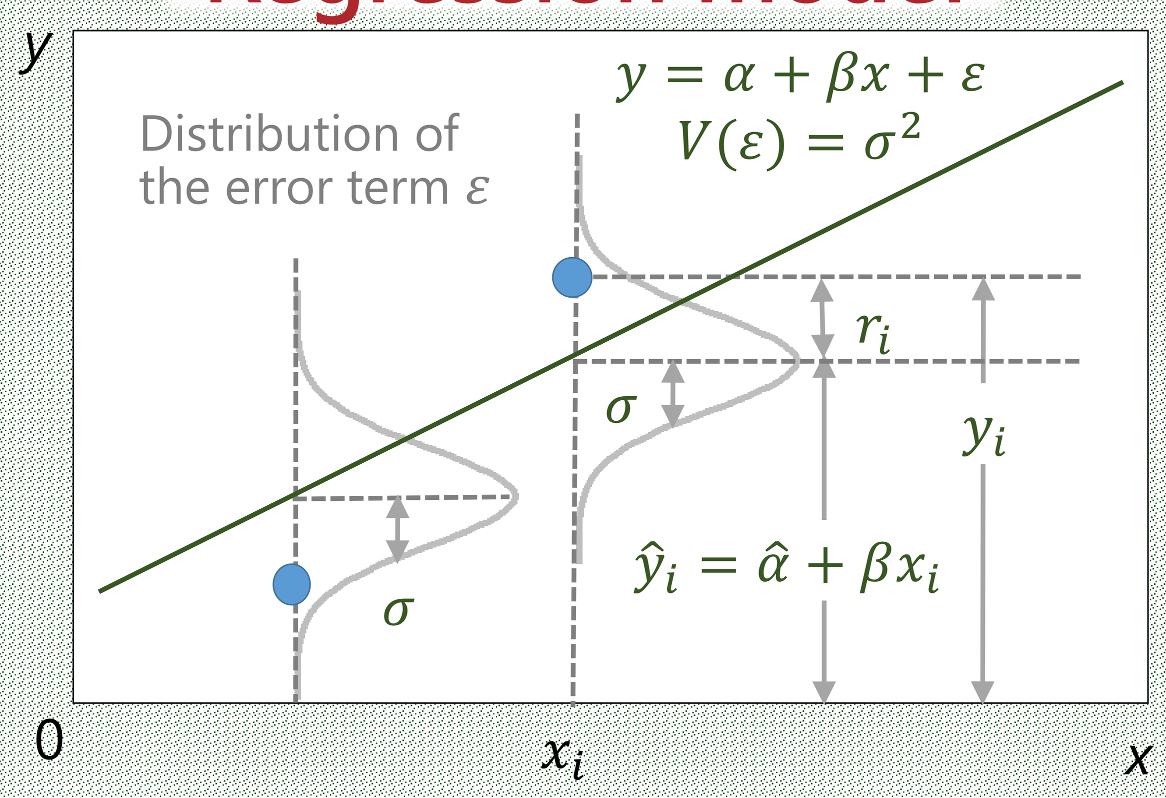
New Techniques and Technologies for Statistics (NTTS) Conference, Brussels, Belgium, 14-16, Mar. 2017.

