# A PROCEDURE FOR ROBUST ESTIMATION IN ECONOMETRIC MODEL

〔計量経済モデルの頑健推定の一方法〕

KEIICHIRO HARADA

This article proposes an idea of a practical procedure for obtaining the parameters of econometric model in more robust way than least squares, which is nonrobust in the sense that arbitrarily small departures of the error term from normality may cause arbitrarily large asymptotic variances and/or biases of the estimator. The robust estimator adopted in this article belongs to the class of general M-estimators, and is reduced to the iteratively reweighted least squares. Using this algorithm, it is necessary to find the initial estimate and the weight function. The least median of squares estimate is recommended for the initial estimate, and the robust diagnostic quantity for detecting the influential observations on the least squares estimate is devised for defining the weight function.

## CONTENTS

# 1 Introduction

Linear regression model is useful for specifying the economic relationships. In this model, the error term is assumed to be the sum of innumerable independent elements, and the random variable "approximately" normal distributed by virtue of the central limit theorem. And, the method of least squares is a valuable instrument for estimating the parameters of the econometric model, because this estimator is uniformly minimum variance unbiased under the normal error assumption. The economic relationship estimated by the least squares is considered to be stable and reliable, under the normal error.

Now, let $\{ (y_i, x_i) : 1 \leq i \leq n \}$ be a sequence of independent indentically distributed (p+1)-dimensional vectors of observations, satisfying the regression model of the form

$$y_i = x_i \beta + u_i, \tag{1.1}$$

where $\beta$ is a $(p \times 1)$-vector of unknown regression parameters. The error term $u_i$ is independent of $x_i$, and indenpendent indentically distributed with distribution function $F(u)$. This error model is assumed to be expanded to some neighbourhood of the normal distribution $N(0, \sigma^2)$, which should be regarded as the "central model", in the sense that for $0 < \varepsilon < 1$

$$F = (1 - \varepsilon) N + \varepsilon H, \tag{1.2}$$

where H is an arbitrary distribution. Then the distribution of $(y_i, x_i)$ is given by

$$d(y, x) = F(u)G(x). \tag{1.3}$$

In empirical works, the least squares is criticized for being not robust, in the sense that although it is optimal when u is normally distributed, its efficiency for non-normal u may be arbitrarily low, and its bias arbitrarily large, even for u arbitrarily close to the normal.

In practice, the main reason for departure from the normal error assumption is due to the erroneous data which is occured by some sources for gross error such as blunders in meaning, wrong decimal points, input errors, miss copying and so on. Erroneous data shows up as outlier which is far away from the bulk of data. The term of outlier is used to mean an observation which does not appear to be from the same distribution as the remainder of the data. Even a single outlier in a large sample is able to have a drastic effect on the estimate.

The gross error may happen in both of y- and x-data. Outliers in y-data often possess large positive or negative residuals. The diagnostic quantities for detecting outliers in y-data are generally based on the residual from least squares fit. When outliers exist in given y-data, the estimated parameters by least squares are biased. This matter affects the residuals from least squares fit, hence detecting outliers. The residual mean squared error is inflated over many data points, and where the residuals from least squares fit are prone to stain the effect of outliers. Using the least squares residual, outliers make their detection more difficult. It is, therefore, inevitable to robustify the residual.

Outlying vectors in x-data set have more opportunities to make a large effect on the least squares estimate. A single outlying vector of x-data will completely blow up the fitted value.

Hoaglin and Welsch (1978) proved that diagonal elements of the projection matrix of the form $X(X^TX)^{-1}X^T$,

$$h_i = x_i\,(X^TX)^{-1}x_i^T,\ 1 \leq i \leq n, \tag{1.4}$$

are diagnostic quantites for detecting extreme vectors of x, where $X = (x_i\ x_2 \cdot\cdot\cdot\cdot\ x_n)^T$ and superscript T denotes transposition.

In using the least squares, it is necessary to test whether the error term fulfills the normality or not. Since every observation could be a gross error, the normal error assumption is not always valid. When the error term is non-normal, the problem is to estimate the parameters of the model (1.1) using data $(y_i, x_i)$ drawn from the distribution (1.3). It is, in this case, proper to take a robust estimator which supplements and modifies the conventional methods, by adding the aspect of robustness. The robust estimation is concerned with obtaining the estimate with low sensitivity to aberrant observations, such as outliers, in exchange for an arbitrarily small loss in efficiency.

