2024年11月18日
公的統計ミクロデータ利活用に関する研究集会
「匿名データ・オーダーメード集計の利用による研究報告」

公的統計における外れ値を考慮した欠測データの新たな代入法の提案

長崎大学
情報データ科学部 准教授
高橋 将宜
博士（理工学）

概要

- 序論
- 欠測データの問題点と対処法
- これまでの研究の紹介
- 新たな頑健比率代入法の紹介
- モンテカルロ・シミュレーション：設定
- モンテカルロ・シミュレーション：結果
- 結語

序論

経済データの特徴

- 公的経済統計のデータ
  - 調査票によって収集
  - 無回答（欠測）が発生する
- 経済統計の対象
  - 中小企業から大企業まで多様な規模の企業
  - 一般的に，経済データの分布は右に裾が長い
  - 分散が均一でない
  - 外れ値の考慮

本研究の目的

- さまざまなパターンの外れ値があっても高い精度で欠測値を処理できる方法を提案する.

- 提案した新たな方法をモンテカルロ・シミュレーションで検証する際に，全国消費実態調査（2004年）の匿名データを参考にしてシミュレーションの設定をした.
  - 匿名データの利用による研究報告

本研究成果：書誌情報

- Takahashi, M. (2022) A new robust ratio estimator by modified Cook's distance for missing data imputation. *Japanese Journal of Statistics and Data Science* 5 (2), pp.783–830.
  - Impact Factor: 1.1
  - 統計関連学会連合の機関誌
  - DOI：https://doi.org/10.1007/s42081-022-00164-0
  - 無料閲覧用URL：https://rdcu.be/dZC9G

欠測データの問題点と対処法

完全データ

企業	売上高	従業員数
A	100	4
B	300	15
C	200	11

$$\overline{\text{売上高}} = \frac{100 + 300 + 200}{3} = 200$$

欠測データ

企業	売上高	従業員数
A	100	4
B	欠測値	15
C	200	11

$$\overline{\text{売上高}} = \frac{100 + \text{欠測値} + 200}{3} = \frac{300 + \text{欠測値}}{3}$$

