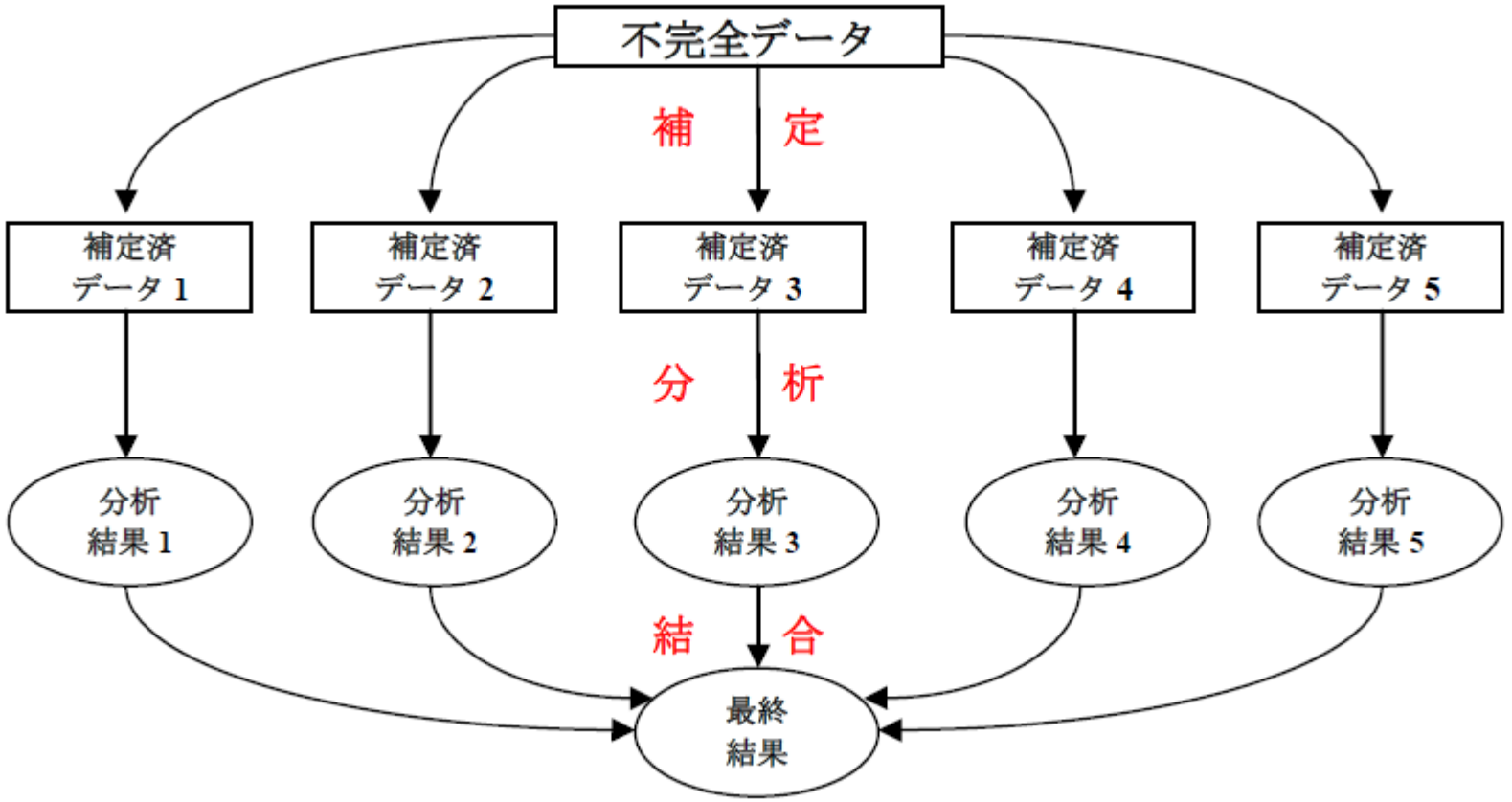
課題

- これまでは、EDINETデータを利用して検証してきた
- 最終的な実用性検証として、経済センサス - 活動調査の実データを用いた検証を行う予定

多重代入法とは

- 補定 = Imputation
- 単一代入法 = Single Imputation
- 多重代入法 = Multiple Imputation
  - 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

