uRos (Use of R in Official Statistics)2019, 7th international conference
Bucharest, Romania, 20th-, 21th, May 2019

# Generalisation and robustification of a ratio model

**Kazumi Wada†, Keiichiro Sakashita†, Hiroe Tsubaki‡**

† National Statistics Center (NSTAC), Japan

‡ The Institute of Statistical Mathematics (ISM)

# Research objective

**Robustification** and **generalization** of the **ratio model**

↓

**Goal: Improve ratio imputation**

# Contents

1. Conventional ratio model
2. Robustification of linear regression model
3. Solve the problem of heteroscedasticity to robustify the ratio model
4. Generalisation of the ratio model
5. Robust estimation of the generalised ratio model
6. Application to Economic Census data

# 1. Conventional ratio model

# Conventional ratio model

$$y_i = rx_i + \epsilon_i$$

$y$: objective variable

$x$: explanatory variable which has a high correlation with the objective variable

$r$: the ratio of $y$ and $x$

$\epsilon_i$: error term, $V(\epsilon) = x\sigma^2$

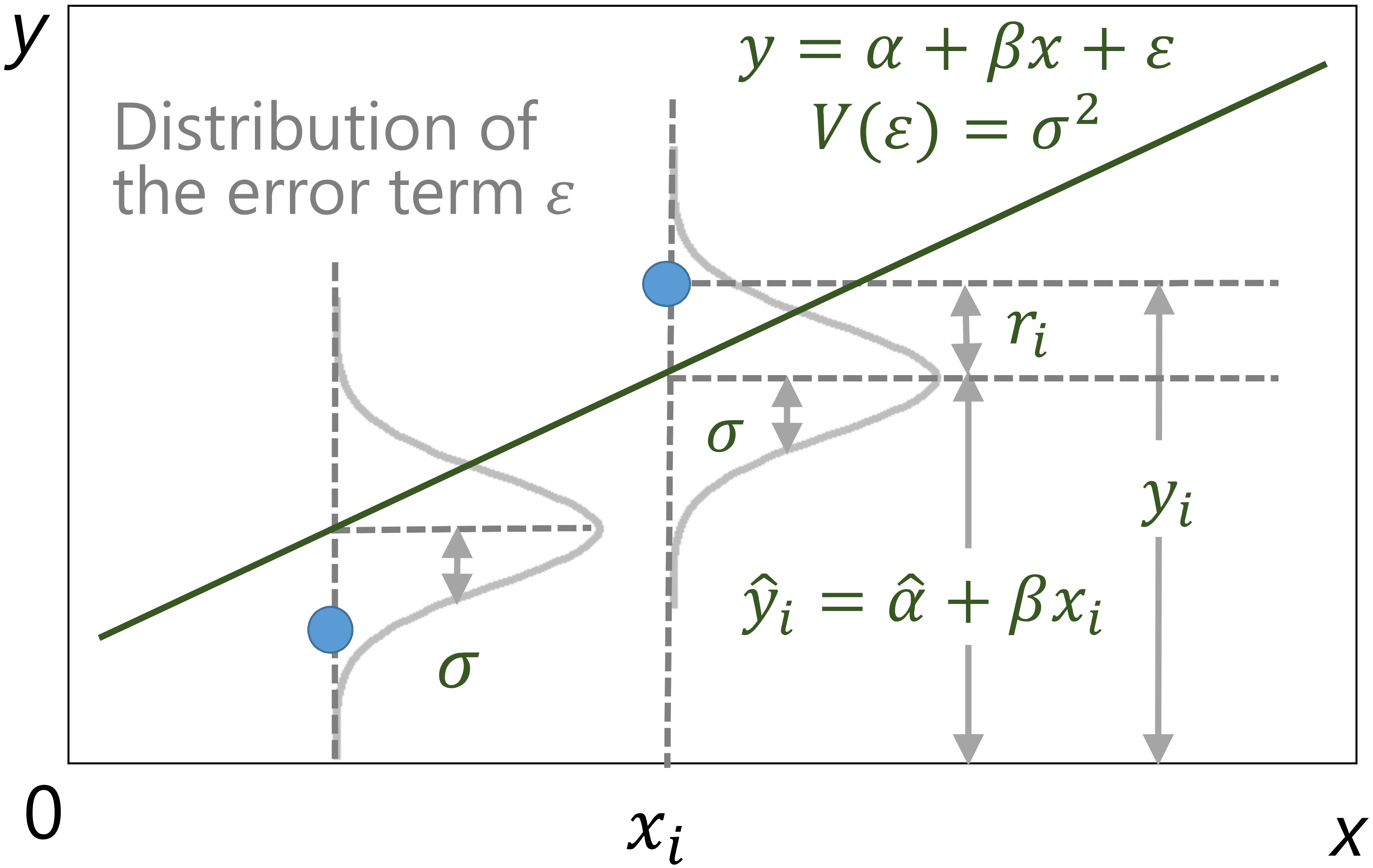
When the model is used for imputation of $y$, unknown $r$ due to missingness is estimated with complete data of $x_i$ and $y_i$.