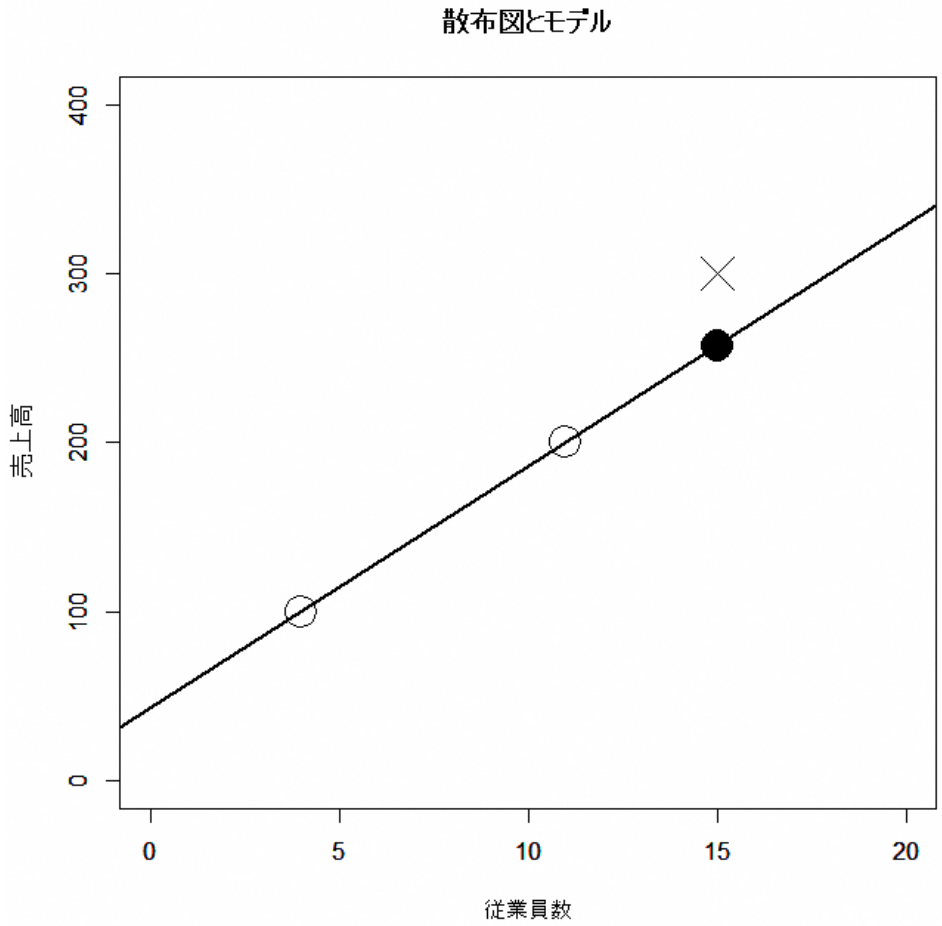
リストワイズ除去

企業	売上高	従業員数
A	100	4
C	200	11

$$\overline{\text{売上高}} = \frac{100 + 200}{2} = 150$$

3社の平均200と異なる

散布図とモデル



$$売上高 = 42.86 + 14.29従業員数$$

代入済みデータ

企業	売上高	従業員数
A	100	4
B	257	15
C	200	11

$$\overline{\text{売上高}} = \frac{100 + 257 + 200}{3} = 186$$

3社の平均200に近づく

詳しくは

今日の話題は単一代入法
母集団における合計や平均値などの点推定値の推定に有効

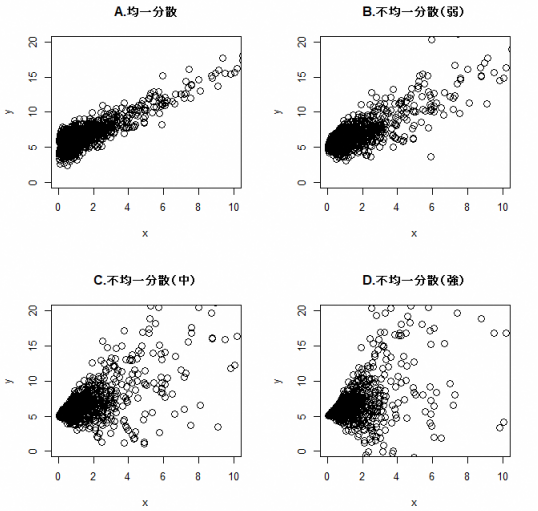
多重代入法
標準誤差を使った母数の推定に有効
詳しくは，拙著『欠測データ処理』をご覧ください.

https://www.kyoritsu-pub.co.jp/book/b10003896.html

これまでの研究の紹介

経済データの分布の特徴

- 現実の経済データの分布
  - 右に裾が長い
- 図B～図D
  - 横軸の変数xが増加すると縦軸の変数yのばらつきが増加



- 年収と食費の例
  - 年収が増えると，外食したり，質素にしたりするため，年収は食費の分散に影響
  - 誤差項の分散が不均一であることを含意
- 図A以外
  - 通常の最小二乗法は適切ではない