NSTAC

Generalized robust ratio estimator for imputation

Kazumi Wada, Keiichiro Sakashita

Regression model



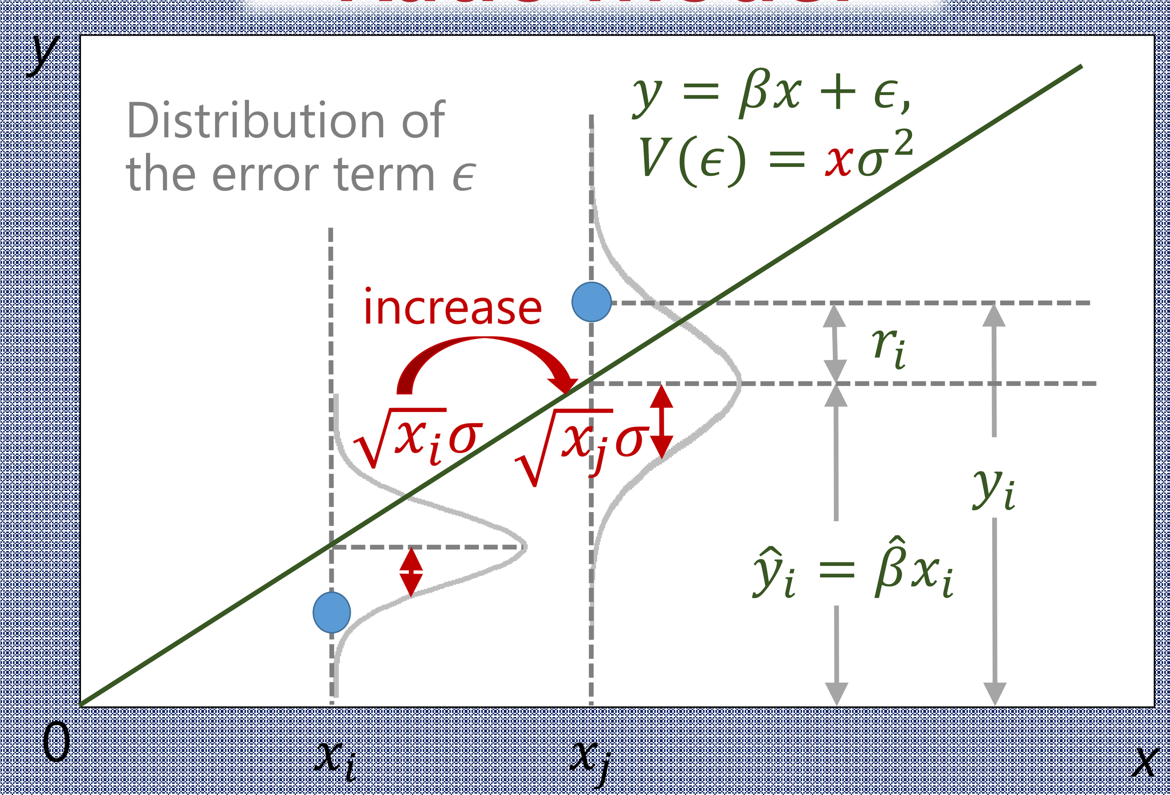
Features in common

- The error term tends to have longer tails regarding survey data
  ⇒ Existence of outliers
- Outliers can be very influential to the estimation

Differences

- No intercept
- Heteroscedastic error term proportional to $\sqrt{x_i}$

Ratio model



Robust estimation of regression model

Iterative Reweighted Least Squares (IRLS)

Regression by OLS

Outliers may have considerable influence

Robust Regression (M-estimators)

IRLS controls the influence of outliers by down weight

[advantage]
- easy to calculate => frequently used in practice

[disadvantages]
- the breakdown point is 1/n as same as the OLS
- not robust for outliers in explanatory variables

Holland & Welsch (1977) Robust Regression Using Iteratively Reweighted Least-Squares, Communication in Statistics –Theory and Methods, A6(9), 813-827.

Generalization

Error term proportional to $x_i^{1/2}$ ➡ x_i^{γ}

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

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

robustification
$$\hat{\beta}_{rob} = \frac{\sum w_i y_i x_i^{1-2\gamma}}{\sum w_i x_i^{2(1-\gamma)}}$$

※ γ : *an arbitral constant*

Robustification of the ratio estimator

1. Making the error term homoscedastic

$$\epsilon \sim N(0, x\sigma^2):\ y_i = \beta x_i + \epsilon_i$$
$$\varepsilon \sim N(0, \sigma^2):\ y_i = \beta x_i + \varepsilon_i \sqrt{x_i}$$
$$\because\ \varepsilon_i = \frac{\epsilon_i}{\sqrt{x_i}}$$