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

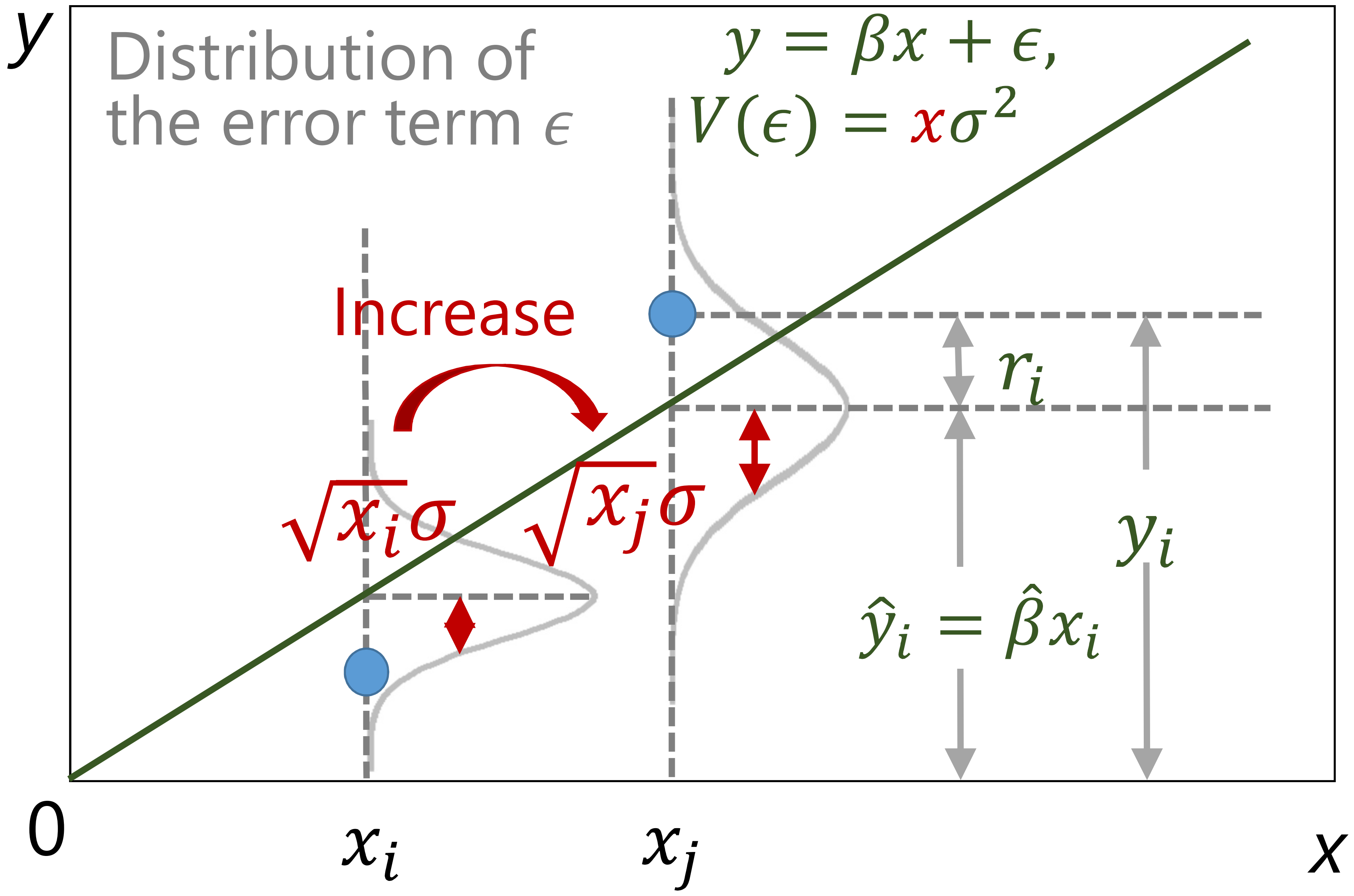
obs: complete observations regarding these two variables

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

# Linear regression model


$y$
$y = \alpha + \beta x + \varepsilon$
$V(\varepsilon) = \sigma^2$
Distribution of the error term $\varepsilon$
$r_i$
$\sigma$
$\sigma$
$y_i$
$\hat{y}_i = \hat{\alpha} + \beta x_i$
0
$x_i$
$x$

# Conventional ratio model


$y$
Distribution of the error term $\epsilon$
$y = \beta x + \epsilon,$
$V(\epsilon) = x\sigma^2$
Increase
$\sqrt{x_i}\sigma$
$\sqrt{x_j}\sigma$
$r_i$
$y_i$
$\hat{y}_i = \hat{\beta} x_i$
0
$x_i$
$x_j$
$x$

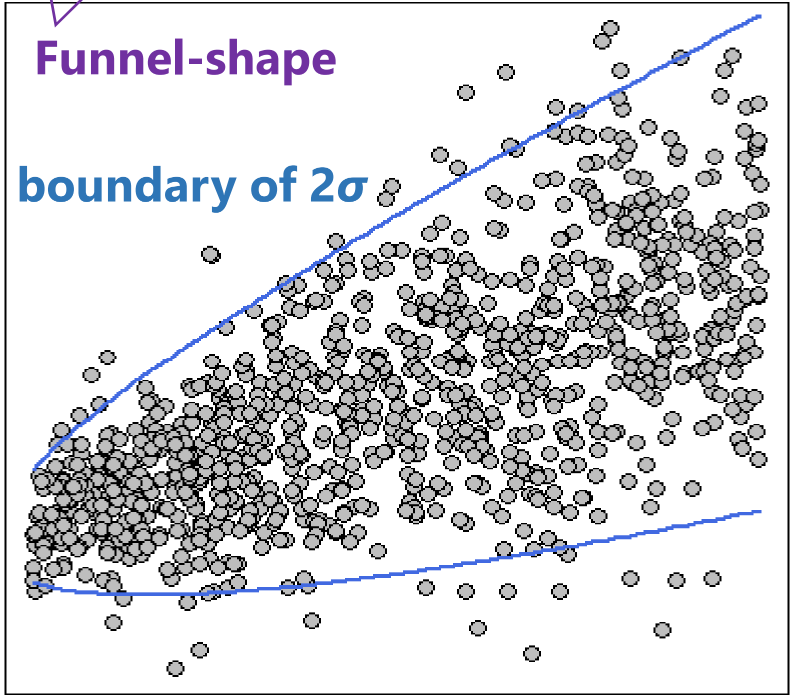
# The difference between the ratio model and a linear regression model

## Ratio model

$$y_i = r x_i + \epsilon_i$$

***Heteroscedastic***

$$\epsilon_i \sim N(0, \sigma^2 x_i)$$

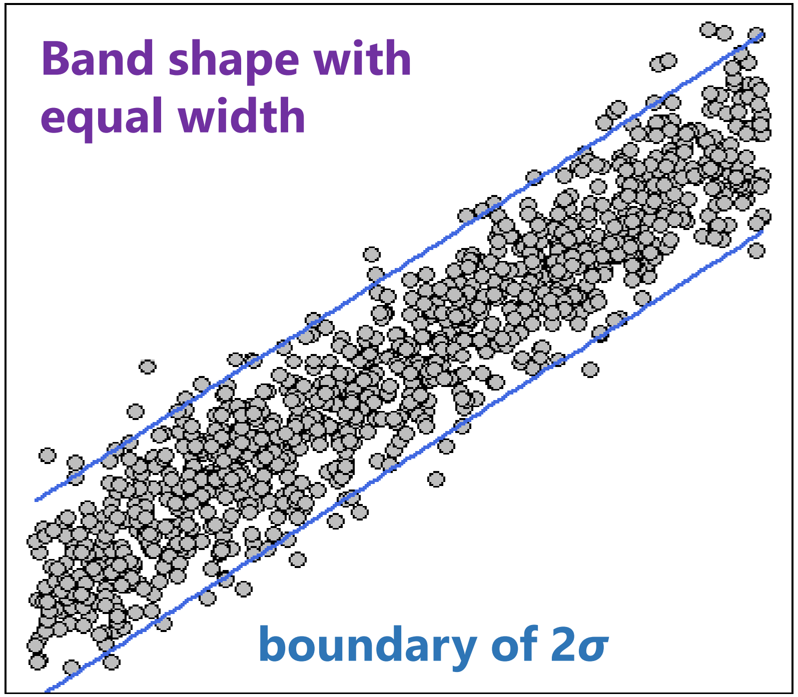


## Regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

***Homoscedastic***

$$\varepsilon_i \sim N(0, \sigma^2)$$



# Two candidate models for the funnel shaped data

1. A regression model with transformation
2. The ratio model without transformation

In case of 1, estimation of the mean and total becomes unstable.