Later chapters provide a practical procedure for obtaining the parameters of econometric model (linear regression model) in robust way, with aid by the results which were presented in the recent development of the robust regression methodology.

## 2 Assessment of Robustness and Robust Estimator for Regression

In robustifying estimator, it is need to consider two important robustness properties. The first is to provide a reliable estimate which is still staying in some neighbourhood of the true parameter value, even if given data is contaminated by gross error. The second is to react as little as possible to any single observation blown up by gross error. This low sensitivity to aberrant observation means a small bias or a stability of the estimator. The second property, however, is in conflict with the efficiency requirement which is the low asymptotic covariance of estimator. The more robustness in the sense of low sensitivity to any observation, the less efficiency.

The degree of robustness in the sense of the first property may be measured by the concept of breakdown point which was introduced by Hampel (1971). Donoho and Huber (1983) defined the sample version of this concept. The finite sample breakdown point is the maximum fraction of aberrant observations which given data may include, without causing the estimate to take an arbitrarily large value.

According to Huber (1973), the basic idea of robust estimator for regression is defined as the solution of objective function, which is a function of residual other than the sum of squares, of the form

$$Q(\hat{\beta}, \hat{\sigma}) = \sum_{i=1}^{n} \rho((y_i - x_i \hat{\beta})/\hat{\sigma}) = \text{minimum}, \tag{2.1}$$

where $\rho$-function is symmetric with unique minimum at zero. If it has derivative $\phi$, $\phi = \rho'$, then (2.1) is equivalent to

$$\sum_{i=1}^{n} \phi((y_i - x_i \hat{\beta})/\hat{\sigma})\, x_i^T = 0. \tag{2.2}$$

The $\phi$-function in (2.2) is chosen to limit the influence of grossly erroneous observation of y-data. The scale estimate $\hat{\sigma}$ may be independently obatined in robust way. This type of estimator is belongs to M-estimators.

M-estimators are statistically efficient, however, the bias of these estimators is not robust. Its breakdown point is 1/n, because the $\phi$-function is not effective to bound the influence of outlying x-vector.

An attempt to correct the drawback of M-estimators is the proposal of general M-estimators defined as the solution of the form

$$\sum_{i=1}^{n} \phi(x_i, (y_i - x_i \hat{\beta})/\hat{\sigma})\, x_i^T = 0, \tag{2.3}$$

where the $\phi$-function is chosen to bound the influence of aberrant observations in both y- and x-data. Maronna, Bustos and Yohai (1979) proved that general M-estimates have a breakdown point of 1/p, which tends to zero when the number of explanatory variables, p, increases. This class of estimators contains the optimal bounded influence estimates obtained by Krasker (1980), and Krasker and Welsch (1982).

Recently, several regression estimates with high breakdown point were introduced, i.e. Siegel (1982) presented the repeated median (RM) estimate, Rousseeuw (1984) proposed the least median of squares (LMS) estimate and the least trimmed squares (LTS) estimate. However, all of these estimtes have very low efficiency when all observations of variables satisfy the regression model with normal error.

Yohai (1987) presented a class of estimates, which he named MM-estimate, having simultaneously high breakdown point and high efficiency under the normal error. Yohai and Zamar (1988) presented an alternative class of estimates, which was defined by minimizing a new estimate of the scale of residuals, having same properties as Yohai (1987).

## 3 General M-estimate with High Breakdown Point

One procedure is proposed, in this section, for obtaining the general M-estimate with high breakdown point.

Generally the equation (2.3) is a set of non-linear equations, therefore an iterative method is required. It is indispensable for iterative algorithms to give an initial estimate. Yohai (1987) proved that, if an initial regression estimate with high breakdown point but not necessarily efficient is computed in the iterative method for M-estimate, then under reasonable conditions the M-estimate has high breakdown point. An initial estimate with high breakdown point is necessary to get a high breakdown general M-estimate.

The iteratively reweighted least squares (IRLS) algorithm with fixed scale parameter was proposed by Beaton and Tukey (1974) as an iterative technique for obtaining M-estimate. The connection between IRLS and M-estimator depends on choosing the weight function.

