UNECE Workshop on Statistical Data Editing 31 Aug. – 4 Sep. 2020

Robust Tools for Statistical Data Editing and Imputation

in R

WADA Kazumi (Tsuda University) and
TSUBAKI Hiroe (The Institute of Statistical Mathematics)

Outline

I. M-estimators for regression (with two weight functions and scale parameters)
II. M-estimators for generalised ratio model (with Tukey's biweight function and average absolute deviation (AAD) for scale parameter)
III. Modified Stahel-Donoho estimator

I. M-estimators for regression

Implementation of an R function based on Bienias et al (1997) [Tukey's biweight function with AAD scale] together with other variations as follows:

- Tukey's biweight function with MAD* scale
- Huber's weight with AAD** scale and MAD scale

* MAD (Median Absolute Deviation)
** AAD (Average Absolute Deviation)

M-estimators for regression

Model :

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

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

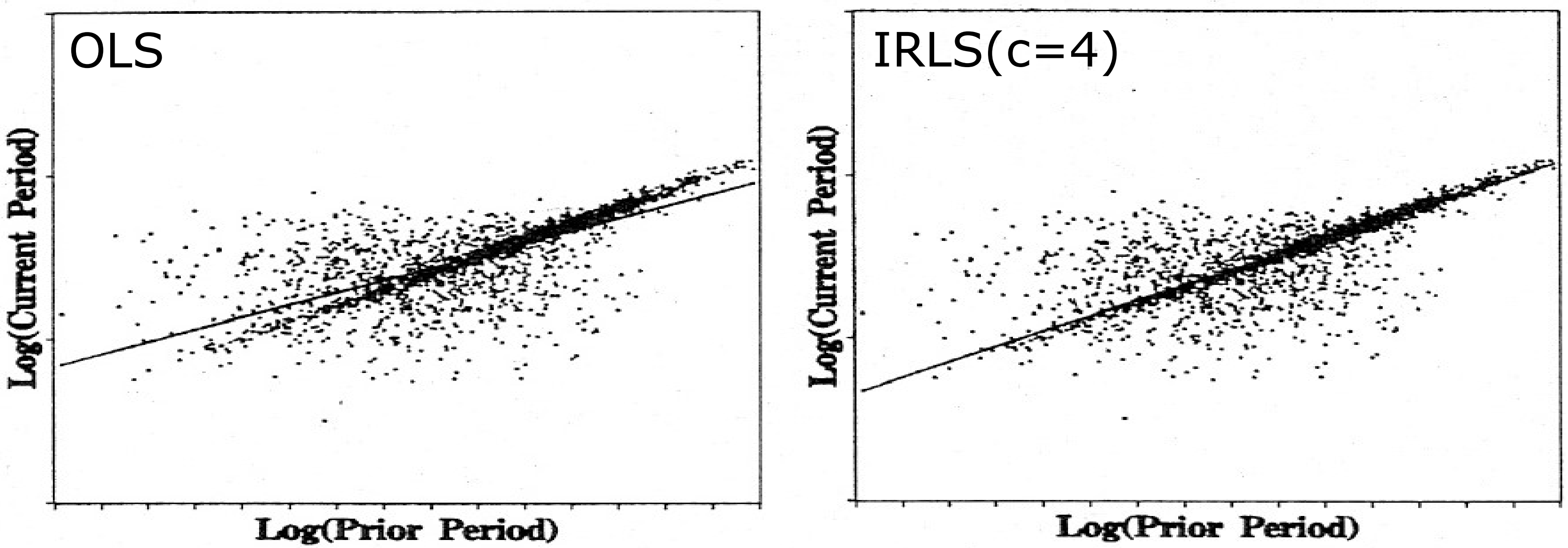
Estimation equation of $\boldsymbol{\beta}$:

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

where the weight function, $w_i = w(e_i)$ is based on the standardized residuals $e_i = (\hat{y}_i - y_i)/\hat{\sigma} = r_i/\hat{\sigma}$.

An example

Statistical Data Editing, Vol.2 – Theory and Methods

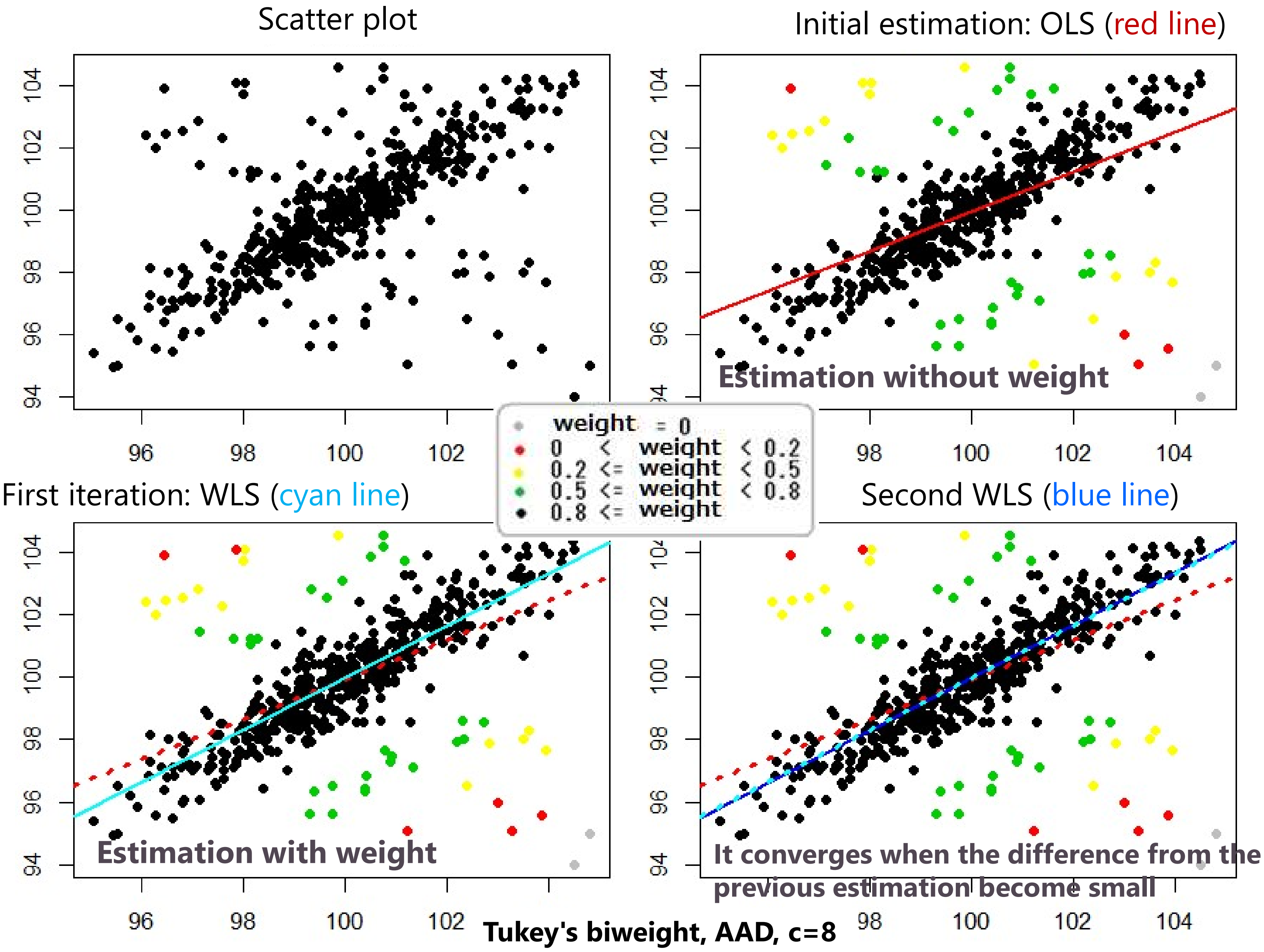


The Monthly Wholesale Trade Survey
– inventory data (U.S. Census Bureau)

