2016年度 統計関連学会連合大会 [H28.09.07]
企画セッション22「公的統計の精度改善に関する多角的取り組み」

平成28年経済センサス - 活動調査のための
ロバストな比率補定の方法について

(独)統計センター　統計技術研究課
和田　かず美
坂下　佳一郎

研究の目的

平成28年経済センサス - 活動調査における企業の主要経理項目（売上、費用、給与）の欠測値について、比率を用いた補定(imputation)」を行うため、平成27年度から以下の二点について研究を行っている

研究成果であるロバスト比率補定推定量と、補定のドメイン設定方法は、平成28年調査の集計で採用

- 外れ値の影響緩和 ← **本日の発表**
- 補定に最適なドメイン（補定値の推定を行う単位）の設定

[参考]　経済センサス - 活動調査研究会（第4回）

http://www.stat.go.jp/info/kenkyu/e-census/katsuken/sidai04.htm

目　次

I. 補定(imputation)のための比推定量

I-1　比率補定について

I-2　外れ値の影響緩和

I-3　ロバスト化の効果

I-4　極端に大きな値を排除する効果

比率補定のモデル

$$y_i = rx_i + \epsilon_i$$

y は目的変数（補定対象項目）、x は説明変数（補定対象項目と相関が高い項目）で、r は y と x との比率

通常 r は未知なので、x_i と y_i の両方の値が存在するデータを用いて、以下のように推定する

$$\hat{r} = \frac{\sum_{k \in \mathrm{obs}} y_i}{\sum_{k \in \mathrm{obs}} x_i}$$

obs:欠測なくセットで計測されている観測値

De Waal et al. (2011) *Handbook on Statistical Data Editing and Imputation*, Wiley handbooks in survey methodology, John Wiley & Sons, p. 244-245.

誤差項の形

通常の比率補定のモデルの場合、誤差項ϵ_iの分散は、x_i に比例:　$\epsilon_i \sim N(0, \sigma^2 x_i)$

ロバスト化のために、誤差項を x_i と関係を持たない $\varepsilon_i \sim N(0, \sigma^2)$ という形で定式化したい

$\varepsilon_i = \epsilon_i / \sqrt{x_i}$ なので、モデル式を$\sqrt{x_i}$ で割り、

$$\frac{y_i}{\sqrt{x_i}} = r\sqrt{x_i} + \varepsilon_i$$

$$y_i = r x_i + \varepsilon_i \sqrt{x_i}$$

Cochran, W. G. (1977) *Sampling Techniques*, 3rd ed., John Wiley & Sons.

さらに一般化

誤差項が x_i のべき乗 x_i^{β} に比例すると仮定

※ βは任意の定数

x_iと関係を持たない等分散の誤差項

モデル

$$\frac{y_i}{x_i^{\beta}} = r x_i^{(1-\beta)} + \varepsilon_i$$

推定量

$$\hat{r} = \frac{\sum y_i x_i^{1-2\beta}}{\sum x_i^{2(1-\beta)}}$$

誤　差

$$\check{\varepsilon}_i = \frac{y_i - \hat{r} x_i^{\beta}}{x_i^{\beta}}$$

β=1のとき:

$$\frac{y_i}{x_i} = r + \varepsilon_i, \qquad \varepsilon_i = \frac{y_i}{x_i} - r \sim N(0, \sigma^2)$$

$$y_i = rx_i + \varepsilon_i x_i, \qquad \hat{r} = \frac{1}{n}\sum \frac{y_i}{x_i}$$

A′

β=1/2のとき: 通常の比率補定の推定量のモデル

$$\frac{y_i}{\sqrt{x_i}} = r\sqrt{x_i} + \varepsilon_i, \quad \varepsilon_i = \frac{y_i}{\sqrt{x_i}} - r\sqrt{x_i} \sim N(0, \sigma^2)$$

B′

$$y_i = rx_i + \varepsilon_i\sqrt{x_i}, \quad \hat{r} = \frac{\sum y_i}{\sum x_i}$$

β=0のとき: 切片のない単回帰モデル

$$y_i = rx_i + \varepsilon_i, \quad \varepsilon_i = y_i - rx_i \sim N(0, \sigma^2)$$

C′

$$\hat{r} = \frac{\sum y_i x_i}{\sum x_i^2}$$

推定量の特徴

推定量(A′)

- 各観測データの比の平均なので、値の大きい特定の観測値にひきずられにくい
- $\hat{r}$の推定が荒れる可能性がある

対策1: ロバスト化

対策2: 規模が大きすぎる観測値は推定から除外する

推定量(B′)

- 各観測データの和の比なので、値の特に大きな観測値だけで比率の推定値が決まる
- $\hat{r}$の推定値が安定する

I. 補定(imputation)のための比推定量

I-1　比率補定について

I-2　外れ値の影響緩和方法

I-3　ロバスト化の効果

I-4　極端に大きな値を排除する効果

外れ値の影響緩和方法

外れ値の考え方:
　誤差項の裾が正規分布よりも長いときの裾部分

外れ値の影響緩和の方法:
　誤差項が大きい観測値に加重し、推定量に与える影響を調整する

計算アルゴリズム: IRLS（繰返し加重最小二乗法）
　計算が簡便で収束が速い

回帰M-推定量を比率補定に拡張

Holland, P. W. and Welsch, R. E. (1977) Robust Regression Using Iteratively Reweighted Least-Squares, Communications in Statistics – Theory and methods, A6(9), pp.813-827

和田(2012) 多変量外れ値の検出～繰返し加重最小二乗(IRLS)法による欠測値の補定方法～, 統計研究彙報, 第69号, pp.23-52, 総務省統計研修所

ロバスト化推定量

$$\hat{r} = \frac{\sum y_i x_i^{1-2\beta}}{\sum x_i^{2(1-\beta)}} \quad \Rightarrow \quad \hat{r} = \frac{\sum w_i y_i (w_i x_i)^{1-2\beta}}{\sum (w_i x_i)^{2(1-\beta)}}$$

β=1のとき:

$$\hat{r}_{robA} = \frac{1}{n} \sum \frac{w_i y_i}{w_i x_i}$$

A

β=1/2のとき:

$$\hat{r}_{robB} = \frac{\sum w_i y_i}{\sum w_i x_i}$$

B

ウェイト関数: Tukeyのbiweight

$$w\left(\frac{\breve{\varepsilon}}{\sigma}\right) = w(e) = \begin{cases} \left[1 - \left(\frac{e}{c}\right)^2\right]^2 & |e| \le c \\ 0 & |e| > c. \end{cases}$$

