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 can be very influential to the estimation

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

Robustification of the ratio estimator
1. Making the error term homoscedastic
   $e_i \sim N(0, \sigma^2)$: $\hat{y}_i = \beta x_i + e_i$
   $\hat{y}_i = \beta x_i + e_i \sqrt{x_i}$
2. Robustification
   $\hat{\beta}_{rob} = \sum w(y_i) \sum w(x_i)$
- Quasi-residual: $e_i = \frac{2}{n} \sum x_i - \hat{\beta}_{rob} x_i$
- Weight function: Tukey’s biweight ($c=8$)
- Scale parameter: $d_{AAD} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$
to standardize residuals: $e_i = \frac{e_i}{d_{AAD}}$

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

Examples: Random data following the model $y = \beta x + \epsilon x^T$ with different $\gamma$

$\gamma = 0$: regression without intercept

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

$\gamma = 1$: Heteroscedastic error proportional to $x$

How IRLS works (regression model)

Scatter plot
- Initial estimation: OLS (red line)
- Iterative Reweighted Least Squares (IRLS)

Features of the Tukey’s biweight

Tukey’s biweight
- Standardized residual $e$
- Robust weight

Features in common
- Outliers can be very influential to the estimation
- Outliers can be very influential to the estimation

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 (disadvantages)
- The breakdown point is 1/n as same as the OLS

Advantages
- Easy to calculate $\Rightarrow$ frequently used in practice
- The breakdown point is 1/n as same as the OLS
- Not robust for outliers in explanatory variables

Standardized residuals
- Tukey’s biweight eliminates influence of outliers with very large residuals
- Huber weight eliminates the influence of outliers but never eliminate it

Examples:
- Random data following the model $y = \beta x + \epsilon x^T$ with different $\gamma$
- Ordinary ratio estimator
- Heteroscedastic error proportional to $\sqrt{x_i}$

Generalization

Error term proportional to $x_i^{1/2} \Rightarrow x_i^{\gamma}$

Robustification

$\hat{\beta}_{rob} = \sum w(y_i) \sum w(x_i)$

Weight function:
- Tukey’s biweight ($c=8$)

Scale parameter:
- $d_{AAD} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$
to standardize residuals: $e_i = \frac{e_i}{d_{AAD}}$

Afterwards:
- To estimate $\hat{\beta}$ as
- $\hat{\beta}_{rob} = \sum w(y_i) \sum w(x_i)$
- $d_{AAD} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$
to standardize residuals: $e_i = \frac{e_i}{d_{AAD}}$

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

Features of the Tukey’s biweight

Tukey’s biweight
- Standardized residual $e$
- Robust weight

Features in common
- Outliers can be very influential to the estimation
- Outliers can be very influential to the estimation

Robustification of the ratio estimator
1. Making the error term homoscedastic
   $e_i \sim N(0, \sigma^2)$: $\hat{y}_i = \beta x_i + e_i$
   $\hat{y}_i = \beta x_i + e_i \sqrt{x_i}$
2. Robustification
   $\hat{\beta}_{rob} = \sum w(y_i) \sum w(x_i)$
- Quasi-residual: $e_i = \frac{2}{n} \sum x_i - \hat{\beta}_{rob} x_i$
- Weight function: Tukey’s biweight ($c=8$)
- Scale parameter: $d_{AAD} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$
to standardize residuals: $e_i = \frac{e_i}{d_{AAD}}$

Afterwards:
- To estimate $\hat{\beta}$ as
- $\hat{\beta}_{rob} = \sum w(y_i) \sum w(x_i)$
- $d_{AAD} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$
to standardize residuals: $e_i = \frac{e_i}{d_{AAD}}$