2. Robustification

$$\hat{\beta}_{rob} = \frac{\sum w_i y_i}{\sum w_i x_i}$$

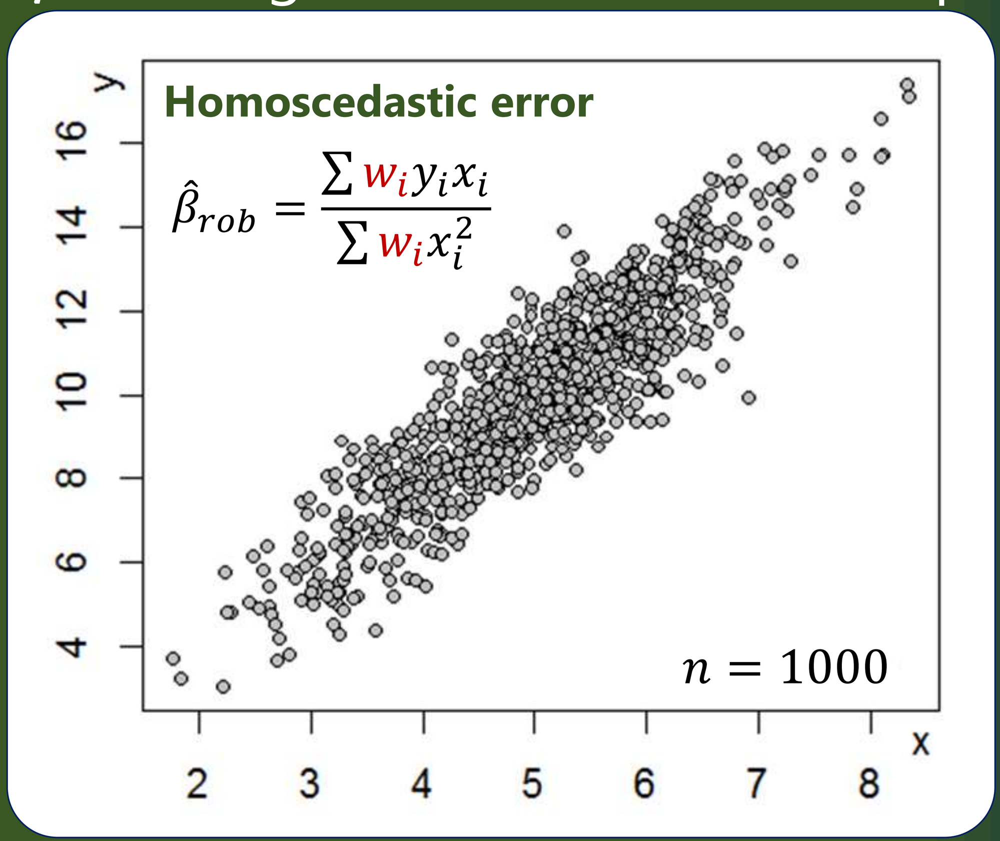
- Quasi-residual: $\breve{\varepsilon}_i = \frac{y_i \sqrt{w_i}}{\sqrt{x_i}} - \hat{\beta}_{rob} \sqrt{w_i x_i}$
- Weight function: Tukey's biweight (c=8)
- Scale parameter: $\hat{\sigma}_{\mathrm{AAD}} = \frac{1}{n} \sum_{i=1}^{n} |\breve{\varepsilon}_i|$