残差

$$\breve{\varepsilon}_i = \frac{y_i}{x_i} - \hat{r}_{robA}$$

A

予稿の誤り
(誤差計算は加重しない)

$$\breve{\varepsilon}_i = \frac{y_i}{\sqrt{x_i}} - \hat{r}_{robB}\sqrt{x_i}$$

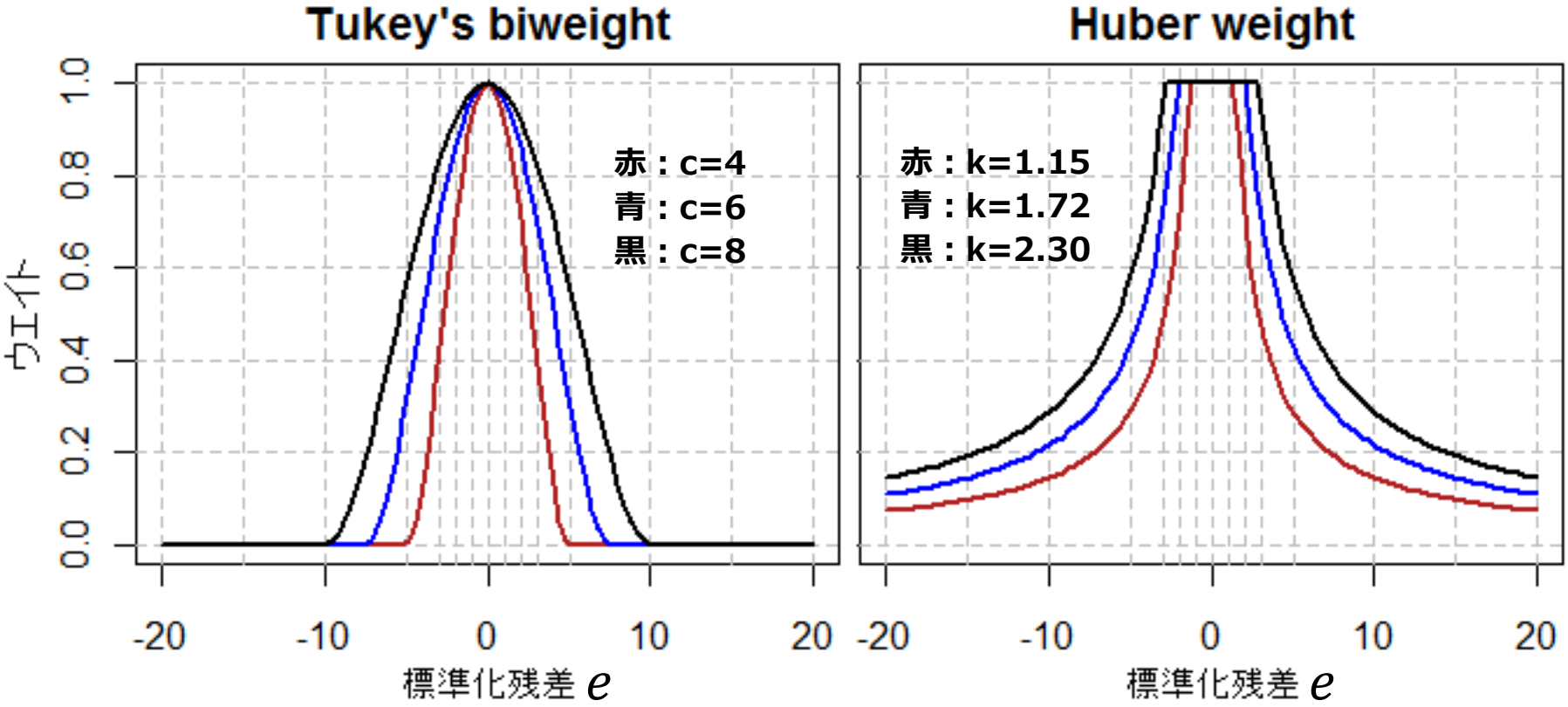
B

残差の尺度パラメータ

$$\sigma_{\mathrm{AAD}} = \frac{1}{n}\sum_{i=1}^{n} |\breve{\varepsilon}_i|$$

調整定数c：8（通常3～8 の間で任意に設定）

ウェイト関数の特徴



$$w(e)=\begin{cases}\left[1-\left(\frac{e}{c}\right)^2\right]^2 & |e|\le c\\ 0 & |e|>c\end{cases}$$

$$w(e)=\begin{cases}1 & |e|\le k\\ \frac{k}{|e|} & |e|>k\end{cases}$$

ある程度中心部から遠い観測値の影響を完全排除できる

中心部から非常に遠い観測値でも、その影響を完全には排除しない

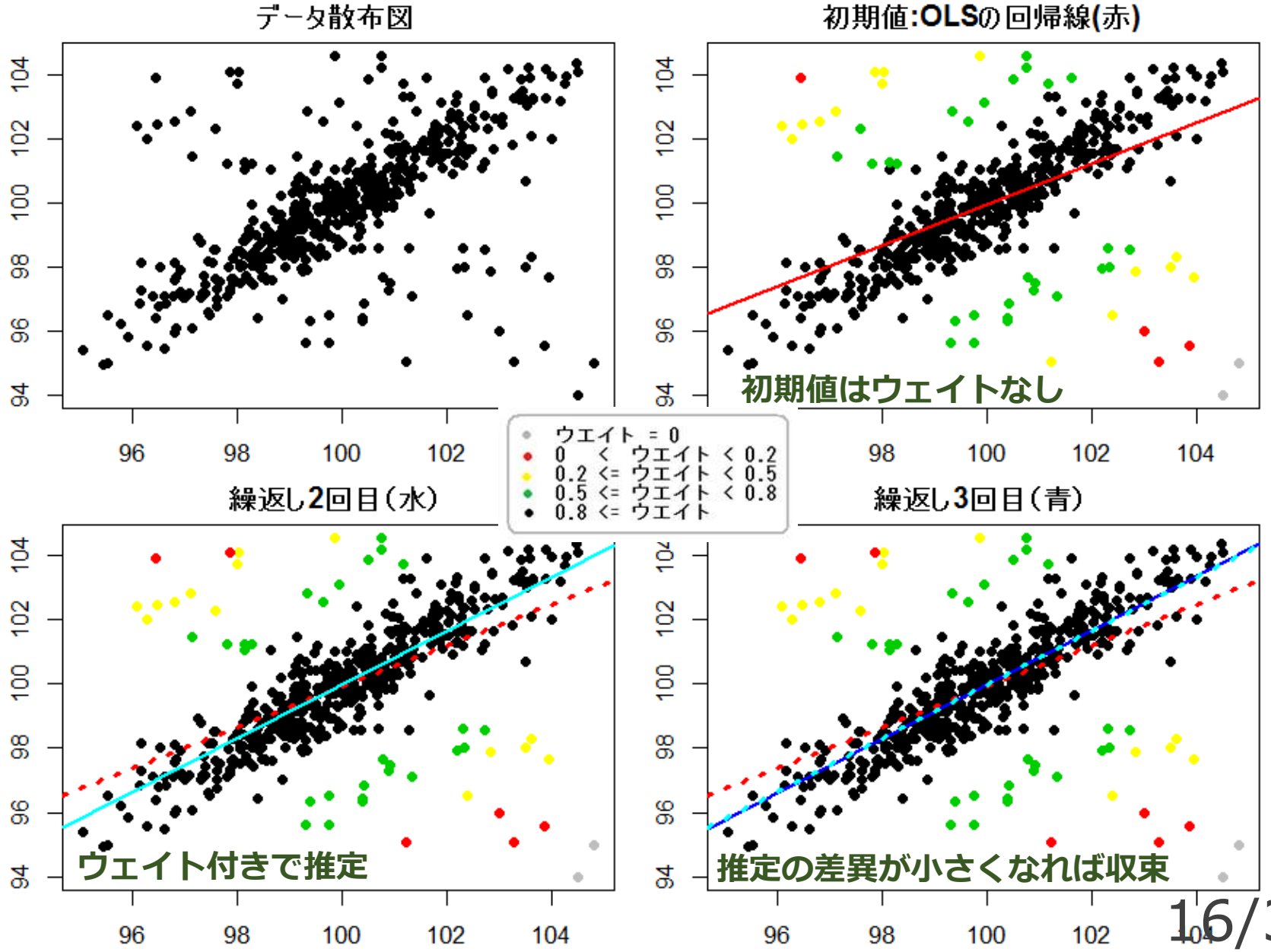
計算アルゴリズム

回帰推定に用いる繰返し加重最小二乗法（IRLS: Iterative Reweighted Least Squares）の仕組みを、回帰の最小二乗法ではなく比推定に適用

