# Multivariate Outlier Detection for Regression

## - Imputation and Aggregation Weight Calibration by IRLS -

Wada, Kazumi
*Senior Researcher*
*National Statistics Center*
*19-1, Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan*
*E-mail: kwada@nstac.go.jp*

Abe, Yutaka
*Researcher*
*National Statistics Center*
*19-1, Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8668, Japan*
*E-mail: yabe3@nstac.go.jp*

**ABSTRACT**

*Outliers occur very frequently in survey data. Some are corrected if they are error, but some are not if they are true. The latter may spoil regression imputation by Ordinary Least Squares (OLS) and those with large aggregation weight may distort the figures in tabulation. In this paper, a comparison of Iterative Reweighted Least Squares (IRLS) and OLS is made regarding regression imputation which explains the enterprise sales by the number of employees. Aggregation weight calibration by the IRLS weight is also discussed. The algorithm of IRLS is easy to calculate, robust to outliers in the dependent variable and therefore, estimated figures for imputation are more stable than those of OLS with existence of influential outlier. In addition to values for imputation, IRLS provides a set of data weight which reflects deviation from the regression model. We would like to propose adjusting aggregation weight with the IRLS weight so that the aggregation weight takes outlyingness of each observation into account. It prevents over-representation of rare extreme observations in statistical tables.*

## 1. Introduction

Non-responses in survey data are generally imputed to avoid bias in compiling official statistics. As for the enterprise sales data, regression imputation by the number of employees may be applied among the variety of imputation methods. OLS is generally used for the regression; however, it is well known that the existence of outliers makes its estimation unreliable and real data often contain them. Such outliers have to be eliminated from the regression estimation by OLS, and their aggregation weight may also require an adjustment so that the rare extreme values do not have an excessive influence on statistical tables especially when the weight is large. In this paper, a robust regression method called IRLS is used to accommodate those problems.

In Section 2, we extend the IRLS algorithm so that the aggregation weight is considered. Section 3 describes the dataset used, fitting of imputation model, and the results of estimation for imputation. Section 4 explains the aggregation weight calibration.

## 2. Methodology

### 2.1 IRLS with aggregation weight

Based on Fox and Weisberg (2010), we describe the M-estimation with aggregation weight regarding

the linear model

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$$

for the *i*-th of *n* observations. Given an estimator $\mathbf{b}$ for $\boldsymbol{\beta}$, the fitted model is

$$\hat{y}_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} = \mathbf{b}'\mathbf{x}_i$$

and the residuals are given by $e_i = y_i - \hat{y}_i$.

The M-estimator with aggregation weight $g_i$ is shown as follows with IRLS weight function $w_i = w(e_i)$.

$$\sum_{i=1}^{n} w_i g_i (y_i - \mathbf{b}'\mathbf{x}_i)\mathbf{x}_i' = \mathbf{0}$$

Computing of the estimator takes the following iterative steps:

1) Compute initial estimate $\mathbf{b}^{(0)}$ by OLS as follows where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$, $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{G} = \mathrm{diag}\{g_i\}$.

$$\mathbf{b}^{(0)} = [\mathbf{X}'\mathbf{G}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{G}\mathbf{y}$$

2) At each iteration *j*, calculate residuals $e_i^{(j-1)}$, its mean absolute deviation $s^{(j-1)}$ and associated IRLS weight $w_i^{(j-1)}$ according to the weight function $w\left(e_i^{(j-1)}\right)$.

3) Solve for new weighted least squares estimates where $\mathbf{W}^{(j-1)} = \mathrm{diag}\{w_i^{(j-1)}\}$.

$$\mathbf{b}^{(j)} = \left[\mathbf{X}'\mathbf{G}\mathbf{W}^{(j-1)}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{G}\mathbf{W}^{(j-1)}\mathbf{y}$$

Steps 2) and 3) are repeated until the estimation converges. We follow the proposal of Bienias et al. (1997) to stop iterating when $s^{(j-1)}/s^{(j-2)}$ becomes less than 0.01. The final IRLS weight $w_i^{(j-1)}$ can be regarded as a scale of outlyingness and will be used for the aggregation weight calibration in Section 4.

## 2.2 Weight function

The following two weight functions are used for the analysis. Constant *c* of the Tukey's biweight function is 4 to 8 according to the setting of Bienias et al. (1977). Corresponding Huber's *k* is 1.15 to 2.30 calculated from the tuning constant shown in Holland and Welsch (1977),.