  to standardize residuals: $e_i = r_i / \hat{\sigma}$

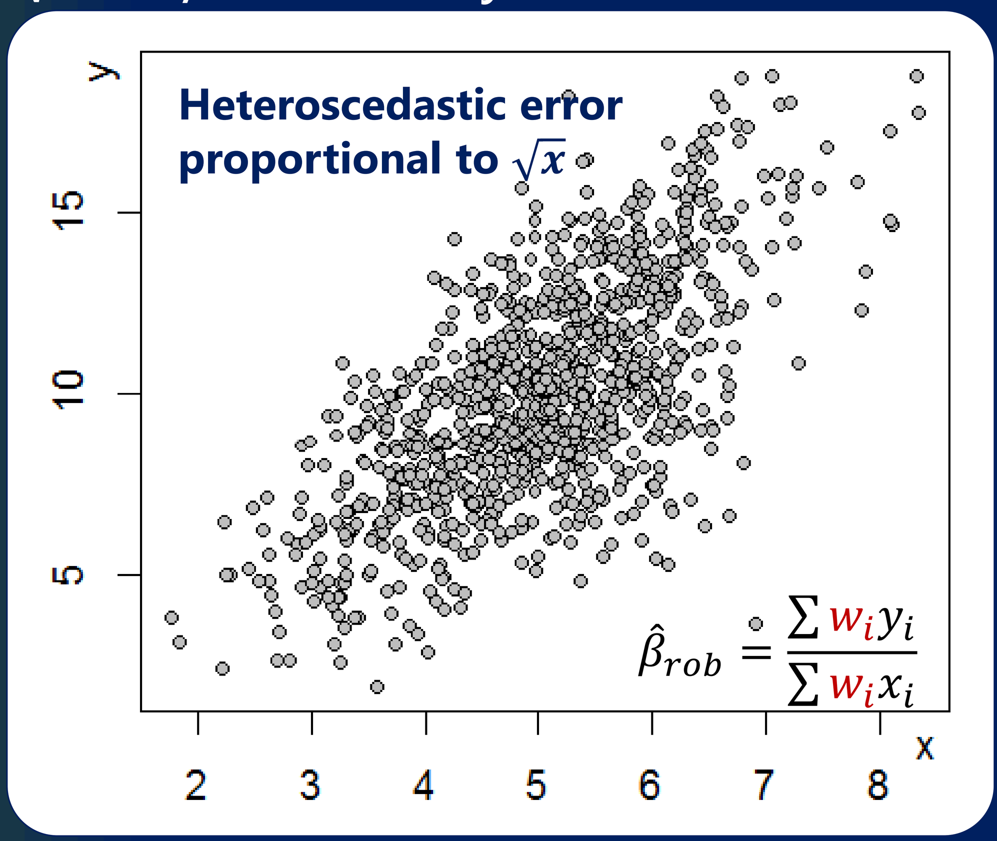
AAD: *Average Absolute Deviation*

Examples: Random data following the model $y = \beta x + \varepsilon x^{\gamma}$ with different γs

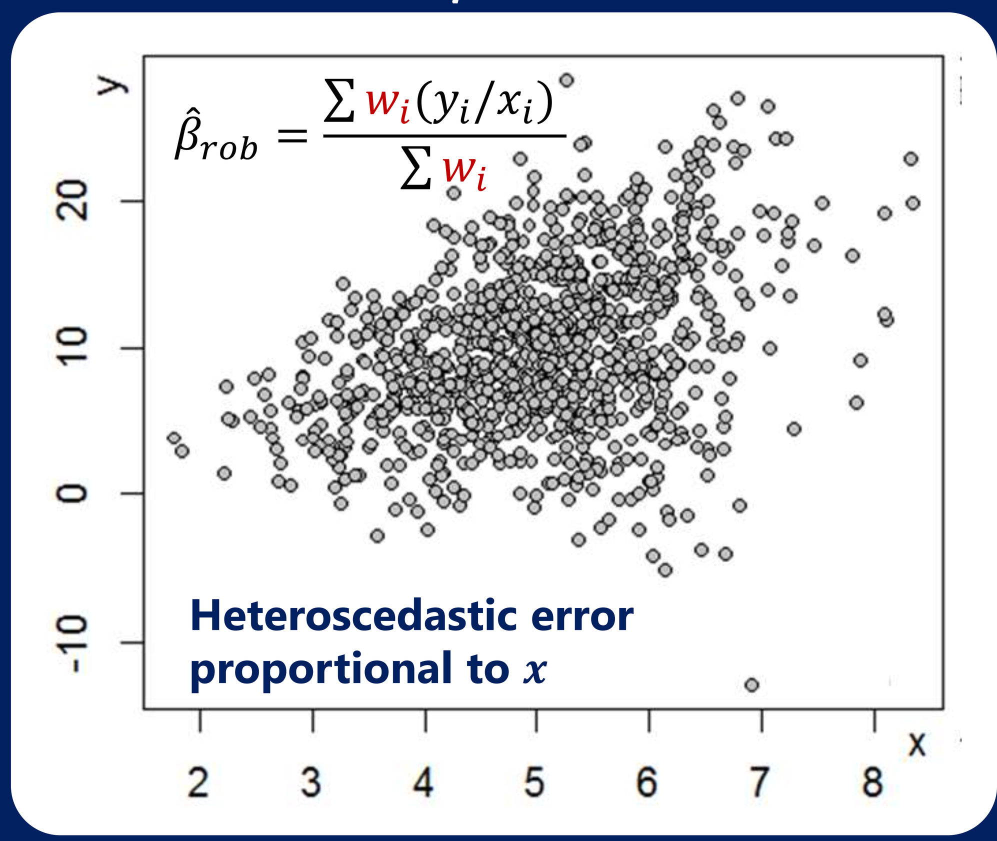
$\gamma = 0$: regression without intercept



