平成28年経済センサス - 活動調査

欠測値の補定方法について

(独)統計センター
統計情報・技術部
統計技術研究課

研究目的

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

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

目　次

I. 補定(imputation)のための比推定量

NSTAC

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、

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

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

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

x_iと関係を持たない誤差項

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

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

※ βは任意の定数

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

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

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

B'

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
- 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
- 結果数値が非常に安定する

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

NSTAC

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

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

$\beta = 1$のとき:

$$\hat{r}_{robA} = \frac{\sum w_i (y_i / x_i)}{\sum w_i}$$

A

$\beta = 1/2$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

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

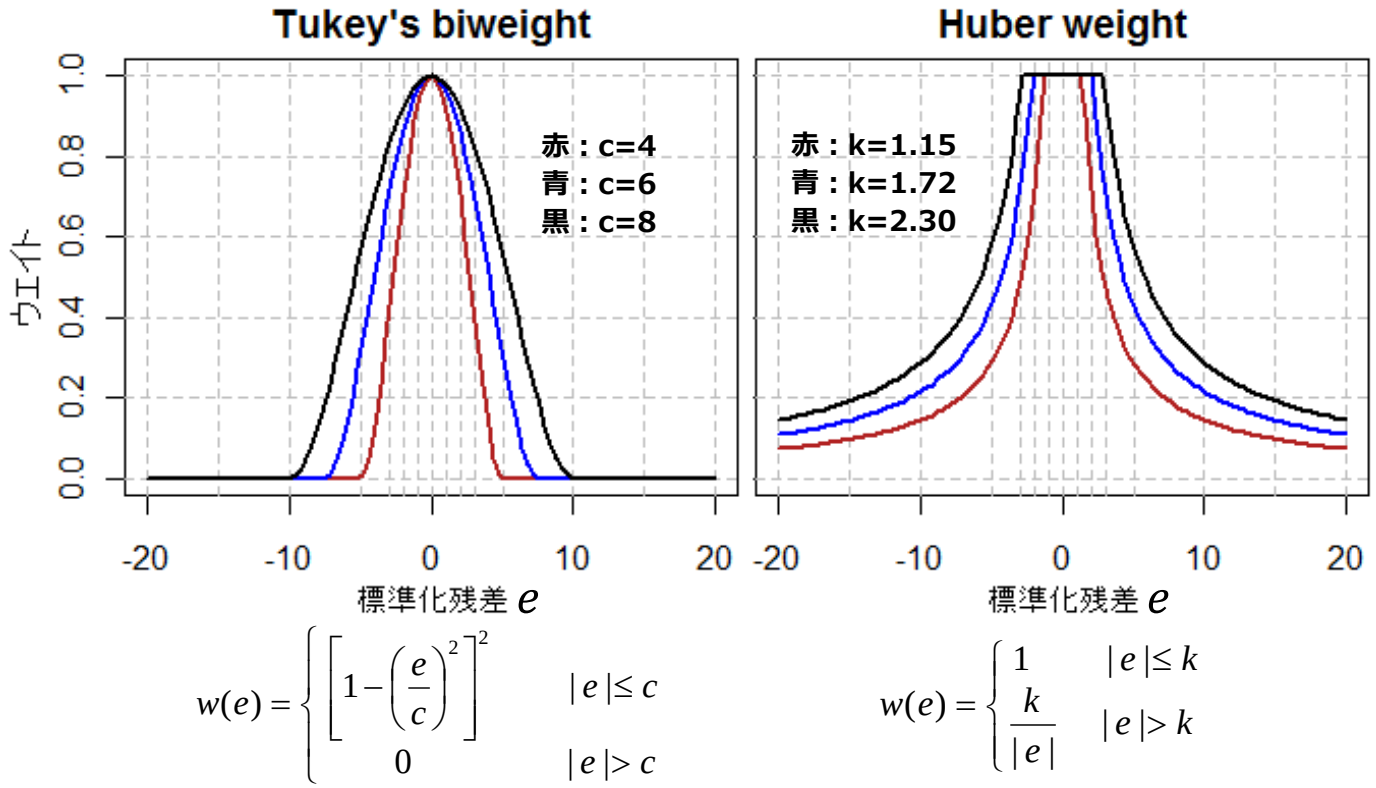
$$\breve{\varepsilon}_i = \frac{y_i}{\sqrt{x_i}} - \hat{r}_{robB}\sqrt{x_i} \qquad \text{B}$$