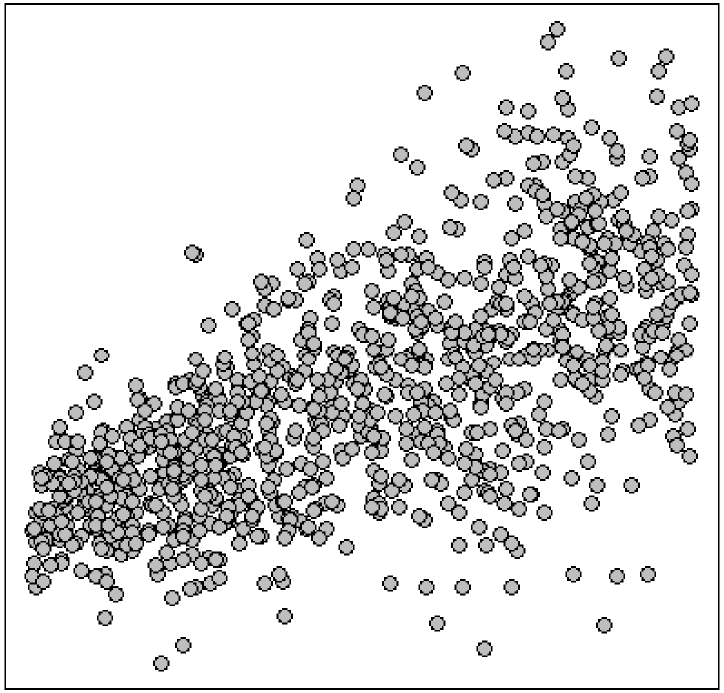
**e.g. Log transformation**

$$\log y_i = \alpha + \beta \cdot \log x_i + \varepsilon_i$$

$$\hat{y}_i = exp(\alpha + \beta \cdot \log x_i) \cdot exp\left(\frac{1}{2} \cdot \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-p}\right)$$

**The ratio model has an advantage for imputation.**

**Survey data often have heteroscedastic variance**



# 2. Robustification to cope with outliers

# A robustification in regression model

**Definition of outlier**:
Observations in the tails of error term distribution

**Robustification**: M-estimator which reduce the weight of the observations with large error

**Computation**: IRLS (Iteratively reweighted least squares) algorithm which is easy to implement and converges very fast

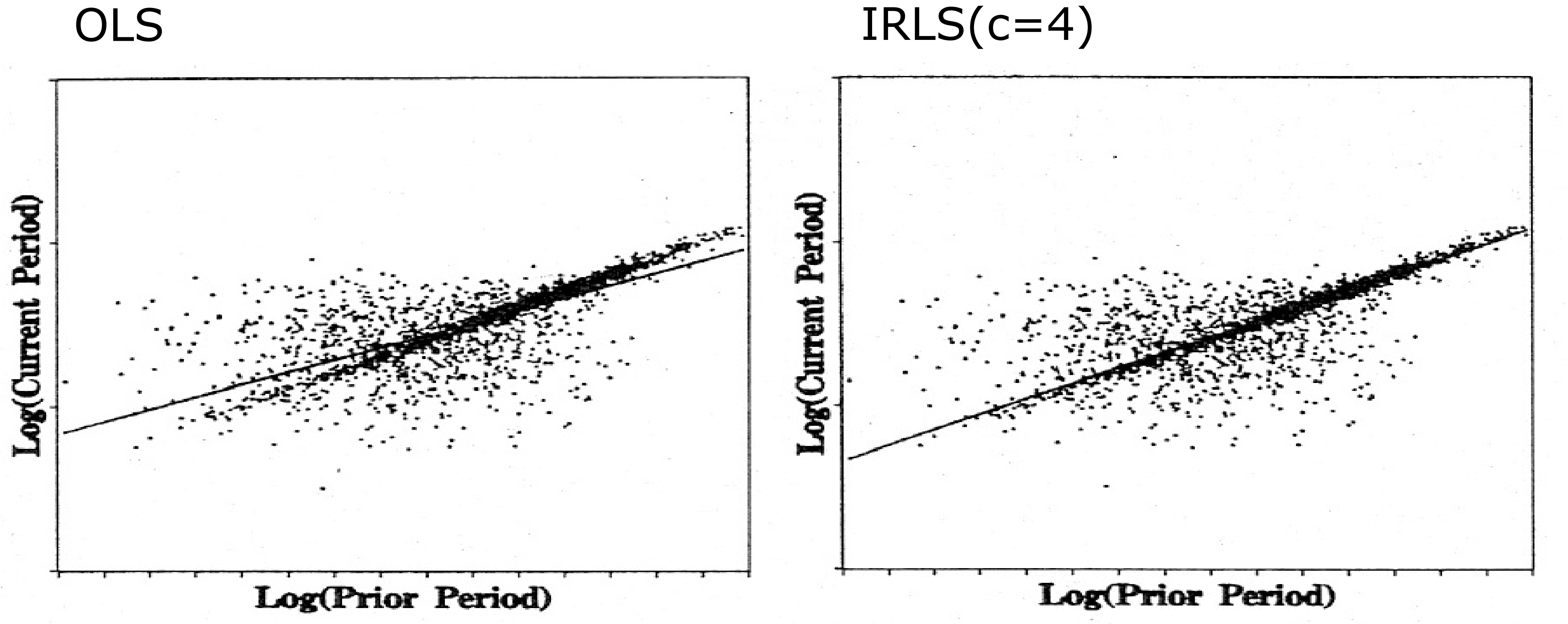
Bienias et al. (1997) Improving Outlier Detection in Two Establishment Surveys, Statistical Data Editing 2, Methods and Techniques. (UNSC and UNECE eds.), 76-83.

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

Wada, K. (2012) Detection of Multivariate Outliers : Regression imputation by the Iterative Reweighted Least Squares, Research memoir of the Statistics, Vol.69, pp.23-52, Statistical Research and Training Institute, Ministry of internal affairs and communications. [in Japanese.]

# Application to real survey data

The Monthly Wholesale Trade Survey
U.S. Census Bureau inventory data
[Bienias et al. (1997)]



OLS

IRLS(c=4)

# M-estimators

Regression model :

$$y_i = \beta_o + \beta_1 x_{i1} + \cdots \beta_p x_{ip} + \varepsilon_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \text{ ,}$$

where $\boldsymbol{x}_i = \left(1, x_{i1}, \dots, x_{ip}\right)^\top, \boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots, \beta_p\right)^\top$, $V(\varepsilon_i) = \sigma^2$.