- ✓ 計算が簡単で非常に収束が速い
- ✓ 推定パラメータが一つなので、ウェイト関数にTukeyのbiweightを使ってもループしない

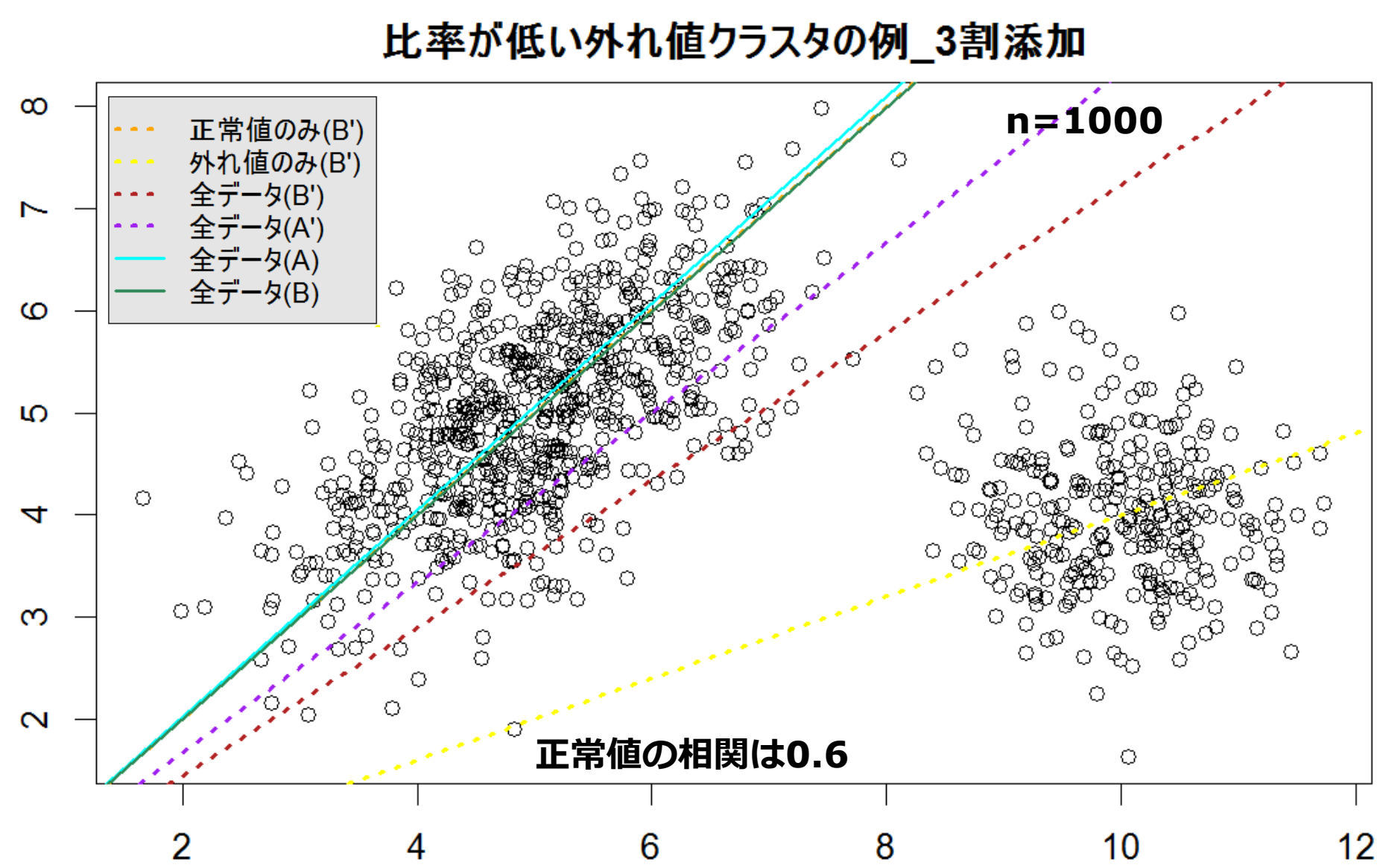
IRLSの仕組み

繰返し計算によるパラメータ推定（単回帰の例）



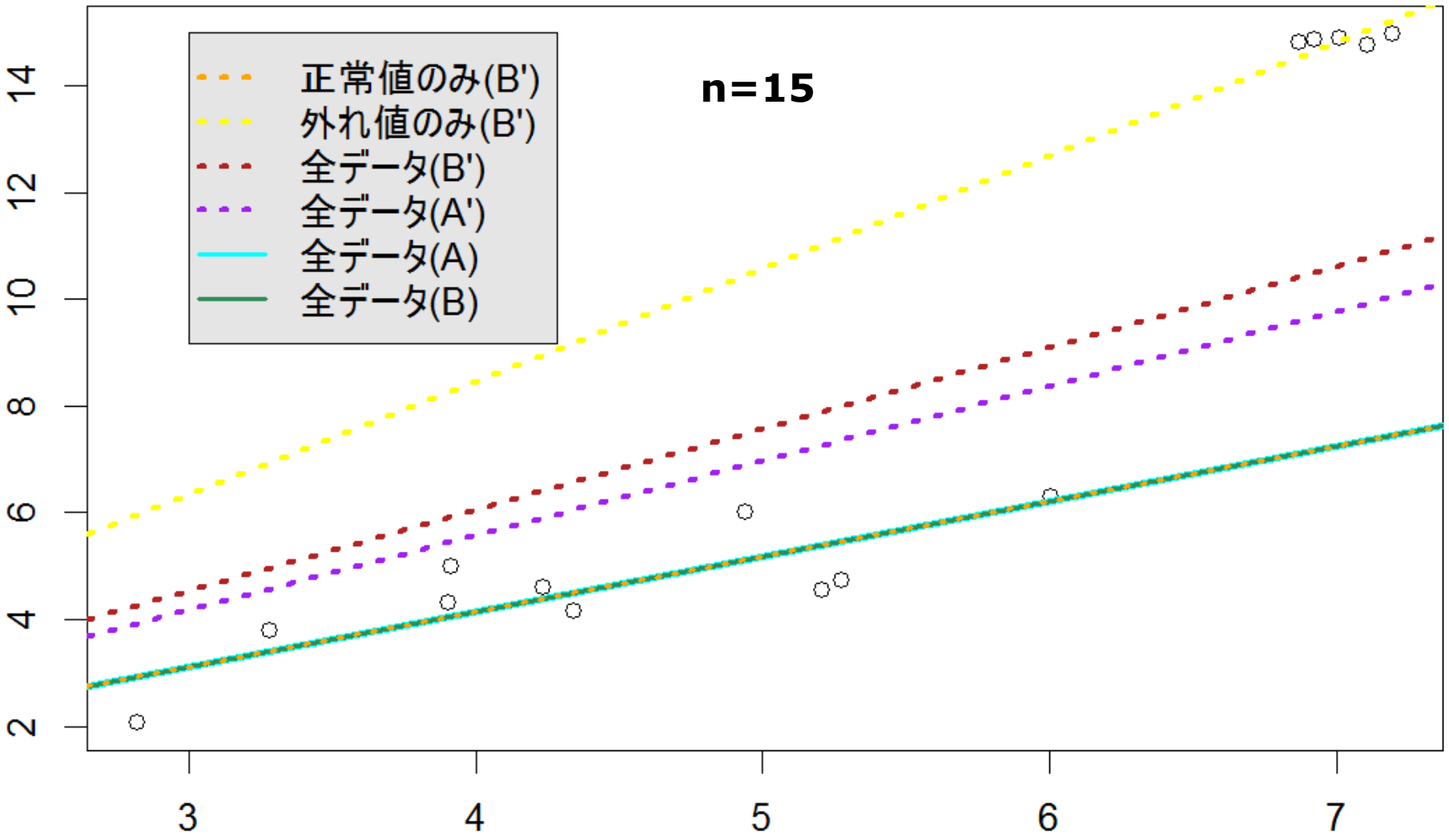
I. 補定(imputation)のための比推定量

I-1　比率補定について
I-2　外れ値の影響緩和方法
I-3　ロバスト化の効果
I-4　極端に大きな値を排除する効果

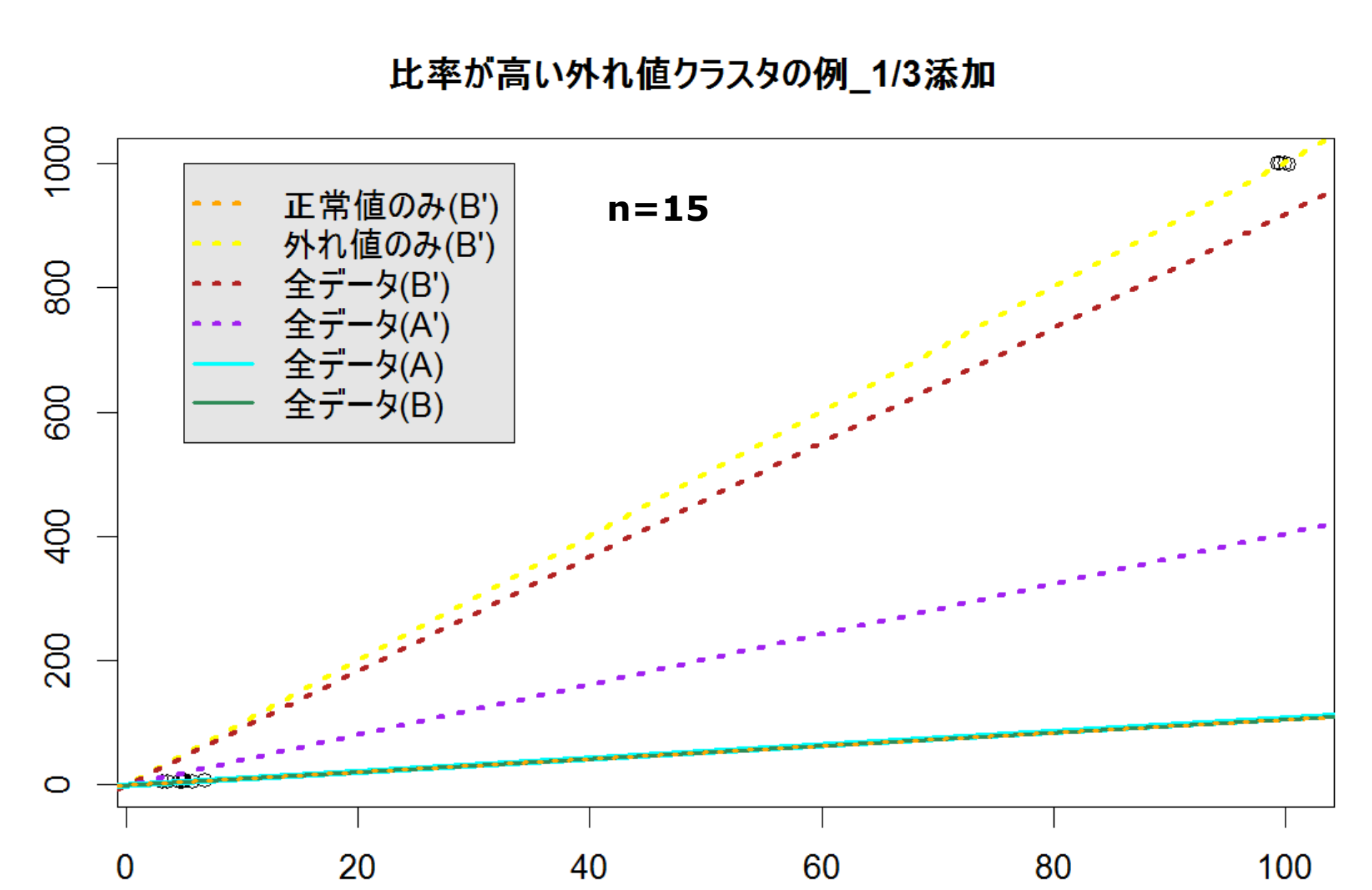


人為的に3割の外れ値を加えても、正常値のみによる比推計に近い結果を得ることができる

比率が高い外れ値クラスタの例_1/3添加



・データ量が少ない場合でも、推定量AとBは外れ値の影響を受けにくい
・推定量A’よりもB’の方が規模の大きい外れ値の影響を受けやすい

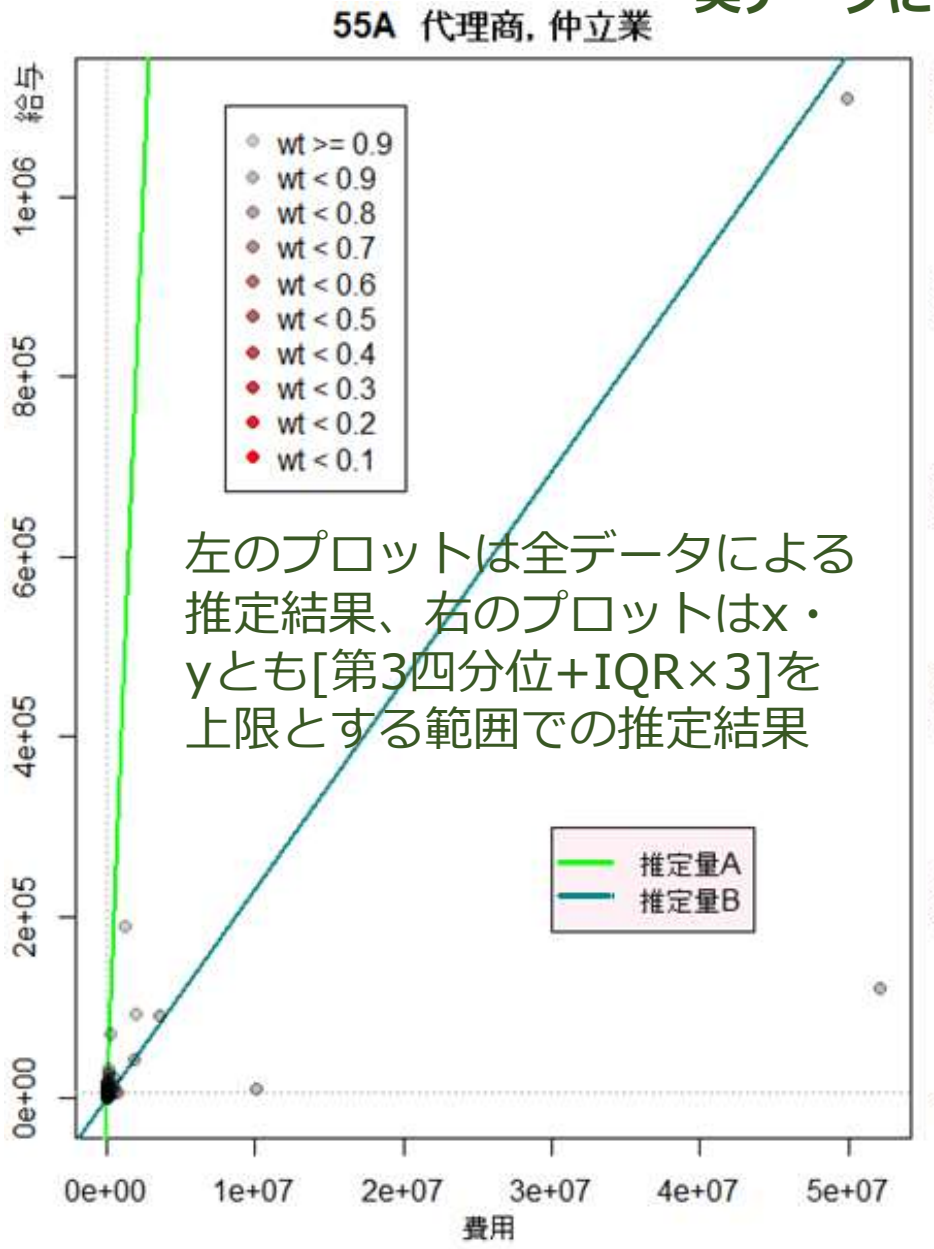