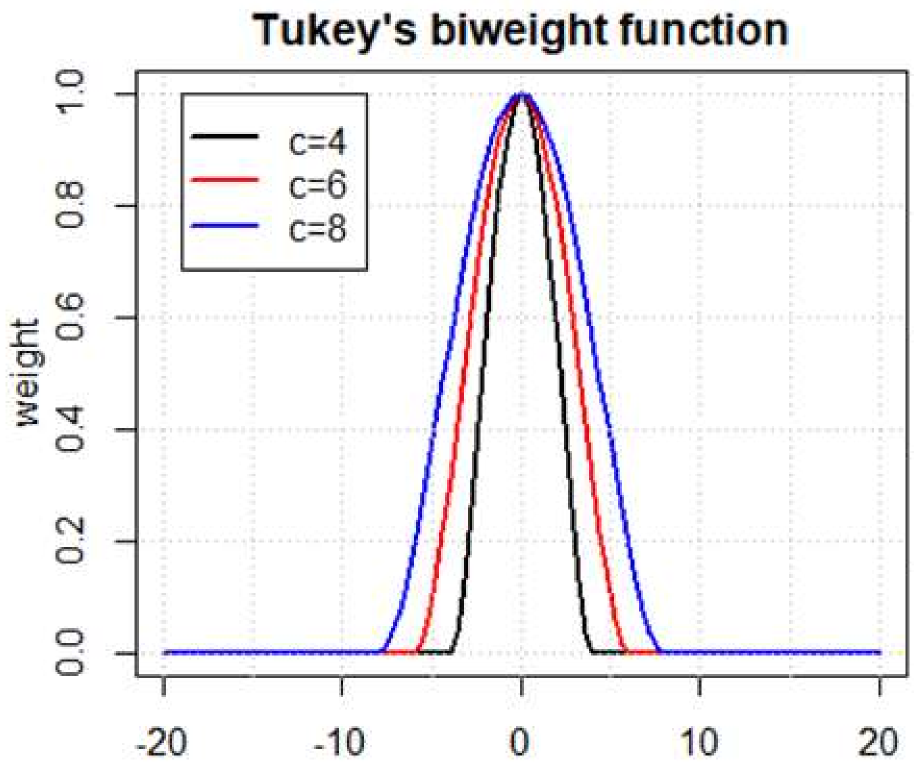
IRLS algorithm

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 $\boldsymbol{b}^{(\boldsymbol{j})}$ by WLS, obtain $r_i^{(j)}, \hat{\sigma}^{(j)}$ and $w_i^{(j+1)}$
v. [convergence condition] $\hat{\sigma}^{(j)}/\hat{\sigma}^{(j-1)} \approx 1$

How IRLS works



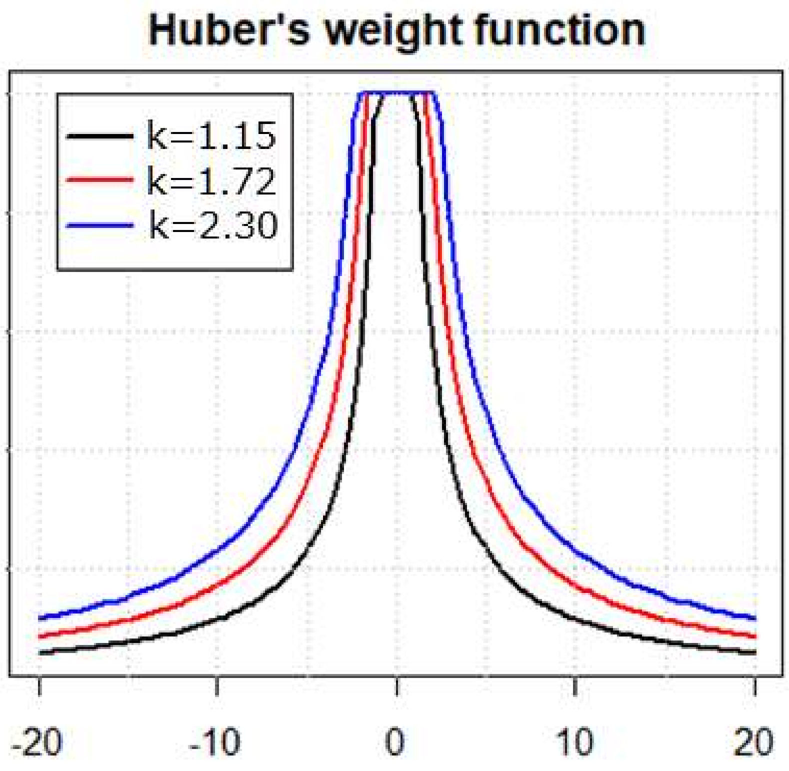
Popular weight functions



Standardized residual e

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

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



Standardized residual e

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

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

Relation between the two tuning constants

$$\frac{\sigma_{\mathrm{AAD}}}{\sigma_{\mathrm{SD}}} = \frac{E|e|}{\sqrt{E(e^2)}} = \sqrt{\frac{2}{\pi}} \approx 0.80, \qquad \text{and}$$

$$\sigma_{\mathrm{SD}} = 1/\Phi^{-1}\left(\frac{3}{4}\right) \cdot \sigma_{\mathrm{MAD}} \approx 1.4826 \cdot \sigma_{\mathrm{MAD}}$$

Tuning constants for 95% asymptotic efficiency with different measures of scale

Tuning constant	95% asymptotic efficiency		
	σ_{SD}	σ_{MAD}	σ_{AAD}
c for Tukey	4.685	3.160	3.738
k for Huber	1.345	0.907	1.073

Tuning constants scaled for a comparison

Bienias et al. (1997)

Tuning constant	Range of tuning constant for σ_{SD}			Range of tuning constant for σ_{MAD}			Range of tuning constant for σ_{AAD}		
c for Tukey	5.01	7.52	10.03	7.43	11.15	14.87	4.00	6.00	8.00
k for Huber	1.44	2.16	2.88	2.13	3.20	4.27	1.15	1.72	2.30

Software

R functions available at https://github.com/kazwd2008/IRLS

File name	R function	Weight function	Scale parameter
Tirls.r	Tirls.aad	Tukey	AAD
	Tirls.mad		MAD
Hirls.r	Hirls.aad	Huber	AAD
	Hirls.mad		MAD

Settings are based on Bienias et al.(1997) for Tukey's biweight function with AAD scale

References and related works

- **IRLS and Tukey's biweight function**

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

- **Huber's weight function**

  Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics*, **35**(1), 73-101.

- **Algorithm implemented**

  Bienias, J. L., Lassman, D. M. Scheleur, S. A. & Hogan H. (1997). Improving outlier detection in two establishment surveys. *Statistical Data Editing* 2 - *Methods and Techniques*. (UNSC and UNECE eds.), 76-83.

- **Implementation and evaluation**

  Wada, K. (2012). Detection of multivariate outliers: Regression imputation by the iteratively reweighted least squares - (**in Japanese**)." *Research Memoir of Official Statistics*, **69**, 23-52. URL https://www.stat.go.jp/training/2kenkyu/ihou/69/pdf/2-2-692.pdf. .

  Wada, K. & Noro, T. (2019). Consideration on the Influence of Weight Functions and the Scale for Robust Regression Estimator (**in Japanese**). *Research Memoir of Official Statistics*, **76**, 101-114. URL https://www.stat.go.jp/training/2kenkyu/ihou/76/pdf/2-2-767.pdf.