**Tukey's biweight**

$$w_i = \begin{cases} \left(1 - \left(\dfrac{e_i}{cs}\right)^2\right)^2 & if \quad |e_i| \le cs, \\ 0 & if \quad |e_i| > cs. \end{cases}$$

**Huber weight**

$$w_i = \begin{cases} 1 & if \quad |e_i| \le ks, \\ \dfrac{ks}{|e_i|} & if \quad |e_i| > ks. \end{cases}$$

# 3. Data and Imputation model

## 3.1 Dataset and aggregation weight

The dataset is derived from the financial statements database of Tokyo Shoko Research, Ltd. as of December 2003 using random stratified sampling by industry and category of number of employees. The dataset contains industry, number of employees at the latest accounting period, and enterprise sales for the last three periods. The enterprise sales are adjusted so that all the figures represent the 12 months period. The number of complete observations and its aggregation weight are shown in table 1. The classification of industry is shown in table 2.

***Table 1: Number of complete data and aggregation weight by strata***

| Industry | Number of employees | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 50-99 | | 100-299 | | 300-499 | | 500- | | Total | |
| | | Weight | | Weight | | Weight | | Weight | | Weight |
| D | 20 | 3.10 | 20 | 1.80 | 2 | 1.00 | 12 | 1.08 | 54 | 113 |
| E | 41 | 96.34 | 37 | 44.89 | 37 | 6.81 | 296 | 1.00 | 411 | 6159 |
| F | 41 | 336.32 | 44 | 200.70 | 38 | 40.00 | 1733 | 1.02 | 1856 | 25912 |
| G | 17 | 3.71 | 19 | 2.42 | 5 | 1.20 | 26 | 1.00 | 67 | 141 |
| H1 | 36 | 124.75 | 34 | 89.50 | 35 | 11.89 | 444 | 1.03 | 549 | 8408 |
| H2 | 25 | 1.12 | 23 | 1.26 | 13 | 1.00 | 21 | 1.14 | 82 | 94 |
| I1 | 45 | 203.96 | 62 | 91.68 | 40 | 25.58 | 886 | 1.04 | 1033 | 16803 |
| I3 | 31 | 24.06 | 36 | 11.69 | 35 | 2.37 | 54 | 1.00 | 156 | 1304 |
| J | 25 | 11.46 | 34 | 7.71 | 33 | 2.48 | 363 | 1.04 | 455 | 1006 |
| K | 35 | 18.20 | 37 | 10.86 | 35 | 2.34 | 92 | 1.04 | 199 | 1217 |
| L | 41 | 199.90 | 43 | 126.28 | 40 | 23.48 | 1379 | 1.08 | 1503 | 16058 |
| Total | 357 | 41427 | 389 | 25845 | 313 | 4418 | 5306 | 5526 | 6365 | 97659207 |

***Table 2: Classification of Industry***

| | |
|---|---|
| **D** | **Mining** |
| **E** | **Construction** |
| **F** | **Manufacturing** |
| **G** | **Electricity, gas, heat supply and water** |
| **H1** | **Transport** |
| **H2** | **Information and communications** |
| **I1** | **Wholesale and retail trade** |
| **I3** | **Eating and drinking places, accommodations** |
| **J** | **Finance and insurance** |
| **K** | **Real estate** |
| **L** | **Services** |

### 3.2 Model selection

In this paper, imputation of the enterprise sales $\mathbf{y} = (y_1, \ldots, y_n)'$ by the number of employees $x = (x_1, \ldots, x_n)'$ is considered. Due to the heteroscedasticity of the error term, the candidate models are as follows:

A. Linear model with logarithmic transformation : $\log y = \alpha + \beta \log x + \varepsilon$
B. Linear model with square root transformation : $\sqrt{y} = \alpha + \beta\sqrt{x} + \varepsilon$
C. Ratio estimation (without transformation) : $y/x = \alpha + \varepsilon$
D. Ratio estimate with square root transformation : $\sqrt{y}/\sqrt{x} = \alpha + \varepsilon$

First, normality of $y/x$, $\log y/\log x$ and $\sqrt{y}/\sqrt{x}$ are compared to choose the data transformation. In addition to the Shapiro-Wilk test, a few goodness-of-fit tests based on empirical distribution such as the Anderson-Darling and the Lillieforce test are used. Moment tests are not suitable for the purpose since one outlier makes the p-value considerably small. Then, the scatter plot with regression line/curve and the residual plot are examined to decide a fit model for each industry. The result is included in table 3.

### 3.3 Imputation by IRLS

According to the fit model by industry, estimated value $\hat{y}_i^{(t)}$ is calculated for the three year periods respectively as follows where *p* is number of regression parameters.

