Least Sine Squares and Robust Compound Regression Analysis

A Dissertation Presented

by

Hao Han

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

December 2011

Stony Brook University

The Graduate School

Hao Han

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Professor, Deputy Chair, Department of Applied Mathematics and Statistics

Xiangmin Jiao - Chairperson of Defense
Associate Professor, Department of Applied Mathematics and Statistics

Jiaqiao Hu
Assistant Professor, Department of Applied Mathematics and Statistics

Ellen Li
Professor, Department of Medicine, Stony Brook University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Least Sine Squares and Robust Compound Regression Analysis

by

Hao Han

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2011

The errors-in-variables (EIV) regression model, being more realistic by accounting for measurement errors in both the dependent and the independent variables, is widely used in econometrics, chemistry, medical, and environmental sciences, etc. The traditional EIV model estimators, however, can be highly biased by outliers and other departures from the underlying assumptions. In this work, we propose two novel nonparametric estimation approaches - the least sine squares (LSS) and the robust compound regression (RCR) analysis methods for the robust estimation of EIV models.

The RCR method, as a natural extension and combination of the new LSS method and the compound regression analysis method developed in our own group (Leng and Zhu 2009), provides the robust counterpart of the entire class of the traditional maximum likelihood estimation (MLE) solutions of the EIV model, in a 1-1 mapping. The advantages of both new approaches lie in their intuitive geometric interpretations, their distribution free property, their independence to the ratio of the error variances, and most importantly their robustness to outlier contamination and other violations of distribution assumptions. Monte Carlo studies are conducted to compare these new robust EIV model estimation methods to other nonparametric regression analysis methods including the least squares (LS) regression analysis method, the orthogonal regression (OR) analysis method, the geometric mean regression (GMR) analysis method, and the robust least median of squares (LMS) regression analysis method. Guidelines on which regression methods are suitable under what circumstances are provided through these simulation studies as well. Real-life examples are provided to further illustrate and motivate these new approaches.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost I am grateful to my advisor Prof. Wei Zhu, not only for her consistent guidance on my research but also for her help to me in every possible way at Stony Brook. She has taught me, both consciously and unconsciously, the necessary attributes a good statistician should have. Whenever I encountered some bottle-neck of my work, she always enlightens me with the thought-provoking questions and her warm smiles always cheer me up.

I am grateful for the guidance from Dr. Yeming Ma in my four years Ph.D. pursuit, especially his help to begin my research career at the Brookhaven National Lab. I will never forget the late nights you spent with me to solve urgent problems at the CEWIT building. I am thankful for the excellent example he has provided as a successful research scientist working on a variety of fields. I also want to thank my colleague and friend Dr. Kith Pradhan. I really learned a lot from you when we are working together at BNL.

For this dissertation I would like to thank my committee members: Prof. Xiangmin Jiao, Prof. Jiaqiao Hu, and Prof. Ellen Li for their time, interest, and insightful questions. I am especially grateful for the suggestions from Prof. Jiao on the development of fast algorithms for the least sine squares regression in higher dimensions.

Lastly, I would like to thank all my friends who make my life enjoyable at Stony Brook. I truly thank my family for all their love, encouragement, and support. And my special thanks go to my girlfriend Ying whose self-sacrifice and faithful support are so appreciated.

Hao Han
Stony Brook University
December 2011

Chapter 1

Introduction

Regression analysis includes many different approaches for modeling and analyzing the relationship between the dependent and the independent variables of interest. Among all approaches available, however, there is little doubt that the least squares (LS) regression analysis method, also referred to as the ordinary least squares method, is the most widely used one despite its often unrealistic and untenable assumption that only the dependent variable is a random variable while the independent variables are not. To take into account the randomness of the independent variables, the orthogonal regression (OR) analysis method, the geometric mean regression (GMR) analysis method, and the more general errors-in-variables (EIV) modeling approach were subsequently introduced. A good monograph on traditional EIV models and results is provided by Fuller (1987), and the more recent development and applications of the EIV model can be found in Van Huffel and Lemmerling (2002).

Furthermore, both the classic regression with the dependent variable random only and the traditional EIV model estimating methods, can behave poorly in the presence of contaminated data or violations of the underlying assumptions (Brown 1982, Carroll and Gallo 1982, Zamar 1989, Cheng and Vanness 1992). Of note, non-degenerate EIV models have a sharply increased chance to be subject to contaminations because the independent variables are also random. Therefore, robust EIV model estimation approaches are highly desired. In this work, we have developed two novel nonparametric approaches – the least sine squares (LSS) regression and the robust compound regression (RCR) analysis methods for the robust estimation of EIV models. These robust regression approaches also apply to the classic regression models with non-random independent variables as they can be treated as a special case of the EIV models.

The rest of this thesis is organized as follows. In Chapter 2, we introduce the existing traditional non-robust and robust estimation approaches for the EIV models, as well as the existing approaches for the computations of OR and GMR in high dimensional case. In Chapter 3, by introducing the robust 'least angle' criterion, we propose the first new robust regression approach – the LSS regression, and derive the multivariate LSS estimator. In Chapter 4, we present the second new robust regression approach – the RCR analysis method, which as a natural extension and combination of the new LSS method and the compound regression analysis method developed in our own group (Leng and Zhu 2009), provides the robust counterpart of the entire class of the traditional MLE solutions of the EIV model, in a 1-1 mapping. A new robust regression efficiency concept and the high breakdown outlier diagnostics based on our new approaches are also presented there. In chapter 5, the novel LSS and RCR estimators are compared to other classic as well as robust regression estimators, and the performance of the RCR estimator is thoroughly calibrated through a series of Monte Carlo studies. Three real-life examples are provided to further illustrate and motivate our new approaches in Chapter 6, and finally in Chapter 7 we conclude and discuss the advantages and disadvantages of our new approaches on the robust estimation of EIV models and propose the future work directions.

Chapter 2

Current EIV Model Estimation Approaches

2.1 Traditional Regression Analysis Methods

The classical simple linear regression model is defined by $Y = \alpha + \beta X + \varepsilon$, where the predictor X is considered as non-random and only the response variable Y is subject to random error ε. It is well known that when this assumption is satisfied, the least squares (LS) estimator is the best linear unbiased estimator (BLUE) by the Gauss-Markov theorem.

The objective of the LS regression of the response variable Y on the predictor X is to minimize the sum of squared Y variable prediction errors, which is to minimize

$$SS_{LS\_YonX} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Here Y_i is the observed response value, and $\hat{Y}_i$ is the corresponding predicted value of Y for a given X_i. Similarly, the LS regression of X on Y, where X is assumed to be random but Y fixed, aims to minimize the sum of squared prediction errors of the response variable X

$$SS_{LS\_XonY} = \sum_{i=1}^{n} (X_i - \hat{X}_i)^2$$

However, the ordinary LS estimators are biased and inconsistent under the circumstances of regression with errors in variables (Fuller 1987). Instead of minimizing sum of squared vertical (or horizontal) distances, the orthogonal regression (OR) takes the middle ground by minimizing the sum of squared orthogonal distances,

$$SS_{OR} = \sum_i d_{OR_i}^2 = \sum_i \frac{(Y_i - \hat{Y}_i)^2}{1 + \beta^2}$$

where d_{OR_i} is the distance from the observed data points to the regression line or to the hyper-plane in higher dimensions. The resulting slope estimator is as follows (Jackson and Dunlevy 1988)

$$\hat{\beta}_{OR} = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}}{2S_{XY}}$$

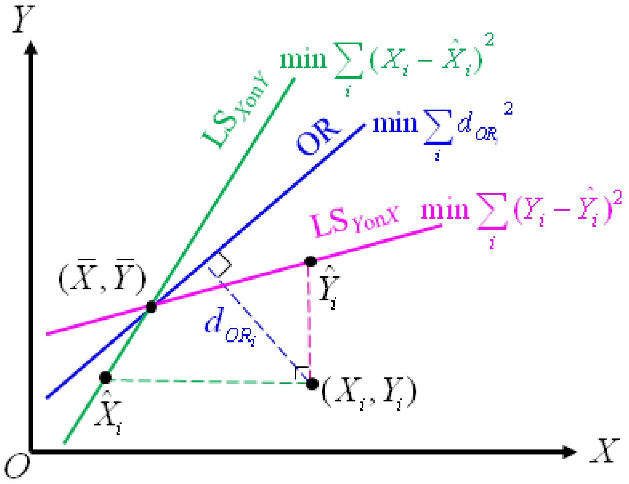


Figure 2.1 Geometrically, the LS regression with *Y* (or *X*) as the response variable minimizes the squared vertical (or horizontal) distances, while the OR takes the middle ground by minimizing the sum of squared orthogonal distances from the observation point to the regression line.

Another traditional approach is the geometric mean regression (GMR) approach. The slope estimator of GMR is the geometric mean of the slope estimate from the LS regression of *Y* on *X* and the reciprocal of the slope estimate from the LS regression of *X* on *Y*.

$$\hat{\beta}_{GMR} = sign(S_{XY})\sqrt{\hat{\beta}_{OLS,Y\ on\ X} * (\hat{\beta}_{OLS,X\ on\ Y})^{-1}} = sign(S_{XY})\sqrt{\frac{S_{YY}}{S_{XX}}}$$

In bivariate case, the GMR minimizes the sum of the triangular areas formed by connecting the measured data points to the estimated line with lines parallel to the coordinate axes (Harvey and Mace 1982, Barker et al. 1988). That is to minimize

$$SS_{GMR} = -\frac{1}{2}sign(\hat{\beta})\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(X_i - \hat{X}_i)$$

One advantage of the GMR solution is its natural symmetry if the roles of *X* and *Y* are reversed, and this symmetry is obvious as interchange of the *X* and *Y* axes leaves the areas unchanged (Draper 1992). The GMR also has its downside in the sense that it is not easy to conduct tests or construct confidence intervals on the parameters (Creasy 1956), and the GMR estimator is known to be inconsistent when the sample size is large.

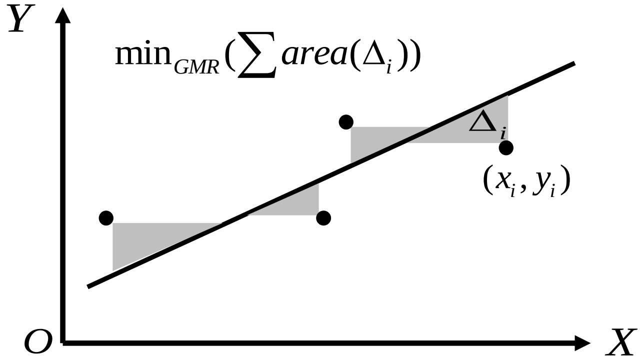


Figure 2.2 Geometrically, the GMR minimizes sum of the triangular area bounded by the regression line and the vertical and horizontal lines through each observation point.

2.2 OR and GMR in Higher Dimensions

It is known that there is a close relationship between the OR estimation approach and the principle component analysis (PCA) (Jackson and Dunlevy 1988). Since the largest eigenvalue of the sample covariance matrix $S = \begin{bmatrix} S_{XX} & S_{XY} \\ S_{XY} & S_{YY} \end{bmatrix}$ of regression variables (X, Y) is $\lambda = \frac{S_{XX}+S_{YY}+\sqrt{(S_{XX}+S_{YY})^2-4(S_{XX}S_{YY}-S_{XY}^2)}}{2}$, and the eigenvector corresponding to this eigenvalue is $(S_{XY}, \frac{S_{YY}-S_{XX}+\sqrt{(S_{XX}+S_{YY})^2-4(S_{XX}S_{YY}-S_{XY}^2)}}{2})$, the slope of the first principle component is $\frac{S_{YY}-S_{XX}+\sqrt{(S_{XX}+S_{YY})^2-4(S_{XX}S_{YY}-S_{XY}^2)}}{2S_{XY}}$, which is exaclty the same as the slope estimator from the OR approach. Intuitively, the first principle component is the line passing through the greatest dimension of the concentration ellipse and it explains the maximum variance of a scatter of points (Morrison 1976). Of course, this connection between OR and PCA also holds in the high dimensional case.