  Wada, K. (2020) Outliers in official statistics, to be appeared in *Japanese Journal of Statistics and Data Science* (JJSD)

II. M-estimators for generalised ratio model

i. Difference between the ratio model and a regression model
ii. Robustification and generalisation
iii. Further challenge (simultaneous estimation)

Conventional ratio model

$$y_i = \beta x_i + \epsilon_i$$

y: objective variable

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

β : the ratio of y and x

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

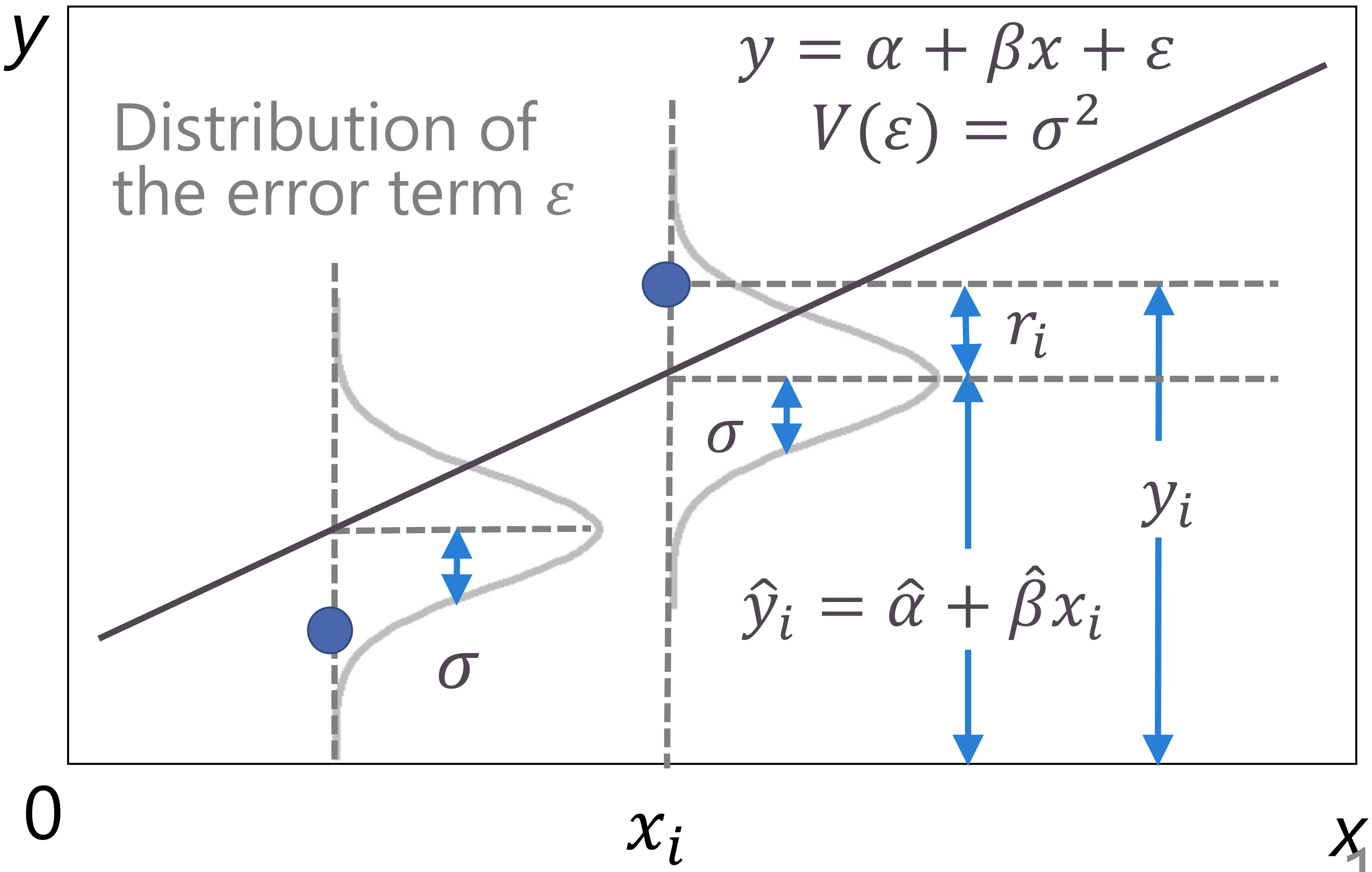
When the model is used for imputation of y, unknown r due to missingness is estimated with complete data of x_i and y_i.

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

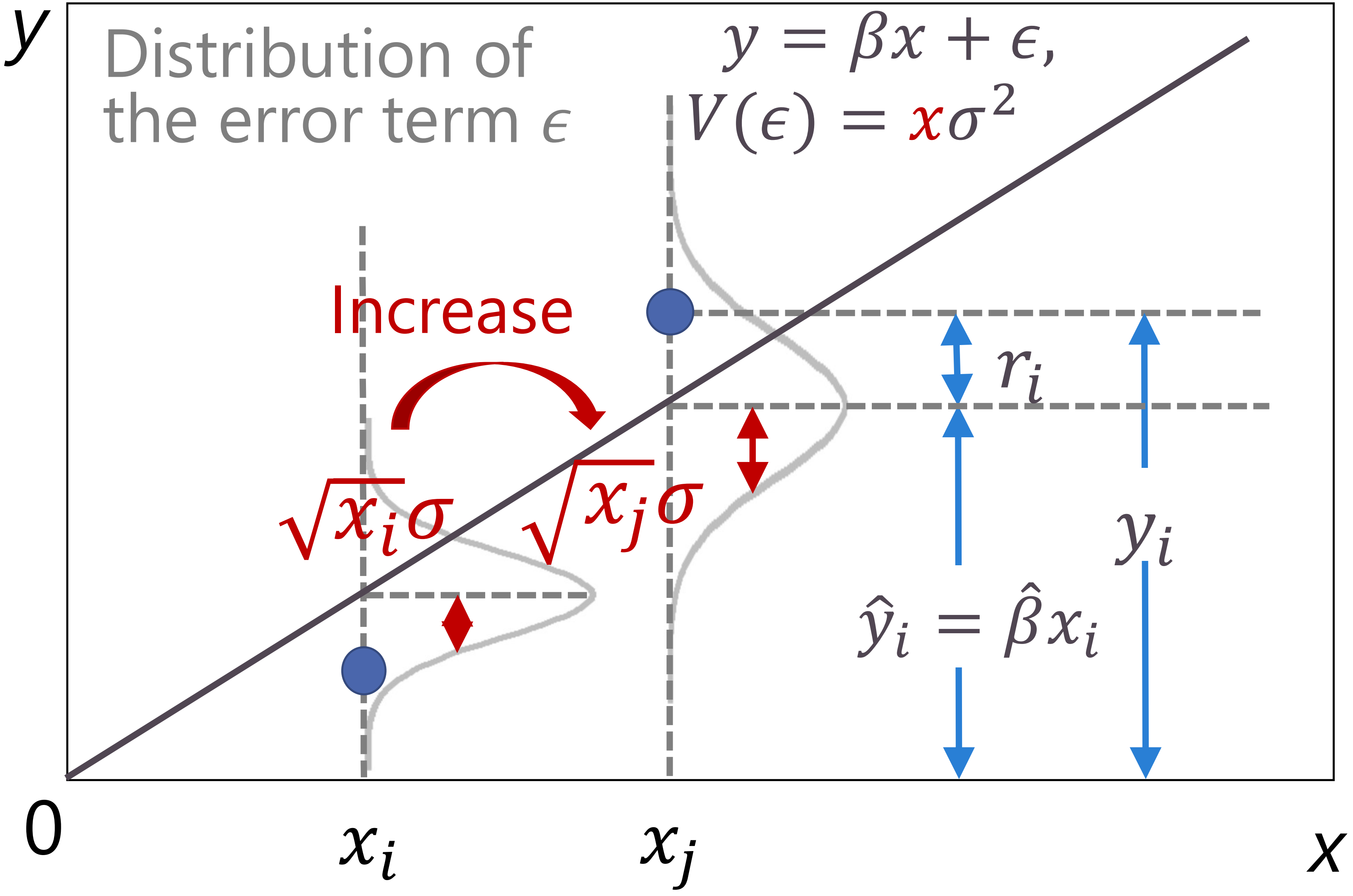
obs: complete observations regarding these two variables

e.g. De Waal et al. (2011)