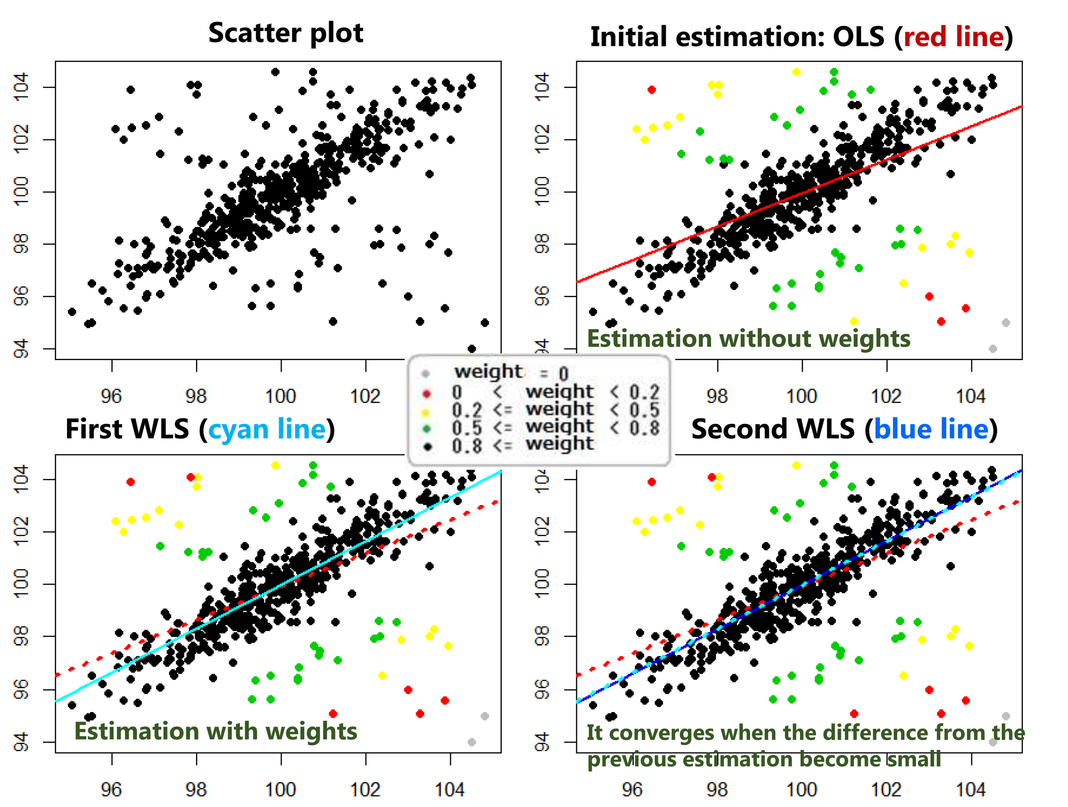
$\gamma = 1/2$: ordinary ratio estimator



$\gamma = 1$

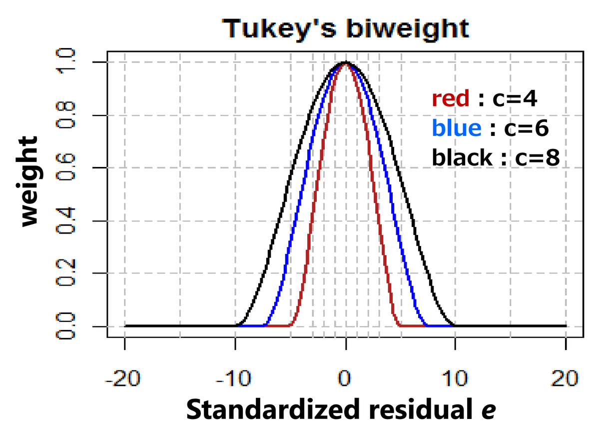


How IRLS works (regression model)



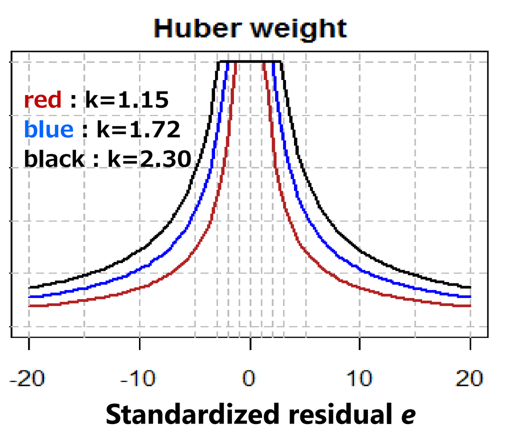
OLS: *Ordinary Least Squares;* **WLS:** *Weighted Least Squares*

Features of the Tukey's biweight



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

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



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

Huber weight alleviate the influence of outliers but never eliminate it

Practical application: 2016 Economic Census for Business Activity

Imputation of the major corporate accounting items such as sales, salaries, and expenditures

Objectives of the Census:
- Identify the structure of establishments and enterprises in all industries on a national and regional level by investigating their economic activity
- Obtain fundamental information for conducting various statistical surveys

Date of Census: 1 Jun. 2016

Coverage: All establishments and enterprises in Japan