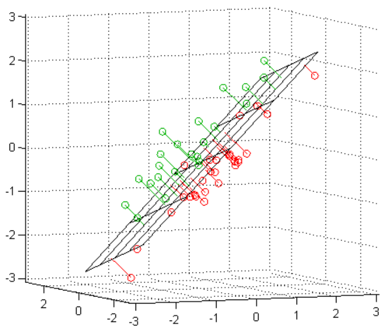


Figure 2.3 OR plane in 3D with orthogonal distance deviations. The OR aims to minimize the sum of squared orthogonal distances from each observation to the regression plane.

An alternative and computationally more efficient way to compute the OR estimate in higher dimensions is by using the singular value decomposition (SVD) (Golub and Van Loan 1996), though the computation of OR via PCA in high dimensional case is analytically tractable as well. In the context of matrix computations, the linear regression model estimation problem can be formed as solving the equation $X\beta = Y$ for β, where X is n-by-p and Y is n-by-1. Here n is the number of observations and p is the number of predictor variables.

That is, we seek to find β that minimizes error matrices E and F for X and Y respectively:

$$argmin_{E,F}||[E\ F]||_f, \text{ s.t. } (X + E)\ \beta = Y + F$$

where $[E\ F]$ is the augmented matrix with E and F side by side and $||.||_f$ is the Frobenius norm, the square root of the sum of the squares of all entries in a matrix. This can be rewritten as

$$[(X + E)\ (Y + F)] \begin{bmatrix} \beta \\ -I_k \end{bmatrix} = 0$$

where I_k is the k x k identity matrix. The goal is then to find $[E\ F]$ that reduces the rank of $[X\ Y]$ by k. Define $[U][\Sigma][V]'$ to be the SVD of the augmented matrix $[X\ Y]$.

$$[X\ Y] = [U_X\ U_Y]\begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{bmatrix}\begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}' = [U_X\ U_Y]\begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{bmatrix}\begin{bmatrix} V_{XX}' & V_{YX}' \\ V_{XY}' & V_{YY}' \end{bmatrix}$$

where V is partitioned into blocks corresponding to the shape of X and Y.

The rank is reduced by setting some of the singular values to zero. That is, we want

$$[(X + E)\ (Y + F)] = [U_X\ U_Y]\begin{bmatrix} \Sigma_X & 0 \\ 0 & 0_{kxk} \end{bmatrix}\begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}',$$

so by linearity, we have

$$[E\ F] = -[U_X\ U_Y]\begin{bmatrix} 0_{nxn} & 0 \\ 0 & \Sigma_Y \end{bmatrix}\begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}',$$

We can then remove blocks from the U and Σ matrices, simplifying to

$$[E\ F] = \text{-}U_X\, \Sigma_Y \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix}' = -[X\ Y] \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix} \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix}'$$

This provides E and F so that

$$[(X + E)\ (Y + F)] \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix} = 0$$

Now if V_{YY} is nonsingular, which is not always the case (note that the behavior of OR when V_{YY} is singular is not well understood yet), we can then right multiply both sides by $-V_{YY}^{-1}$ to bring the bottom block of the right matrix to the negative identity.

$$[(X + E)\ (Y + F)] \begin{bmatrix} -V_{XY} & V_{YY}{}^{-1} \\ -V_{YY} & V_{YY}{}^{-1} \end{bmatrix} = [(X + E)(Y + F)] \begin{bmatrix} \beta \\ -I_k \end{bmatrix} = 0$$

and so $\beta = -V_{XY}V_{YY}^{-1}$ is the orthogonal regression coefficient estimates, which extends the OR estimation approach to the higher dimensions.

The computation of GMR estimates in higher dimensions is finally realized by solving a quadratic convex programming problem (Draper and Yang 1997). By defining the squared vertical deviation in the Y direction $D_{Y_i}^2 = (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji} - Y_i)^2$ and in the X_l direction $D_{X_{li}}^2 = \frac{(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji} - Y_i)^2}{\beta_l{}^2}$, we consider a composite measure of deviations in all directions.

$$G_i = (D_{Y_i}^2 D_{X_{1i}}^2 D_{X_{2i}}^2 \ldots D_{X_{pi}}^2)^{\,1/_{k+1}} = V_i{}^{2/_{k+1}}$$

Algebraically, it is the geometric mean of all the squared deviations; and geometrically, it depends on the volume V_i created by drawing, from each point, lines parallel to all the axes to the fitted hyper-plane.

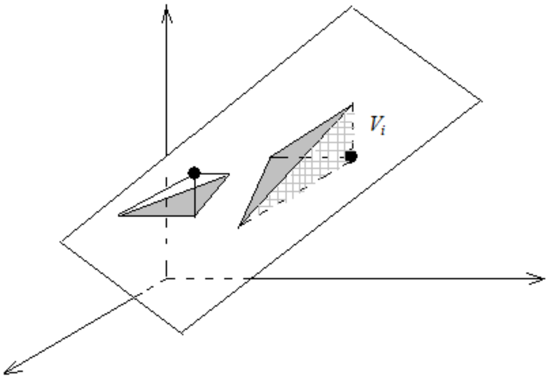


Figure 2.4 GMR plane in 3D with volume deviations. The volume V_i is created by drawing, from each point, lines parallel to all the axes to the fitted plane.

When $k = 1$, $G_i = V_i$ is equivalent to the right-angled triangle area, which means the generalized criterion is equivalent to the definition of GMR in simple linear regression. In fact when $k > 1$, G_i is the extension of this area in higher dimensions. The minimization of $\sum_i G_i$ can be further simplified to be a quadratic problem as follows

$$\min \boldsymbol{\tau}'\boldsymbol{S}\boldsymbol{\tau} \quad \text{subject to } \tau_0\tau_1 \ldots \tau_p = 1 \ \& \ \tau_i > 0$$

where $\boldsymbol{S}$ is the sample covariance matrix, and $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_p)'$ is a column vector with $\tau_j = \frac{\beta_j}{(\prod_{j=1}^{p} \beta_j)^{\frac{1}{p}}}$.

2.3 Structural EIV Modeling Approach

The simple linear EIV model assumes there exists an underlying linear relationship $\eta = \alpha + \beta\xi$ between two latent variables η and ξ. Instead of observing the latent variables, one observes $Y = \eta + \varepsilon$ and $X = \xi + \delta$, where the corresponding measurement errors $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\delta \sim N(0,$

σ_δ^2) are independent to the unobserved variable ξ (Sprent 1969). The variables enter the EIV model in a symmetric manner from a distributional point of view, but we generally choose to identify one variable as the response variable *Y*.

There are two well-established EIV modeling approaches - the functional and the structural approaches. The difference between them is whether to consider the underlying variable ξ as a non-random (functional approach) or random variable with mean μ and variance τ^2 (structural approach). In this work, we will mainly focus on the more general structural approach, in which situation *X* and *Y* will follow a joint bivariate normal distribution.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \alpha + \beta\mu \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma_\delta^2 & \beta\tau^2 \\ \beta\tau^2 & \beta^2\tau^2 + \sigma_\epsilon^2 \end{pmatrix}\right)$$

Subsequently, we can obtain the MLE of the slope coefficient, which depends on the ratio of the two error variances $\lambda = \sigma_\varepsilon^2/\sigma_\delta^2$ (Lindley 1947, Wong 1989).

$$\hat{\beta} = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2}}{2S_{XY}}$$

It is known that, given the bivariate normality assumption, the three traditional nonparametric regression methods - LS, OR and GMR, are special cases of the MLE solutions, each corresponding to a unique value of λ. The LS regression of *X* on *Y* is the MLE solution of structural EIV model when $\lambda = 0$ (error variance of *Y* is 0, i.e. *Y* is not random); while the LS of *Y* on *X* is equivalent to the MLE solution when $\lambda = \infty$ (error variance of *X* is 0, i.e. *X* is not random). The orthogonal regression (OR) is the MLE solution when $\lambda = 1$ which means that the OR is suitable when the error variances are equal. The geometric mean regression (GMR) is equivalent to the MLE solution when $\lambda = S_{YY}/S_{XX}$ (Sprent and Dolby 1980). This means that the

GMR approach is suitable when the randomness from X and Y are from the random errors only. That is, when we take the functional analysis approach by assuming that ξ is not random (Leng et al. 2007).

Unfortunately, the MLE approach for the EIV model estimation is unattainable, because it depends on the normality assumption, and furthermore the λ ratio of the error variances, which is usually unknown and unable to be estimated from the data alone. If the λ is known, the use of MLE solution is probably best, but it is suggested to use the GMR estimator otherwise, if the assumption that $\lambda = S_{YY}/S_{XX}$ is unreasonable (Draper 1992, Draper and Smith 1998).

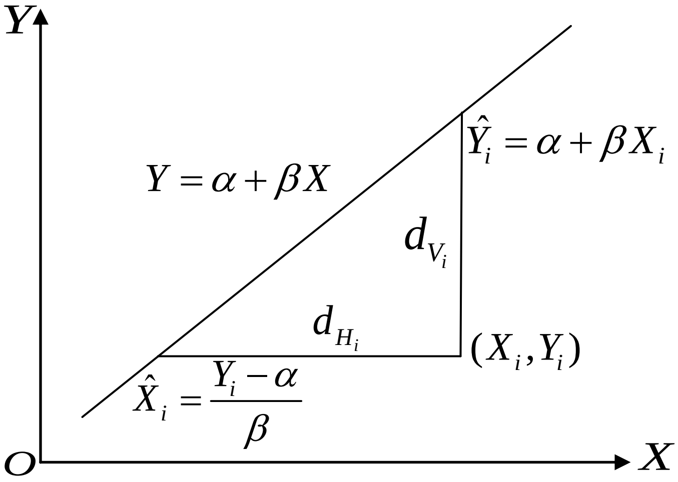
The multiple linear regression EIV model, analogous to the simple linear EIV model, assumes a linear relationship between p latent variables $\sum_{j=1}^{p} \beta_j \xi_j = \alpha$, which is uniquely defined by imposing the constraint $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. One only obverses $X_j = \xi_j + \varepsilon_j$ with the random errors $\varepsilon_j \sim N(0,\ \lambda_j \sigma^2)$, $j = 1, 2, ..., p$, where the non-negative ratios of the error variances λ_j are specified. As a result, the error vector is $\boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \sigma^2\boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = diag(\lambda_1,\ \lambda_2, \dots, \lambda_p)$.

Similar to the simple structural EIV model defined earlier, the multivariate structural EIV approach assumes $\boldsymbol{\xi} = (\xi_1,\ \xi_2, \dots, \xi_p)'$ being independent of $\boldsymbol{\varepsilon}$ will follow a p-variate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the MLE solution of the structural EIV model with $\boldsymbol{\Lambda}$ non-singular is given by

$$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\omega}}_1' \boldsymbol{\Lambda}^{-1} \widehat{\boldsymbol{\omega}}_1)^{-\frac{1}{2}} \boldsymbol{\Lambda}^{-\frac{1}{2}} \widehat{\boldsymbol{\omega}}_1$$

where $\widehat{\boldsymbol{\omega}}_1$ is the normalized eigenvector associated with the smallest eigenvalue of $\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{S} \boldsymbol{\Lambda}^{-\frac{1}{2}}$, and $\boldsymbol{S}$ is the sample covariance matrix of the observations $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ (Patefield 1981).

2.4 Compound Regression Analysis Method



Y
$Y = \alpha + \beta X$
$\hat{Y}_i = \alpha + \beta X_i$
d_{V_i}
d_{H_i}
(X_i, Y_i)
$\hat{X}_i = \frac{Y_i - \alpha}{\beta}$
O
X

Figure 2.5 Geometric demonstration of the compound regression approach.