Estimation equation of $\boldsymbol{\beta}$ :

$$\sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \boldsymbol{x}_i = 0,$$

where the standardised residuals $e_i = r_i/\hat{\sigma},$ and the weight function, $w_i = w(e_i)$.

e. g. Huber, P. J. & Ronchetti, E. M. (2009). Robust Statistics. 2nd ed., Wiley, New Jersey., p.171.

# IRLS algorithm

i. Initial estimation of $\boldsymbol{\beta}$ by OLS : $\boldsymbol{b}^{(\mathbf{0})}$

ii. Compute the scale parameter $\hat{\sigma}^{(0)}$ and the initial weights $w_i^{(1)}$ from the residuals $r_i^{(0)}$

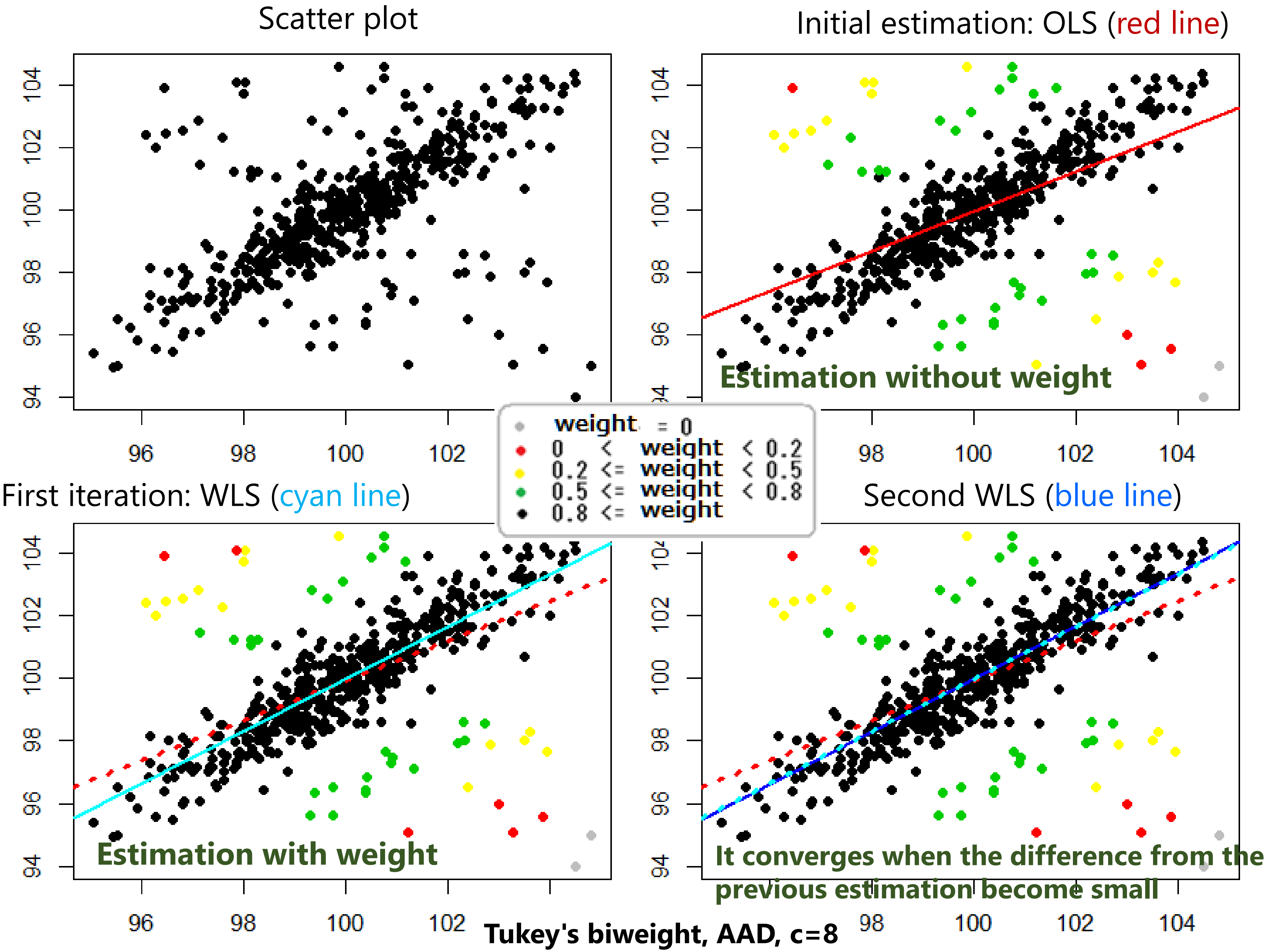
iii. [first iteration] Estimate $\boldsymbol{b}^{(\mathbf{1})}$ by WLS, obtain $r_i^{(1)}, \hat{\sigma}^{(1)}$ and $w_i^{(2)}$

iv. [jth iteration] Estimate $\boldsymbol{b}^{(j)}$ by WLS, obtain $r_i^{(j)}, \hat{\sigma}^{(j)}$ and $w_i^{(j+1)}$

v. [convergence condition] $\hat{\sigma}^{(j)}/\hat{\sigma}^{(j-1)} \approx 1$

Beaton, A. E. & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, **16**, 147-185.

# The mechanism of IRLS



Scatter plot

Initial estimation: OLS (red line)

Estimation without weight

First iteration: WLS (cyan line)