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



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

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

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

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

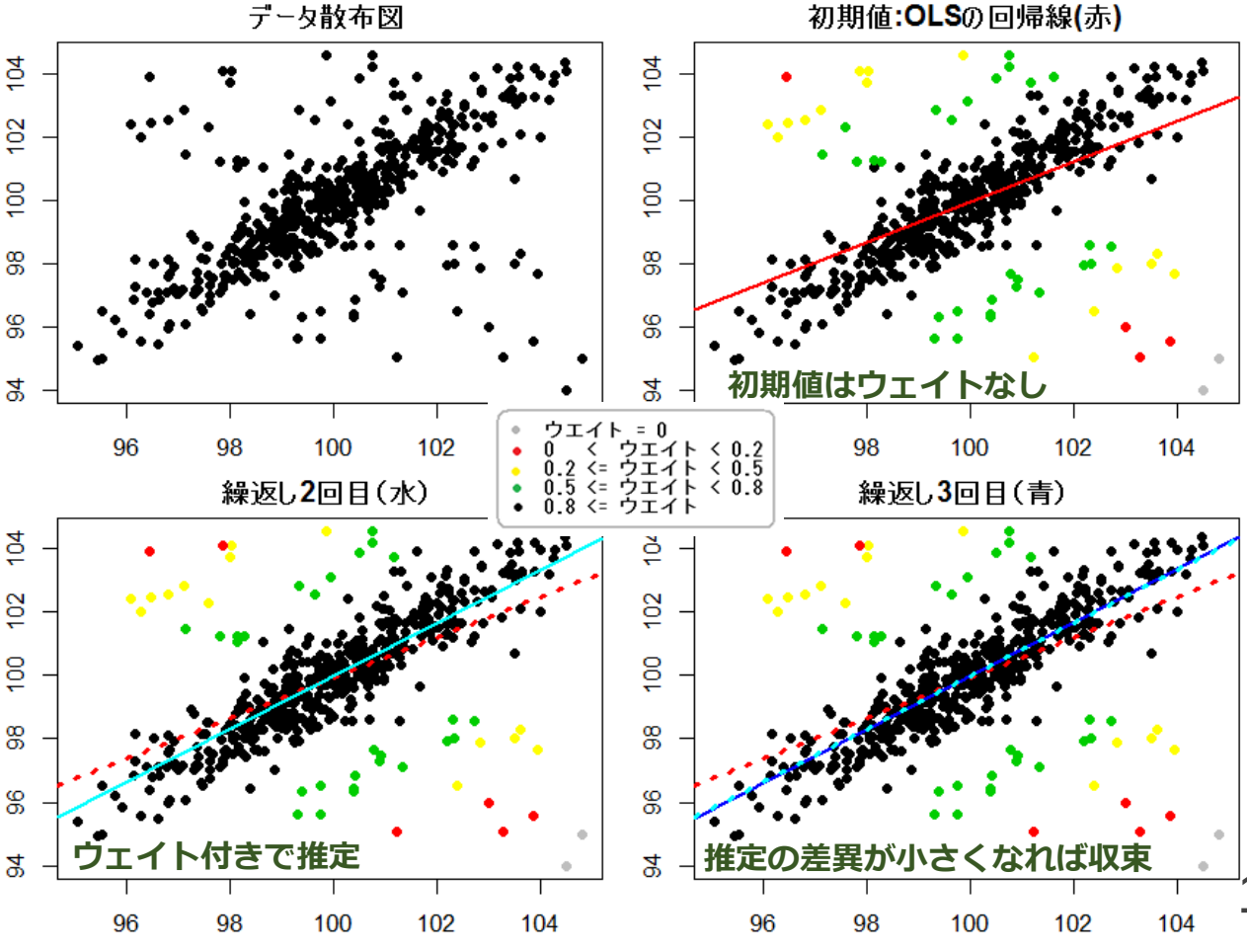
計算アルゴリズム

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

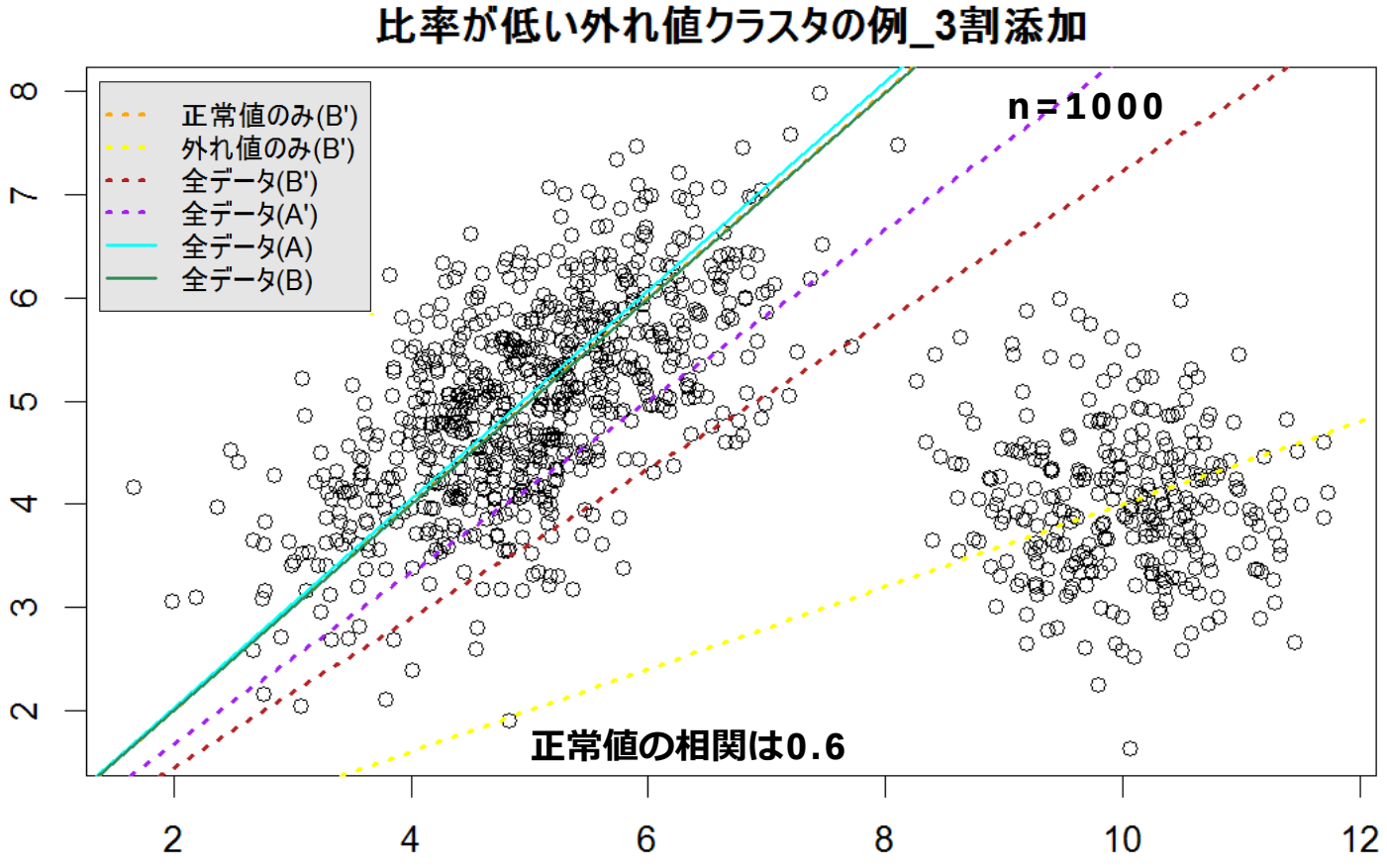
IRLSの仕組み