多重代入法の概念図



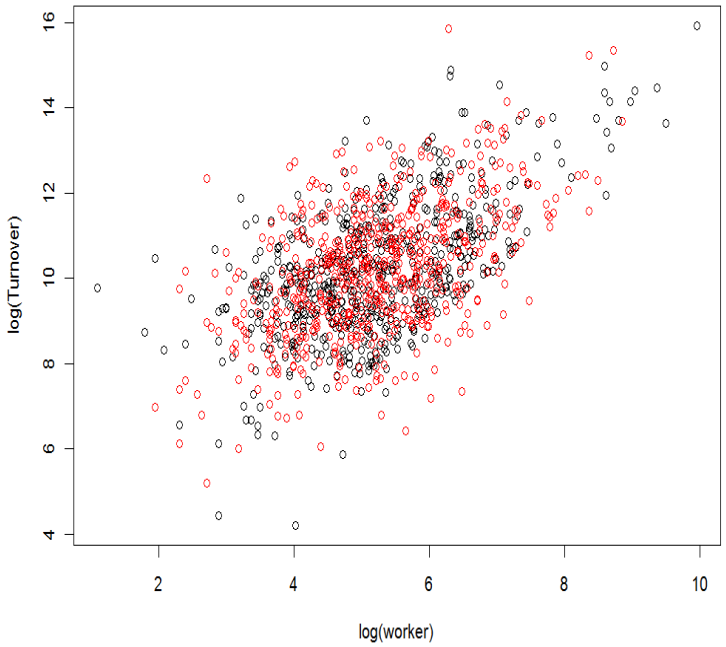
M = 5の場合

多重代入法と単一代入法の比較

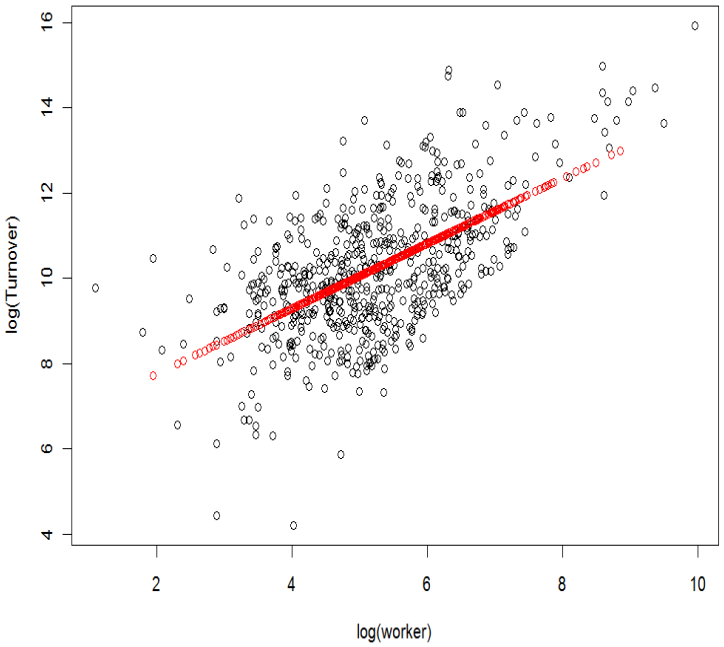
- 高橋, 伊藤(2013a)
  - 点推定値：多重代入法と単一代入法で差なし
  - 標準偏差、標準誤差、分布：多重代入法の優位