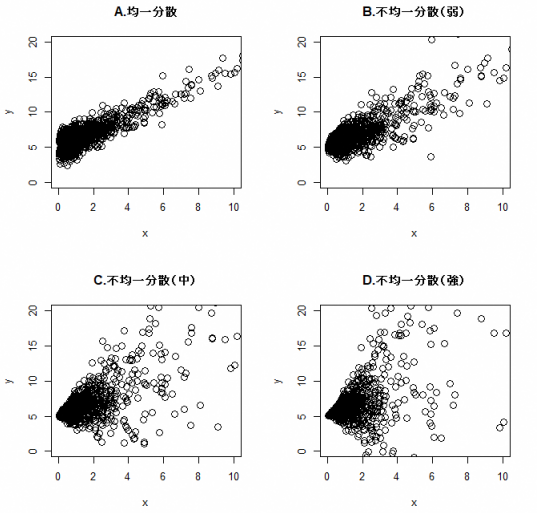
高橋（2017）

- 国連欧州経済委員会の加盟20か国に聞き取り調査
  - 公的経済統計の欠測値処理について，比率代入法が最もよく用いられる．
- 比率代入法

$$Y_i = \beta X_i + \varepsilon_i$$

$$\hat{\beta} = \frac{\bar{Y}}{\bar{X}}$$

  - 図Bの不均一分散に有効

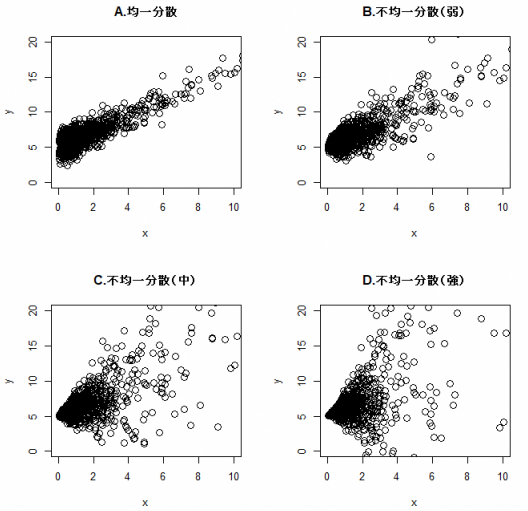


Takahashi, Iwasaki, and Tsubaki（2017）

- 重み付き最小二乗法
  - 比率代入法モデルも，通常の最小二乗法モデルも，1つのモデルに一般化できる．

$$Y_i = \beta X_i + \varepsilon_i, \varepsilon_i \sim N\left(0, X_i^{2\theta}\sigma^2\right)$$

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}}$$



- 図A～図Dのように，誤差項の不均一分散の程度に応じて，適切なモデルが複数あることを明らかにした．

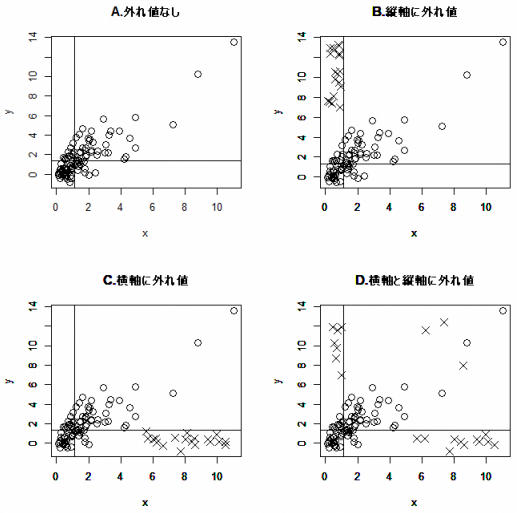
本研究の問い

- 誤差項の分散が不均一な経済データにおいて，欠測値の適切な処理方法をどのように決定するか？
- Takahashi, Iwasaki, and Tsubaki（2017）
  - 重み付き最小二乗法の枠組みを用いて比率代入法を一般化した.
  - ここから拡張すべき事項

→次のスライド

拡張すべき事項

- 欠測値を処理する際に外れ値の影響をどのように考慮するか？
- 国連欧州経済委員会の加盟14か国に聞き取り調査
  - 6か国が外れ値の影響に対処していた
  - 中央値，トリム平均値，繰り返し加重最小二乗法，ウィンザー化平均値など，統一的な見解は存在しない
  - これらの伝統的手法は，図Bの縦軸側の外れ値にのみ対応可能



解決策

- Takahashi, Iwasaki, and Tsubaki（2017）
  - 重み付き最小二乗法の枠組みにより比率代入法モデルを一般化した残差を導出した.
  - この残差は，通常の最小二乗法による残差とは異なって，不均一分散の度合いに応じて変化する特殊な形をしている.

解決策（続き）

- クックの距離
  - 縦軸方向の外れ値を検出するスチューデント化残差と，横軸方向の外れ値を検出するテコ比（ハット値）により構成される.
- Takahashi, Iwasaki, and Tsubaki（2017）において導出した残差を活用することで拡張できると考えた.