繰返し計算によるパラメータ推定



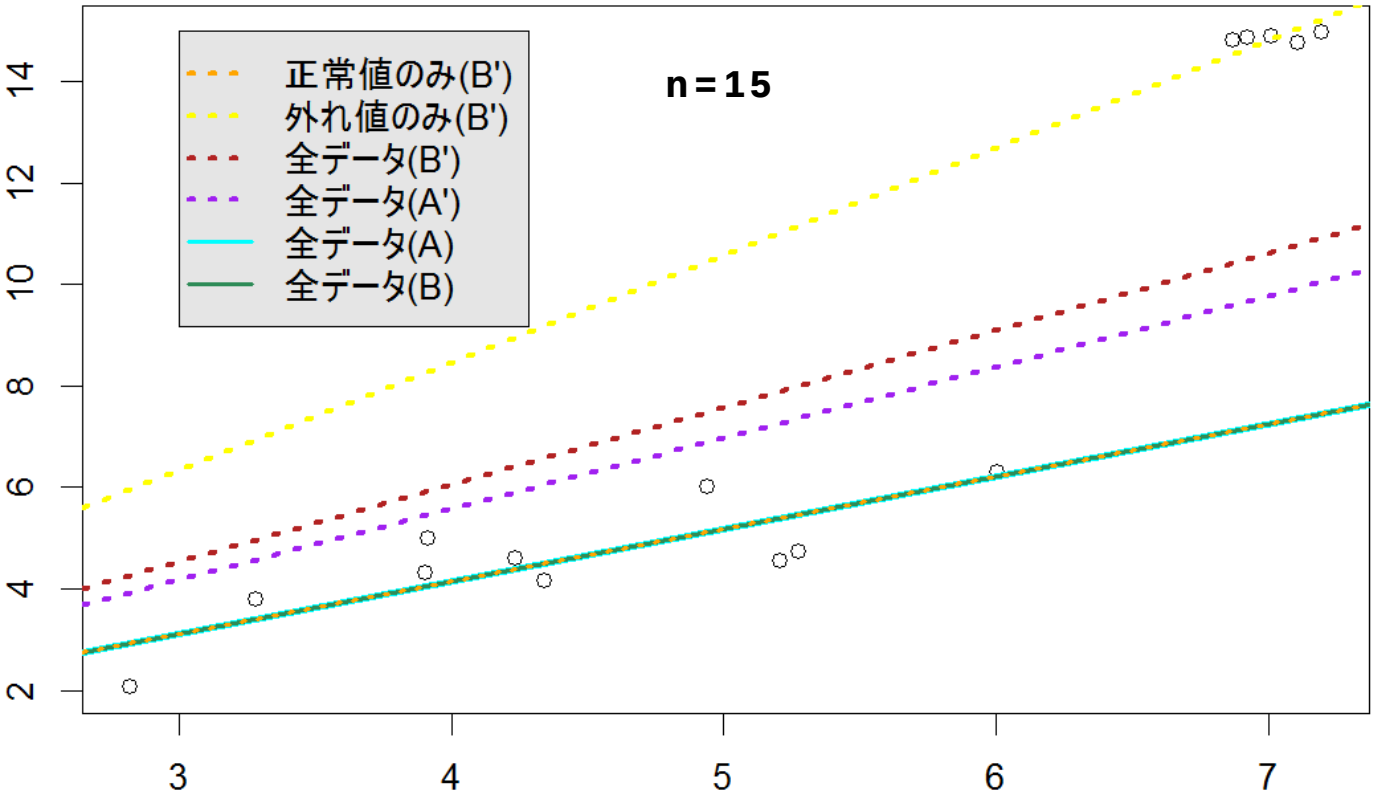
Ⅰ. 補定(imputation)のための比推定量

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



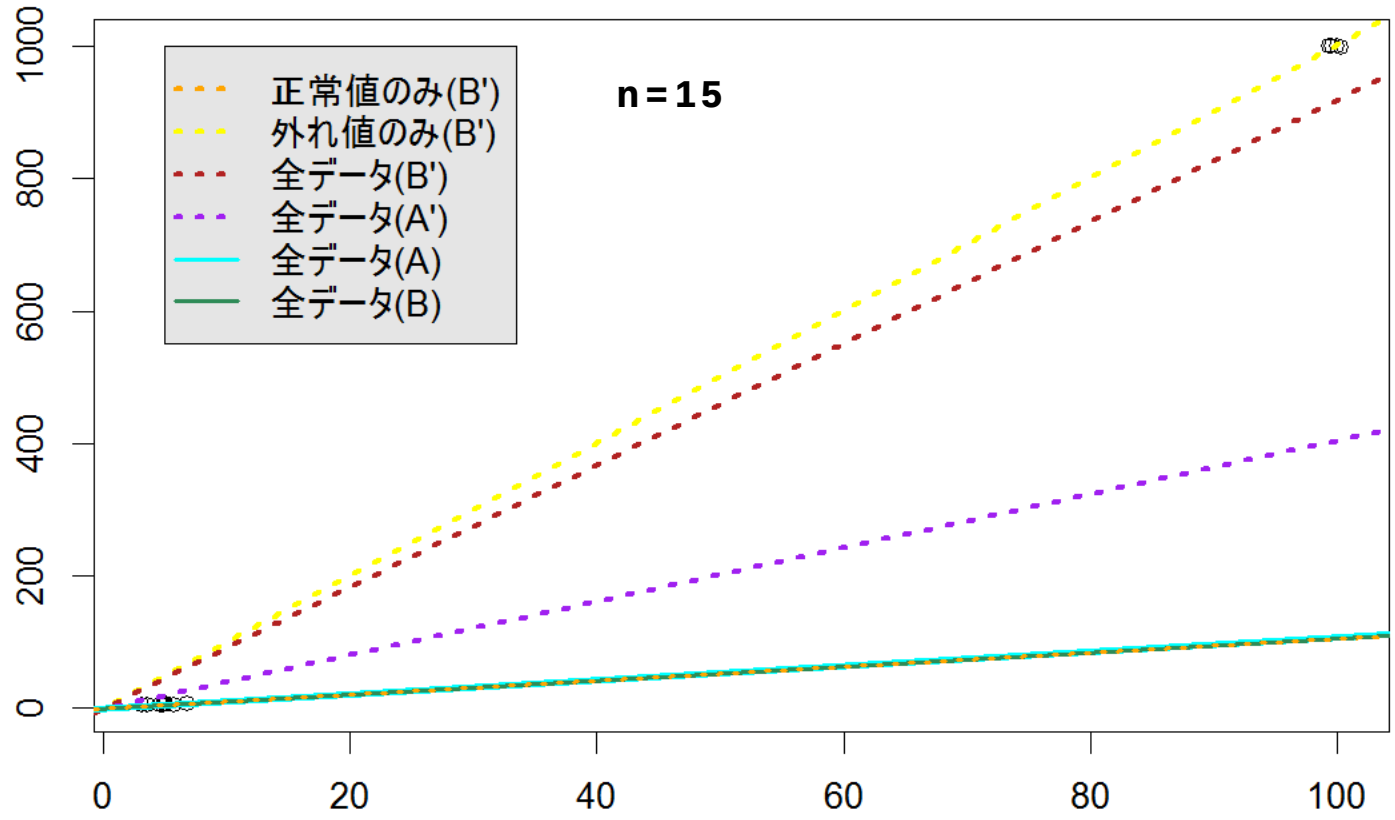
人為的に3割の外れ値を加えても、正常値のみによる比推計に近い結果を得ることができる