When both Y and X are random, one would naturally wish to find a regression line that will minimize variations in both directions. The compound regression (CR) (Leng and Zhu 2009), being a direct generalization of the ordinary least squares method, is designed to minimize the weighted average of the sum of squared vertical and horizontal distances as follows:

$$SS_\gamma = \gamma \sum_i (Y_i - \hat{Y}_i)^2 + (1-\gamma) \sum_i (X_i - \hat{X}_i)^2$$

$$= \gamma \sum_i (Y_i - \alpha - \beta X_i)^2 + (1-\gamma) \sum_i (X_i - \tfrac{Y_i - \alpha}{\beta})^2 \qquad \text{where } 0 \le \gamma \le 1$$

In the same manner as the traditional regressions being special cases of the MLE solution of EIV models (Casella and Berger 2002), the three classic regressions are reduced to special cases of the compound regression model as well. At the two extremes, the CR model is equivalent to the LS regression of Y on X when $\gamma = 1$, while $\gamma = 0$ corresponds to the LS regression of X on Y. For the OR and GMR approaches, the corresponding γ's in the CR framework are determined from each specific dataset.

For each $\gamma \in [0,\ 1]$, the minimization of SS_γ requires solving the equations $\frac{\partial \mathrm{SS}_\gamma}{\partial \alpha} = 0$ and $\frac{\partial \mathrm{SS}_\gamma}{\partial \beta} = 0$ simultaneously. Straight-forward derivation shows that the estimators $\hat{\alpha}$ and $\hat{\beta}$ would satisfy:

$$\alpha = \bar{Y} - \beta \bar{X} \qquad (1)$$

$$\gamma S_{XX}\beta^4 - \gamma S_{XY}\beta^3 + (1-\gamma)S_{XY}\beta - (1-\gamma)S_{YY} = 0 \qquad (2)$$

The solutions can be obtained using any standard numerical software.

It is shown that there exists a monotonic relationship between γ and $\hat{\beta}$, and a monotonic relationship between λ and $\hat{\beta}$ as well. Moreover, since the Equation (2) can be rewritten as $\gamma = \frac{\widetilde{S_{YY}} - \widehat{\beta}\widetilde{S_{XY}}}{\widetilde{S_{YY}} - \widehat{\beta}\widetilde{S_{XY}} + \widehat{\beta}^4\widetilde{S_{XX}} - \widehat{\beta}^3\widetilde{S_{XY}}}$, and the MLE solution of structural EIV approach presented in the last section can be expressed in the form of $\lambda = \frac{\widehat{\beta} S_{XY} - S_{YY}}{S_{XY} - S_{XX}\widehat{\beta}}\hat{\beta}$, it has been further shown that there is a 1-1 correspondence between each CR regression line and each MLE solution of the structural EIV regression model in that each particular value of γ corresponds to a unique value of λ, and vice versa.

In high dimensional case, the multivariate CR takes account of all the prediction errors with different weights, and aims to minimize the following weighted average of sum of squares function

$$SS_\gamma = \gamma_0 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \gamma_1 \sum_{i=1}^{n} (X_{1i} - \hat{X}_{1i})^2 + \cdots + \gamma_p \sum_{i=1}^{n} (X_{pi} - \hat{X}_{pi})^2$$

The equivalence between the CR approach and the structural EIV approach in higher dimensions, and the fact that the OR is the same as the MLE when all the errors are equal ($\Lambda = I$), have been proven as well (Leng and Zhu 2009).

2.5 Robust Estimation of EIV Models

Since both the MLE approach for EIV models and its nonparametric counterpart - the compound regression analysis method weigh each observation equally, and thus are very sensitive to outliers and influential observations, it is necessary to develop robust estimators of EIV models that will not only down-weigh observations with large residuals but also account for measurement errors in both dependent and independent variables.

We will firstly introduce several major robust ordinary regression estimators that are not originally developed for EIV models, but we should be aware that even the robust ordinary regressions are useful in providing a robust estimation of the EIV models (Ketellapper and Weisbeek 1983, Ketellapper and Ronner 1984).

The least absolute deviations (LAD) regression also called L_1–norm regression leads to minimizing the sum of absolute deviations in the response variable Y direction $\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$, while the general class of L_p–norm regressions will minimize the sum of the pth power of the absolute deviations $\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|^p$. It is suggested that a value of p = 1.5 could be a good choice for the LAD method (Forsythe 1972) as it is known to be non-robust to gross outliers. The 'maximum likelihood type' M-estimators (Maronna 1976) suppose the errors are independently and identically distributed as $f(\varepsilon)$, and then the MLE of β is given by maximizing the likelihood function $\prod_{i=1}^{n} f(Y_i - X_i'\beta)$. However, it is known that the M-estimators are not resistant to outliers in the explanatory variable, which are so called leverage points in analogy to the notion of leverage in mechanics. The following figure will graphically interpret the difference between the two basic types of outliers in simple linear regression.

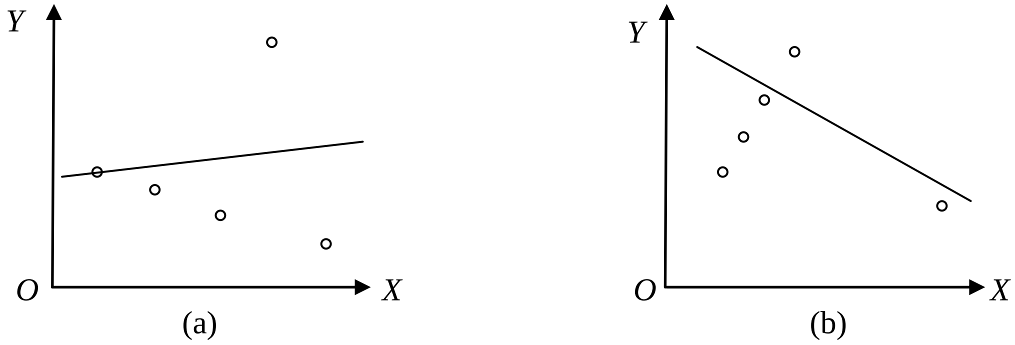


Figure 2.6 (a) An outlier in the *Y*-direction has its *Y* coordinate lying out of the *Y* range of the main bulk of 'good' points; (b) an outlier in the *X*-direction (leverage point) has its *X* coordinate lying out of the *X* range of the main bulk of 'good' points.

The best known and most widely used robust ordinary regression technique is the least median of squares (LMS) regression (Erickson et al. 2006). The LMS regression aims to find the regression line or hyper-plane that will minimize the median of squared response variable *Y* residuals, i.e., to minimize $Median_{i=1:n}(Y_i - \hat{Y}_i)^2$ (Rousseeuw 1984). Graphically, by finding the narrowest strip (in *Y*-axis direction) that covers "half" of the observations, the simple LMS line lies at the center of this band as shown in the figure below. One important property about the LMS is that, different from the parametric or nonparametric regression approaches we mentioned earlier, the LMS line or hyper-plane is not determined to pass through the 'center of mass' of the target dataset.

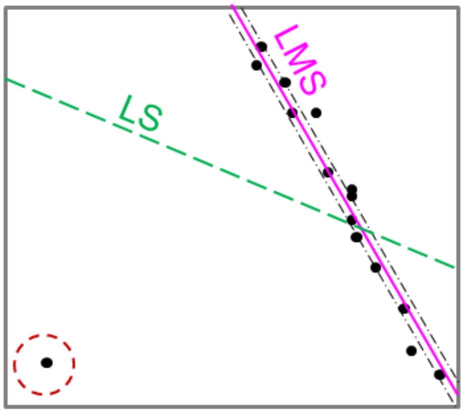


Figure 2.7 LMS's advantage over LS on robustness to leverage points. The LMS line lies at the center of the narrowest strip (in vertical direction) covering half of the observations. Unlike the traditional methods, the LMS line does not necessarily pass through the center of the whole dataset.

The most important property of the LMS lies in its robustness to outliers with the highest possible breakdown point of 50%. In the context of robust regression, the definition of a breakdown point is as follows. Let *Z* represent the sample of *n* data points (z_1, ..., z_n), and let *T* be a regression estimator. Consider *Z*' be all possible corrupted samples that are obtained by substituting any subset of size *m* of the original data points by arbitrary values, the maximum bias that can be caused from such a contamination is denoted by $bias(m, T, Z) \equiv sup_{Z'} \|T(Z) - T(Z')\|$, where the supremum is taken over all possible *Z*'. Then the finite-sample breakdown point is defined as $\varepsilon_n^* = \min[\frac{m}{n}; bias(m, T, Z) = \infty]$ (Donoho and Huber 1983), which does not depend on the underlying distribution. Intuitively, the ε_n^* is at most 50%, if it is larger than 50%, one could build a configuration of outliers which is just a translation image of the "good" data points (Rousseeuw and Leroy 1987).

In addition, the LMS estimator is regression-, scale-, and affine- equivariant, which corresponds to the three basic types of equivariance for regression estimators, ranked from higher to lower priority. An estimator is called regression equivariant if

$$T(\{(X_i, Y_i + X_i V); i = 1, \dots, n\}) = T(\{(X_i, Y_i); i = 1, \dots, n\}) + V$$

where *V* is any column vector. This ensures the validity of the phrase ‘without loss of generality, let $\boldsymbol{\beta} = \mathbf{0}$’ at the beginning of many proofs of asymptotic properties or descriptions of Monte Carlo studies. An estimator is said to be scale equivariant if

$$T(\{(X_i, cY_i); i = 1, \dots, n\}) = cT(\{(X_i, Y_i); i = 1, \dots, n\})$$

for any constant *c*. It implies that the fit is essentially independent of the choice of measurement unit for the response variable Y. And a regression estimator is called affine equivariant if

$$T(\{(X_{\boldsymbol{i}} A, Y_i); i = 1, \dots, n\}) = A^{-1} T(\{(X_{\boldsymbol{i}}, Y_i); i = 1, \dots, n\})$$

for any nonsingular square matrix *A*. This allows us to use another coordinate system for the explanatory variables, without affecting the estimated $\widehat{Y}_i$ (Rousseeuw and Leroy 1987). The ordinary LS estimator, on which the LMS is based, is also regression-, scale-, and affine-equivariant, but the OR does not show any of these equivariances, while the GMR is scale-equivariant only (see Theorem 1 in the Appendix).

On the other hand, gold cannot be pure and the LMS method cannot be perfect. First of all, there exists the high computing complexity to compute the LMS estimate (Erickson et al. 2006), which makes the computational time of the LMS to be abnormally more than that of the traditional regressions. Secondly, the median of squared residuals lacks a smooth squared residual function and takes a long time to converge, and the asymptotic convergence rate of the

LMS as of $n^{-\frac{1}{3}}$, which is much slower than that of the usual approaches as of $n^{-\frac{1}{2}}$ (Martin 2002). Last but not the least, the LMS performs poorly for small samples.

Although the robust ordinary regression methods discussed above are not designed for the situations of regression with errors in variables, the approaches to obtain the robust estimates of the EIV models are in many ways analogous to these approaches. The robust *w*-estimator was first developed by applying robust ordinary regression techniques to the EIV model (Brown 1982). For the detection of influential observations, Kelly (1984) derived the influence function of the orthogonal regression. Zamar (1989) proposed the robust orthogonal regression M-estimators (ORM) and S-estimators that are adapted to the EIV problems, and it was shown his methods outperforms the corresponding robust ordinary regression methods. Later on, the robust orthogonal generalized M-estimators (Cheng and Vanness 1992) were proposed to generalize the methods proposed by Zamar (1989). Recently, a quantile regression (QR) approach was provided by He and Liang (2000) to account for the presence of heavier-tailed errors rather than the Gaussian errors in a class of linear EIV models. By combining the orthogonal regression with the robust M-, S-, and MCD- estimators, a class of robust weighted orthogonal regression estimators was newly developed as well (Fekri and Ruiz-Gazen 2004).

Additionally, when Rousseeuw (1984) made the groundbreaking work by proposing his distinctive LMS regression, he also pointed out the idea to generalize the LMS by minimizing the median of the squared orthogonal residuals (Rousseeuw and Leroy 1987), but failed to provide a thorough development of this idea. Several researchers seized the importance of this insight to the robust estimation of EIV models, and developed the orthogonal least median of squares regression (Hartmann et al. 1997, Sarabia et al. 1997) and the analogous orthogonal least trimmed squares regression (Jung 2007). It has been viewed that the performance of these new approaches is still largely restricted by the computational low efficiency of the underlying LMS kernel, and we ought to acknowledge that all robust estimators sacrifice some of their efficiency in order to reduce their sensitivity to contamination or violation of the assumptions.