新たな頑健比率代入法の紹介

研究の内容

- クックの距離とトリム平均値の考え方を応用した新たな頑健比率代入法を提案
  - 横軸の外れ値に対しても，縦軸の外れ値に対しても頑健

クックの距離：通常の最小二乗法

$$D_i = \frac{e_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

クックの距離

$$e_i' = \frac{e_i}{s\sqrt{1-h_i}}$$

スチューデント化残差
Y軸方向の外れ値

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$$

ハット値
X軸方向のテコ比（外れ値）

比率代入法（$\theta = 0.5$）における残差

$$Y_i = \beta X_i + \varepsilon_i$$ 母集団モデル

$$\varepsilon_i \sim N\left(0, \sigma^2 X_i^{2\theta}\right)$$ 誤差項

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}}$$ 重み付き最小二乗法

$$e_{Ri} = \frac{Y_i - \hat{\beta}_{WLS} X_i}{X_i^{\theta}}$$ 残差

提案した内容1：頑健比率代入法のアルゴリズム

- 平均値の比率により$\hat{\beta}$を推定

$$\hat{\beta} = \frac{\bar{Y}}{\bar{X}}$$

- 残差を算出

$$e_{Ri} = \frac{Y_i - \hat{\beta} X_i}{X_i^{0.5}}$$

- スチューデント化残差を算出

$$e'_{Ri} = \frac{e_{Ri}}{s\sqrt{1 - h_i}}$$

- クックの距離を算出

$$D_i = \frac{e'^2_{Ri}}{k+1} \times \frac{h_i}{1 - h_i}$$

- クックの距離に基づいて，データを降順に並べて外れ値をトリミング
- 再度，平均値の比率により$\hat{\beta}_{cook}$を推定し，$\hat{Y}_i = \hat{\beta}_{cook} X_i$によって代入値を生成

提案した内容2：外れ値の自動検出方法

- 外れ値を削除した場合，決定係数は上昇する傾向がある．外れ値を削除すればするほど，決定係数は上昇し続けるものの，大きな影響力を持つ外れ値を削除し終えると，決定係数の上昇率はほぼ一定になってくる．この性質を利用して，外れ値の自動検出も実装した．