A. $$\hat{y}_i^{(t)} = \exp(a^{(t)} + b_1^{(t)} \log x_i^{(t)}) \times \exp\left( \frac{1}{2} \cdot \frac{\sum_{i=1}^{n} \left(e_i^{(t)}\right)^2}{(n-p)} \right)$$

D. $$\hat{y}_i^{(t)} = \left( \left(a^{(t)}\right)^2 + \frac{\sum_{i=1}^{n} \left(e_i^{(t)}\right)^2}{(n-p)} \right) x_i^{(t)}$$

Then the mean of $\hat{y}_i^{(t)}$ for each period $\overline{y_{est.}^{(t)}}$ is calculated. The figure shown in table 3 is the standard deviation $\sigma$ of $\overline{y_{est.}^{(t)}}$ for the three periods and divided by 1000 to reduce the number of the printed digits. Total is the square root of sum of $\sigma^2$.

$$\overline{y_{est.}^{(t)}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i^{(t)} \quad , \quad \sigma = \sqrt{\frac{1}{(T-1)} \sum_{t=1}^{T} \overline{y_{est.}^{(t)}}}$$

In the table, "TK8" means Tukey's biweight of $c$=8 and "HB8", Huber weight of $k$=2.30. Since the dataset has relatively longer tail than the normal distribution, the result of largest $c$ and $k$ are printed. The figures of IRLS tend to be more stable than those of OLS regardless of the weight function and the standard deviation improves by using aggregation weight in IRLS estimation for most industries. The total figure of "TK8/OLS" without aggregation weight is 9.8% and "HB8/OLS", 22.8%.

***Table 3: Fit model and standard deviation of the estimated mean***

| Ind. | Fit model | Number of data | Standard deviation of the estimated mean | | | | |
|---|---|---|---|---|---|---|---|
| | | | OLS/1000 | TK8/1000 | TK8/OLS | HB8/1000 | HB8/OLS |
| D | A | 54 | 280 | 287 | 102.4% | 306 | 109.5% |
| E | D | 411 | 1629 | 684 | 42.0% | 685 | 42.1% |
| F | A | 1856 | 973 | 573 | 58.9% | 690 | 70.9% |
| G | A | 67 | 7708 | 7850 | 101.8% | 7895 | 102.4% |
| H1 | A | 549 | 403 | 220 | 54.5% | 270 | 67.1% |
| H2 | A | 82 | 234801 | 14434 | 6.1% | 49889 | 21.2% |
| I1 | A | 1033 | 492 | 473 | 96.2% | 460 | 93.5% |
| I3 | D | 156 | 278 | 227 | 81.9% | 221 | 79.8% |
| J | A | 455 | 11110 | 5301 | 47.7% | 6232 | 56.1% |
| K | A | 199 | 4779 | 1602 | 33.5% | 2132 | 44.6% |
| L | A | 1503 | 729 | 560 | 76.8% | 544 | 74.6% |
| Total | - | 6365 | 235249 | 17382 | 7.4% | 50954 | 21.7% |

## 4. Calibration of the aggregation weight

The mean of the enterprise sales $\overline{y^{(t)}}$ using the aggregation weight $g_i$ is calculated as follows:

$$\overline{y^{(t)}} = \sum_{n=1}^{i} (g_i \cdot y_i^{(t)}) / \sum_{n=1}^{i} g_i$$

Usually aggregation weight $g_i$ is the inverse of sampling probability. We propose following new aggregation weight $g_i^*$ which is adjusted $g_i$ with the final IRLS weight $w_i$ so that it reflects both sampling probability and outlyingness of observations. The weight $g_i^*$ is calculated by each stratum, i.e. by industry and category of number of employees regarding the dataset used here.

$$g_i^* = g_i \cdot w_i \cdot \frac{n}{\sum_{i=1}^{n} w_i}$$

Proportional change of standard deviation $\sigma'$ of the mean $\overline{y^{(t)}}$ by weight calibration is shown in table 4. The figures are divided by 1000.