かなり極端な外れ値を添加しても、推定量AとBは影響を受けにくい

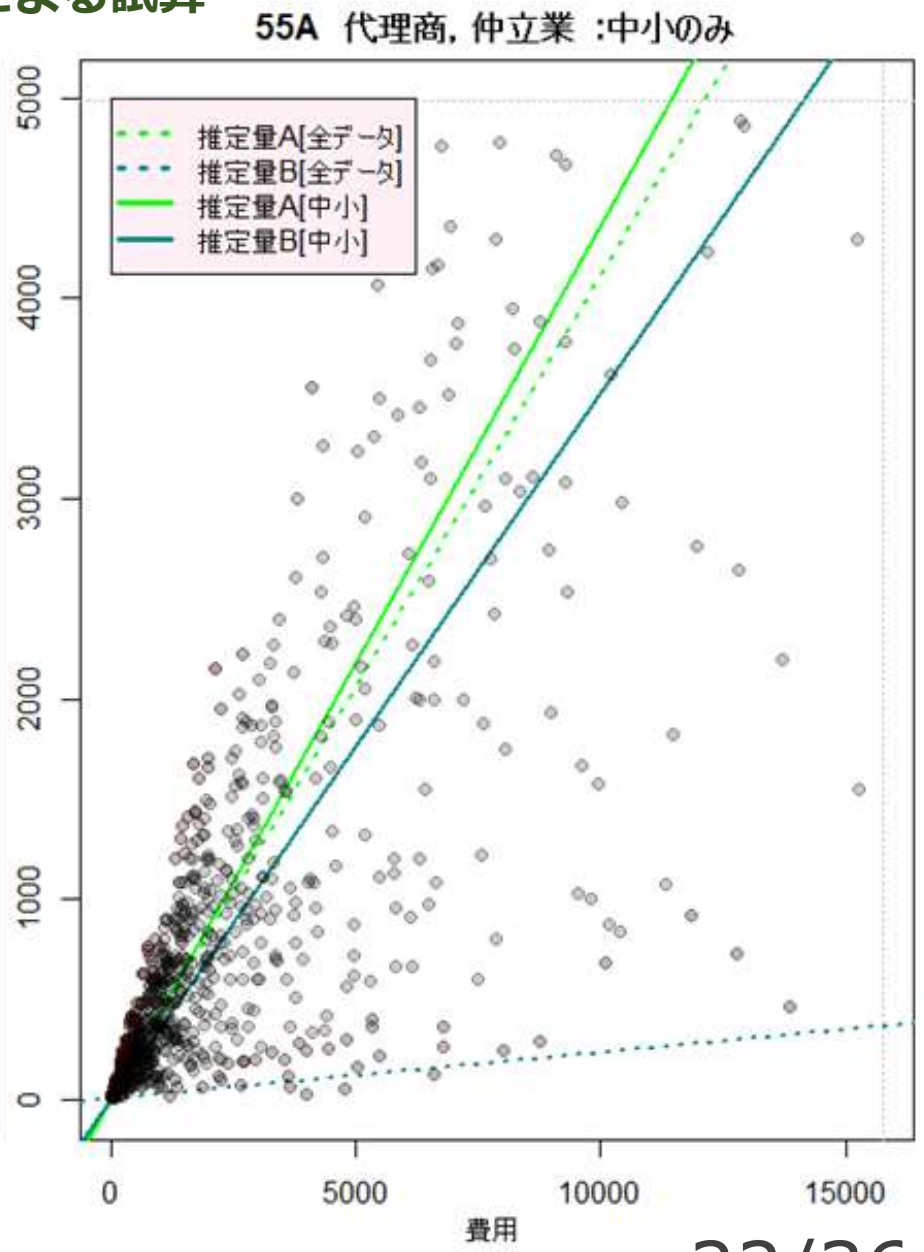
I. 補定(imputation)のための比推定量

I-1　比率補定について
I-2　外れ値の影響緩和方法
I-3　ロバスト化の効果
I-4　極端に大きな値を排除する効果

実データによる試算

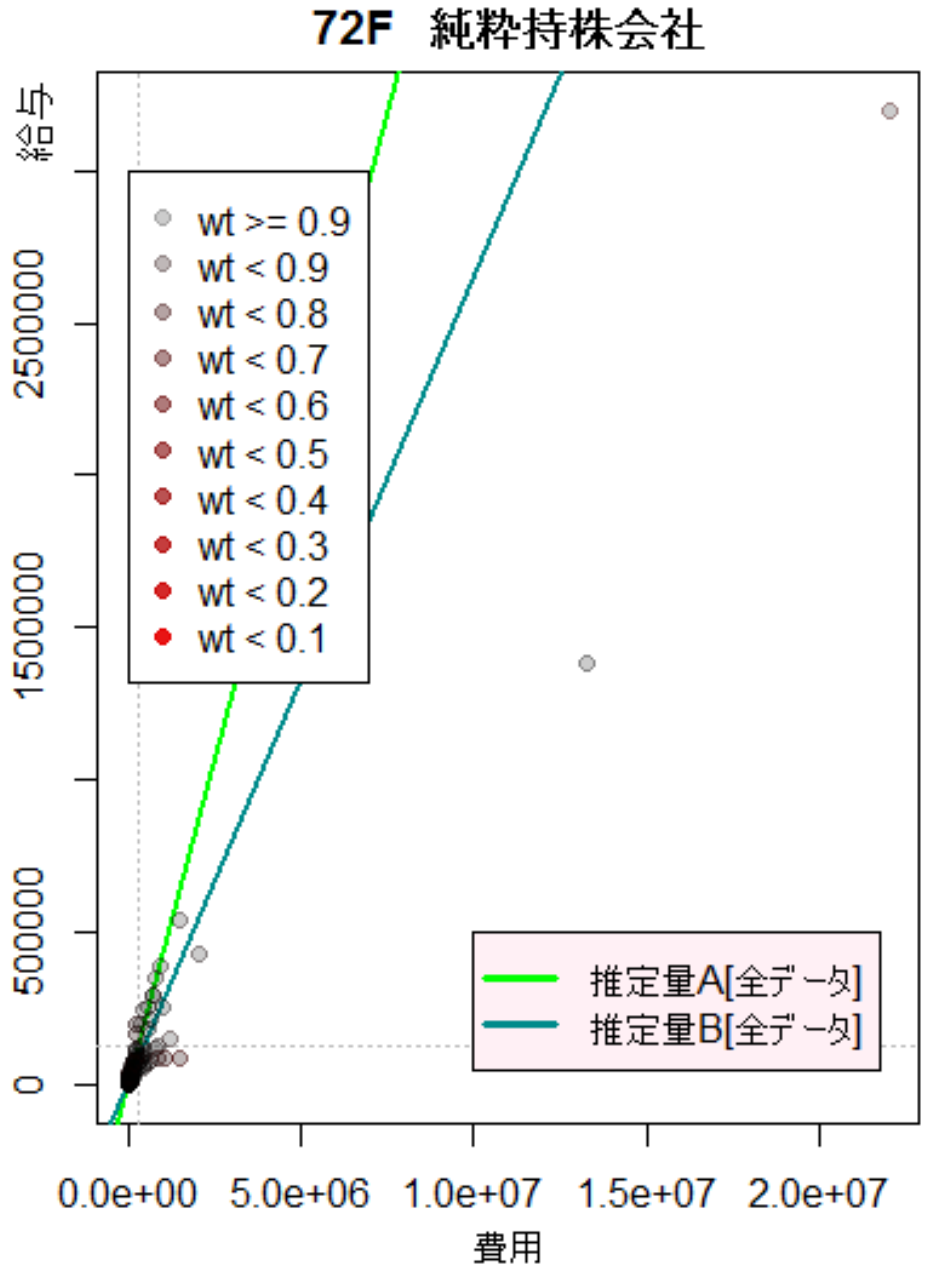


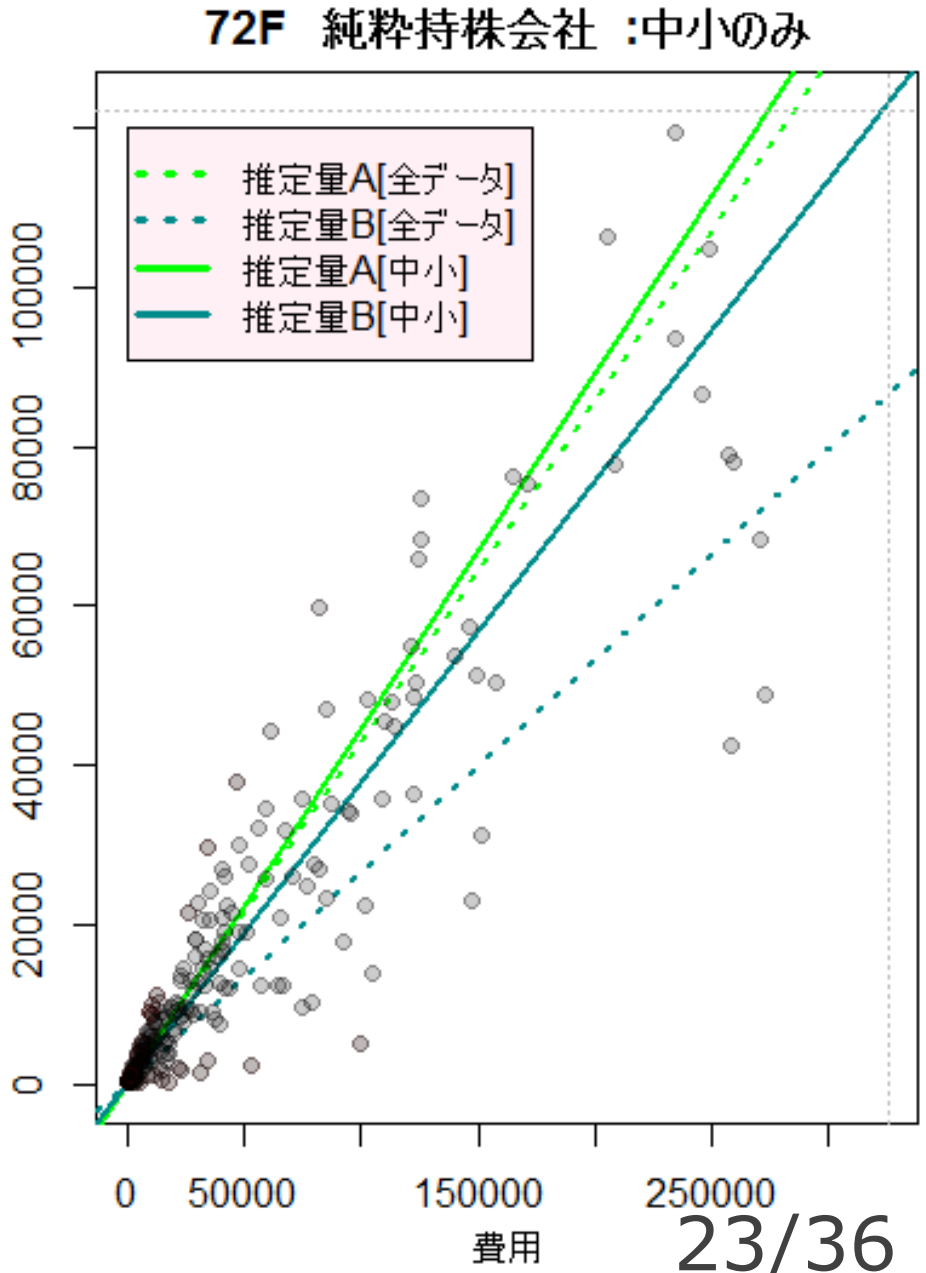
左のプロットは全データによる推定結果、右のプロットはx・yとも[第3四分位+IQR×3]を上限とする範囲での推定結果



推定量Bは、その性質上大きな数値の影響をとても強く受ける

実データによる試算





極端に大きな数値を外すと、推定量Bの問題点を改善することができる

Ⅱ．モデル選択

候補のデータのモデルは二つ

β=1のとき:　比率の平均タイプの推定量

$$\frac{y_i}{x_i} = r + \varepsilon_i, \qquad \varepsilon_i = \frac{y_i}{x_i} - r \sim N(0, \sigma^2)$$