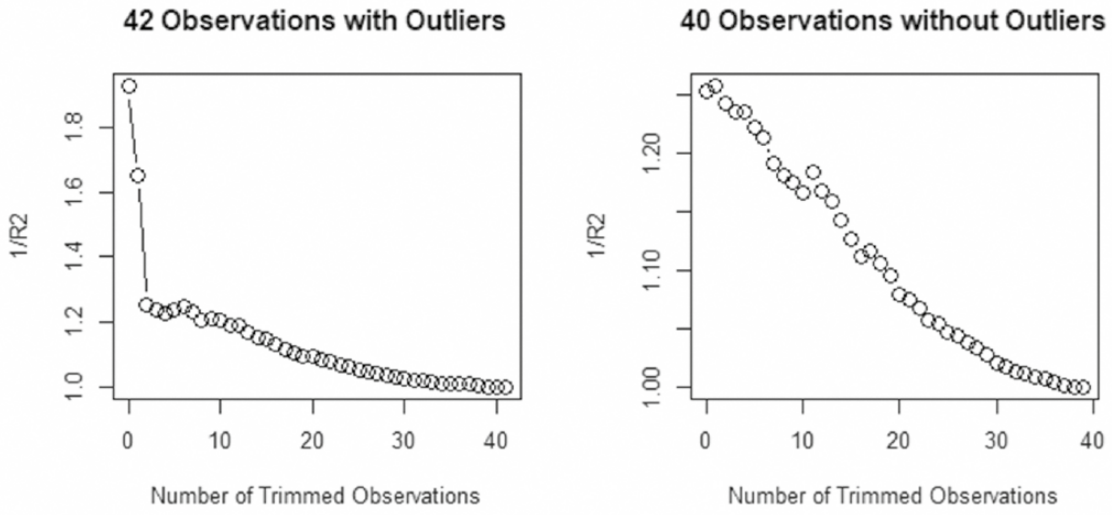


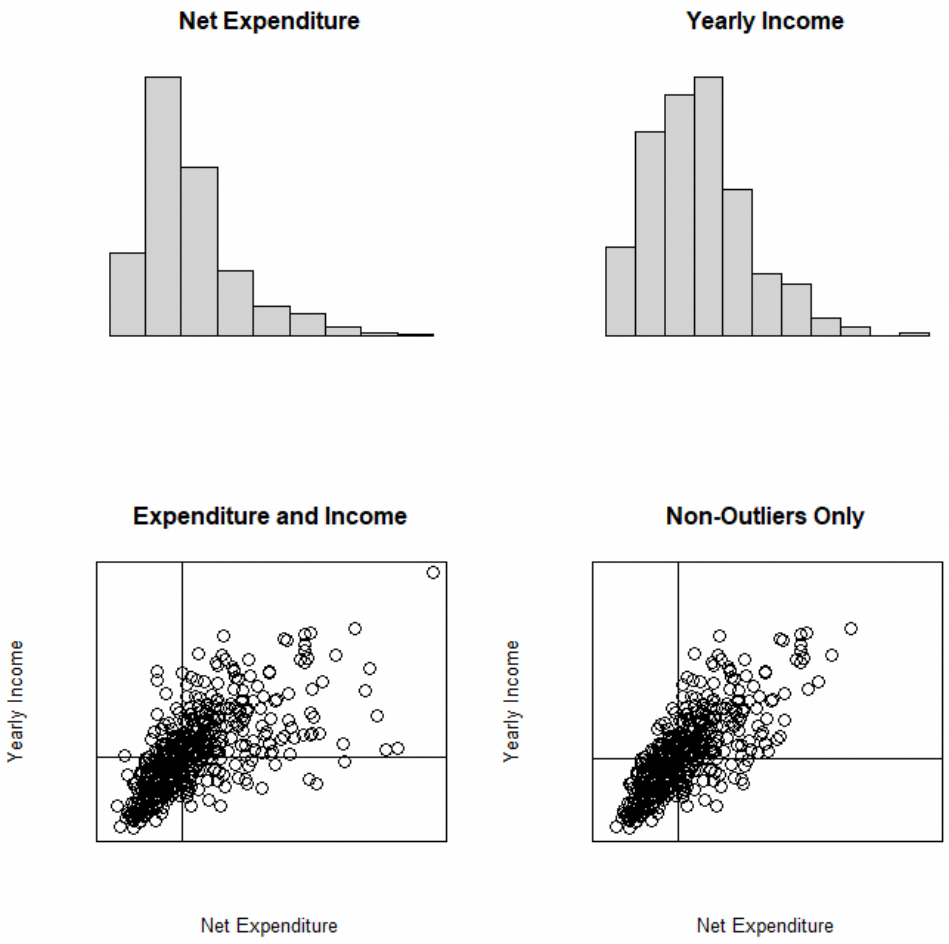
Fig. 1 Examples of the scree-like plot to detect the number of potential outliers for Table 1

TC-ratio estimator

- ratio estimator with trimming based on Cook’s distance

モンテカルロ・シミュレーション：設定

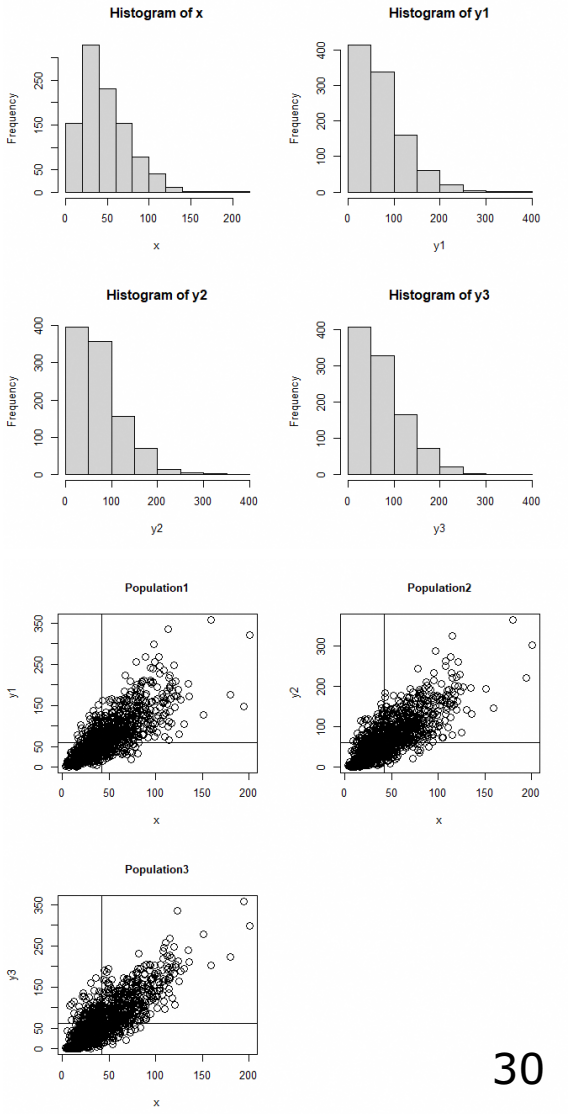
具体例：全国消費実態調査（2004年）の匿名データ



注：秘匿の目的のため，軸は意図的に非表示としている

母集団モデル1～3：ガンマ分布

- 変数x：ガンマ分布で生成
  - $x \sim gamma(3, 48)$
- $n = 1000$
- $\mu_y = \beta x$
  - $\beta = 1.5$
- $\sigma_y^2 = d^2 x^{2g}$
  - 母集団モデル1：$d = 1.84,\ g = 0.75$
  - 母集団モデル2：$d = 5.13,\ g = 0.50$
  - 母集団モデル3：$d = 13.78,\ g = 0.25$
- 参考文献
  - Lee et al. 1994, p.236
  - Rao and Sitter 1995, p.455
  - Sitter and Rao 1997, p.69
  - Haziza and Valée 2020