Linear regression model



Conventional ratio model



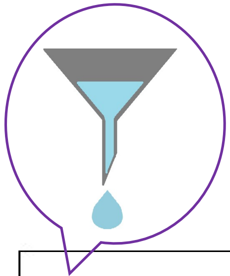
Difference between the ratio model and a linear regression model

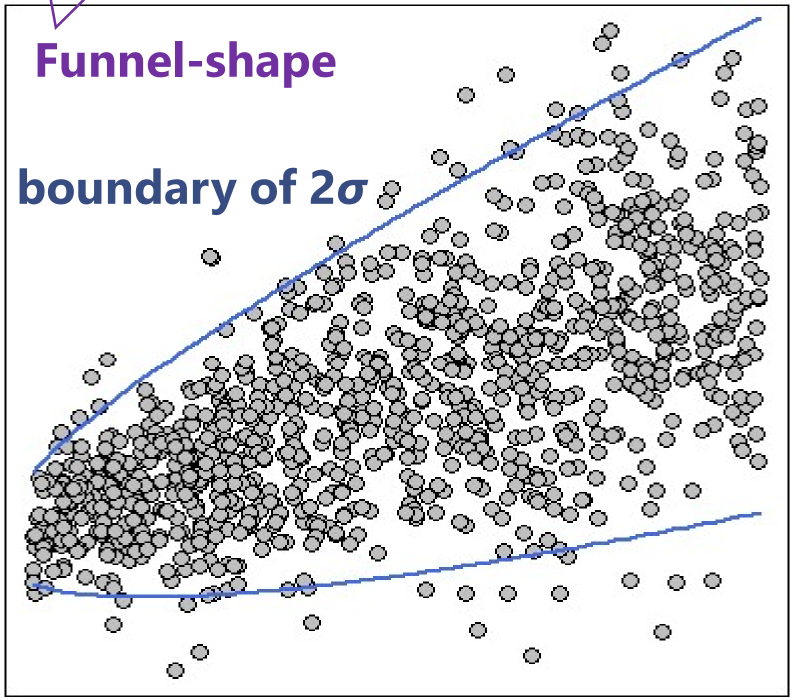
Ratio model

$$y_i = \beta x_i + \epsilon_i$$

Heteroscedastic

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



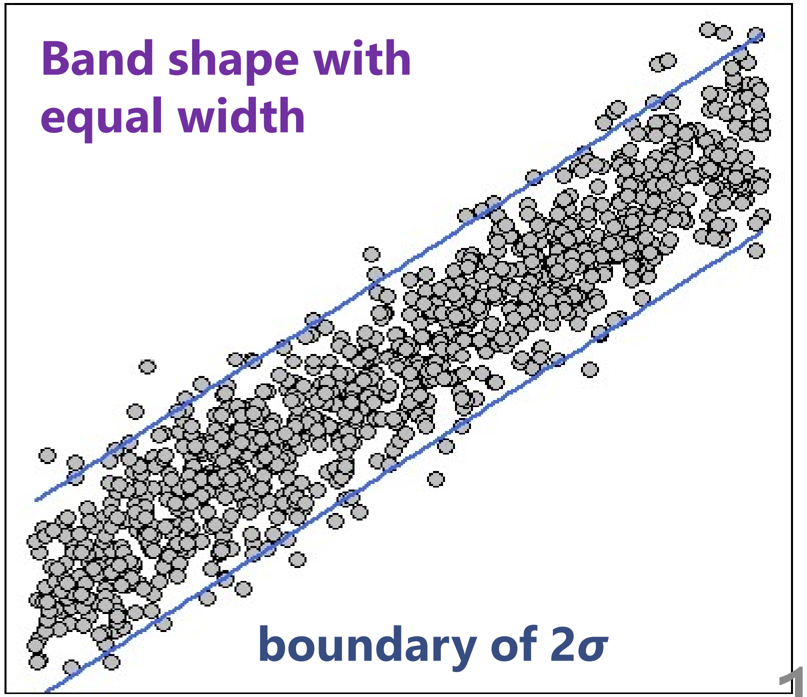


Regression model

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

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

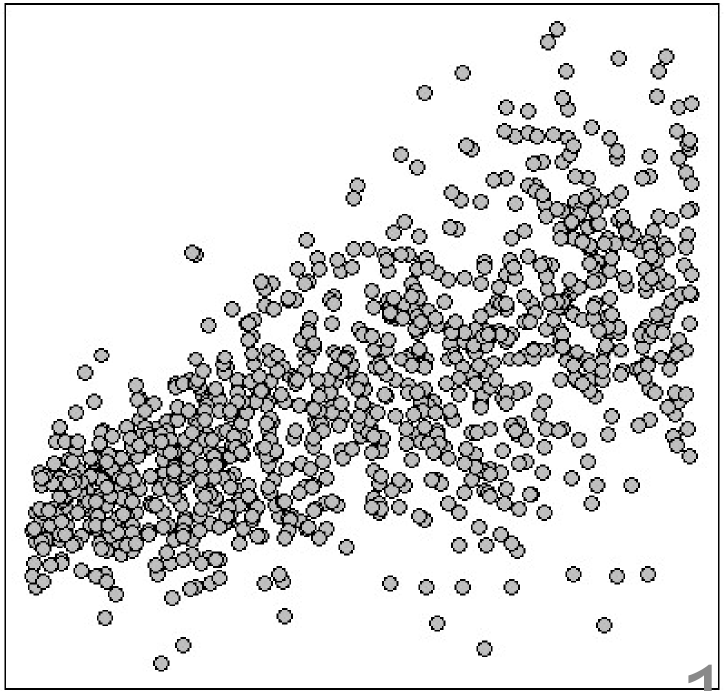
$$\log y_i = \alpha + \beta \cdot \log x_i + \varepsilon_i$$

Bias correction

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

The ratio model has an advantage for imputation.

Survey data often have heteroscedastic variance



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 $\sqrt{x_i}$

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

$$\underline{y_i = \beta x_i + \varepsilon_i \sqrt{x_i}}$$

Modify the ratio model with homoscedastic error

$$y_i = \beta x_i + \epsilon_i$$
$$y_i = \beta x_i + \varepsilon_i \sqrt{x_i}$$

Now M-estimation is possible as same as a regression model.

Generalisation

Assume the original error term ϵ_i is proportional to x_i^{γ}

Homoscedastic error

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

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

※ γ is an arbitral constant. When $\gamma = 1/2$, the model is identical to the conventional ratio model.

Generalisation accommodates different models

γ=1:

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

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

A′

γ=1/2: Conventional ratio model

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

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

B′

γ=0: Single regression model without intercept

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

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

C′

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.