$$y_i = rx_i + \varepsilon_i x_i, \qquad \hat{r} = \frac{1}{n}\sum \frac{y_i}{x_i}$$

A′

β=1/2のとき: 通常の比率補定の推定量のモデル

$$\frac{y_i}{\sqrt{x_i}} = r\sqrt{x_i} + \varepsilon_i, \quad \varepsilon_i = \frac{y_i}{\sqrt{x_i}} - r\sqrt{x_i} \sim N(0, \sigma^2)$$

$$y_i = rx_i + \varepsilon_i \sqrt{x_i}, \qquad \hat{r} = \frac{\sum y_i}{\sum x_i}$$

B′

※ β=0 の切片のない回帰モデルが候補とならないのは散布図から明らか

モデルの選択方法:
モンテカルロシミュレーション

- 平成24年経済センサス - 活動調査データを使用
- 完全データについて、実際の欠測率に応じてx及びyが[第３四分位+IQR×3]を上限とする範囲内でランダムに選んだレコードを欠測とみなし、補定値を計算する
- 補定値$\hat{y}$の真値からの乖離の絶対値の合計を比較
- 対象項目は、売上(費用)、費用(売上)、給与(費用)の三つ　※ 括弧内は説明変数

結果:

➡ 全て推定量B

	真値との乖離の和が最小の回数が最も多い区分								
	売上			費用			給与		
推定量	(A)	(B)	(B)′	(A)	(B)	(B)′	(A)	(B)	(B)′
3.5桁	**20**	**122**	**15**	**48**	**106**	**55**	**74**	**131**	**37**
小	**23**	**115**	**18**	**39**	**105**	**54**	**63**	**115**	**32**
中	**5**	**109**	**22**	**34**	**93**	**32**	**32**	**70**	**30**
1.5桁	**4**	**138**	**7**	**22**	**102**	**17**	**40**	**65**	**52**