母集団モデル4～5：一様分布と正規分布

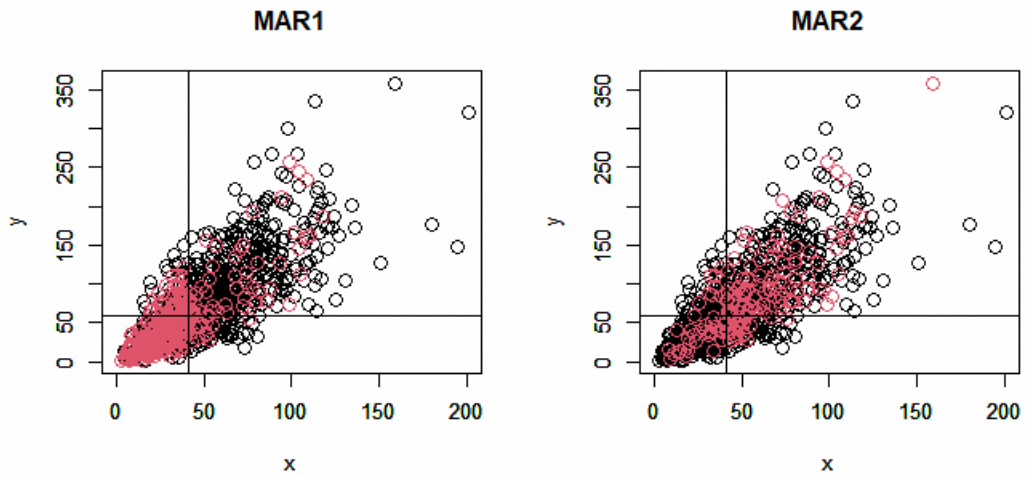
- 変数x：一様分布で生成
  - Zou et al. 2010, p.871
  - Wada and Sakashita 2017, p.3

- 変数x：正規分布で生成
  - Zou et al. 2010, p.871
  - Lui 2020, p.140

欠測メカニズム：2種類

注1：MAR = Missing At Random
注2：外れ値は欠測させない

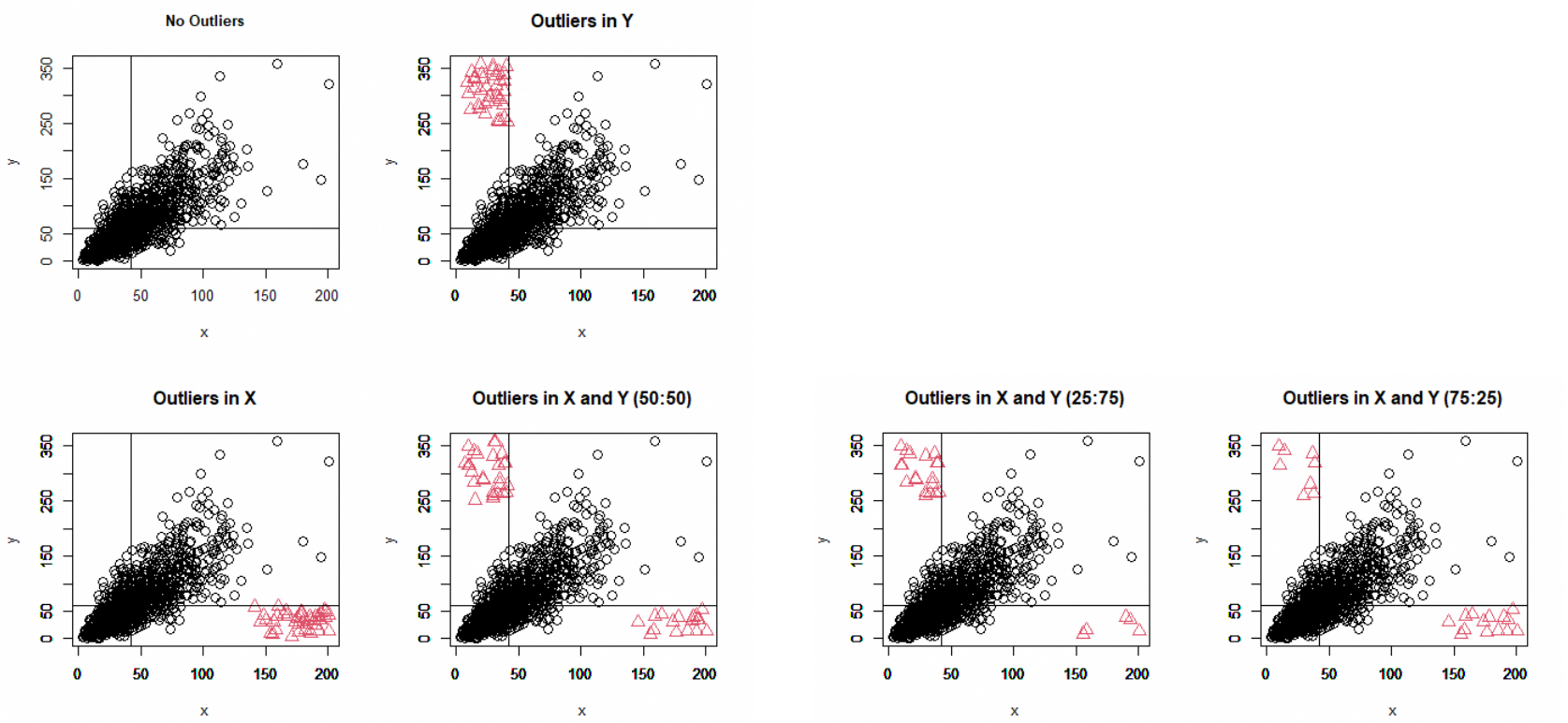
- MAR1：xが小さいとき，yの欠測が発生しやすい
- MAR2：xが大きいとき，yの欠測が発生しやすい