Robustified estimators

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

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

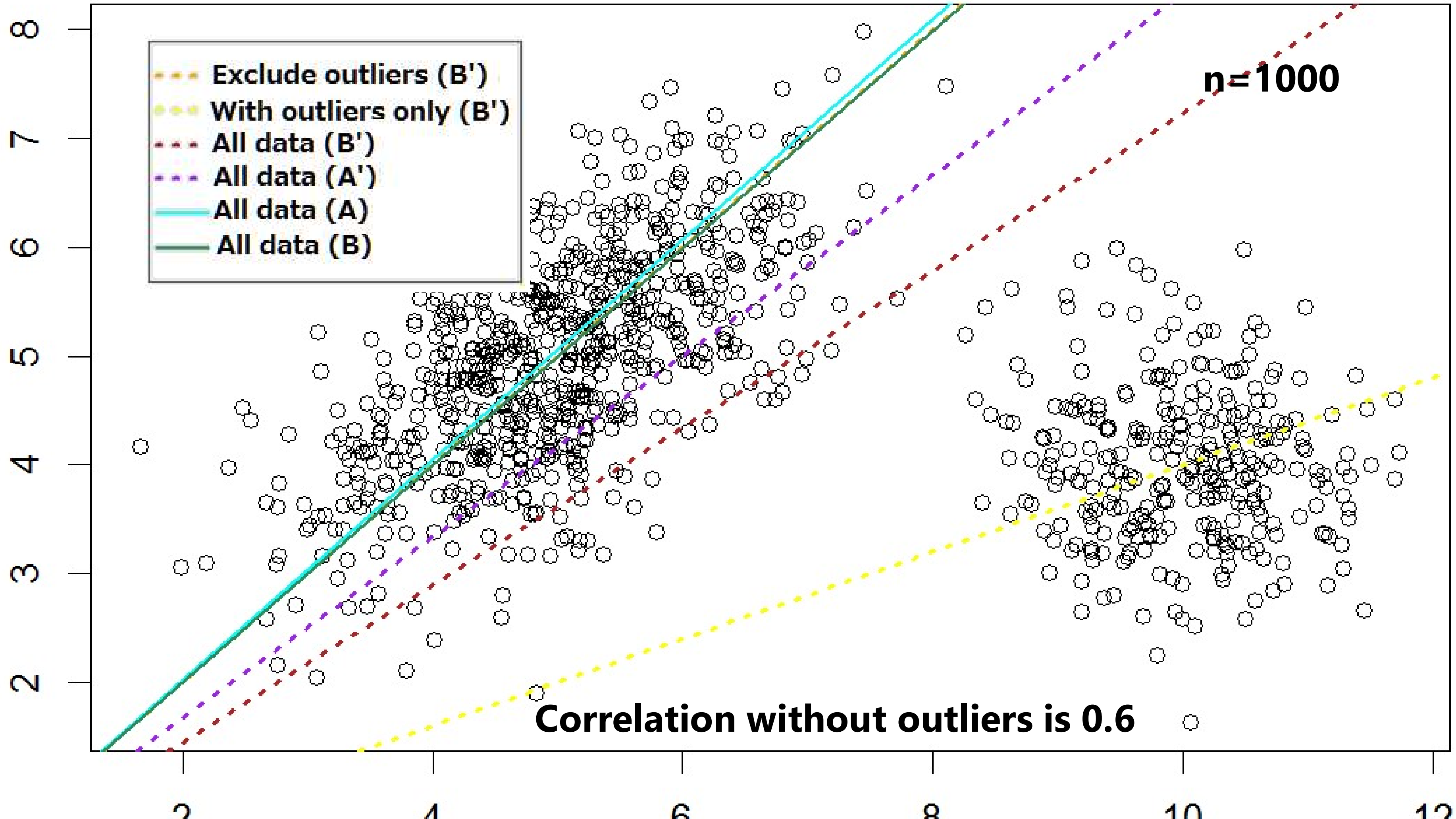
Scale parameter of quasi-residuals

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

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

The effect of the robust estimation(1)

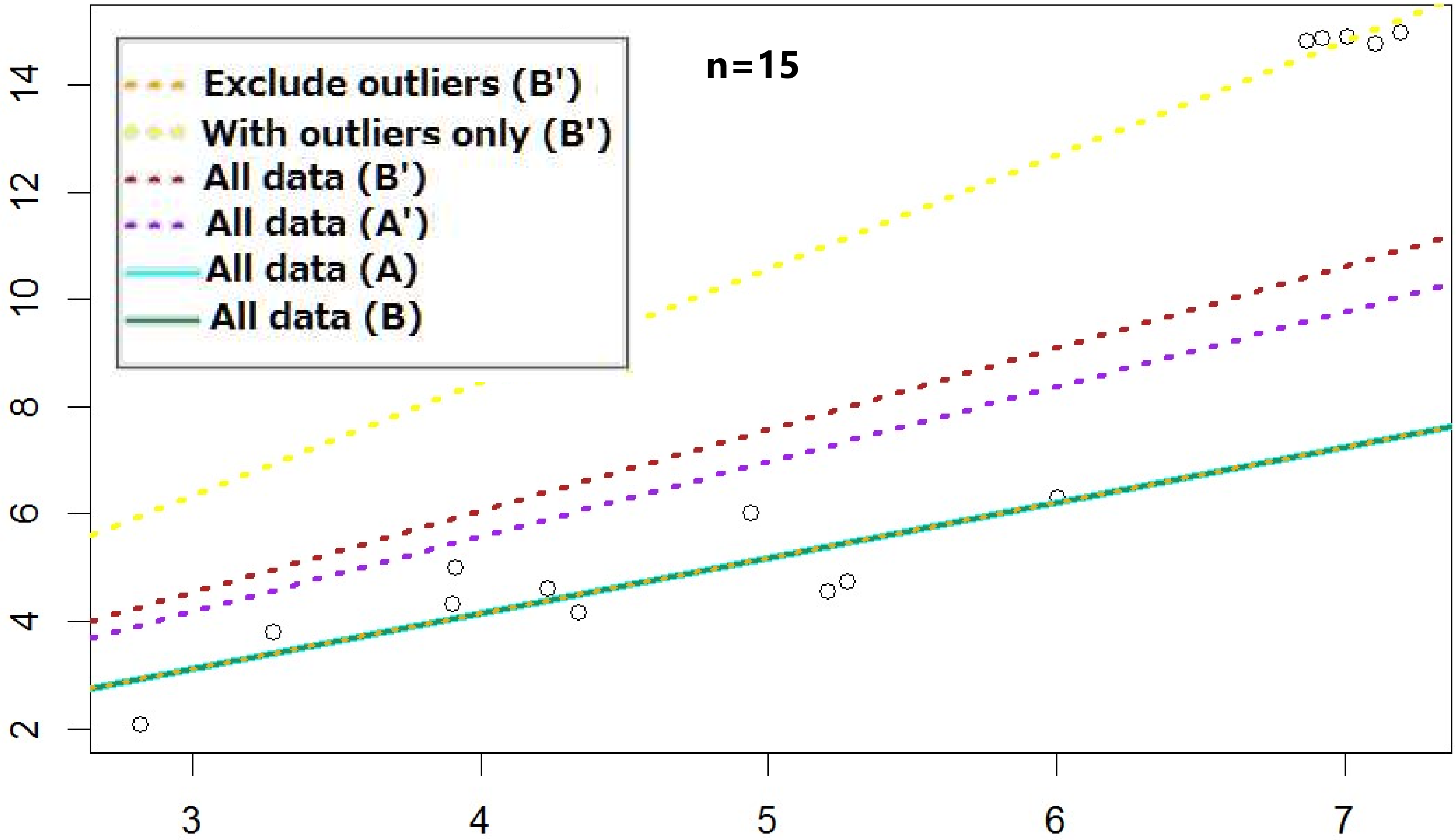
Contamination of 30 % outliers with lower rate



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

The effect of the robust estimation(2)