Chapter 3

Least Sine Squares Regression

3.1 Introduction of Angular Regression

Instead of using different kinds of distance measures as adopted in the existing regression methods to gauge the discrepancy between the fitted data from model and the original data, the angular regression (AR), from a novel point of view, makes use of the angular measure to build up its best-line-of-fit criterion as illustrated by Figure 3.1.

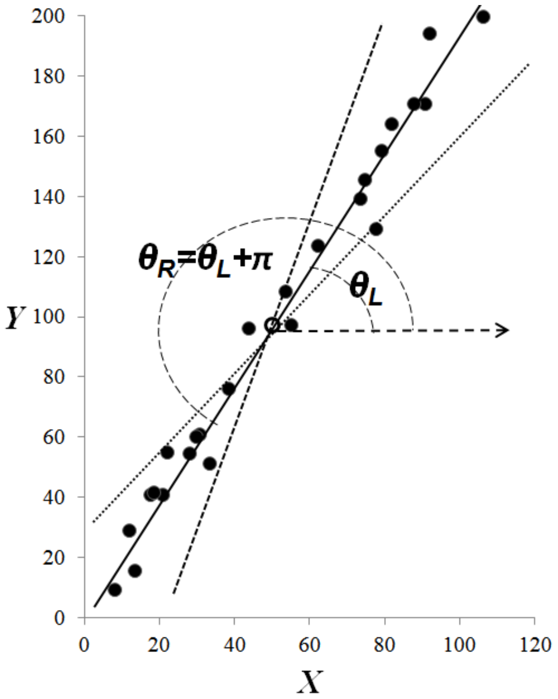


Figure 3.1 Geometric illustration of angular regression. The primary goal of linear regression analysis is to find the best regression line with slope $\beta = tan\theta$, or equivalently to find the

corresponding best angle θ relative to the X-positive direction to represent the underlying data. Intuitively, the neighborhood around the best angle shall cover the highest density of data points to make the fitted line the most representative. To search for the best angle, suppose a radar station is set up at the center of dataset, when the radar beam sweeps its angle θ from 0 to π, we will find one angle denoted as θ_L having the highest density of data points under the radar beam coverage. In the same manner, as searching the angle from π to 2π, we will observe another peak at angle θ_R. As the difference between θ_R and θ_L is approximately π, the estimated angle of the AR line is defined to be $\hat{\theta} = \frac{\theta_L + \theta_R - \pi}{2}$.

The conceptually simple AR approach is not only distribution free, but also robust to outliers and influential observations. The AR line is more determined by the data points near the true regression line, but less sensitive to noisy points that are relatively far away.

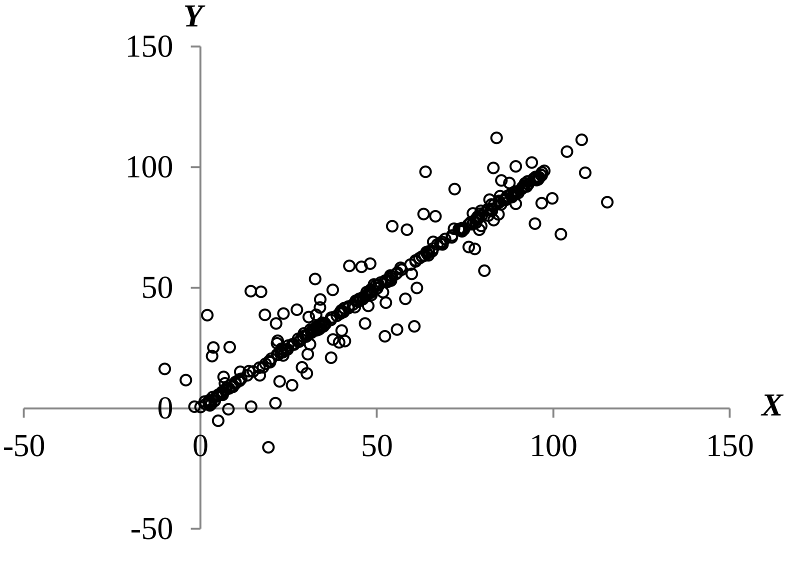


Figure 3.2 Scatter plot of the two normal mixture data. The slope $\beta = 1$, $\xi \sim U(0, 100)$, and the data is a mixture of $n_1 = 100$ points with random errors ε_X & $\varepsilon_Y \sim N(0, 100)$, and $n_2 = 200$ points with random errors ε_X & $\varepsilon_Y \sim N(0, 0.25)$.

To have a better understanding on the robustness property of AR, a motivational simulation study is provided in the following part. We set up the simple linear EIV model with the true slope $\beta = 1$ and observations $Y = \xi + \varepsilon_Y$ and $X = \xi + \varepsilon_X$, where $\xi \sim U(0, 100)$ distribution. Then the data with a sample size of $n_1 + n_2$ is generated as a mixture of two components. In the more noisy component of $n_1 = 100$ points, the random errors ε_Y and ε_X both follow $N(0, 100)$ distribution, while for the less noisy component of $n_2 = 200$ points, the random errors both come from the $N(0, 0.25)$ distribution. Figure 3.2 is the scatter plot of the two normal mixture data set generated, and the empirical probability density functions (EPDF) estimated from each component and their mixture are shown in the Figure 3.3.

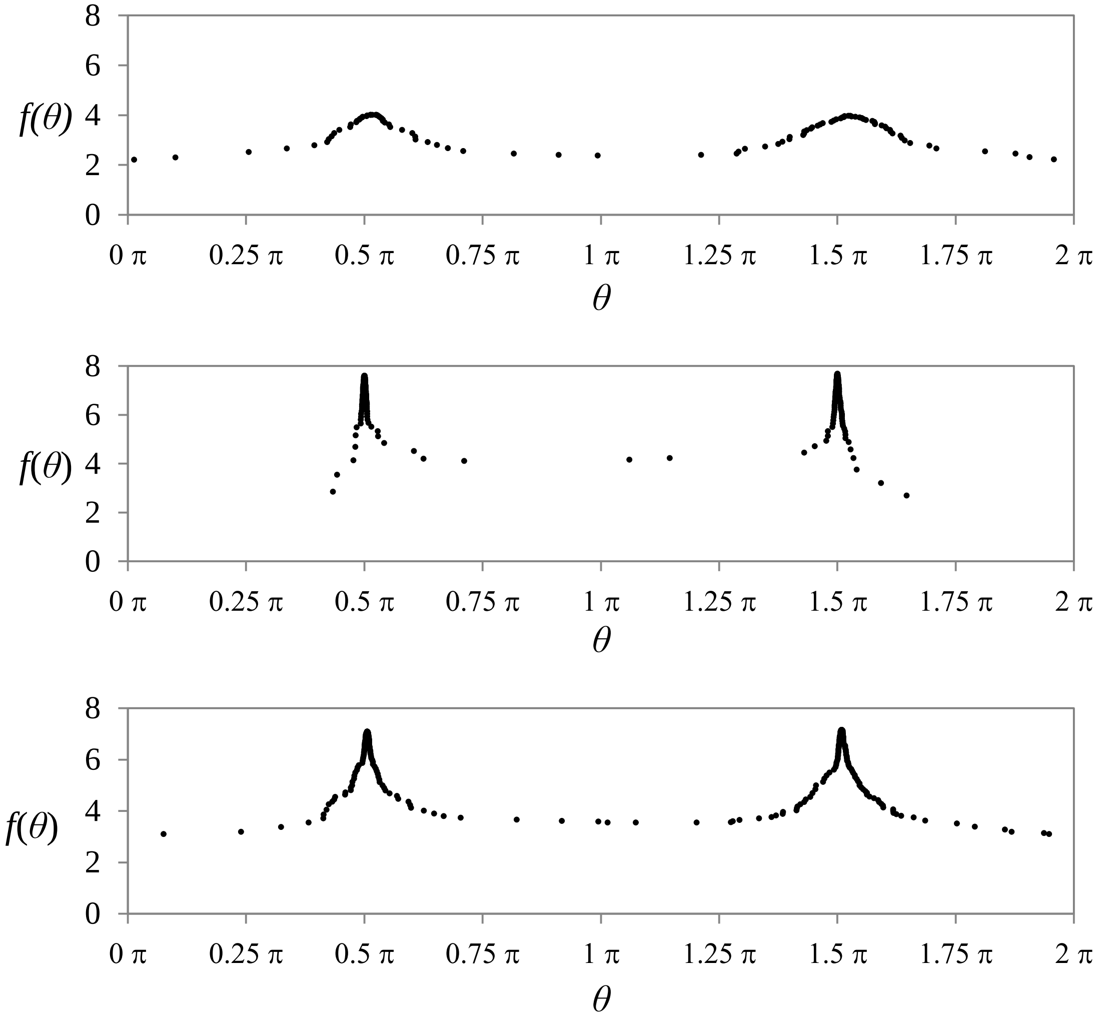
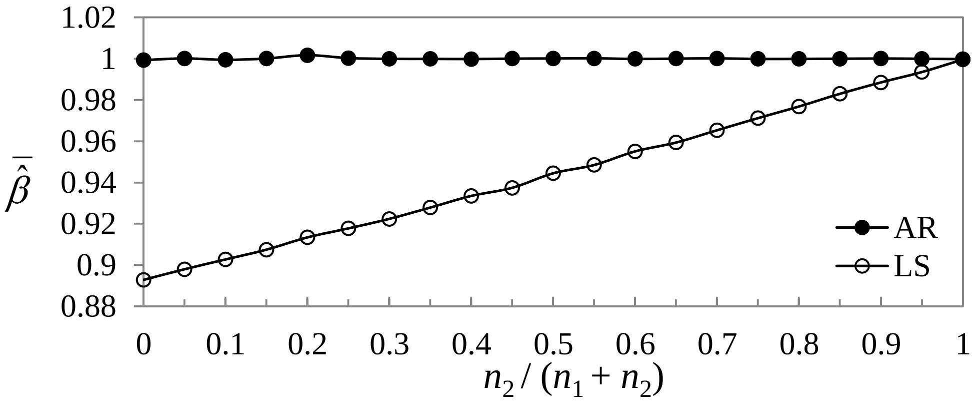


Figure 3.3 EPDF of each normal component and their mixture. The vertical-axis represents the number of data points covered by every one degree of the angle θ when sweeping from 0 to 2π, which serves as an EPDF with respect to the angle θ. Intuitively, for any data with a linear trend

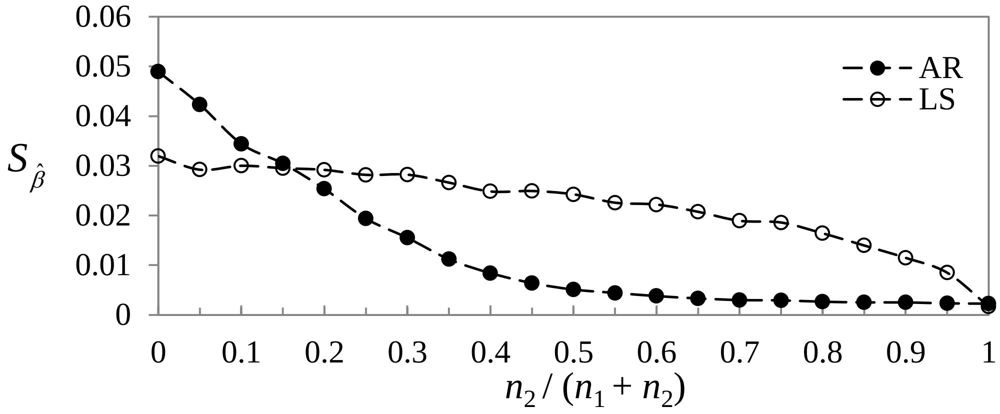
in general positions, its EPDF has only two peaks as can be seen. Top panel: the smoothed EPDF of $n_1 = 100$ points with random errors follow $N(0, 100)$; Middle panel: the smoothed EPDF of $n_2 = 200$ points with noise follow $N(0, 0.25)$; Bottom panel: the smoothed EPDF of the mixture of the same $n_1 = 100$, $n_2 = 200$ points.