***Table 4: Effect of the weight calibration by sampling domain***

| Ind. | Tukey, c=8 [%] | | | | | Huber, k=2.30 [%] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Number of employees | | | | | Number of employees | | | | |
| | 50-99 | 100-299 | 300-499 | 500- | Total | 50-99 | 100-299 | 300-499 | 500- | Total |
| D | 254.1 | 46.0 | 61.0 | 99.4 | 70.3 | 291.2 | 56.1 | 58.2 | 97.6 | 84.6 |
| E | 99.4 | 90.5 | 115.6 | 117.7 | 115.2 | 100.4 | 89.3 | 104.2 | 135.9 | 117.5 |
| F | 139.1 | 42.2 | 90.6 | 101.3 | 76.6 | 141.5 | 56.9 | 97.4 | 101.5 | 83.4 |
| G | 77.8 | 290.4 | 100.9 | 110.8 | 113.0 | 75.1 | 267.9 | 100.0 | 107.0 | 109.0 |
| H1 | 101.6 | 68.2 | 95.3 | 115.8 | 99.8 | 110.9 | 69.2 | 87.1 | 101.7 | 98.6 |
| H2 | 88.5 | 88.5 | 80.9 | 104.0 | 104.7 | 87.6 | 89.8 | 100.0 | 103.6 | 103.3 |
| I1 | 82.2 | 89.6 | 92.8 | 66.5 | 96.0 | 88.6 | 102.0 | 100.2 | 58.9 | 93.7 |
| I3 | 104.8 | 100.1 | 125.8 | 96.1 | 102.5 | 100.0 | 104.8 | 118.4 | 100.8 | 104.1 |
| J | 1.1 | 8.6 | 379.2 | 86.2 | 67.9 | 24.3 | 30.7 | 412.8 | 87.6 | 73.4 |
| K | 88.2 | 106.7 | 94.2 | 42.8 | 84.4 | 90.2 | 100.0 | 100.1 | 43.6 | 86.7 |
| L | 94.9 | 76.6 | 91.7 | 107.3 | 86.8 | 108.8 | 74.0 | 99.9 | 104.8 | 86.7 |
| Total | 9.6 | 62.1 | 69.4 | 103.4 | 99.9 | 26.1 | 66.0 | 74.9 | 103.0 | 99.3 |

Wada and Tsubaki (2011) describes there are two different factors which affect stability of the aggregated mean. Mild outliers moving around marginal area across the periods slightly increase the standard deviation $\sigma'$ with $g_i^*$ compared to that with $g_i$ since any move of the IRLS weight through the periods differentiate the aggregation weight $g_i^*$, too. On the other hand, the figure with $g_i$ soars compared to that with $g_i^*$ in existence of any extreme outlier(s) especially when $g_i$ is large. Although the former negative effect seems relatively restrictive, the latter favorable effect of $g_i^*$ increases according to the influence of outliers. The figures in table 4 become less than 100 when the latter effect goes over the former. The last row of the table shows a tendency of the weight calibration that it is effective especially in the domains with large aggregation weight.

Figure 1 shows a scatter plot of industry "J" (Finance and Insurance) for example. The weight calibration shows the largest effect in domain #1 (employees 50 to 99) since there are two extreme outliers (No.389 and 466). The original aggregation weight for this domain is 11.46 as shown in table 1. On the contrary, the calibration makes the standard deviation about 4 times larger in domain #3. It is because the standard deviation with $g_i^*$ of this domain has extremely small figure as shown in table 5. The figures are divided by 1000. It is caused by the large fluctuation of extreme outliers of those move contradicts the trend of the majority.

***Table 5: Standard deviation of the mean (Finance and Insurance [J])***

| | Sampling domain (by number of employees) | | | | |
|---|---|---|---|---|---|
| | 50-99 | 100-299 | 300-499 | -500 | Total |
| No calibration | 6116 | 1820 | 120 | 12012 | 5794 |
| Tukey, c=8 | 341 | 685 | 804 | 931 | 455 |
| Percentage | 1.1% | 8.6% | 379.2% | 86.2% | 67.9% |