- 欠測率：30%
  - 参考：Schenker et al.(2006, p.925)
  - National Health Interview Surveyの収入と所得の欠測率は約30%

外れ値

- 外れ値の割合：0%, 1%, 5%, 10%



その他の設定と評価方法

- シミュレーション回数：10,000回
- 推測対象：完全データの$\bar{Y}$

母集団モデル：5
欠測メカニズム：2
外れ値の割合：3
外れ値の位置：5
5*2*3*5+10=160

- 偏り

$$\mathrm{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

  - Schafer and Graham (2002, p.157)
  - 標準誤差の推定値の1/2よりも大きい場合，偏りは問題視される.
  - 今回のデータでは，標準誤差は1.7なので，0.85未満であれば問題なしと判断する.

- RMSE

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2}$$

比較対象の手法

- Comp：完全データ
- LD：リストワイズ除去（欠測データを削除）
- Ratio：通常の比率代入法（頑健ではないもの）
- M-1：繰り返し加重最小二乗法を応用した平均値の比率による代入法（less robust）
- M-2：繰り返し加重最小二乗法を応用した平均値の比率による代入法（more robust）
- Med：中央値の比率による代入法
- Trim：トリム平均値の比率による代入法(5%)
- Wins：ウィンザー化平均値の比率による代入法(5%)
- C1：クックによる頑健比率代入法1（提案手法）
- C2：クックによる頑健比率代入法2
  - （外れ値の閾値=4/(n-k-1)）