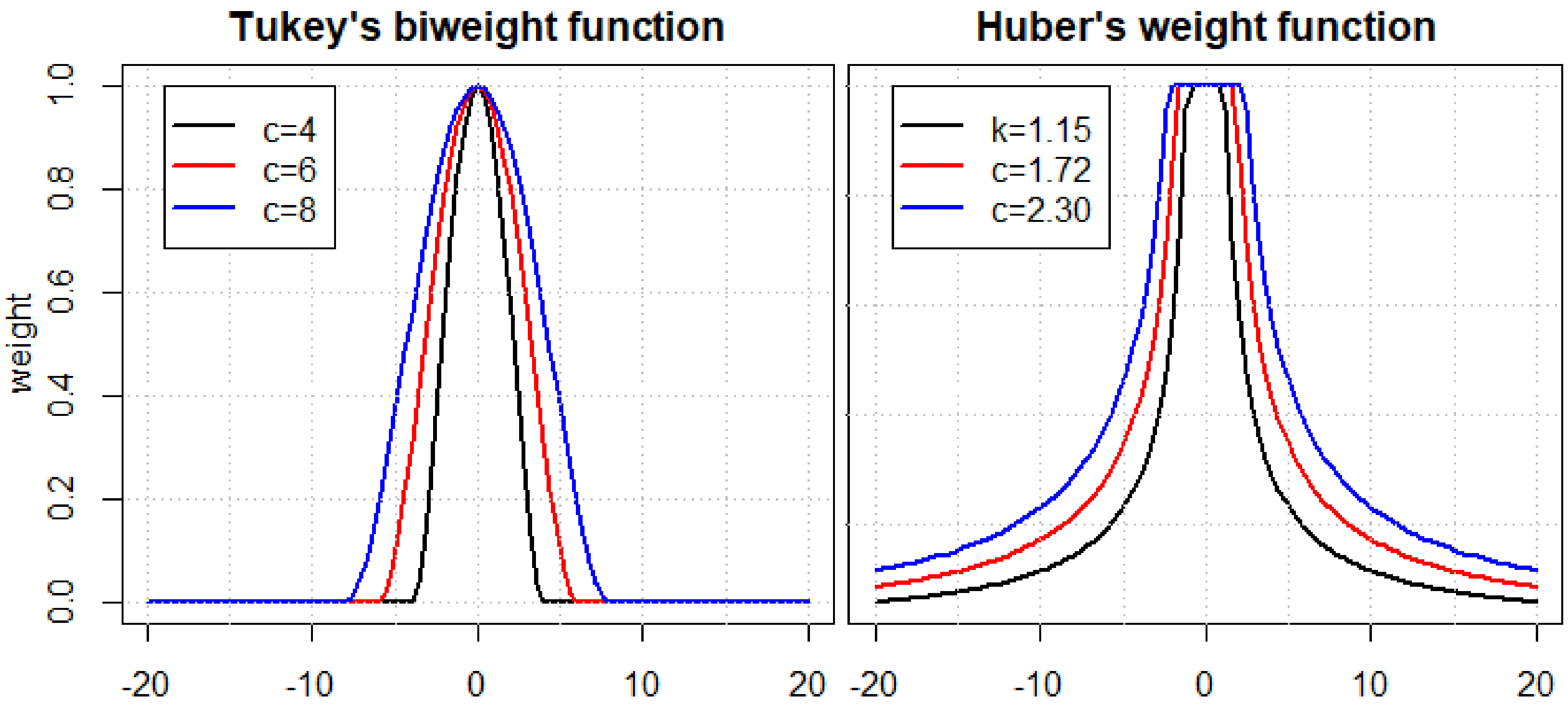
Estimation with weight

Second WLS (blue line)

It converges when the difference from the previous estimation become small

Tukey's biweight, AAD, c=8

# Popular weight functions



Standardised residual $e$

Standardised residual $e$

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

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

Tukey’s biweight eliminates influence of outliers with very large residuals.

Huber weight alleviate the influence of outliers but not eliminate it.

# An obstacle for the ratio model

**Heteroscedastic error term** $\epsilon_i \sim N(0, \sigma^2 x_i)$

**Homoscedastic error term** $\varepsilon_i \sim N(0, \sigma^2)$ is necessary for M-estimation

The relation of these error terms:

$$\varepsilon_i = \epsilon_i / \sqrt{x_i}$$

Divide the model equation by $\sqrt{x_i}$

$$y_i / \sqrt{x_i} = r\sqrt{x_i} + \epsilon_i / \sqrt{x_i}$$

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

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

# Modify the ratio model with homoscedastic error

$$y_i = r x_i + \epsilon_i$$

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

It becomes straight forward to make a robust estimation by means of M-estimation.

# 3. Generalisation of the ratio model

# Generalisation

Assume the original error term $\epsilon_i$ is proportional to $x_i^{\gamma}$

Homoscedastic error

**Model** $$y_i = rx_i + \varepsilon_i x_i^{\gamma}$$

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

※ $\gamma$ is an arbitral constant

# Generalisation accommodates different models

**γ=1:**

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

A′

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

**γ=1/2: Conventional ratio model**

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

B′

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

**γ=0: Single regression model without intercept**

C′

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

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

# Features of the estimator A' and B'

## Estimator A’

☺ $$\hat{r} = \frac{1}{n}\sum \frac{y_i}{x_i}$$

← Regardless of the magnitude of each variable, the rate of each observation is averaged.

☹ The variance can be very large.

## Estimator B’

☺ $$\hat{r} = \frac{\sum y_i}{\sum x_i}$$

← The very large values in each variable have great influence to the estimand.

☺ The variance is small.

# 4. Robustification of the generalised ratio model

# Robustified estimators

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

**γ=1:**

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

A

**γ=1/2:**

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

B

# An implementation

## Weight function : Tukey's biweight

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

## Quasi-residuals

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

A

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

B

## Scale parameter of quasi-residuals

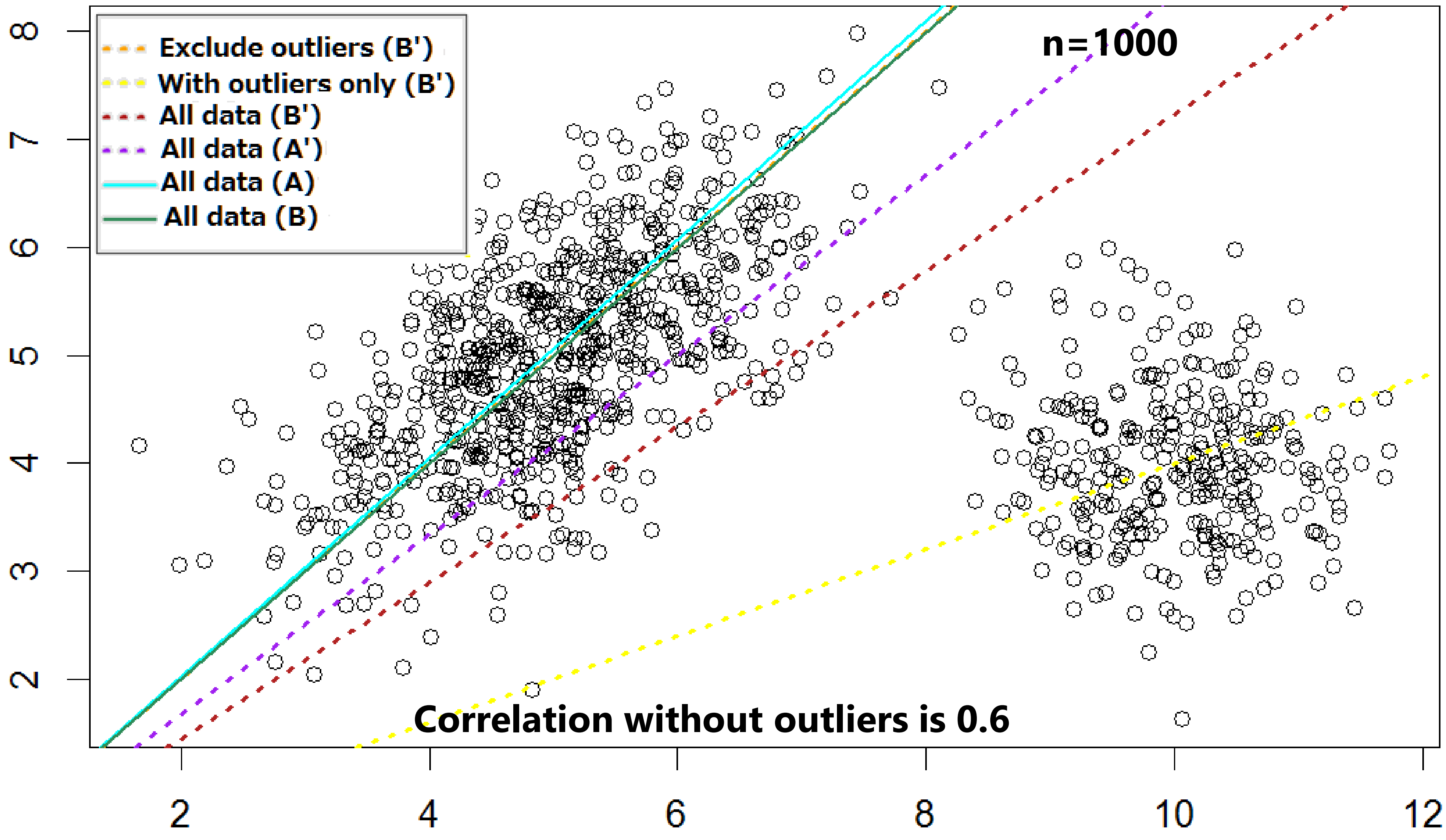
$$\sigma_{\mathrm{AAD}} = \frac{1}{n}\sum_{i=1}^{n}|\breve{\varepsilon}_i|$$