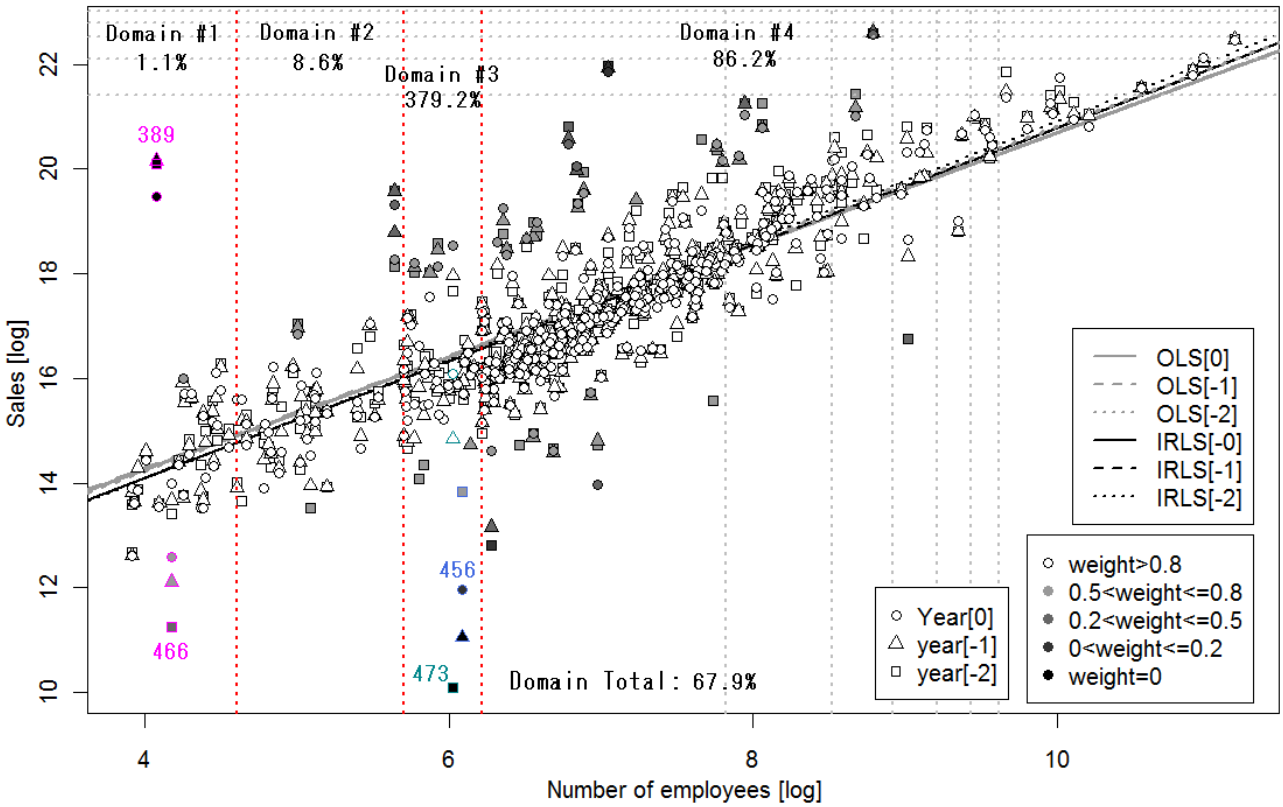


***Figure 1: Finance and Insurance [J]***

## REFERENCES

Anscombe, F. J. and Guttman, I. (1960) Rejection of outliers, Technometrics **2**, 123-147

Bienias, J. L., Lassman, D. M. Scheleur, S. A. and Hogan H. (1997) Improving Outlier Detection in Two Establishment Surveys. *Statistical Data Editing **2** - Methods and Techniques.* (UNSC and UNECE eds.) 76-83.

Fox, J. and Weisberg S. (Oct. 2010) Robust Regression, Appendix to *An R Companion to Applied Regression*. Sage, Thousand Oaks, CA, second edition, 2011.

Holland, P. W. and Welsch, R. E. (1977), Robust Regression Using Iteratively Reweighted Least-Squares, *Communications in Statistics – Theory and Methods* **A6(9)**, 813-827

Huber, P. J. (1964) Robust Estimation of a Location Parameter, *Annals of Mathematical Statistics* **35(1)**, 73-101

Wada, K. and Tsubaki, H. (2011), Aggregation Weight Calibration of Outliers Using Robust Regression (in Japanese), To appear in *Proceedings of Annual Conference of Japanese Society of Applied Statistics*,4. Jun. 2011, Osaka.

## RÉSUMÉ

*Outliers arrive très fréquemment dans les données d'enquête. Quelques-uns sont corrigés s'ils sont l'erreur, mais quelques-uns ne sont pas s'ils sont vrais. Le dernier peut gâter l'imputation de régression par moindres carrés ordinaire (MCO) et ceux-là avec le grand poids d'agrégation peut déformer des figures dans la tabulation. Dans ce papier, nous comarons MCPI avec MCO sur le poids d'agrégation qui explique les ventes d'entreprise par le nombre d'employés. Nous discutons aussi le calibrage de poids d'agrégation par le poids de MCPI. MCPI est facile à calculer, robuste à outliers dans la variable dépendante et donc, les valeurs estimées pour l'imputation sont plus d'écurie que ceux-là de MCO avec l'existence d'outlier influent. En plus des valeurs d'imputation, MCPI fournit aussi du poids de données qui est une échelle de l'outlyingness. Ce poids de MCPI utilise à ajuster le poids d'agrégation pour que les valeurs extrêmes n'ont pas l'influence excessive dans les tables statistiques.*

**NOTE**: The opinions expressed in this paper do not necessarily reflect those of organization to which the authors belong.