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

比率が高い外れ値クラスタの例_1/3添加

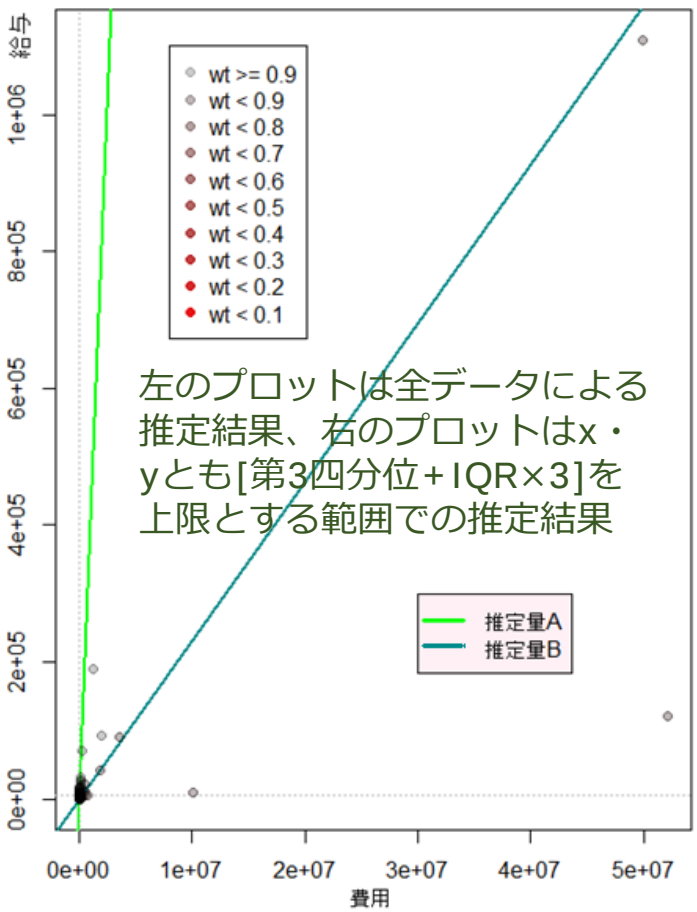


かなり極端な外れ値を添加しても、推定量AとBは影響を受けにくい

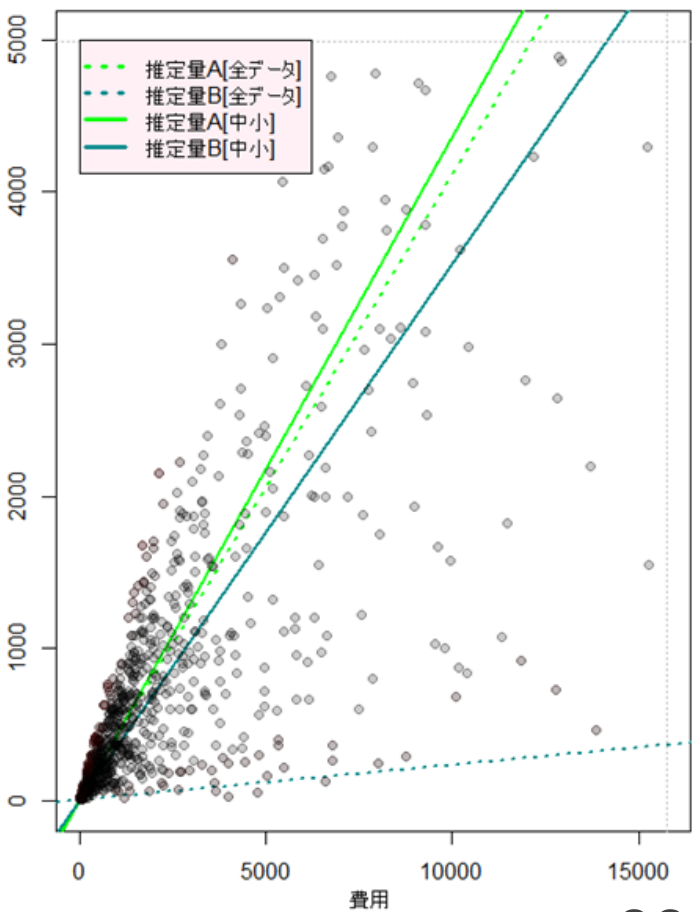
I. 補定(imputation)のための比推定量

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

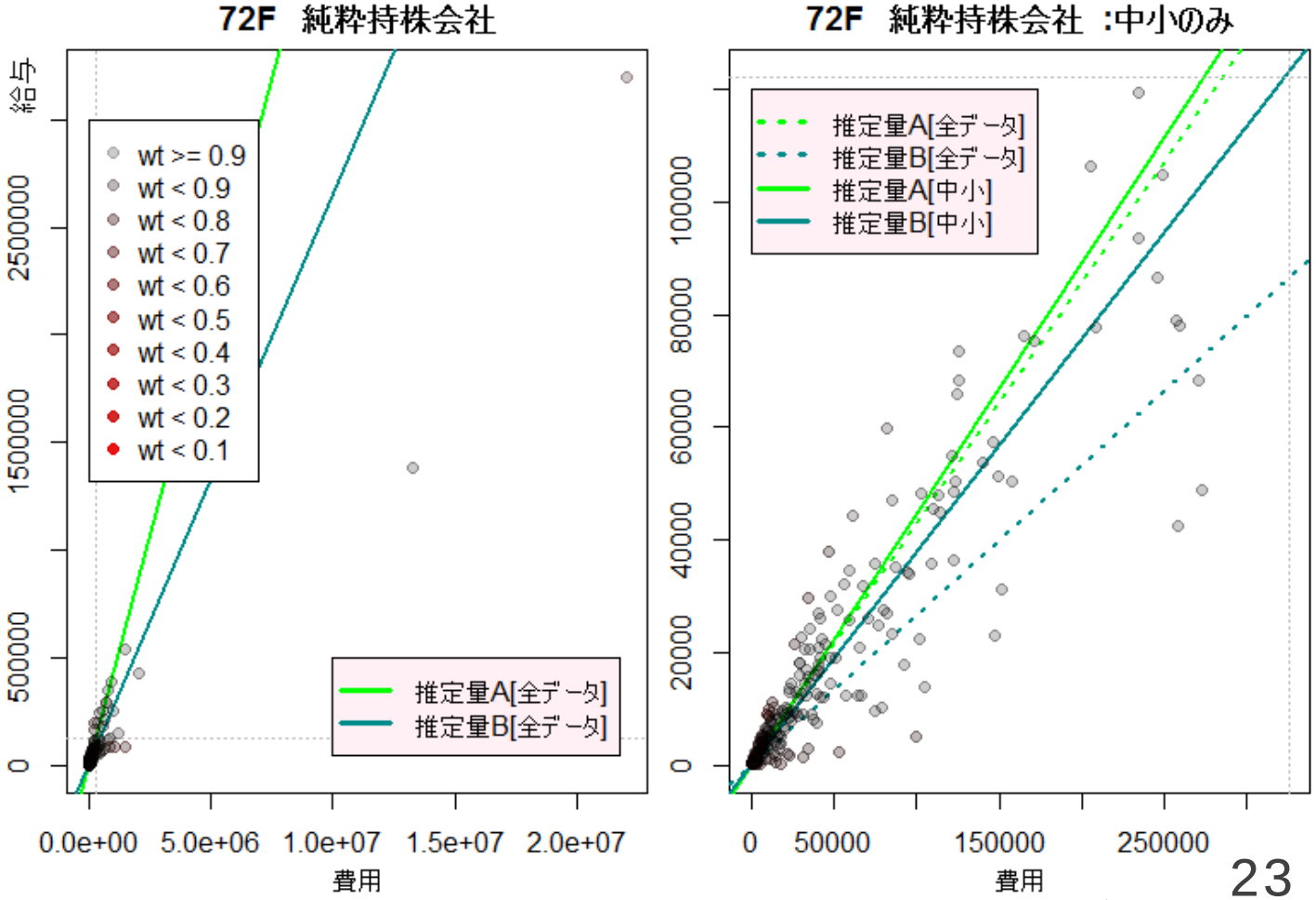


左のプロットは全データによる推定結果、右のプロットはx・yとも[第3四分位+IQR×3]を上限とする範囲での推定結果



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

NSTAC

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**

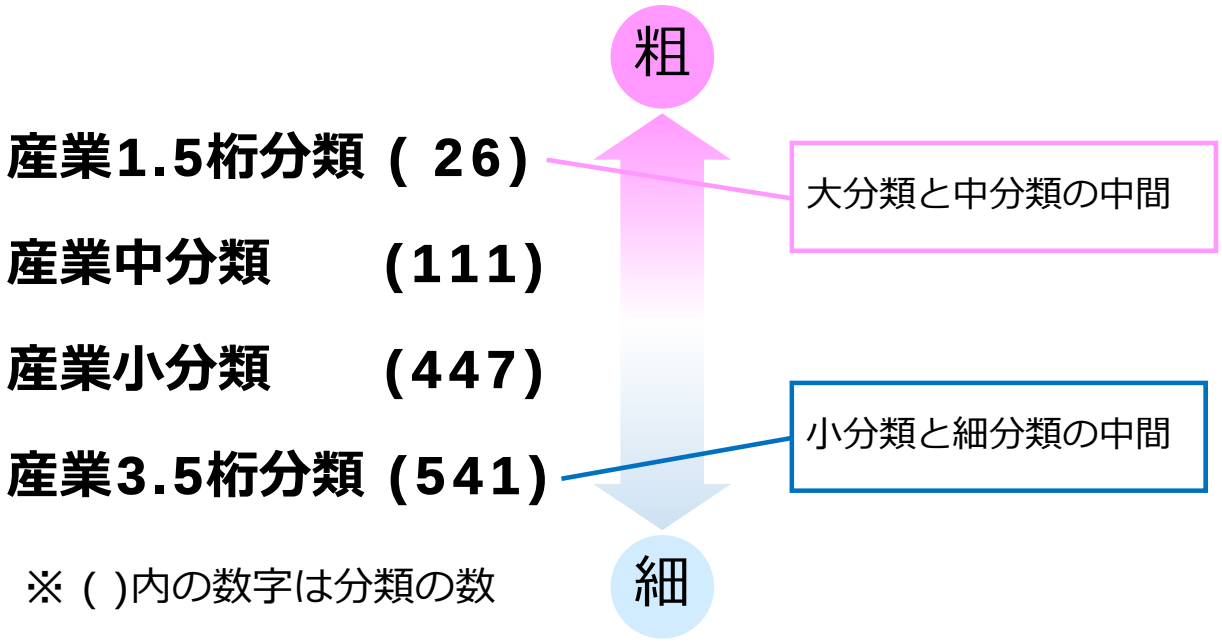
Ⅲ. 補定ドメインの設定方法

Ⅲ-1　産業分類
Ⅲ-2　産業分類以外の調査項目
Ⅲ-3　nが不足するとき

NSTAC

産業分類（確報）

補定を行うために、最も適切な産業区分を特定する



産業分類（速報）

速報データの産業分類は1.5桁分類のみ　**最も粗い分類**

➡ 前回調査と1.5桁分類が同じであれば、前回調査の産業分類の利用可能性を検討する

ドメインの選択肢	各レコードの状況		
	旧分類なし	新旧1.5桁 不一致	新旧1.5桁 一致