**Tuning constant *c*** : 8（Usually users are supposed to choose from 3 to 8.)

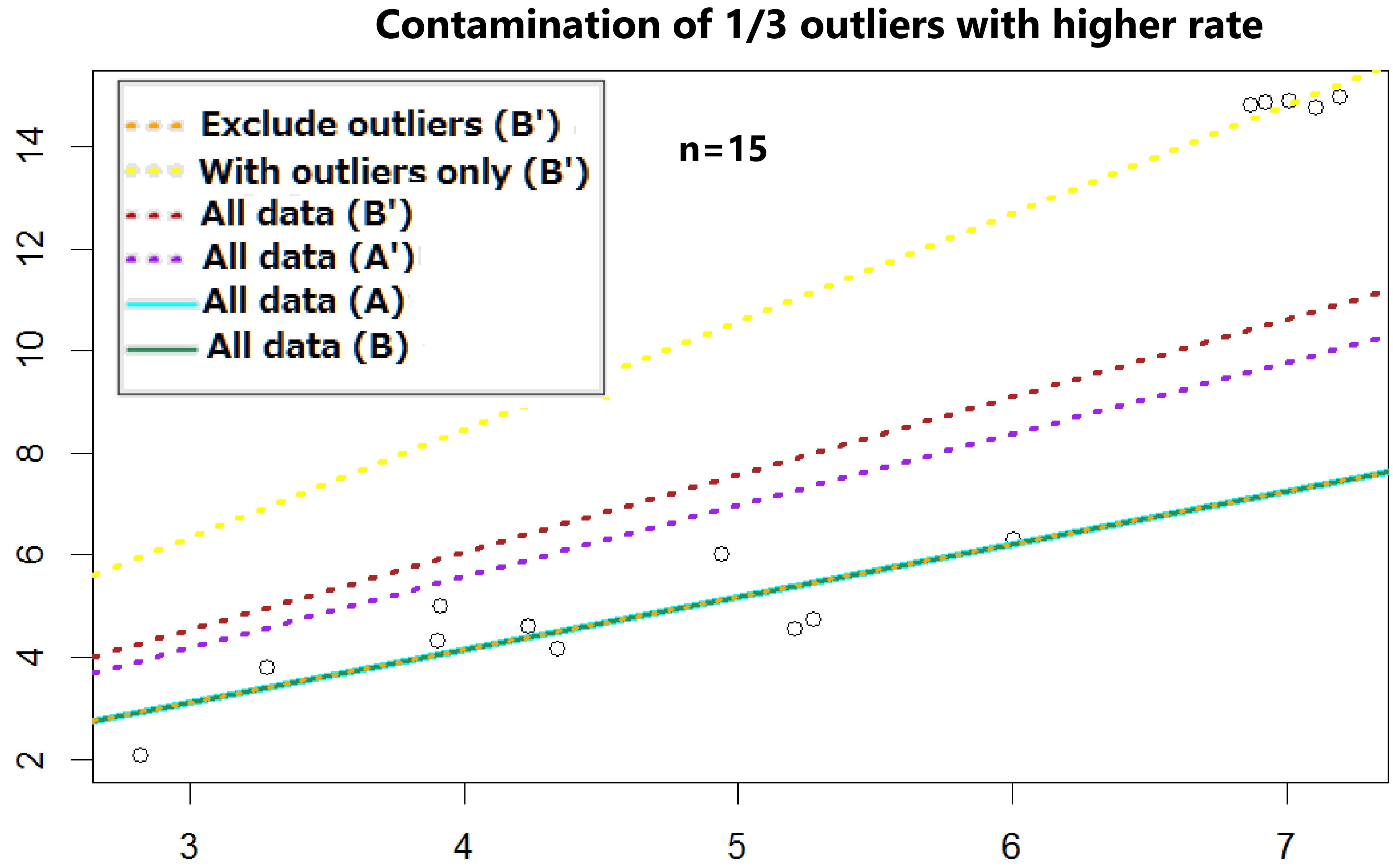
# 5. The effect of the robustification

# The effect of the robust estimation

Contamination of 30 % outliers with lower rate

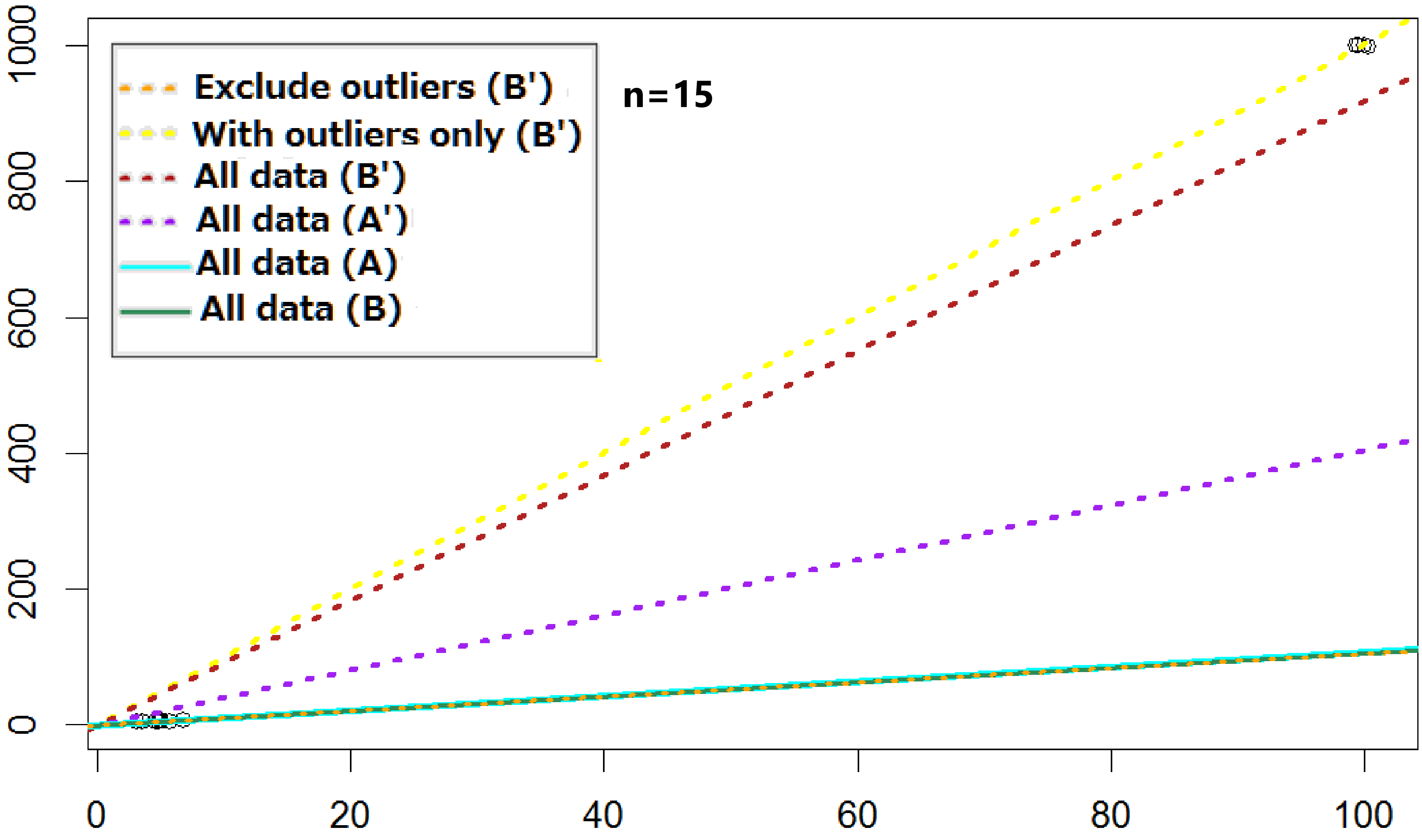


Robust estimation tolerates 30% contamination of outliers with lower rate.



- Estimator A and B are still robust with smaller size of dataset.
- Estimator B' are more influenced than A' by extremely large outliers.

# Contamination of 1/3 extreme outliers with higher rate



Estimator A and B are hardly affected by the extreme outliers.

# 6. Application to Economic Census data

# The 2016 Economic Census for Business Activity

The Census is conducted by the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry as of June 1, 2016.  It aims to

- identify the structure of establishments and enterprises in all industries on a national and regional level, and
- obtain basic information to conduct various statistical surveys by investigating the economic activity of these establishments and enterprises.

# Application

Imputation of the major corporate accounting items for the 2016 Economic Census for Business Activity