In the case of general M-estimator, if the weight function is defined as

$$w_i (x_i, (y_i - x_i \hat{\beta})/\hat{\sigma}) = \phi (x_i, (y_i - x_i \hat{\beta})/\hat{\sigma})/( (y_i - x_i \hat{\beta})/\hat{\sigma}), \tag{3.1}$$

then the equation (2.3) becomes the weighted least squares estimator

$$\sum_{i=1}^{n} w_i (x_i, (y_i - x_i \hat{\beta})/\hat{\sigma}) (y_i - x_i \hat{\beta}) x_i^T = 0. \tag{3.2}$$

Since the weight function (3.1) depends on $\hat{\beta}$ and $\hat{\sigma}$, parameter estimates can not be immediately obtained by solving equation (3.2). The equation suggests to consider the IRLS for $\hat{\beta}$.

If current value $\hat{\beta}_t$ and scale estimate $\hat{\sigma}$ (fixed) have been obtained, at each step substituting the current value into the weight function, then holding it fixed, it is possible to solve the equation (3.2) to get next value $\hat{\beta}_{t+1}$. It follows that

$$\begin{aligned} \hat{\beta}_{t+1} &= \hat{\beta}_t + \sum_{i=1}^{n} w_{i,t} (y_i - x_i \hat{\beta}_t) x_i^T / \sum_{i=1}^{n} w_{i,t} x_i x_i^T, \\ w_{i,t} &= w_i (x_i, (y_i - x_i \hat{\beta}_t)/\hat{\sigma}). \end{aligned} \tag{3.3}$$

Thus the general M-estimator is reduced to IRLS. This algorithm is able to use the existing least squares computer package other than to compute the weight function.

Huber (1981, p. 184) investigated the convergence conditions of IRLS. However, it is difficult to arrange the strict conditions for its convergence. This algorithm is expected to converge to a local minimum, by using the bounded monotone decreasing weight function. Therefore, it is crucial to get an initial estimate having good property.

Rousseeuw (1984) proved that the least median of squares (LMS) estimate, which is defined as the value satisfying

$$\underset{\hat{\beta}}{\text{minimize}} \; \underset{i}{\text{median}} \, (y_i - x_i \hat{\beta})^2, \tag{3.4}$$

has the finite sample breakdown point of $\{(n/2)-p+2\}/n$. This is asymptotically 50%. The breakdown point of fifty percent is the highest value one can expect (for large amount of contaminated observations in a sample, it is impossible to distinguish between the 'good' and 'bad' parts of the sample). The LMS estimate is acceptable 50% of contaminated obaservations in a sample without spoiling the estimate completely.

Therefore, the LMS estimate is proper for the initial estimate in IRLS to obtain the general M-estimate with high breakdown point. It is, however, mathematically impossible for multi-dimensional case to derive a straightforward formula of LMS estimate from the definition in (3.4). The computation method of LMS was devised by Leroy and Rousseeuw (1985).

In addition to the initial estimate of regression parameters, it is need to estimate the scale parameter in robust way. Rousseeuw (1984) proposed the scale estimate based on the equation

$$S_R = 1.483 \, C(n,p) \, \{\underset{\hat{\beta}}{\text{min.}} \; \underset{i}{\text{med.}} \, (y_i - x_i \hat{\beta})^2\}^{1/2}, \tag{3.5}$$

where the value of 1.483 is an asymptotic correction factor for the normal error. The constant $C(n,p)$ is a finite sample correction factor to be determined empirically, and necessary to make $S_R$ approximately unbiased when the simulating samples with normal distribution. According to Leroy and Rousseeuw (1985), the constant is given by

$$C(n,p) = \{1+5/(n-p)\}. \tag{3.6}$$

## 4 Sensitivity and Weight Function

Robust estimators are required to have low sensitivity mentioned in section 2. In order to define the weight function of (3.2), which makes the sensitivity of IRLS lower, the sensitivity of ordinary least squares is considered.