	新設	存続 産業転換あり	存続 産業転換なし
1.5桁		○	
旧中分類	×		○
旧小分類	×		○
旧3.5桁	×		○

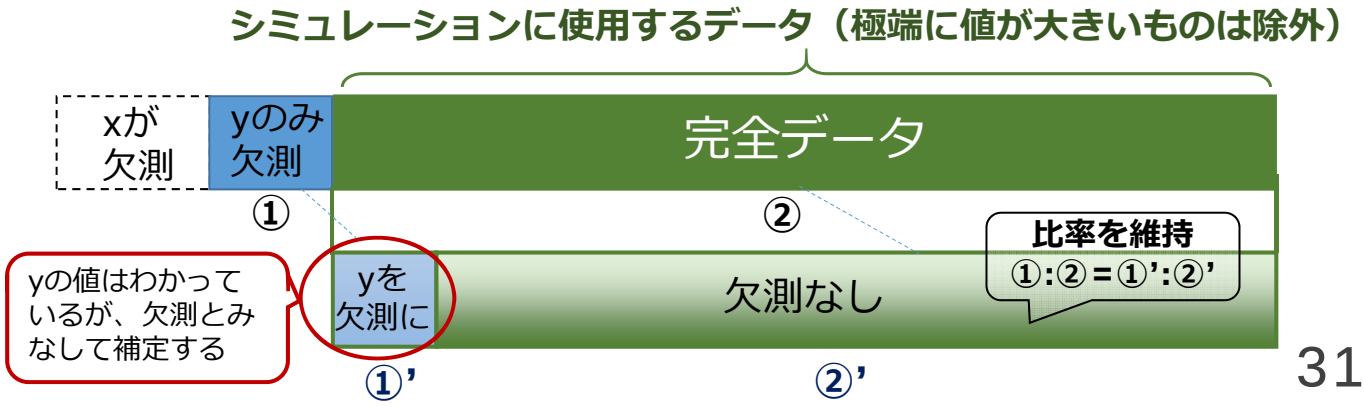
方法論

① 平成24年経済センサス - 活動調査データでのシミュレーション

② 欠測のない完全データのうち、 x及びyが[第３四分位+四分位範囲×3] を上限とするものについて、ランダムに選んだレコードを欠測とみなし、ドメイン候補別に補定を行う

③ 補定対象とする経理項目の合計値について、真値との乖離が最も小さいドメインを選ぶ

④ 評価の基準は、速報は1.5桁分類（確報は3.5桁分類）



評価方法

- 産業分類のあるレベル（速報では1.5桁分類、確報では3.5桁分類）に着目
- 補定前のドメインの合計値（真値）と、補定後の合計値の差の絶対値で評価
- 各回で異なるレベルでのドメイン別の結果を比較する
  - ✓ 差の絶対値が最小となる回数
  - ✓ 差の絶対値が最大となる回数
  - ✓ 各シミュレーション毎の平均絶対偏差と平均偏差

産業分類

○ **速報の旧産業使用とドメイン設定**

若干の例外はあるが、より詳細な旧分類を使用したほうが良い

○ **確報の産業ドメイン設定**

若干の例外はあるが、より詳細な分類を使用したほうが良い

Ⅲ. 補定ドメインの設定方法

Ⅲ-1　産業分類
Ⅲ-2　産業分類以外の調査項目
Ⅲ-3　nが不足するとき

NSTAC

産業分類以外の調査項目

- 24年経済センサス - 活動調査データを使用
- まず産業1.5桁分類で分割
- 候補となる項目は、28年経済センサス - 活動調査の速報データに含まれるものから選択