### Requirements

- **Ratio model**
- **Alleviation of outliers' influence**

# Compared models

* Reprint of slide 25

## Weight function : Tukey's biweight

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

## Quasi-residuals

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

A

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

B

## Scale parameter of quasi-residuals

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

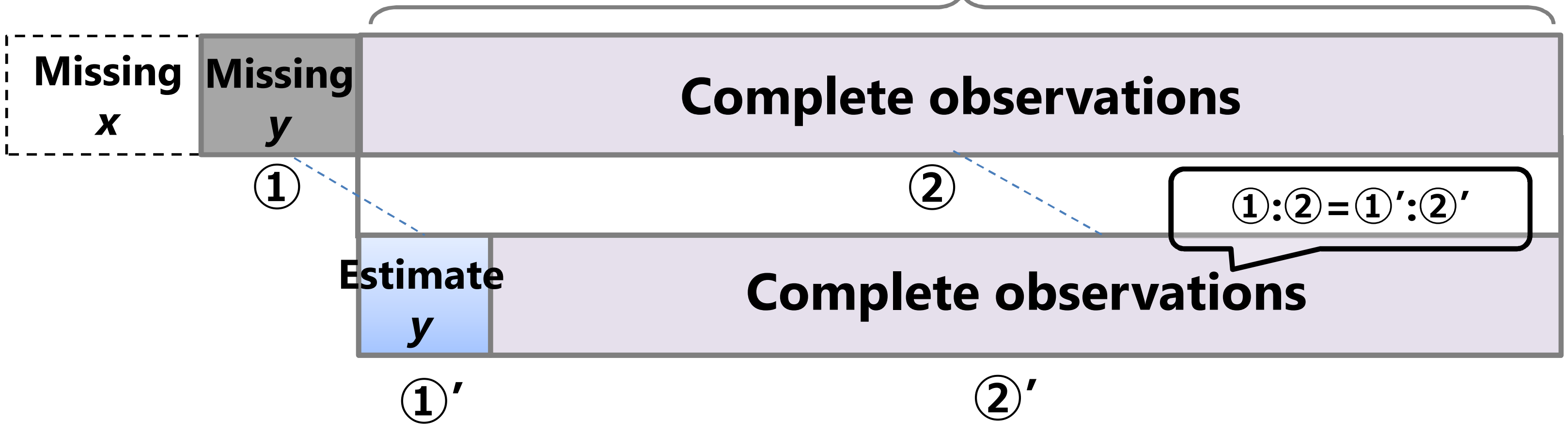
**Tuning constant *c*** : 8 （Usually users are supposed to choose from 3 to 8.）

# Monte Carlo simulation

- Using previous 2012 Census data [about 3 million records]
- Imputed variables : sales (by expenditure), expenditure (by sales), salary (by expenditure)

1. Extract complete observations regarding the objective and explanatory variables
2. Estimate values to impute regarding randomly selected observations according to the actual rate of missing data [* the observations assumed to be missing are limited to the thresholds **below Q3 + IQR × 3** regarding both *x* and *y*]
3. Compare the total sum of the absolute deviation between the real and estimated value