By comparing the three panels presented in Figure 3.3 simultaneously, it is not hard to find that the bottom panel is an approximate superposition of the above two, and the peaks can be easily detected even after the combination of the more noisy component (top panel) and the less noisy component (middle panel), which indicates the peaks of the EPDF of the mixture is mainly determined by the more peaky less noisy component. Put it in another way, we can draw a conclusion that the critical angles θ_L and θ_R to obtain the AR estimate are mainly determined by the less noisy component, which dramatically enhances the robustness of the AR estimator.

To further illustrate the robustness of the AR estimator, we will compare it with the most widely used LS estimator. The simulation results based on 1000 replications of generated data with a sample size of $n_1 + n_2 = 200$ are summarized in Figure 3.4, where the mixture proportion of less noisy component ranges from 0 to 100 percent.



(a)



(b)

Figure 3.4 Illustration of the advantage of the AR compared to the LS in terms of the (a) mean $\bar{\hat{\beta}}$ and (b) standard error $S_{\hat{\beta}}$ of their slope estimates for a dataset of $n_1 + n_2 = 200$ points with a changeable mixture portion of the less noisy component $n_2 / (n_1 + n_2)$.

Our simulation studies indicate that the AR estimator is always less biased to the true slope $\beta = 1$ compared to the LS estimator. Meanwhile, the standard error of the AR slope estimate is smaller than that of the LS when the mixture proportion of less noisy component exceeds 15% in this simulation case. The simulation results again substantiate our argument that more concerns on the angular regression concept are deserved in robust regression analysis.

3.2 LSS Regression

As can be seen, the AR approach represents a promising direction in developing new robust regression methods from the angular measure point of view. Since the AR estimator is not analytically tractable, in order to increase its efficiency, we hereby propose the novel least sine squares (LSS) regression analysis method. The objective of the LSS is to minimize the sum of squared *sine* distances:

$$SS_{LSS} = \sum_{i=1}^{n} sin^2\varphi_i = \sum_{i=1}^{n} \frac{d_{OR_i}^2}{R_i^2}$$

where φ_i is the angle between the line connecting a data point with the 'center of mass' $(\bar{X}, \bar{Y})$ and the fitted regression line. d_{OR_i} represents the orthogonal distance from each point to the regression line, and R_{i} is the distance from each observation to the 'center of mass'.

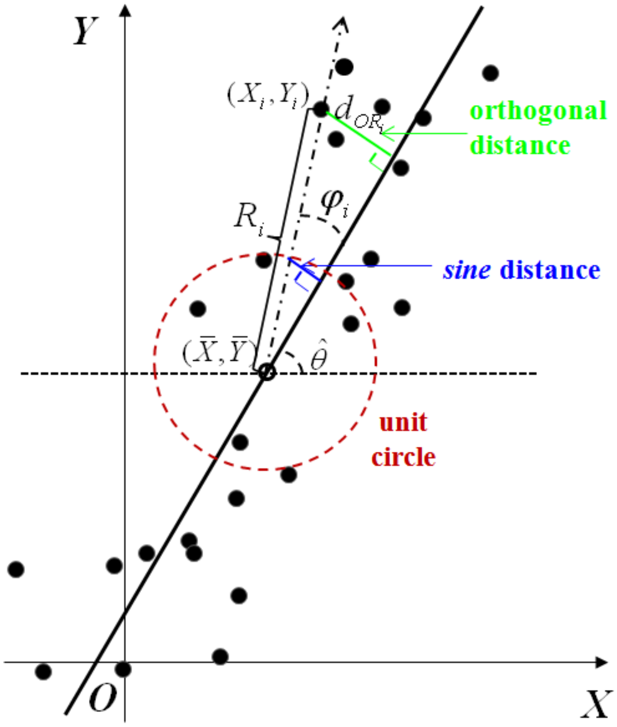


Figure 3.5 Geometric interpretation of the least sine squares (LSS) regression approach. The LSS projects each orthogonal distance d_{OR_i} into the corresponding *sine*(φ_i) distance, the distance from the intersection point of the dotted arrow line and the unite circle centered at the mean $(\bar{X}, \bar{Y})$ to the regression line. The LSS minimizes the sum of squared *sine* distances $\sum_i sin^2\varphi_i = \sum_i \frac{d_{OR_i}^2}{R_i^2}$, where φ_i is the acute angle formed by the fitted regression line and the line connecting the mean $(\bar{X}, \bar{Y})$ and the observation (X_i, Y_i), and R_i denotes the distance between the *i*th observation and the mean.

The intuitive geometric representation of the LSS approach is illustrated through Figure 3.5. Each observation (X_i, Y_i) is firstly projected, along the R_i directing from each observation to the mean $(\bar{X}, \bar{Y})$, onto the unit circle centered at the mean, and the orthogonal distance is then transformed into the corresponding *sine*(φ_i) distance from the projection point to the regression line, where φ_i is the acute angle formed by the fitted regression line and the line connecting the mean $(\bar{X}, \bar{Y})$ and the observation (X_i, Y_i). According to the angular regression concept, the line of best fit should be the one having as many small $\varphi_i{}'s$ as possible, and equivalently that is the line rendering the quantity of sum of $sin^2\varphi_i$ to be a minimum.

As one observes that the objective of the OR approach is to minimize $SS_{OR} = \sum_i d_{OR_i}^2$, while the LSS regression aims to minimize $SS_{LSS} = \sum_{i=1}^{n} \frac{d_{OR_i}^2}{R_i^2}$, the LSS can be regarded as the weighted OR with a weighting parameter $\frac{1}{R_i^2}$ imposed to each orthogonal residual. Consequently, the LSS as a robust weighted version of the OR not only accounts for measurement errors from both the dependent and independent variables but also down-weighs observations with large orthogonal residuals.

Inspired from this point of view, the generalized LSS (GLSS) is defined to minimize $SS_{GLSS-k} = \sum_{i=1}^{n} \frac{d_{OR_i}^2}{R_i^k}$ (k is any non-negative integer), where $d_{OR_i} = \frac{|Y_i - X_i\beta - \alpha|}{\sqrt{1+\beta'\beta}}$ is the orthogonal distance from each observation point to the regression plane, and $R_i = \sqrt{(X_i - \bar{X})(X_i - \bar{X})' + (Y_i - \bar{Y})^2}$ is the distance from each observation to the 'center of mass'. When k = 0 it is the objective function of the OR, which works well for data without outliers, and when k = 2 it is equivalent to the LSS. The specific case with k = 1 is so called GLSS-1 regression, which is a compromise of the OR and LSS approaches. Further investigations are needed to settle down a general good choice of *k* in balancing the robustness and efficiency of the GLSS-*k* estimator.

3.3 LSS in Higher Dimensions

LSS also has an extension to the high dimensional case. The problem is formed as we are fitting a regression model of a response *Y* in terms of the predictors X_1, X_2, …, X_p. Since the orthogonal distance from the *i*th observation (X_{1i}, X_{2i}, …, X_{pi}, Y_i) to the regression hyper-plane $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$ is $d_{OR_i} = \frac{\left|Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji}\right|}{\sqrt{1+\sum_{j=1}^{p} {\beta_j}^2}}$ and the distance from the *i*th observation to the data set mean is $R_i = \sqrt{\sum_{j=1}^{p} \left(X_{ji} - \bar{X}_j\right)^2 + (Y_i - \bar{Y})^2}$, the LSS regression will minimize

$$\sum_{i=1}^{n} \frac{{d_{OR_i}}^2}{R_i^2} = \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji})^2}{\left(1+\sum_{j=1}^{p} {\beta_j}^2\right)[\sum_{j=1}^{p} \left(X_{ji} - \bar{X}_j\right)^2 + (Y_i - \bar{Y})^2]}$$

From the insight of a close relationship between the OR approach and the PCA, we have shown that the LSS estimate is the eigenvector associated to the smallest eigenvalue of the robust weighted sample covariance matrix:

$$\tilde{S} = \begin{bmatrix} \tilde{S}_{X_1X_1} & \cdots & \tilde{S}_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \tilde{S}_{X_pX_1} & \cdots & \tilde{S}_{X_pX_p} \end{bmatrix}_{p\mathrm{x}p}$$

where $\tilde{S}_{X_jX_k} = \sum_{i=1}^{n} \frac{(X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k)}{{R_i}^2}$, *j*, *k* = 1, …, *p* (see Theorem 2 in the Appendix).

For instance in simple linear regression, assume the robust sample covariance matrix $\tilde{S}$ for the regression of two variables X and Y is

$$\tilde{S} = \begin{bmatrix} \tilde{S}_{XX} & \tilde{S}_{XY} \\ \tilde{S}_{XY} & \tilde{S}_{YY} \end{bmatrix} = \begin{bmatrix} \sum_i \frac{(X_i - \bar{X})^2}{R_i^2} & \sum_i \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{R_i^2} \\ \sum_i \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{R_i^2} & \sum_i \frac{(Y_i - \bar{Y})^2}{R_i^2} \end{bmatrix}$$

Depending on the robust sample covariance matrix $\tilde{S}$, the LSS estimator

$$\hat{\beta}_{LSS} = \frac{\tilde{S}_{YY} - \lambda \tilde{S}_{XX} + \sqrt{(\tilde{S}_{YY} - \lambda \tilde{S}_{XX})^2 + 4\lambda \tilde{S}_{XY}^2}}{2\tilde{S}_{XY}}$$

has exactly the same form as the OR estimator which is based on the usual sample covariance matrix S.

Furthermore, analogous to the computation of OR in higher dimensions, the more computationally efficient way to compute the LSS hyper-plane in high dimensional case is through the SVD of the robust weighted augmented matrix $[\tilde{X}\ \tilde{Y}]$:

$$[\tilde{X}\ \tilde{Y}] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p, \tilde{Y}] = \begin{bmatrix} \frac{X_{11} - \bar{X}_1}{R_1} & \cdots & \frac{X_{p1} - \bar{X}_p}{R_1} & \frac{Y_1 - \bar{Y}}{R_1} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{X_{1n} - \bar{X}_1}{R_n} & \cdots & \frac{X_{pn} - \bar{X}_p}{R_n} & \frac{Y_n - \bar{Y}}{R_n} \end{bmatrix}_{n\mathrm{x}(p+1)}$$

Define $[\tilde{U}][\tilde{\Sigma}][\tilde{V}]'$ to be the SVD of the augmented matrix $[\tilde{X}\ \tilde{Y}]$

$$[\tilde{X}\ \tilde{Y}] = [\tilde{U}_X\ \tilde{U}_Y] \begin{bmatrix} \tilde{\Sigma}_X & 0 \\ 0 & \tilde{\Sigma}_Y \end{bmatrix} \begin{bmatrix} \tilde{V}_{XX} & \tilde{V}_{XY} \\ \tilde{V}_{YX} & \tilde{V}_{YY} \end{bmatrix},$$

It is easy to have $\hat{\beta}_{LSS} = -\tilde{V}_{XY} \tilde{V}_{YY}^{-1}$ as the LSS estimator in higher dimensions.

Chapter 4

Robust Compound Regression

4.1 RCR Approach

Robust compound regression (RCR), as a natural extension and combination of the new LSS approach and the compound regression analysis method, is defined to minimize the weighted average of the sum of squared weighted vertical and horizontal distances as follows:

$$SS_\gamma = \gamma \sum_i \frac{(Y_i - \hat{Y}_i)^2}{R_i^2} + (1-\gamma) \sum_i \frac{(X_i - \hat{X}_i)^2}{R_i^2}$$

$$= \gamma \sum_i \frac{[Y_i - \alpha - \beta X_i]^2}{R_i^2} + (1-\gamma) \sum_i \frac{(X_i - \frac{Y_i - \alpha}{\beta})^2}{R_i^2} \qquad (0 \le \gamma \le 1)$$

Analogous to the compound regression situation, when γ = 1 the RCR model reduced to the robust LS (RLS) regression of *Y* on *X*, while γ = 0 is equivalent to the RLS regression of *X* on *Y*. Just as the compound regression (CR) contains the OR and GMR as special cases, the RCR contains the robust OR (i.e. the LSS) and robust GMR (RGMR) as special cases too.