Hampel (1974) introduced the influence function, IF $(y,x)$, which measures the degree of the bias robustness of an estimate when the distribution of the central model is subject to an infinitesimal contamination. And, he defined the sensitivity by the norm of the influence function:

$$\gamma^* = \sup_{y,x} |IF(y,x)| = \sup_{y,x} \{IF(y,x)^T IF(y,x)\}^{1/2}. \tag{4.1}$$

The sensitivity measures approximately the maximum influence of a single observation on the estimate $\hat{\beta}$. Krasker and Welsch (1982) presented the alternative definition of the sensitivity of estimator for regression such that

$$\gamma^* = \sup_{y,x} \{IF(y,x)^T V^{-1} IF(y,x)\}^{1/2}, \tag{4.2}$$

where V is the covariance matrix of the influence function.

Instead of the influence function of least squares estimate, next quantity is considered at the central model,

$$b - b\,(i) = \frac{(X^T X)^{-1} x_i^T}{1 - h_i} (y_i - x_i\, b) \tag{4.3}$$

where b denotes the least squares estimate of $\beta$ and b(i) is its estimate without i-th observation. This quantity measures the influence of i-th observation on the least squares estimate.

Quantity (4.3) is devided by the scale estimate of b, which is $\{S^2(X^T X)^{-1}\}^{1/2}$ comparable to V and S denotes the estimate of $\sigma$. The quantity, $(b-b\,(i))/\{S^2\ (X^TX)^{-1}\}^{1/2}$, remains unchanged when the observations are shifted (rescaled). Then, this quantity is squared to convert it to the scalar. Since $(b-b\,(i))^T\ (b-b\,(i))$ is linear combination of p elements, the squared quantity may be shown as the form of

$$D\,(i) = \frac{(y_i - x_i b)^2}{pS^2} \cdot \frac{h_i}{(1-h_i)^2}\,. \tag{4.4}$$

In order to robustify the quantity (4.4), it is necessary to replace b with $\hat{\beta}$ estimated by LMS method and S with $S_R$ obtained in the equation (3.5). Thus, the quantity (4.4) is modified such as

$$\text{ROBUST D(i)} = \frac{(y_i - x_i\,\hat{\beta})^2}{pS^2_R} \cdot \frac{h_i}{(1-h_i)^2}\,. \tag{4.5}$$

In the sense of equation (4.2), the quantity (4.5) may be regarded as the 'squared' sensitivity of the least squares estimate, and useful diagnostic quantity for detecting influential observations on the least squares estimate. Furthermore, this quantity is followed by F{p, (n−p)} distribution at the central model.[1]

As far as given observations keep finite value, ROBUST D(i) is finite. If an observation is aberrant, it takes a large value. However, it is possible to make the high sensitivity of least squares more lower, by making use of the weighted least squares. The sensitivity of the weighted least squares estimate to i-th observation may be expressed by the form of

$$\gamma_w(i) = w_i(y_i, x_i, \hat{\beta}) \cdot \text{ROBUST } D(i), \tag{4.6}$$

where $w_i(y_i, x_i, \hat{\beta})$ is the weight to make the sensitivity lower.

In the weighted least squares, the weight should put an observation less weight than one only if its influence to the estimate would exceed the maximum allowance value, and all other observations would be given a weight of one. The maximum allowance value is set, by making use of the F-statistic, such that

$$D^* = F\{\alpha{:}p, (n-p)\} \tag{4.7}$$

with significance level $\alpha$.

The weight function for IRLS must satisfy the condition, $\gamma_w(i) \leq D^*$. By (4.5), (4.6) and (4.7), the weight function for IRLS at t-th step is defined as follows:

$$w_{i,t}(y_i, x_i, \hat{\beta}_t) = \begin{cases} 1 & \text{for ROBUST } D_t(i) \leq D^* \\ D^*/\text{ROBUST } D_t(i) & \text{for ROBUST } D_t(i) > D^* \end{cases},$$

$$\text{ROBUST } D_t(i) = \frac{(y_i - x_i\hat{\beta}_t)^2}{pS_R^{\,2}} \cdot \frac{h_i}{(1-h_i)^2}, \tag{4.8}$$

where $\hat{\beta}_{t=0}$ is the initial estimate of regression parameter.

For ROBUST $D_t(i) > D^*$, this weight function is bounded and monotone decreasing.

## 5 Concluding Remarks

For the purpose of feasibility of the general M-estimator, it is reduced to the method of iteratively reweighted least squares. Instead of defining the $\phi$-function, the weight function which is effective in making the influence of aberrant observations lower is directly given for this procedure. In order that the IRLS estimate of regression parameters may have high breakdown point, the least median of squares estimate is taken as its initial value. The initial estimate with high breakdown point is very crucial to make an iterative procedure provide a robust estimate.