多重代入法と単一代入法の比較

多重代入法



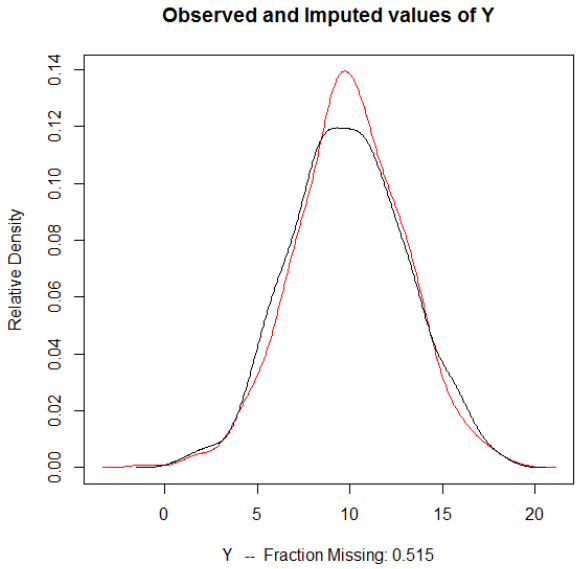
単一代入法：確定的補定



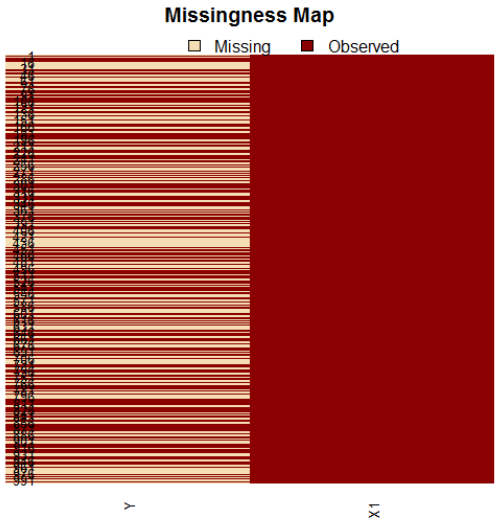
注：欠測メカニズムMCARの場合

多重代入法の診断手法

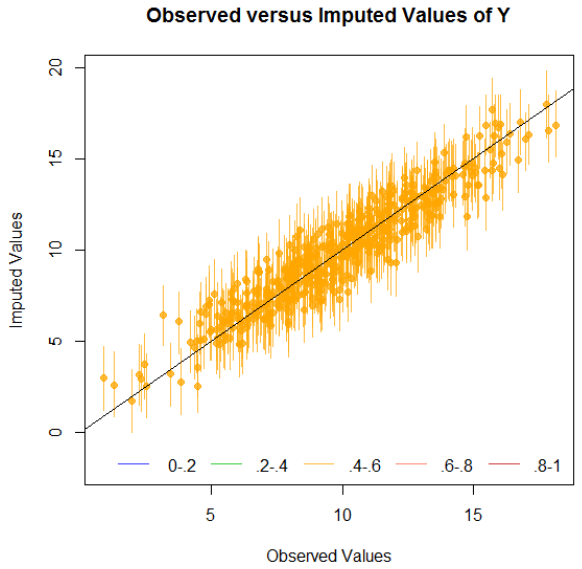
観測値と補定値の密度



欠測地図



過剰補定

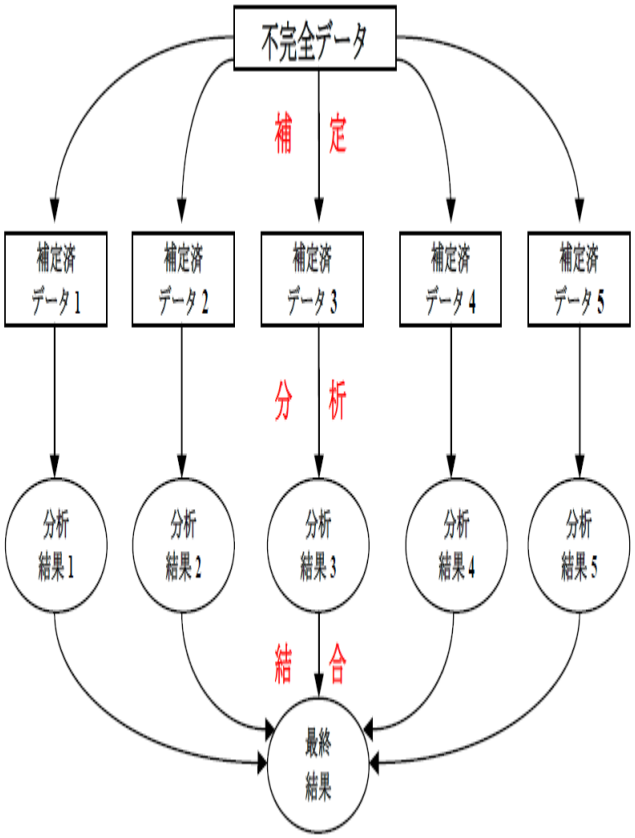


注：欠測メカニズムMCARの場合

様々な多重代入法アルゴリズム

- Rubin (1978)
  - 多重代入法の理論的概念が発案
  - 数十年の時が経過
- 事後分布からの無作為抽出の実装
  - 計算上、難しい

→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



マルコフ連鎖モンテカルロ法(MCMC)

- Markov chain Monte Carloの略
- 1980年代にDonald B. Rubinによって提唱されたオリジナルの多重代入法
- ベイズ統計学の枠組みで構築

マルコフ連鎖モンテカルロ法(MCMC)

- データ拡大法(DA: Data Augmentation)
  - MCMCの計算アルゴリズム
  - 繰り返し手法を用いて推定値を改善していく方法
- ソフトウェア
  - RパッケージNorm 3.0.0
  - SAS PROC MI 9.3