Similar as the generalization from the LSS to the GLSS, the generalized RCR can be defined to minimize $SS_\gamma = \gamma \sum_i \frac{(Y_i - \hat{Y}_i)^2}{R_i^k} + (1-\gamma) \sum_i \frac{(X_i - \hat{X}_i)^2}{R_i^k}$, where k can be any nonnegative integer.

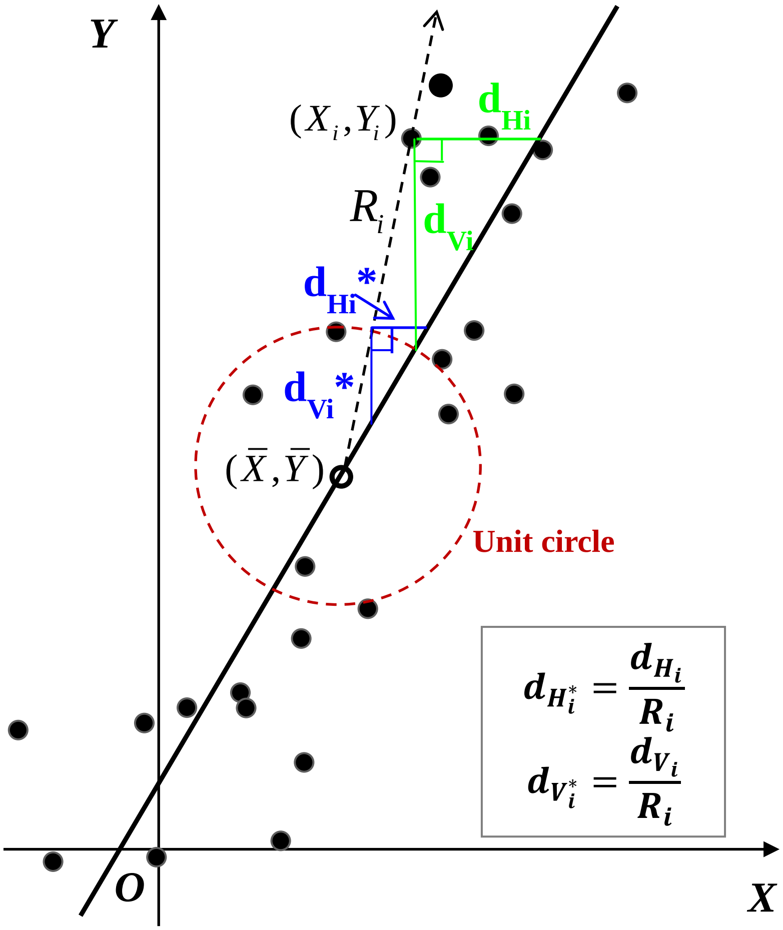


Figure 4.1 Geometric interpretation of the robust compound regression (RCR) approach. The RCR transforms the usual horizontal (d_{H_i}) and vertical residuals (d_{V_i}) into the corresponding transformed residuals $d_{H_i}^*$ and $d_{V_i}^*$ respectively, in the same way as the LSS transforms the usual orthogonal distance into the *sine* distance as described in Figure 3.5.

For each $\gamma \in [0, 1]$, we can obtain the least squares estimators of the regression parameters by solving $\frac{\partial SS_\gamma}{\partial \alpha} = 0$ and $\frac{\partial SS_\gamma}{\partial \beta} = 0$ simultaneously. Straight-forward derivations show that the least squares estimators of the regression coefficients, $\hat{\alpha}$ and $\widehat{\beta}$, would satisfy:

$$\alpha = \bar{Y} - \beta\bar{X} \qquad \text{(a)}$$

$$\gamma\widetilde{S_{XX}}\beta^4 - \gamma\widetilde{S_{XY}}\beta^3 + (1-\gamma)\widetilde{S_{XY}}\beta - (1-\gamma)\widetilde{S_{YY}} = 0 \qquad \text{(b)}$$

where $\tilde{S}_{XX} = \sum_i \frac{(X_i - \bar{X})^2}{R_i^2}$, $\tilde{S}_{XY} = \sum_i \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{R_i^2}$, $\tilde{S}_{YY} = \sum_i \frac{(Y_i - \bar{Y})^2}{R_i^2}$. From Equation (b), for each RCR slope estimator $\hat{\beta}$, we have the corresponding

$$\gamma = \frac{\widetilde{S_{YY}} - \hat{\beta}\widetilde{S_{XY}}}{\widetilde{S_{YY}} - \hat{\beta}\widetilde{S_{XY}} + \hat{\beta}^4\widetilde{S_{XX}} - \hat{\beta}^3\widetilde{S_{XY}}} \qquad (*)$$

If one approach is a special case of the RCR, there should be a corresponding $\gamma \in [0, 1]$ to the estimated $\hat{\beta}$ in Equation (*).

Similar to the generalization of LSS, we have the generalized RCR in the following form:

$$SS_\gamma = \gamma \sum_i \frac{(Y_i - \hat{Y}_i)^2}{R_i^k} + (1-\gamma) \sum_i \frac{(X_i - \hat{X}_i)^2}{R_i^k}$$

where the power *k* can be any nonnegative integer. When *k*=0, it corresponds to sum of squares function of compound regression, while it corresponds to the ordinary robust compound regression when *k*=2.

The multivariate RCR takes account of all the prediction errors with different weight and obtains the estimators of the regression parameters by minimizing the following sum of squares function

$$SS_\gamma = \gamma_0 \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{R_i^2} + \gamma_1 \sum_{i=1}^{n} \frac{(X_{1i} - \hat{X}_{1i})^2}{R_i^2} + \cdots + \gamma_p \sum_{i=1}^{n} \frac{(X_{pi} - \hat{X}_{pi})^2}{R_i^2}$$

subject to $\sum_{j=1}^{p} \gamma_i = 1$, where p is the number of predictor variables. For the higher dimension case, the estimation for RCR can be carried out in the same way. The above objective function can be simplified to

$$SS_\gamma = \gamma_0 \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{R_i^2} + \sum_{j=1}^{p} \frac{\gamma_j}{\beta_j^2} \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{R_i^2}$$

$$= \left(\gamma_0 + \sum_{j=1}^{p} \frac{\gamma_j}{\beta_j^2}\right) \sum_{i=1}^{n} \frac{\left[Y_i - \bar{Y} - \sum_{j=1}^{p} \beta_j \left(X_{ji} - \bar{X}_j\right)\right]^2}{R_i^2}$$

$$= \left(\gamma_0 + \sum_{j=1}^{p} \frac{\gamma_j}{\beta_j^2}\right) (\tilde{S}_{YY} + \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j \beta_k \tilde{S}_{X_j X_k} - 2 \sum_{j=1}^{p} \beta_j \tilde{S}_{X_j Y})$$

Estimators of the RCR regression coefficients can be obtained by solving the system of equations $\frac{\partial SS_{\tilde{\gamma}}}{\partial \beta_j} = 0,$ j = 1, 2, …, p simultaneously, where for any l in j = 1, 2, …, p, and we have:

$$\frac{\partial SS_\gamma}{\partial \beta_l} = -\frac{2\gamma_l}{\beta_l^3} \left(\tilde{S}_{YY} + \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j \beta_k \tilde{S}_{X_j X_k} - 2 \sum_{j=1}^{p} \beta_j \tilde{S}_{X_j Y} \right)$$

$$+ \left(\gamma_0 + \sum\nolimits_{j=1}^{p} \frac{\gamma_j}{\beta_j^2}\right) \left(2 \sum_{\substack{j=1 \\ j \neq l}}^{p} \beta_j \tilde{S}_{X_i X_j} + 2\beta_l \tilde{S}_{X_i X_i} - 2\tilde{S}_{X_i Y} \right) = 0$$

The RCR includes the robust counterpart of the LS, OR, and GMR as special cases. For instance, the RLS regression of Y on X is defined to minimize $SS_{\gamma=1} = \sum_i \frac{(Y_i - \hat{Y}_i)^2}{R_i^2}$, and plugging its corresponding slope estimate $\hat{\beta}_{RLS\_Y} = \frac{\tilde{S}_{XY}}{\tilde{S}_{XX}}$ into (*) will yield $\gamma_{RLS\_Y} = 1$. Similarly, the RLS regression of X on Y will minimize $SS_{\gamma=0} = \sum_i \frac{(X_i - \hat{X}_i)^2}{R_i^2}$, and plugging the slope estimate

$\hat{\beta}_{RLS\_X} = \frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}$ into Equation (*) gives us $\gamma_{RLS\_X} = 0$. Furthermore, it has been shown that there is a monotonic relationship between γ and $\hat{\beta}$ (see Theorem 3 in the Appendix).

As can be seen, the robust OR is the alias of the LSS, which is defined to minimize $\sum_i \frac{d_{OR_i}^2}{R_i^2}$. From Equation (*) we know that the corresponding compound parameter for the LSS is

$$\gamma_{LSS} = \frac{\widetilde{S_{YY}} - \hat{\beta}_{LSS}\widetilde{S_{XY}}}{\widetilde{S_{YY}} - \hat{\beta}_{LSS}\widetilde{S_{XY}} + \hat{\beta}_{LSS}{}^4\widetilde{S_{XX}} - \hat{\beta}_{LSS}{}^3\widetilde{S_{XY}}}$$

By Cauchy-Schwarz inequality, we know that $\tilde{S}_{XY}^2 \leq \tilde{S}_{XX}\tilde{S}_{YY}$. Based on this fact, straight-forward derivations show that $\hat{\beta}_{\text{LSS}}$ is always bounded by $\hat{\beta}_{\text{RLS}\_Y}$ and $\hat{\beta}_{\text{RLS}\_X}$, that is when $\tilde{S}_{XY} \geq 0$ we have $\frac{\tilde{S}_{XY}}{\tilde{S}_{XX}} \leq \hat{\beta}_{LSS} = \frac{\tilde{S}_{YY} - \lambda\tilde{S}_{XX} + \sqrt{(\tilde{S}_{YY} - \lambda\tilde{S}_{XX})^2 + 4\lambda\tilde{S}_{XY}^2}}{2\tilde{S}_{XY}} \leq \frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}$. Otherwise, when $\tilde{S}_{XY} < 0$ we have $\frac{\tilde{S}_{XY}}{\tilde{S}_{XX}} \geq \hat{\beta}_{LSS} \geq \frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}$. Meanwhile, because of the monotonic relationship between γ and $\hat{\beta}$, we must have $0 \leq \gamma_{\text{LSS}} \leq 1$, which indicates the LSS belongs to the RCR framework.

The robust GMR (RGMR) for simple linear regression is defined to minimize $SS_{RGMR} = -\frac{1}{2} sign(\beta) \sum_i \frac{(X_i - \hat{X}_i)(Y_i - \hat{Y}_i)}{R_i^2}$. As we know that the RGMR line must pass through the mean of all data points $(\overline{\text{X}}, \overline{\text{Y}})$, hence the sum of squares can be restated as

$$SS_{RGMR} = -\frac{1}{2} sign(\beta) \sum_i \frac{[(X_i - \bar{X}) - \frac{1}{\beta}(Y_i - \bar{Y})][(Y_i - \bar{Y}) - \beta(X_i - \bar{X})]}{R_i^2}$$

By solving $\frac{\partial SS_{RGMR}}{\partial \beta} = 0$, we obtain the RGMR slope estimate $\hat{\beta}_{RGMR} = sign(\tilde{S}_{XY})\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}}$. From Equation (*), it is easy to find the corresponding γ_{RGMR} as well. Similarly, by Cauchy-Schwarz inequality, we can easily proof that $\frac{\tilde{S}_{XY}}{\tilde{S}_{XX}} \le \hat{\beta}_{RGMR} = \sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}} \le \frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}$ if $\tilde{S}_{XY} \ge 0$, and $\frac{\tilde{S}_{XY}}{\tilde{S}_{XX}} \ge \hat{\beta}_{RGMR} = -\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}} \ge \frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}$ if $\tilde{S}_{XY} < 0$. Therefore, because of the monotonic relationship between γ and $\hat{\beta}$, we know that $0 \le \gamma_{\mathrm{RGMR}} \le 1$, which implies that the RGMR also belongs to the RCR.