The procedure treated in this paper is able to fit a regression line to the majority of given data, then discover the influential observations as those points which have a large value of the ROBUST D(i).

Asymptotic properties of this robust estimator, which are consistent and asymptotically normal, can be obtained from the Theorems of Maronna and Yohai (1981). However, its covariance matrix does not coincide with that of the least squares estimate at the central model, because it sacrifices the efficiency a little in order to make its sensitivity to aberrant observations of both y- and x-variable lower.

Asymptotic normality of this estimator ensures that conventional t-statistics and F-statistics for hypothesis testing of regression parameters are asymptotically valid. These test procedures are only valid asymptotically, so that the precise level of significance is not emphasized.

---

1 As long as u and x are completely independent and x is fixed at $x_1\ x_2 \cdots x_n$, the random variable x does not affect F-distribution statements (Graybill (1976), p.382).

## ACKNOWLEDGMENTS

This research was performed while the author was visiting the Australia-Japan Research Centre, Research School of Pacific Studies, Australian National University. The author is deeply greatful to executive director Peter Drysdale for making the arrangement of excellent research conditions.

An earlier version of this paper was presented at the seminar in A-JRC, 1 August 1988, Seminar Room 4, J.G. Crawford Building, Australian National University. The author would like to thank participants in the seminar for their comments. He also thanks Dr. Colin Mackenzie for reading the preliminary draft of the seminar paper and providing helpful comments.

## REFERENCES

Beaton, A.E. and Tukey, J.W. (1974). The fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectropic Data. *Technometrics* 16, 147-185.

Donoho, D.L. and Huber, P.J. (1983). The Notion of Breakdown Point. In *A Festschrift for Erich L. Lehmann,* Bickel, P.J., Doksum, K.A. and Hodges Jr., J.L. Eds., pp. 157-184. Belmont, Calif.: Wadsworth International Group.

Graybill, F.L. (1976). *Theory and Application of the Linear Model.* Boston, Massachusettes: Duxbury Press.

Hampel, F.R. (1971). A General Qualitative Definition of Robustness. *Annals of Mathematical Statistics* 42, 1887-1896.

———— (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association* 69, 383-393.

———— (1985). The Breakdown Points of the Mean Combined with Some Rejection Rules. *Technometrics* 27, 95-107.

Hampel F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: John Wiley.

Hoagling, D.C. and Welsch, R.E. (1978). The Hat Matrix in Regression and ANOVA. *American Statistician* 32, 17-22.

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

Huber, P.J. (1973). Robust Regression: Asymptotic Conjectures and Monte Carlo. *Annals of Statistics* 1, 799-821.

———— (1981). *Robust Statistics.* New York: John Wiley.

Krasker, W.S. (1980). Estimation in Linear Regression Models with Disparate Data Points. *Econometrica* 48, 1333-1346.

Krasker, W.S. and Welsch, R.E. (1982). Efficient Bounded-Influence Regression Estimation. *Journal of the American Statistical Association* 77, 595-604.

Leroy, A. and Rousseeuw, P.J. (1985). Progress: *A Program for Robust Regression Analysis on the IBM PC.* Report of the Department of Mathematics and Informatics No. 85-14. Delft: Delft University of Technology.

Maronna, R.A., Bustos, O.H., and Yohai, V.J. (1979). Bias- and Efficiency-Robustness of General M-Estimators for Regression With Random Carriers. In *Smoothing Techniques for Curve Estimation* (Lecture Note in Mathematics, 757), Gasser, T. and Rosenblatt, M. Eds., pp. 91-116. Berlin: Springer-Verlag.

Maronna, R.A. and Yohai, V.J. (1981). Asymptotic Behavior of General M. Estimates for Regression and Scale With Random Carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 58, 7-20.

Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association* 79, 871-880.

Siegel, A.F. (1982). Robust Regression Using Repeated Medians. *Biometrika* 69, 242-244.

White, H. (1984). *Asymptotic Theory for Econometricians.* New York: Academic Press.

Yohai, V.J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *Annals of Statistics* 15, 642-656.

Yohai, V.J. and Zamar, R.H. (1988). High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale. *Journal of the American Statistical Association* 83, 406-413.