MCMCの長所と短所

- 長所
  - Rubinのオリジナルの多重代入法を再現
  - Properな補定
- 短所
  - 多変量正規分布を仮定
  - カテゴリカル変数の補定は不得手
  - 大規模データセットに対応できないとされる

完全条件付指定(FCS)

- Fully Conditional Specificationの略
- MCMCの代替法として提唱
- 各々の不完全な変数に対して補定モデルを構築
- それぞれの変数に対して補定値を繰り返し作成

完全条件付指定(FCS)

- ソフトウェア
  - RパッケージMICE 2.13
  - PASW Missing Values 18
  - SOLAS 4.01

FCSの長所と短所

- 長所
  - 各々の変数に対して補定モデルを構築するため、正規分布を仮定しない
  - カテゴリカル変数の補定にも対応可
  - 非常にフレキシブル
- 短所
  - 理論的根拠が希薄であり、Properな補定かどうか、議論の余地がある

EMBアルゴリズム

- 伝統的な期待値最大化法(EM: Expectation-Maximization)にノンパラメトリック・ブートストラップ法を応用
- ブートストラップにより不確実性を反映させ、EMによりパラメータを推定し、補定を行う
- ソフトウェア
  - RパッケージAmelia II (version 1.6.1)

EMBの長所と短所

- 長所
  - コレスキー分解を行う必用がなく、計算効率が高いとされる
  - 大規模データセットの処理が得意とされる
- 短所
  - 多変量正規分布を仮定

比較検証

- 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
- 現在、経済センサスの実データとシミュレーションデータを用い、様々な多重代入法アルゴリズムを比較中
- 詳細は、9月の全国研究大会にて報告予定

適用可能性（ケース1）

- SeleMixにより外れ値（エラー）を検出
- エラーを削除し欠測値と同様に扱い、多重代入法によって補定

適用可能性（ケース2）

- 欠測値の補定を多重代入法によって行う
- 多変量正規分布を想定しているため、外れ値の影響を考慮に入れる必要がある
- SeleMixにより、外れ値を検出して対処した上で多重代入を行なう

データキュレーション(Data Curation)
丸山宏先生（ESTRELA 2013年6月号, p.5）

- データの選択、前処理、クレンジング
- センサーとITの発達により、非常に大規模な生データが生産されている（ビッグデータの時代）
- しかし、ビッグデータには、欠測値や外れ値が含まれている

データキュレーション(Data Curation)
丸山宏先生（ESTRELA 2013年6月号, p.5）

- 異なる条件下で収集された複数のデータセットを統合するには、バイアスを修正する必要がある
  - データのフォーマットや単位系の変換
  - データ項目の意味の関連付け
- 目的に応じ、どのデータに、どのような前処理やクレンジングを施して使うかというノウハウが重要

参考文献（英語）

- de Waal, Ton, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.
- Latouche, Michel and Jean-Marie Berthelot. (1990). "Use of A Score Function for Error Correction in Business Surveys at Statistics Canada," *Proceedings of the International Conference on Measurement Errors in Surveys.*
- Latouche, Michel and Jean-Marie Berthelot. (1992). "Use of A Score Function to Prioritize and Limit Recontacts in Editing Business Surveys," *Journal of Official Statistics* vol.8, no.3: 389-400.
- Nordbotten, Svein. (1955). "Measuring the Error of Editing Questionnaires in a Census," *American Statistical Association Journal* vol.55: pp.364-369.
- 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.
- 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, United Nations Economic Commission for Europe*, Oslo, Norway, 24-26 September 2012.

参考文献（日本語）

- 金融庁. (2012). **EDINET**金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システム. http://info.edinet-fsa.go.jp/. 2013年6月6日アクセス.
- 高橋将宜. (2012). 「諸外国のデータエディティング及び混淆正規分布モデルによる多変量外れ値検出法についての研究」, 『製表技術参考資料17』, 独立行政法人統計センター.
- 高橋将宜, 伊藤孝之. (2013a).「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.
- 高橋将宜, 伊藤孝之. (2013b). 「経済センサスの欠測値補定～様々な多重代入法アルゴリズムの比較～」, 経済統計学会第57回全国研究大会. 報告予定.
- 高橋将宜. (2013).「諸外国における最新のデータエディティング事情及び混淆正規分布モデルによる多変量外れ値検出法の検証」, 『製表技術参考資料23』, 独立行政法人統計センター. 刊行予定.
- 丸山宏. (2013). 「データに基づく意思決定」, *ESTRELA* no.231: pp.2-7.

ご清聴ありがとうございました