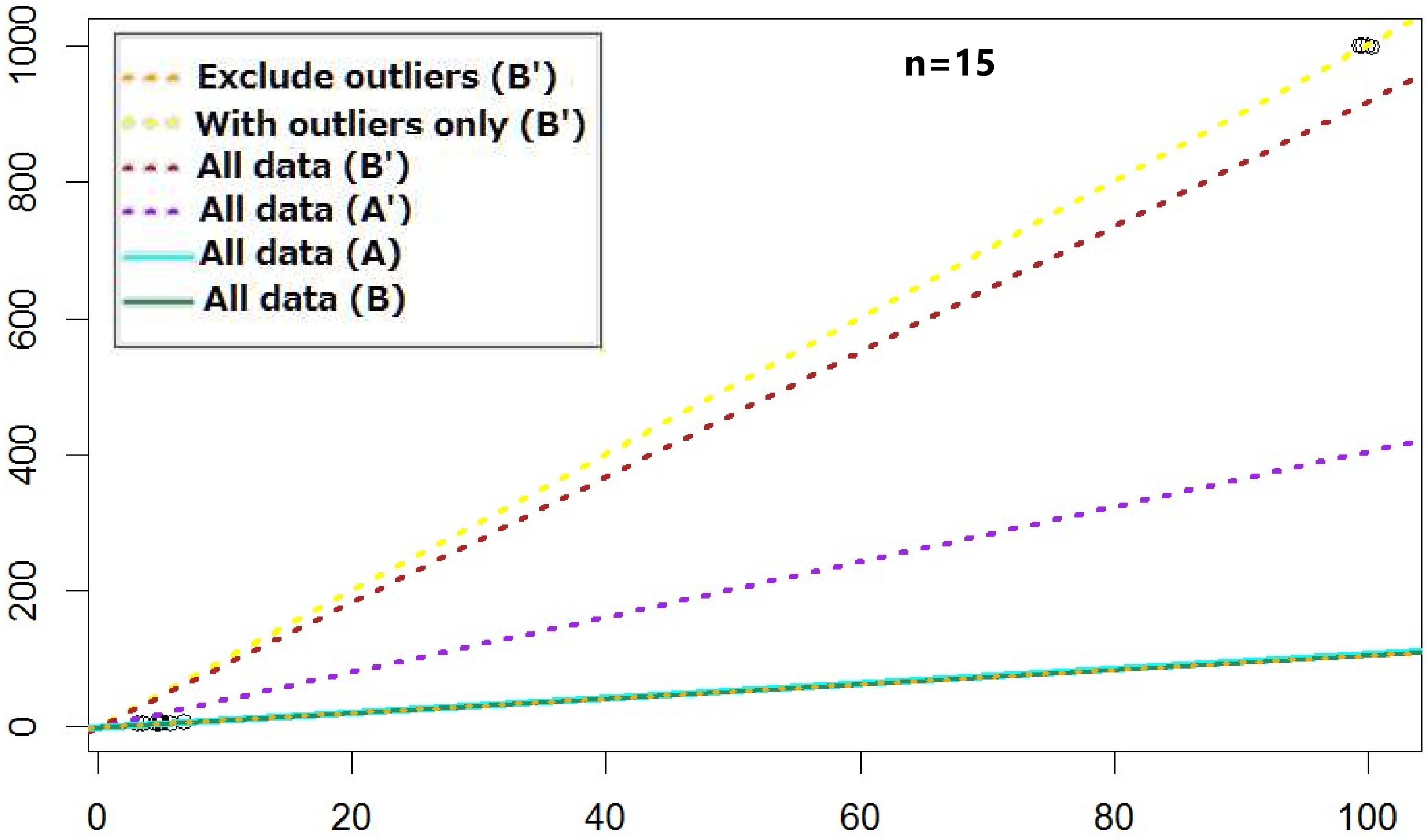
Contamination of 1/3 outliers with higher rate



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

The effect of the robust estimation(3)

Contamination of 1/3 extreme outliers with higher rate



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

Practical application

■2016 Economic Census for Business Activity

Estimator B (γ=1/2) is used for the imputation of major corporate accounting items after model selection

■2019 Unincorporated Enterprise Survey

Estimator B (γ=1/2) is used for the imputation of sales from the previous data of the same enterprise

Further improvement

Automation of the model selection

⇨ **Estimation of γ together with β**

Two different R functions developed.

Two R functions developed

Code A: (computationally efficient version)

1. Estimate $\hat{\beta}^0 = \sum y_i x_i^{1-2\gamma} / \sum x_i^{2(1-\gamma)}$ with an appropriate initial value γ^0

Homoscedastic

2. Compute w_i^0 based on quasi residuals $\breve{r}_i = (y_i - \hat{\beta} x_i)/x_i^{\gamma}$ and then estimate γ^1 by computing $(z^\top z)^{-1} z^\top s$,

Weighted two-stage least squares

$$\text{where} \quad z = \begin{bmatrix} 1 & \log(x_1 w_1) \\ \vdots & \vdots \\ 1 & \log(x_n w_n) \end{bmatrix}, \quad s = \log\left(\left|y_i - \hat{\beta} x_i\right| w_i\right),$$

Heteroscedastic residuals

$$w = (w_1, \ldots, w_n).$$

3. Repeat the followings until the difference of the scale parameter meets the convergence condition
   i. Estimate $\hat{\beta}^k = \dfrac{\sum w_i y_i (w_i x_i)^{1-2\gamma}}{\sum (w_i x_i)^{2(1-\gamma)}}$ with the latest $\hat{\gamma}^{(k-1)}$ and $w_i^{(k-1)}$
   ii. Obtain new weight w_i^k and estimate γ^k

Code B: (stepwise version)

1. Estimate $\hat{\beta} = \sum y_i x_i^{1-2\gamma} / \sum {x_i}^{2(1-\gamma)}$ with an appropriate initial value γ^0
2. Repeat **non robust estimation of β and γ** until the difference of the scale parameter $\hat{\sigma}$ meets the convergence condition
   i. Estimate γ using $\hat{\beta}$ by computing $(z'^{\top} z')^{-1} z'^{\top} s'$, where
   $$z' = \begin{bmatrix} 1 & \log(x_1) \\ \vdots & \vdots \\ 1 & \log(x_n) \end{bmatrix}, \; s' = \log(|y_i - \hat{\beta} x_i|)$$
   ii. Estimate new $\hat{\beta}$ using $\hat{\gamma}$

R function *RBred*

3. Repeat **robust estimation of $\widehat{\beta}$** $= \frac{\sum w_i y_i (w_i x_i)^{1-2\gamma}}{\sum (w_i x_i)^{2(1-\gamma)}}$ where w_i is decided by quasi residuals $\breve{r}_i = (y_i - \hat{\beta} x_i)/x_i^{\gamma}$ based on $\hat{\gamma}$ obtained in the previous step
4. Repeat **robust estimation of γ** by computing $(z^{\top} z)^{-1} z^{\top} s$, where $w = (w_1, \dots, w_n)$.
5. Repeat the **robust estimation of β and γ** as followings until the difference of the scale parameter meets the convergence condition
   i. Estimate $\hat{\beta}^k = \frac{\sum w_i y_i (w_i x_i)^{1-2\gamma}}{\sum (w_i x_i)^{2(1-\gamma)}}$ with the latest $\gamma^{(k-1)}$ and $w_i^{(k-1)}$
   ii. Obtain new weight w_i^k and estimate γ^k

Settings of the functions

- Weight function: Tukey's biweight
- Scale parameter: Average Absolute Deviation (AAD)
- Conversion condition: $\left|1 - \hat{\sigma}^{(k)}/\hat{\sigma}^{(k-1)}\right| < 0.001$

Comparisom with *optim*

Comparison with R ***optim*** function

✓ create a function of the generalized estimator

minimize the sum of absolute deviation of **quasi residuals**

$$\sum \left| \frac{y_i - \hat{\beta} x_i}{x_i^{\gamma}} \right| \xrightarrow{\text{Computationally more efficient}} \sum \left(\frac{y_i - \hat{\beta} x_i}{x_i^{\gamma}} \right)^2$$