	真値との乖離の平均が最小となる区分数								
	売上			費用			給与		
推定量	(A)	(B)	(B)′	(A)	(B)	(B)′	(A)	(B)	(B)′
3.5桁	**10**	**122**	**9**	**38**	**103**	**38**	**40**	**138**	**43**
小	**10**	**113**	**16**	**34**	**108**	**38**	**28**	**125**	**36**
中	**9**	**103**	**33**	**29**	**104**	**30**	**36**	**75**	**39**
1.5桁	**11**	**130**	**14**	**27**	**99**	**34**	**37**	**58**	**51**

	真値との乖離の和が最大の回数が最も多い区分								
	売上			費用			給与		
推定量	(A)	(B)	(B)′	(A)	(B)	(B)′	(A)	(B)	(B)′
3.5桁	**111**	**2**	**8**	**104**	**2**	**15**	**50**	**1**	**6**
小	**107**	**2**	**9**	**105**	**2**	**17**	**58**	**8**	**7**
中	**89**	**5**	**7**	**109**	**12**	**20**	**94**	**32**	**38**
1.5桁	**220**	**7**	**6**	**225**	**21**	**24**	**152**	**15**	**136**

III. 計算効率と推定効率

シミュレーションの設定

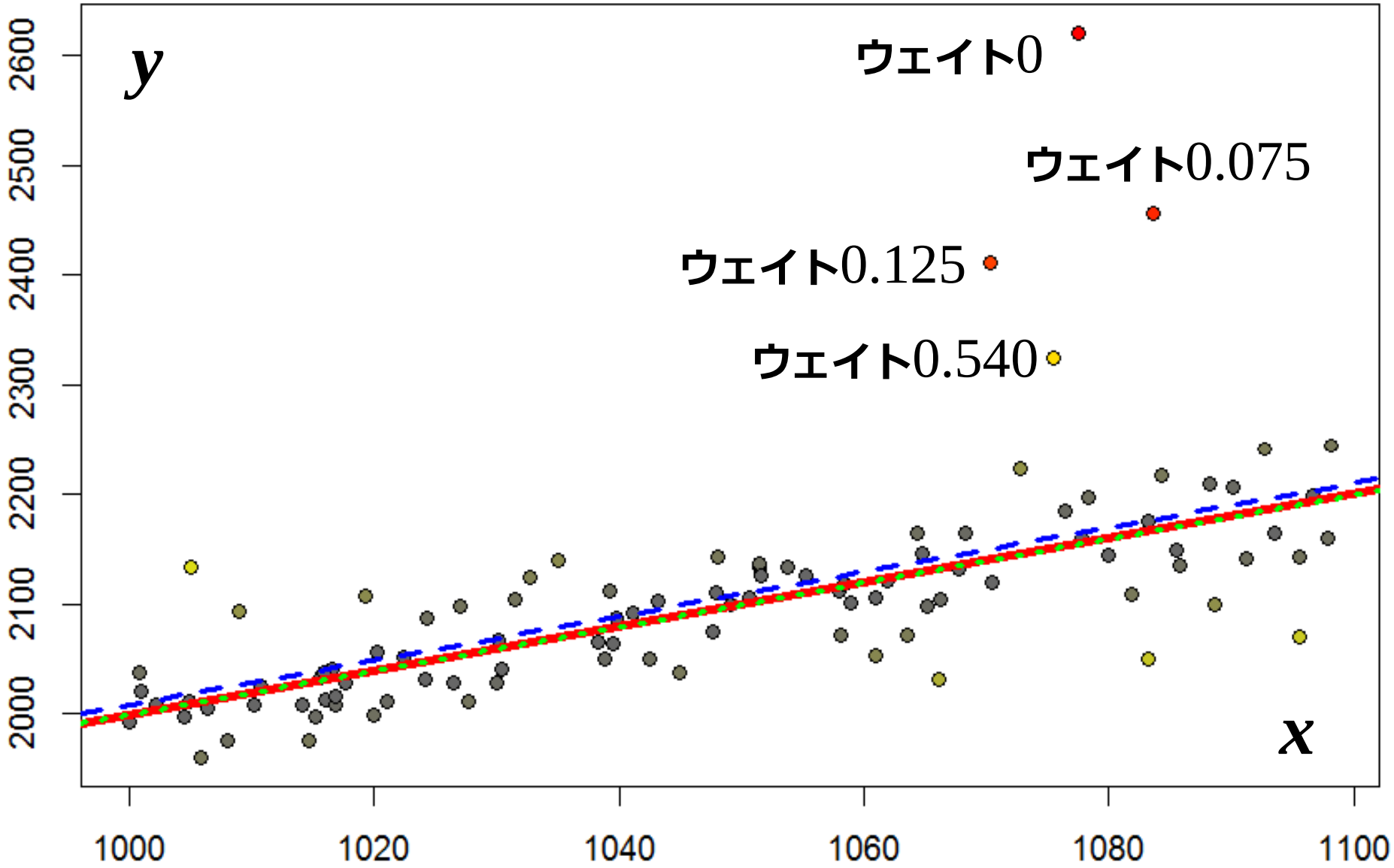
試行回数　：10万回
データサイズ:n=100
説明変数x ：1000から1100の値域の一様乱数
比率R　　：2（真値）
誤差項ε　：自由度1,2,3,5,10,Inf(∞)のt分布
目的変数y ：x, R, εを用いて下式により算出

$$y_i = Rx_i + \varepsilon_i\sqrt{x_i}$$

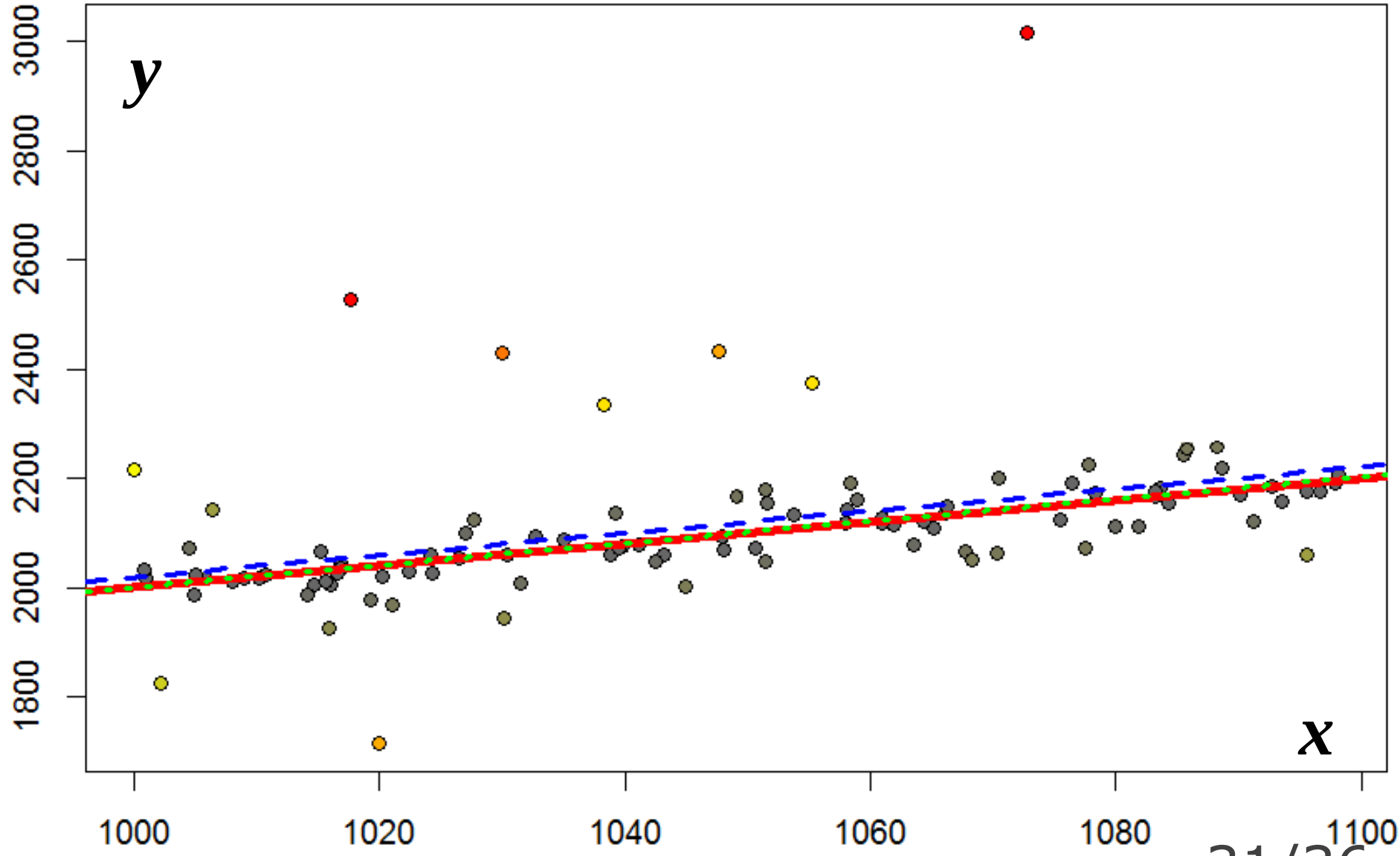
評価基準　：$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