For the higher dimension case, we can prove that both the LSS and the RGMR are special cases of the RCR (see Theorem 2 for the proof on LSS, and Theorem 4 for the proof on RGMR in the Appendix).

4.2 RCR Efficiency and Constrained RCR

The RCR efficiency with respect to a specific regression variable is defined as the ratio of the minimized to the observed sum of robust weighted squared residuals along the corresponding coordinate direction.

$$e_Y = \frac{\min \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{R_i^2}}{\sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{R_i^2}} = \frac{\sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i^{robustLS\_Y})^2}{R_i^2}}{\sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{R_i^2}}$$

$$e_{X_j} = \frac{\min \sum_{i=1}^{n} \frac{(X_{ji} - \hat{X}_{ji})^2}{R_{ji}^2}}{\sum_{i=1}^{n} \frac{(X_{ji} - \hat{X}_{ji})^2}{R_{ji}^2}} = \frac{\sum_{i=1}^{n} \frac{(X_{ji} - \hat{X}_{ji}^{robustLS\_X_j})^2}{R_{ji}^2}}{\sum_{i=1}^{n} \frac{(X_{ji} - \hat{X}_{ji})^2}{R_{ji}^2}}, \quad j = 1, ..., p$$

From the estimates of each regression approach, we can calculate the corresponding regression efficiencies with respect to each regression variable. A general goodness-of-fit criterion we suggest is that, when all the variables involved are considered equally important, the higher the sum of regression efficiencies $e_Y + \sum_{j=1}^{p} e_{X_j}$, the better the fitted model is.

In terms of the RCR efficiency, for a given c∈ [0,1], the constrained RCR will maximize e$_Y$ subject to e$_X \geq c$ if we want the prediction accuracy of X to be no less than a critical threshold. Symmetrically, one may set the desired prediction accuracy of Y to be at least c and obtain the best regression line that will maximize e$_X$ subject to e$_Y \geq c$.

In simple linear regression situation, it has been shown that the RGMR will always yield the equal e$_Y$ and e$_X$, and furthermore the maximum sum of regression efficiencies e$_Y$ + e$_X$ (see Theorem 5 in the Appendix).

4.3 High Breakdown Outlier Diagnostics

There are generally two ways to deal with outliers in regression analysis. For the priori outlier treatment, one may down-weigh the potential influence of outliers by directly applying the robust regression techniques; one could transform the data by using Box-Cox transformation, delta method etc. to achieve linearity, normality and homogeneity assumptions; and one can simply get rid of what he/she believes as the ‘bad’ points, and compare the magnitude of the change of parameter estimates before and after the removal.

On the other hand, for the posterior outlier diagnostics, one point we ought to emphasize here is that the diagnostics through LS residuals is non-robust as the LS always tries to avoid large residuals which may even lead to the misidentification of outliers. Another class of diagnostics is based on the principle of deleting single/multiple cases at a time. Then the difference between the regression coefficient estimates with or without the *i*th case gives the extent to which the presence of the *i*th case affects the regression fit. However, when sample size is large the computations involved are infeasible as there are so many combinations of subsets to be considered.

The high breakdown outlier diagnostics by utilizing the residuals obtained from robust fits is powerful to detect all the outliers present in the data. Our new robust approaches, of course, can be used to formulate two new high breakdown diagnostics. Since the LSS gives a robust fit based on orthogonal residuals $Resid_{OR_i} = \frac{Y_i - \hat{\alpha} - \sum_{j=1}^{p} \hat{\beta}_j X_{ji}}{\sqrt{1+\sum_{j=1}^{p} \hat{\beta}_j{}^2}}$, the outlier diagnostics can be carried out by examining the plot of standardized orthogonal residuals versus the fitted response values of Y_{i}. Moreover, the RGMR diagnostics is based on inspecting the standardized GMR residuals from each observation to the RGMR regression hyper-plane

$$Resid_{GMR_i} = \frac{Y_i - \hat{\alpha} - \sum_{j=1}^{p} \hat{\beta}_j X_{ji}}{(\prod_{j=1}^{p} \hat{\beta}_j)^{\frac{1}{p+1}}}$$

The standardized residual plot lends us a visual screening tool to detect all observations for which the diagnostic exceeds its cutoff. The 95% cut-off $\pm\sqrt{\chi^2_{0.975}(1)^{-1}} = \pm 2.2414$ is usually recommended for the diagnostics of standardized residuals (Rousseeuw and Leroy 1987).

Chapter 5

Simulation Studies

A shortcoming of real data analyses using the EIV models is that it is hard to 'prove' which regression method yields the best fit without knowing the truth. A popular alternative is the Monte Carlo simulations, because at that time one knows the true parameters for the data generated.

The estimated coefficients from a specific regression method may be incorrect if its underlying model assumptions are not met. Two factors in particular that may result in incorrect estimates are: measurement errors of the independent variable and presence of outliers in the data analysis.

To show the property of each method and the advantage of our new approaches, we examine the performance of several estimators on the simulated dataset simultaneously. Since all the analysis methods except the robust LMS has the assumption that the lines will pass through the 'center of mass' $(\bar{X}, \bar{Y})$, the fitted regression lines can be expressed in point-slope form as $Y - \bar{Y} = \hat{\beta}(X - \bar{X})$, we will thus focus solely on the estimation of the slope parameter β.

Simulation studies are conducted by setting up a simple linear structural EIV model with a true linear relationship $\eta = 1 + \xi$ defined on two latent variables η and ξ, and one only observes $Y = \eta + \varepsilon$ and $X = \xi + \delta$ as the observations, where the underlying $\xi \sim N(0, 100)$, the random errors $\delta \sim N(0, \sigma_\delta^2)$, and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. In our simulations, we perform 10,000 replications of a sample size of 200 for all data-generation situations.

5.1 Comparison of the New Approaches to Existing Approaches

In this section, we dedicate the first three cases to examine the performance of each analysis method on uncontaminated data with various ratios of the error variances, and we expect our new approaches should perform fairly well compared to the optimal one in each situation. Then in the last two cases we test the robustness of each estimator when dealing with outliers. Based on the above settings, the Monte Carlo experiments are designed as follows:

Table 5.1 λ ratio settings for Monte Carlo experiments in Section 5.1

Settings	(a)	(b)	(c)
X noise-to-signal $\sigma^2_\delta/\sigma^2_\xi$	20%	10%	5%
Y noise-to-signal $\sigma^2_\varepsilon/\sigma^2_\eta$	(0,5,10,15,20)%	(0,5,10,15,20)%	(5,10,15,20, 25)%
ratio $\lambda = \sigma^2_\varepsilon/\sigma^2_\eta$	$0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$	$0, \frac{1}{2}, 1, \frac{3}{2}, 2$	1, 2, 3, 4, 5

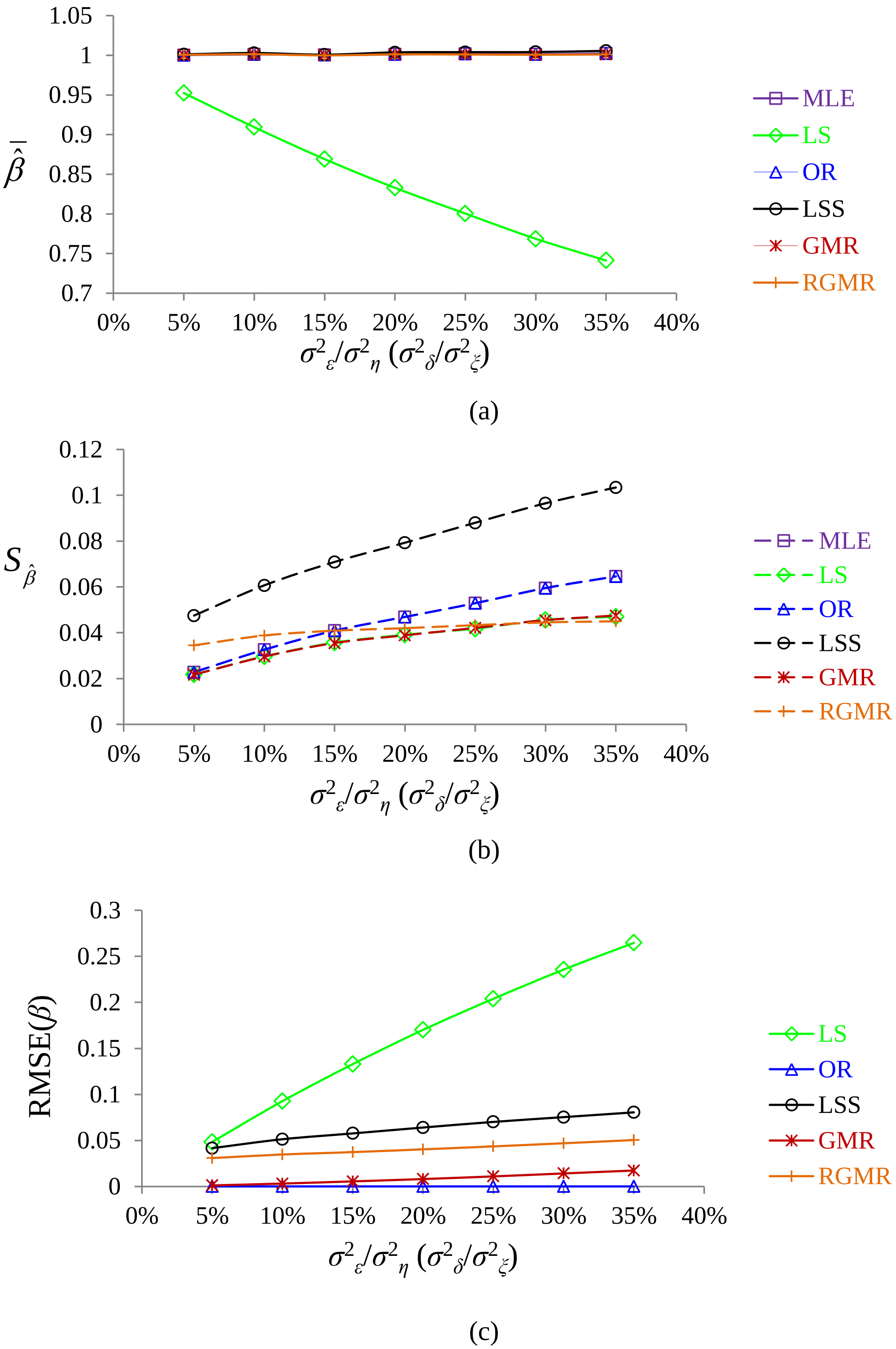
We will compare five different estimates from the LS, OR, GMR, and the LSS and RGMR to the MLE when λ is assumed known for the simulations purposes. (Of course, in real data analysis situation, λ is usually unknown which renders the MLE method obsolete.) The efficiency of each method will be summarized in terms of the mean $\bar{\hat{\beta}} = \frac{1}{r}\sum_{i=1}^{r}\hat{\beta}_i$, the standard error $S_{\hat{\beta}} = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(\hat{\beta}_i - \bar{\hat{\beta}})^2}$, and the root mean squared error (RMSE) based on the MLE $\mathrm{RMSE}(\hat{\beta}) = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\hat{\beta}_i - \hat{\beta}_{\mathrm{MLE}_i})^2}$ of the slope estimate $\hat{\beta}$ over the r = 10,000 runs.