✓ provide initial values and estimate β and γ using the ***optim*** function

Opt1 <- function(x, y, **pm**) (sum((y - **pm[1]** * x)/x^**pm[2]**)^2)
optim(pm, Opt1, x=x1, y=y1)

$\uparrow \beta$ $\uparrow \gamma$

Results

- Code B is the best among the three
- Estimation of γ may not so accurate for both optim and Code B
- Estimation of β by code B is closest to the true value

Remaining issues

Improve estimation of γ

Estimation of γ is not so accurate at this moment, although γ is not used for imputation

Robust estimators to cope with outliers in explanatory variable

GM-estimators and MM-estimators

Software available at https://github.com/kazwd2008/IRLS

<table>
<tr><th>File name</th><th>R function</th><th>Feature</th><th>Weight function</th><th>Scale parameter</th></tr>
<tr><td rowspan="2">Tirls.r</td><td>Tirls.aad</td><td rowspan="4">Robust estimation for Regression model</td><td rowspan="2">Tukey</td><td>AAD</td></tr>
<tr><td>Tirls.mad</td><td>MAD</td></tr>
<tr><td rowspan="2">Hirls.r</td><td>Hirls.aad</td><td rowspan="2">Huber</td><td>AAD</td></tr>
<tr><td>Hirls.mad</td><td>MAD</td></tr>
<tr><td rowspan="6">RrT.r*</td><td>RrTa.aad</td><td rowspan="12">Robust estimation for generalized ratio model with a fixed γ value

* All functions in RrT.r and RrH.r are included in the REGRM package at https://github.com/kazwd2008/REGRM.</td><td rowspan="6">Tukey</td><td rowspan="3">AAD</td></tr>
<tr><td>RrTb.aad</td></tr>
<tr><td>RrTc.aad</td></tr>
<tr><td>RrTa.mad</td><td rowspan="3">MAD</td></tr>
<tr><td>RrTb.mad</td></tr>
<tr><td>RrTc.mad</td></tr>
<tr><td rowspan="6">RrH.r*</td><td>RrHa.aad</td><td rowspan="6">Huber</td><td rowspan="3">AAD</td></tr>
<tr><td>RrHb.aad</td></tr>
<tr><td>RrHc.aad</td></tr>
<tr><td>RrHa.mad</td><td rowspan="3">MAD</td></tr>
<tr><td>RrHb.mad</td></tr>
<tr><td>RrHc.mad</td></tr>
<tr><td rowspan="2">RBreds.r</td><td>RBred</td><td>Simultaneous robust estimation of γ and β</td><td rowspan="2">Tukey</td><td rowspan="2">AAD</td></tr>
<tr><td>Bred</td><td>Simultaneous non robust estimation of γ and β</td></tr>
</table>

References and related works

- **Ratio estimator**

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

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

- **Robustification and generalisation**

  Wada, K. & Sakashita, K. (2017) Generalized robust ratio estimator for imputation. In Proceedings of New Techniques and Technologies for Statistics (NTTS), Brussels, Belgium. URL http://nt17.pg2.at/data/abstracts/abstract_56.html

  Wada, K., Sakashita, K. & Tsubaki, H. (2020) Robust estimation for a generalised ratio model, to be appeared in *Austrian Journal of Statistics*.

- **Simultaneous estimation of β and γ**

  Wada, K., Takata, S. & Tsubaki, H. (2019) An algorithm of generalized robust ratio model estimation for imputation. In JSM Proceedings, Government Statistics Session, Denver, CO: American Statistical Association. 3120-3128.

III. Modified Stahel-Donoho estimator

- Implementation of the improved Euroedit version together with Canada version used for AWRTS
- Paralellisation for higher dimensional data
- Application in the Unincorporated Enterprise Survey for cleaning hot deck imputation donor candidates

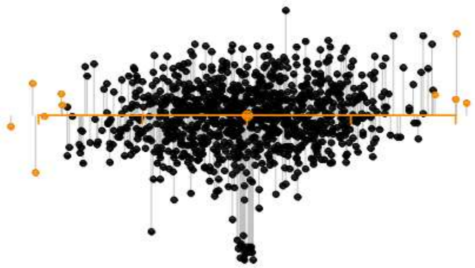
Algorithm (1)

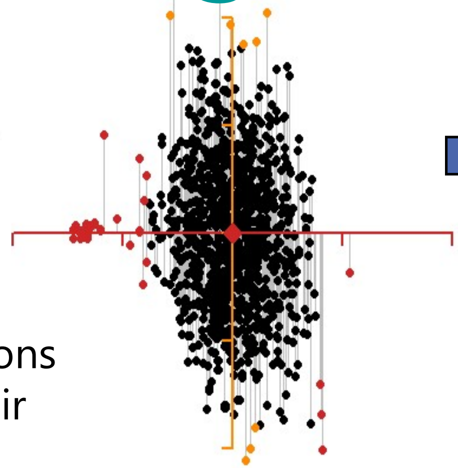
1. Prepare random orthogonal bases.

   number of orthogonal bases : $\exp(2.1328 + 0.8023p)/p$

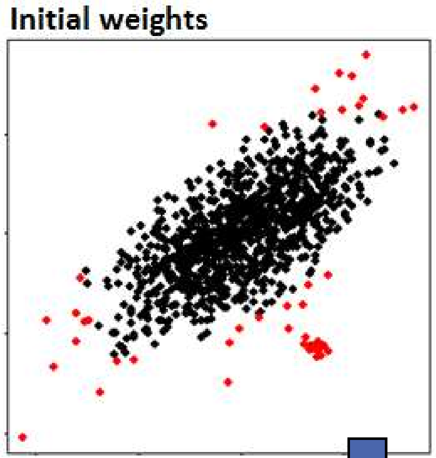
   (p: number of variables)
2. Perform projection into the orthogonal bases.
3. Compute the corresponding weights per vector from the residuals.
4. Multiply weights by each vector to have a set of weights per basis. Then choose the smallest one as the initial weight for each data point.
5. Conduct robust PCA from the estimated initial mean vector and covariance using the initial weights.
6. Perform projection onto the eigenvectors and calculate the secondary weight. Then compare the initial and secondary weight and choose the smaller as the final weight for the final mean vector and the covariance matrix.
7. Compute Mahalanobis distance to decide outliers.

Algorithm (2)

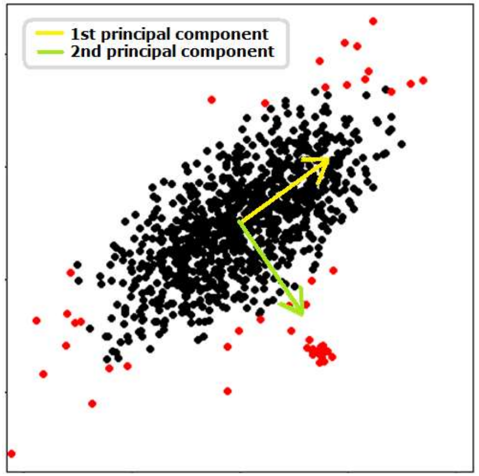




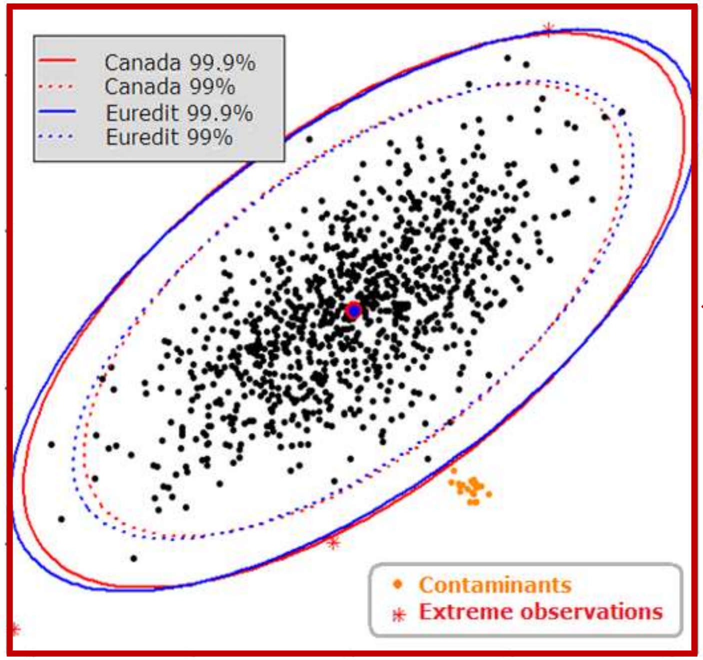
Trim weights of observations at the edge to control their influence on estimation