説明変数
- 経営組織
- 国内常用雇用者
- 国内支所
- 総支所

質的変数
量的変数

→ **CART**
(Classification And Regression Tree)

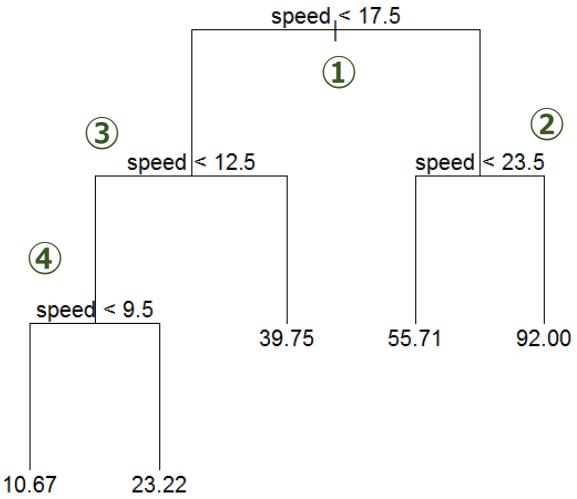
- 目的変数 y_i / x_i

※ *x*及び*y*は[第３四分位＋四分位範囲×3]を上限とする範囲内

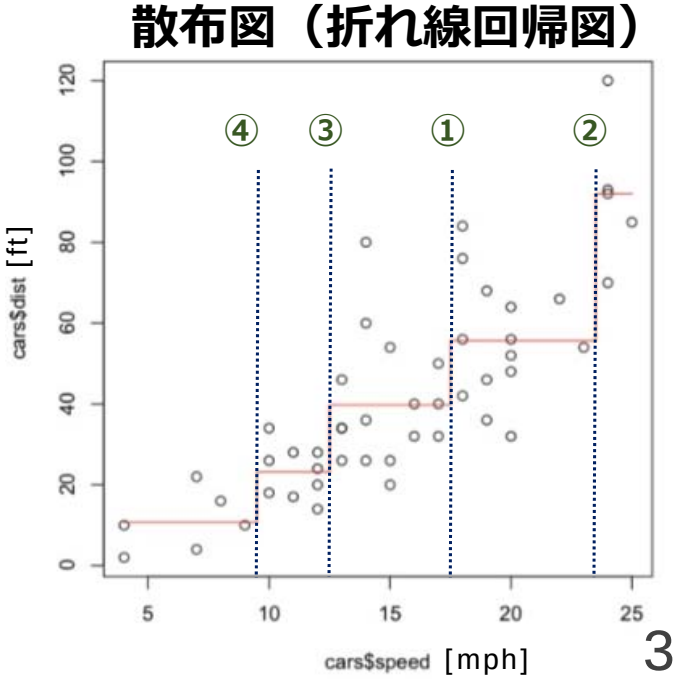
CART(分類木／回帰木)