シミュレーション対象は、推定量B及びB′

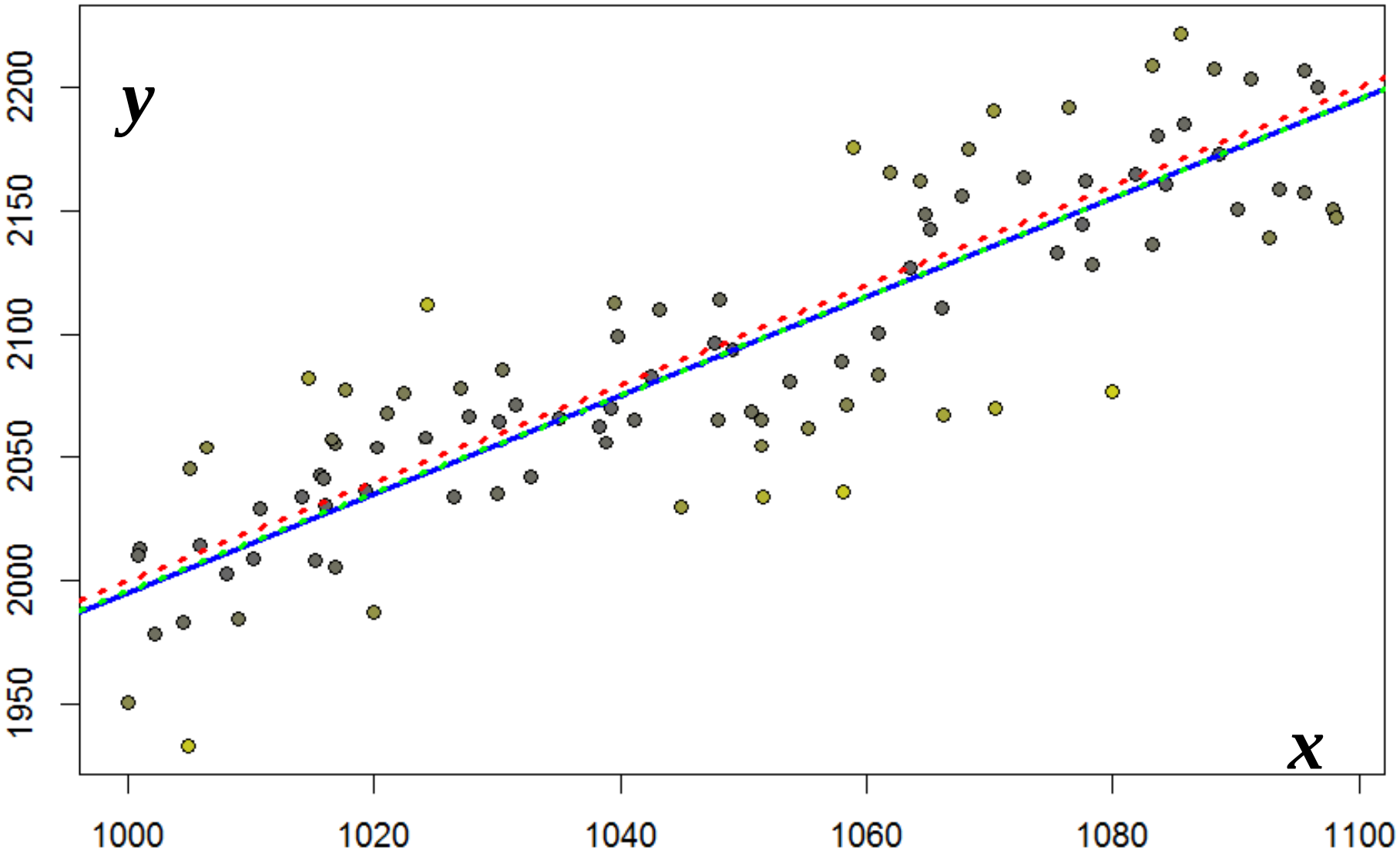
自由度3の誤差項を持つデータの例



自由度1の誤差項を持つデータの例



自由度Infの誤差項を持つデータの例



推定したパラメータの平均値と計算効率

平均値	自由度1	自由度2	自由度3	自由度5	自由度10	自由度Inf
通常版	2.007	2.000	2.000	2.000	2.000	2.000
ロバスト版	2.000	2.000	2.000	2.000	2.000	2.000

繰返計算	自由度1	自由度2	自由度3	自由度5	自由度10	自由度Inf
2回	22006	70557	90232	98180	99818	99995
3回	75142	29435	9768	1820	182	5
4回	2852	8	0	0	0	0
計	100000	100000	100000	100000	100000	100000

- パラメータ推定に偏りはみられない
- 繰返し計算は、ロバストではない比率を初期値とする計算を1回目とカウントするため、最低繰返し回数は2回
- 誤差が従う分布の裾が長くなると、繰返し回数が増える傾向があるが、自由度1のt分布（つまりコーシー分布）でも4回までにすべて収束しており、計算効率が非常に高い

RMSEと相対推定効率

	自由度1	自由度2	自由度3	自由度5	自由度10	自由度Inf
通常の比率補定	90*10^5	1.46	0.28	0.16	0.12	0.10
ロバスト比率補定	0.66	0.24	0.17	0.14	0.12	0.10
相対効率	0.00	0.16	0.61	0.87	0.98	1.05

- 正規分布の誤差項の場合、ロバスト比率補定は通常の比率補定よりも若干推定効率が落ちる
- 誤差項の裾が正規分布よりも長ければ（つまり外れ値が存在すれば）ロバスト比率補定の方が通常の比率補定よりも推定効率が高い

Ⅳ. まとめ

補定に使用する推定量について

- ✓　補定には推定量Bを使用し、外れ値の影響を緩和する
- ✓　推定量Bの場合、非常に大きいxやyの値が比率の推定値に大きく影響するという欠点があるため、規模が大きな企業は個別に指定し、推定対象から除外