モンテカルロ・シミュレーション：結果

ガンマ分布（$d = 1.84, g = 0.75$），MAR1：偏り

%X	%Y	Comp	LD	Ratio	M-1	M-2	Med	Trim	Wins	C-1	C-2
0.00	0.00	-0.017	9.022	-0.012	-0.303	-0.721	-1.425	-0.463	-0.200	-0.326	-0.885
0.00	1.00	-0.019	9.034	1.641	-0.016	-0.152	-0.830	0.577	1.123	0.011	0.076
0.00	5.00	-0.009	9.029	7.888	4.001	1.222	1.859	6.823	8.087	0.386	4.014
0.00	10.00	-0.011	9.026	14.957	11.086	6.781	5.904	14.654	15.438	9.911	10.288
1.00	0.00	-0.010	9.043	-1.066	-1.023	-1.045	-2.008	-1.243	-1.127	-0.398	-1.009
5.00	0.00	0.005	9.048	-3.999	-3.935	-3.505	-4.145	-4.254	-4.284	-0.838	-1.085
10.00	0.00	0.017	9.051	-6.159	-6.281	-6.441	-6.061	-6.591	-6.447	-5.921	-3.479
0.25	0.75	0.014	9.058	0.954	-0.265	-0.316	-1.097	0.123	0.537	-0.124	-0.012
1.25	3.75	-0.022	9.018	4.185	0.779	-0.676	0.126	3.422	4.445	0.186	1.594
2.50	7.50	-0.019	9.024	7.409	3.816	0.597	1.835	7.332	7.660	0.395	4.152
0.50	0.50	0.014	9.062	0.261	-0.520	-0.522	-1.396	-0.341	-0.040	-0.242	-0.125
2.50	2.50	-0.027	9.012	1.029	-1.561	-1.782	-1.439	0.219	0.901	-0.175	-0.787
5.00	5.00	-0.001	9.039	1.746	-1.359	-3.087	-1.427	1.045	1.674	-0.534	-1.178
0.75	0.25	-0.001	9.046	-0.416	-0.775	-0.761	-1.706	-0.801	-0.595	-0.334	-0.320
3.75	1.25	-0.035	9.005	-1.685	-3.032	-2.629	-2.882	-2.419	-2.249	-0.374	-2.060
7.50	2.50	-0.012	9.030	-2.652	-4.665	-4.855	-4.163	-3.673	-2.942	-0.907	-4.721

注1：0.85未満
注2：%X indicates the percentage of outliers in x.
注3：%Y indicates the percentage of outliers in y.

ガンマ分布（$d = 1.84, g = 0.75$），MAR1：RMSE

%X	%Y	Comp	LD	Ratio	M-1	M-2	Med	Trim	Wins	C-1	C-2
0.00	0.00	1.729	9.296	1.852	1.870	1.973	2.326	1.894	1.855	1.878	2.034
0.00	1.00	1.723	9.304	2.515	1.828	1.834	2.011	1.931	2.176	1.841	1.846
0.00	5.00	1.739	9.304	8.302	4.529	2.283	2.685	7.205	8.493	1.903	4.665
0.00	10.00	1.723	9.297	15.404	11.509	7.219	6.267	15.051	15.883	10.687	10.794
1.00	0.00	1.742	9.316	2.122	2.102	2.113	2.712	2.206	2.147	1.897	2.093
5.00	0.00	1.749	9.322	4.395	4.335	3.949	4.515	4.616	4.652	2.044	2.143
10.00	0.00	1.729	9.321	6.413	6.530	6.685	6.301	6.822	6.688	6.199	3.921
0.25	0.75	1.733	9.332	2.119	1.865	1.876	2.148	1.857	1.943	1.860	1.855
1.25	3.75	1.730	9.290	4.733	2.058	1.958	1.881	3.979	4.969	1.869	2.576
2.50	7.50	1.715	9.295	7.880	4.420	2.034	2.661	7.748	8.119	1.895	4.785
0.50	0.50	1.722	9.329	1.862	1.892	1.896	2.292	1.851	1.830	1.847	1.834
2.50	2.50	1.731	9.286	2.239	2.406	2.543	2.329	1.873	2.158	1.853	2.050
5.00	5.00	1.745	9.310	2.787	2.348	3.596	2.328	2.245	2.728	1.934	2.405
0.75	0.25	1.742	9.319	1.885	1.986	1.983	2.497	1.990	1.923	1.873	1.873
3.75	1.25	1.745	9.284	2.519	3.537	3.199	3.402	3.022	2.909	1.895	2.873
7.50	2.50	1.727	9.306	3.259	5.004	5.179	4.527	4.094	3.487	2.085	5.080

注1：%X indicates the percentage of outliers in x.
注2：%Y indicates the percentage of outliers in y.

それ以外の結果

- 母集団モデル：5
- 欠測メカニズム：2
- 外れ値の割合：3
- 外れ値の位置：5
- 5*2*3*5+10=160

Table 6 Summary of the overall results in 160 data patterns

	Comp	LD	Ratio	M-1	M-2	Med	Trim	Wins	C-1	C-2
Unbiased	160	0	39	55	57	39	55	41	*114*	60
RMSE	NA	3	11	22	18	4	8	9	*106*	26

For RMSE comparisons, Comp is excluded; thus, displayed as NA (not applicable)

結語

結語

- クックの距離を比率代入法に応用した新たな頑健比率代入法を提案した.
  - Xにおける外れ値に対しても，Yにおける外れ値に対しても頑健であることを示した.