例) carsデータ: 制動距離(dist)を速度(speed)で説明

樹形図

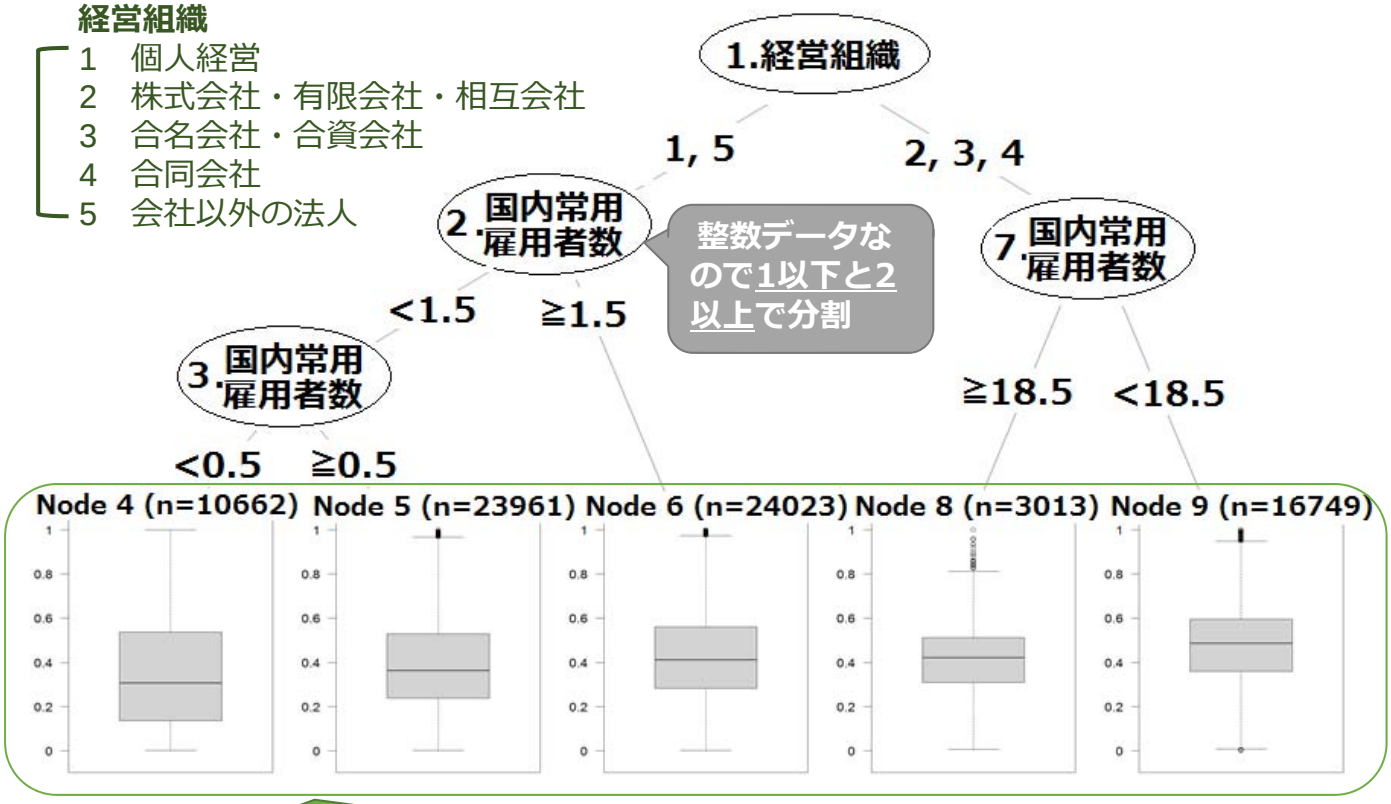


散布図（折れ線回帰図）



CARTによるデータ区分のイメージ

経営組織
1　個人経営
2　株式会社・有限会社・相互会社
3　合名会社・合資会社
4　合同会社
5　会社以外の法人



最終ドメイン別のデータ数と比率の分布

CARTの結果 ： 目的変数 y_i/x_i

1.5桁分類	費用／売上			
	経営組織	常雇(国内)	支所(国内)	支所
A		①		
B		①		
@	②	①		
C		①		
D	①	②		
E	①			
F		①		
G1	①	②		
G2		①		
H	①			
I1	①			
I2	①	②		
J	①	②		

1.5桁分類	費用／売上			
	経営組織	常雇(国内)	支所(国内)	支所
K1	①			
K2	①	②		
L	①			
M1	①	②		
M2	①	②		
N	①			
O1	②	①		
O2		①		
P	①			
Q1		①		
Q2		①		
R1		①		
R2	①			

※　○付数字は区分に有効な変数順（3番目以降は「○」）
※ 使用データは、x及びyが[第３四分位＋四分位範囲×3]を上限とする

CARTの結果 ： 目的変数 y_i/x_i

1.5桁分類	売上／費用			
	経営組織	常雇(国内)	支所(国内)	支所
A	①			
B		①		
@	①	②		
C	①	②		
D	①			
E		①		
F		①		
G1	①			
G2	①			
H	①			
I1	①			
I2	①			
J	①			

1.5桁分類	売上／費用			
	経営組織	常雇(国内)	支所(国内)	支所
K1	①			
K2	①			
L	①			
M1	①			
M2		①		
N		①		
O1		①		
O2		①		
P		①		
Q1		①		
Q2		①	②	
R1		①		
R2	①			

※　○付数字は区分に有効な変数順（3番目以降は「○」）
※ 使用データは、x及びyが[第３四分位+四分位範囲×3]を上限とする

CARTの結果 ： 目的変数 y_i/x_i

1.5桁分類	給与／費用			
	経営組織	常雇(国内)	支所(国内)	支所
A			①	
B		①		
@	①			
C		①		
D	①	②		○
E	○	②		①
F		①		
G1		①		
G2	①		②	
H	①	②		②
I1		①		
I2	①	②	○	
J	①	②		

1.5桁分類	給与／費用			
	経営組織	常雇(国内)	支所(国内)	支所
K1	①	②		
K2				①
L	①	②		②
M1	①	②		
M2	①	②		
N	②	①		○
O1	②	①		
O2	①			
P	①	②		
Q1		①		
Q2				①
R1		①		
R2	①	②	○	

※　○付数字は区分に有効な変数順（3番目以降は「○」）
※ 使用データは、x及びyが[第３四分位+四分位範囲×3]を上限とする

Ⅲ. 補定ドメインの設定方法

Ⅲ-1　産業分類
Ⅲ-2　産業分類以外の調査項目
Ⅲ-3　nが不足するとき

NSTAC

n が不足するとき

✓ 経験的には、おおむねn＝30が補定のための推定を行う際の最小データ数

✓ 産業分類その他の項目で、補定ドメインを設定した場合、nが小さくなる可能性がある

- まず同一産業分類内のドメイン、次に同じ上位産業分類内のドメインの順で、比率の分布が最も近い（U検定のp値が最も大きい）ドメインに統合
- ただし、統合候補のドメインのデータ分布が大きく違う場合は、上位産業分類に統合するか、分散を考慮しつつnが小さくても一つのドメインとする

マン・ホイットニーのU検定
（ウィルコクソンの順位和検定）

- 対応のない二群の母集団が等しいという帰無仮説を検定する

☺正規分布の仮定を必要としない

☺順序統計量に基づくため外れ値の影響を受けにくい

☺正規性を仮定できないとき、平均値に関するt検定の代わりに使用されるが、正規性が仮定できる状況であっても検出力はt検定の95%余（$3/\pi$） [Mood (1954)]

☹帰無仮説は「各グループの分布に差異がない」

☹等分散が前提

- 帰無仮説の採択は、「有意差が認められない」ということで、分布が同じであることを保証できるわけではない
- 複数回の検定には多重比較をする必要があるが、ここでは最も分布が近い統合先候補がわかれば良い

Ⅳ. まとめ

NSTAC

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
- ✓　非常に大きい数値の影響が大きいという欠点があるため、規模の大きい企業は個別に推定対象から除外する

補定の産業分類ドメインの設定

- ✓ 確報については、推定量Bを用いて補定シミュレーションを行い、適切な産業分類ドメインを選択する
- ✓ 速報は、新設ではなく1.5桁分類が前回調査と同じ企業について、1.5桁分類と前回調査時のより詳細な産業分類の中から最適な産業分類ドメインを、同様のシミュレーションにより選択する

産業分類以外のドメインの設定

- ✓ 候補とする調査項目は以下のとおり
  - 経営組織
  - 国内常用雇用者
- ✓ 各産業分類ドメインについて、さらにCARTを用いて適切なドメインに細分化する