**Data set used for the simulation （exclude extremely large data）**



# Result: estimator B for all variables

| Industrial classification | Least sum of deviations from real values (Number of Domains) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sales | | | Expenditure | | | Salary | | |
| Estimator | (A) | (B) | (B)′ | (A) | (B) | (B)′ | (A) | (B) | (B)′ |
| Level 1.5 | **20** | **122** | **15** | **48** | **106** | **55** | **74** | **131** | **37** |
| Level 2 | **23** | **115** | **18** | **39** | **105** | **54** | **63** | **115** | **32** |
| Level 3 | **5** | **109** | **22** | **34** | **93** | **32** | **32** | **70** | **30** |
| Level 3.5 | **4** | **138** | **7** | **22** | **102** | **17** | **40** | **65** | **52** |

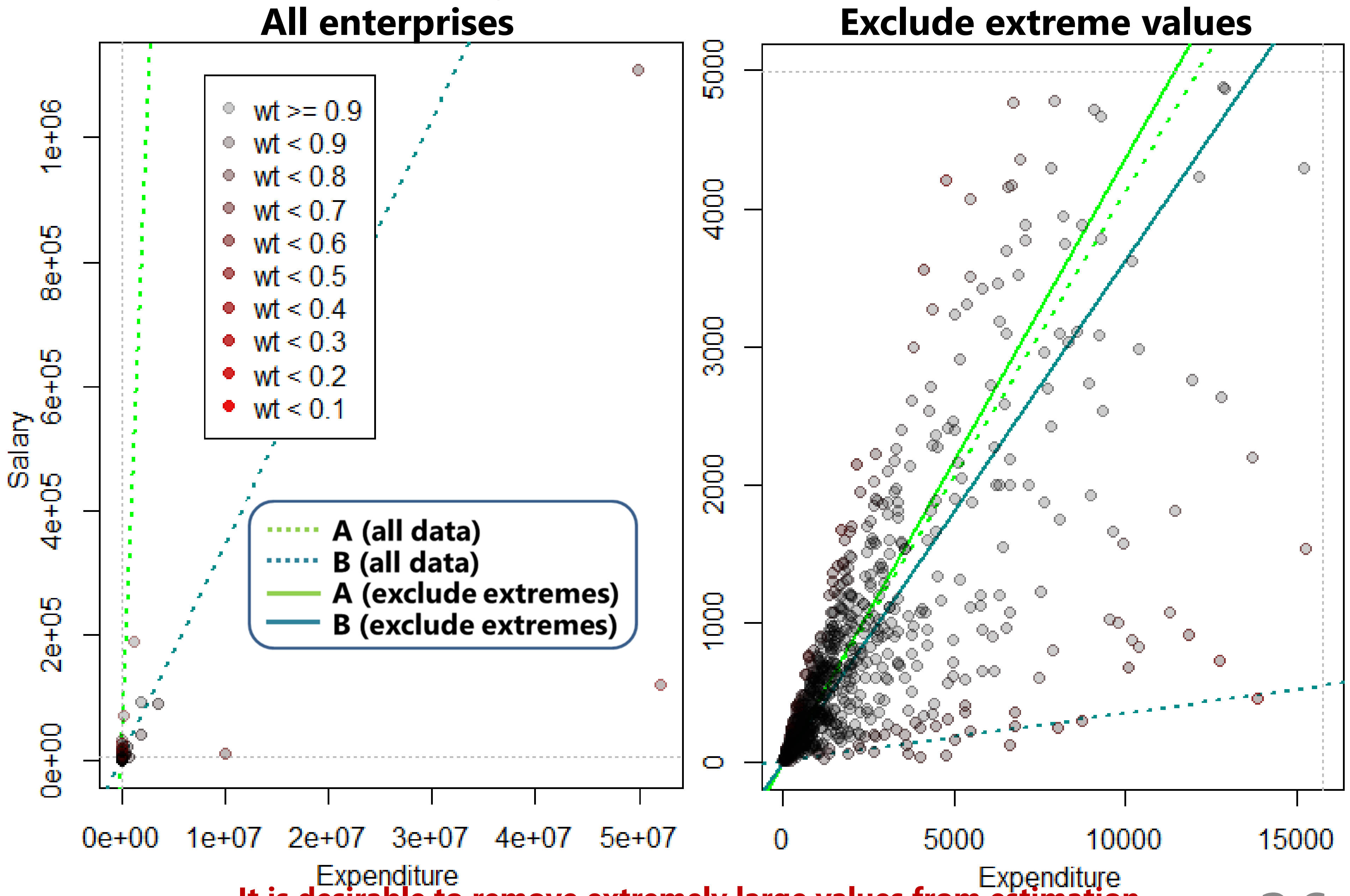
| Industrial classification | Minimum average deviations from real values (Number of Domains) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sales | | | Expenditure | | | Salary | | |
| Estimator | (A) | (B) | (B)′ | (A) | (B) | (B)′ | (A) | (B) | (B)′ |
| Level 1.5 | **10** | **122** | **9** | **38** | **103** | **38** | **40** | **138** | **43** |
| Level 2 | **10** | **113** | **16** | **34** | **108** | **38** | **28** | **125** | **36** |
| Level 3 | **9** | **103** | **33** | **29** | **104** | **30** | **36** | **75** | **39** |
| Level 3.5 | **11** | **130** | **14** | **27** | **99** | **34** | **37** | **58** | **51** |

**Larger figure shows favorable result**

| Industrial classification | Maximum sum of deviations from real values (Number of Domains) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sales | | | Expenditure | | | Salary | | |
| Estimator | (A) | (B) | (B)′ | (A) | (B) | (B)′ | (A) | (B) | (B)′ |
| Level 1.5 | **111** | **2** | **8** | **104** | **2** | **15** | **50** | **1** | **6** |
| Level 2 | **107** | **2** | **9** | **105** | **2** | **17** | **58** | **8** | **7** |
| Level 3 | **89** | **5** | **7** | **109** | **12** | **20** | **94** | **32** | **38** |
| Level 3.5 | **220** | **7** | **6** | **225** | **21** | **24** | **152** | **15** | **136** |

**Larger figure shows unfavorable result**

# 55A Agents and brokers



It is desirable to remove extremely large values from estimation as estimator B is greatly affected by them.

# Practical Application of the robust ratio estimator

The proposed methods are applied for the **2016 Economic Census for Business Activity** in Japan.

All the simulations and analysis are done within the R environment.

**Wada, K. and Sakashita, K. (2017) Generalized robust ratio estimator for imputation, Proceedings of New Techniques and Technologies for Statistics, 14-16, Mar. Brussels, Belgium.**