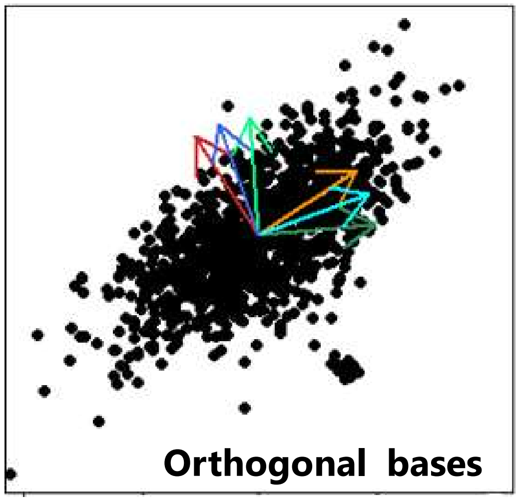
Weighted Principal components analysis



Projection on to principal components to improve initial weights



Estimate the robust mean vector and the covariance matrix by weighting



Prepare random orthogonal bases for projection

Practical application

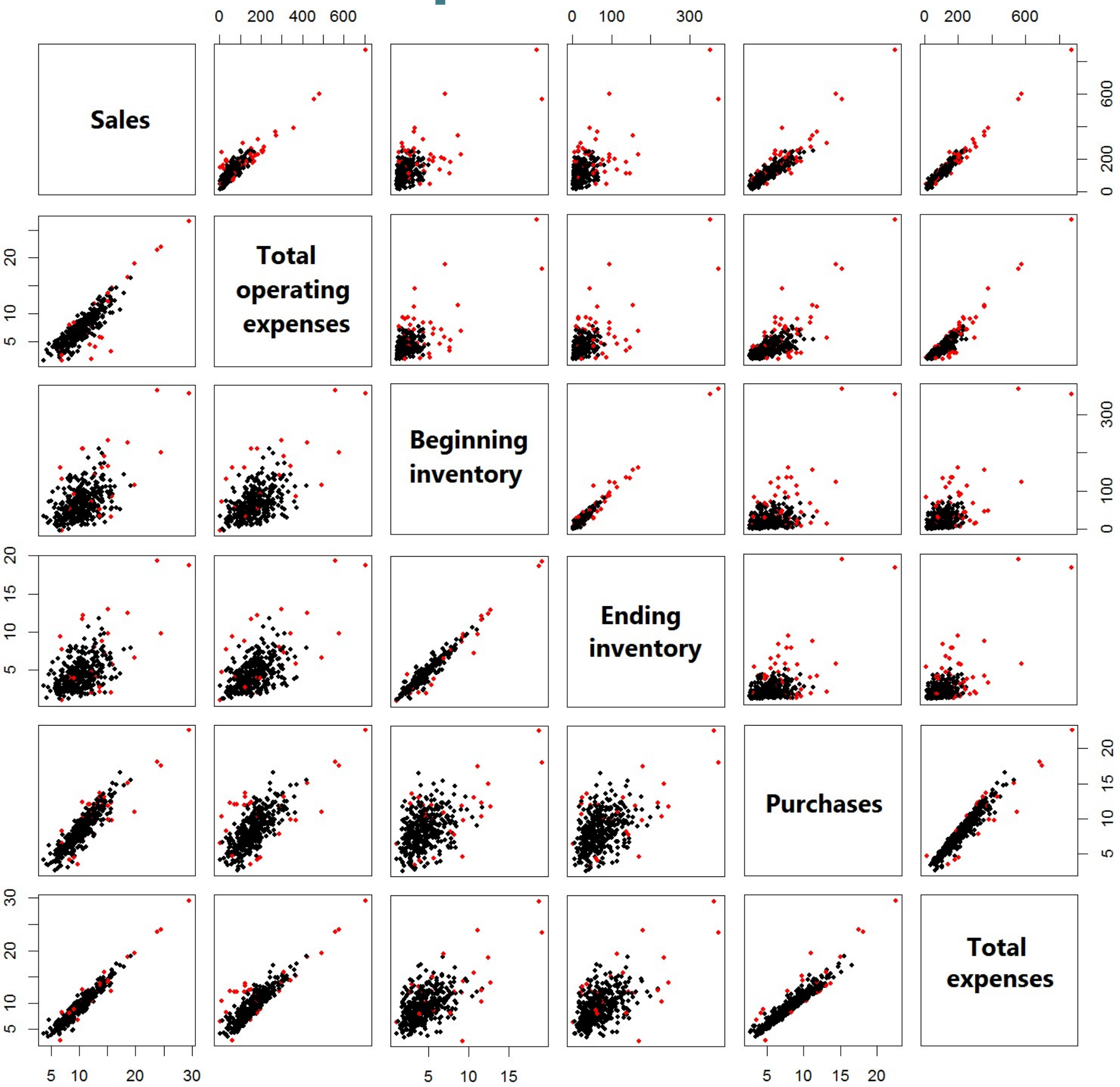
2019 Unincorporated Enterprise Survey

Removal of extreme outliers from the hot deck candidates for major corporate accounting items (single core version for 4 variables)

Software

- R single core version both for used in Statistics Canada and improved version: https://github.com/kazwd2008/MSD
- R paralleled version for high-dimensional data: https://github.com/kazwd2008/MSD.parallel

An example: manufacturing industry



Upper triangular matrix: square root transformation

Lower triangular matrix: biquadratic root transformation

References and related works

- **Practice in Statistics Canada for AWRTS**

Franklin, S. and Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method, *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 186-191.

- **Improvement proposal**

Béguin, C., Hulliger, B., (2003), Robust Multivariate Outlier Detection and Imputation With Incomplete Survey Data, EUREDIT Deliverable D4/5.2.1/2 Part

- **Comparison with other methods (BACON, Fast-MCD, NNVE)**

Wada, K., Kawano, M. & Tsubaki, H. (2020). Comparison of multivariate outlier detection methods for nearly elliptical distributions, Austrian Journal of Statistics, **49**(2), 1-17. URL https://doi.org/10.17713/ajs.v49i2.872

- **Parallelisation**

Wada, K. & Tsubaki, H. (2013). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection. In *Proceedings of 2013 IEEE International Conference on Cloud Computing and Big Data* (*CloudCom-Asia*), 304-311, 16-19, Dec. 2013, Fuzhou, China.

Remaining issues

Tuning of the paralleled version

The paralleled function has a parameter “dv” which indicates the maximum number of elements processed together on the same core regarding data projection.

Finding a good value for the parameter needs trial and error, although the parameter depends on the number of cores, memory capacity, and size of datasets and greatly affects the computational time.

Computational time

The paralleled function has no constraint regarding the number of variables; while high dimensional data consumes a long computational time

Thank you for your attention.

Any questions, comments and suggestions:

k5.wada@soumu.go.jp