Case 1: Test the optimality of OR

To verify the optimality of the OR approach when the two error variances are equal, we tune the noise-to-signal ratios over several critical values from small to large with the constraint λ = 1. The mean $\bar{\hat{\beta}}$, standard error $S_{\hat{\beta}}$, and $\mathrm{RMSE}(\hat{\beta})$ over the 10,000 runs from five different regression approaches are visually summarized through Figure 5.1. We can clearly see that, all estimators except the LS seem unbiased, and meanwhile the small standard error of the LS estimate double confirms its inconsistency when dealing with the EIV problems. In terms of RMSE, the OR as expected is the most efficient one as it is the MLE in this situation. That is, the OR is optimal when the two error variances are equal.

Table 5.2 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.1

Noise-to-signal level	5%	10%	15%	20%	25%	30%	35%
Mean(S_{YY}/S_{XX})	1	1	1	1	1	1	1
Std(S_{YY}/S_{XX})	0.04	0.06	0.07	0.08	0.09	0.09	0.10



(a)

(b)

(c)

Figure 5.1 Comparison of five different estimators with the MLE when λ = 1, β = 1, n = 200 with 10,000 replications. We clearly see that from (a) bias: (LS > LSS > GMR > RGMR > OR), all methods except the LS seem unbiased, and from (b) standard error: (LSS > OR > RGMR >

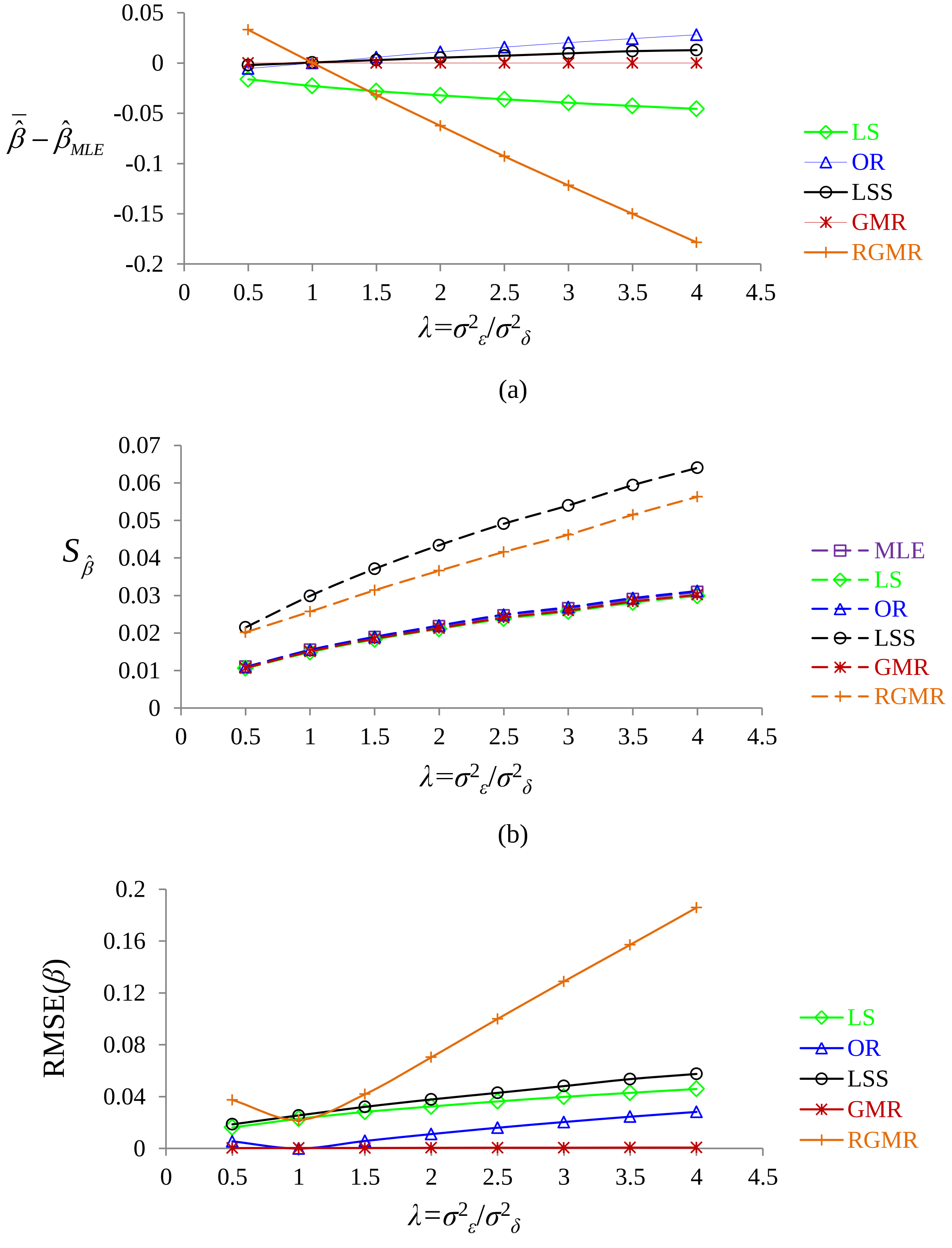
GMR > LS), the small standard error of the LS estimates double confirms its inconsistency in EIV models. In terms of (c) RMSE of $\hat{\beta}$: (LS > LSS > RGMR > GMR > OR), the OR as expected is the most efficient one as it is the MLE in this situation, that is the OR is optimal when the two error variances are equal.

Case 2 : Test the optimality of GMR

To verify the optimality of the GMR approach when the noises of the data come from the random errors only, the functional EIV model is utilized where ξ is not random. In this case, we first generate one set of $\xi \sim U(0, 100)$, and then noise-to-signal settings (b) from Table 5.1 is used to tune the λ ratio from small to large. Of note, the true slope is set to be $\sqrt{\lambda}$ in order to meet our assumption that $\lambda = \frac{S_{YY}}{S_{XX}}$. The bias $\bar{\hat{\beta}} - \hat{\beta}_{MLE}$, standard error $S_{\hat{\beta}}$, and RMSE($\hat{\beta}$) over the 10,000 runs from five different approaches are summarized through Figure 5.2. We can see that, the GMR is unbiased with a small standard error which indicates its consistency. In terms of RMSE, the GMR as expected is the most efficient one as theoretically it is the MLE in this situation, and the table summarizing the true λ ratio and the corresponding sample variance ratio S_{YY}/S_{XX} is as follows.

Table 5.3 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.2

λ	0.5	1	1.5	2	2.5	3	3.5	4
Mean(S_{YY}/S_{XX})	0.5	1	1.5	2	2.5	3	3.5	4
Std(S_{YY}/S_{XX})	0.02	0.03	0.05	0.06	0.08	0.09	0.11	0.12



(c)

Figure 5.2 Comparison of five different estimators with the MLE when λ = S_{YY}/S_{XX} (functional approach ξ is not random), $\beta = \sqrt{\lambda}$, n = 200 with 10,000 replications. We see that from (a) bias: (RGMR > LS > OR > LSS > GMR), the GRM is the most unbiased while the RGMR is highly

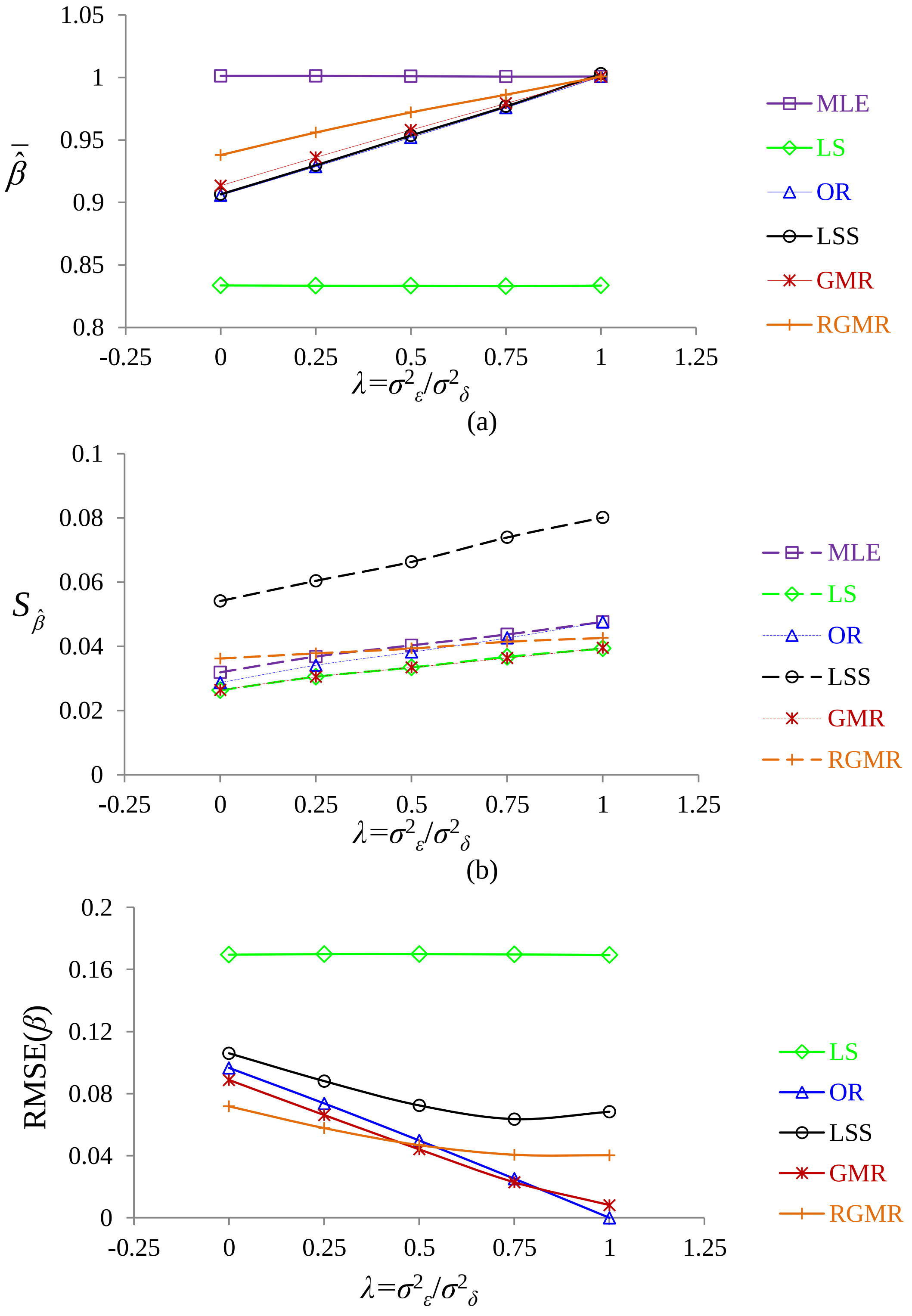
biased in this case, and from (b) standard error: (LSS > RGMR > OR > GMR > LS), the small standard error of the GMR estimator meanwhile indicates its consistency. In terms of (c) RMSE of $\hat{\beta}$: (RGMR > LSS > LS > OR > GMR), the GMR as expected is the most efficient one as theoretically it is the MLE in this situation, that is the GMR is optimal when the noises are from random errors only.

Case 3: Performance on uncontaminated EIV data

To examine the performance of each method in more general situations, we conduct simulations under the structural EIV model approach with various λ ratios of the error variances. Refer to the noise-to-signal Table 5.1, settings (a) for simulations in Figure 5.3, (b) for Figure 5.4, and (c) for Figure 5.5 are used respectively. To sum up these three figures, we notice that the LS estimates are always highly biased and inconsistent in EIV problems, while all other approaches seem unbiased when the λ ratio equals to 1. Furthermore in terms of RMSE, we view that the GMR is generally the optimal one that is good for a wide range of λ in EIV models, and the OR is near optimal when the λ ratio of the error variances is around the range of (0.5, 3). In addition, the RGMR always perform better than the LSS for uncontaminated data. Generally speaking, the results for simulation case 3 suggest one to preferably choose the traditional EIV model estimation methods rather than the robust methods when there are no outliers.

Table 5.4 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.3

λ	0	0.25	0.5	0.75	1
Mean(S_{YY}/S_{XX})	0.83	0.88	0.92	0.96	1
Std(S_{YY}/S_{XX})	0.05	0.06	0.06	0.07	0.08



(a)

(b)

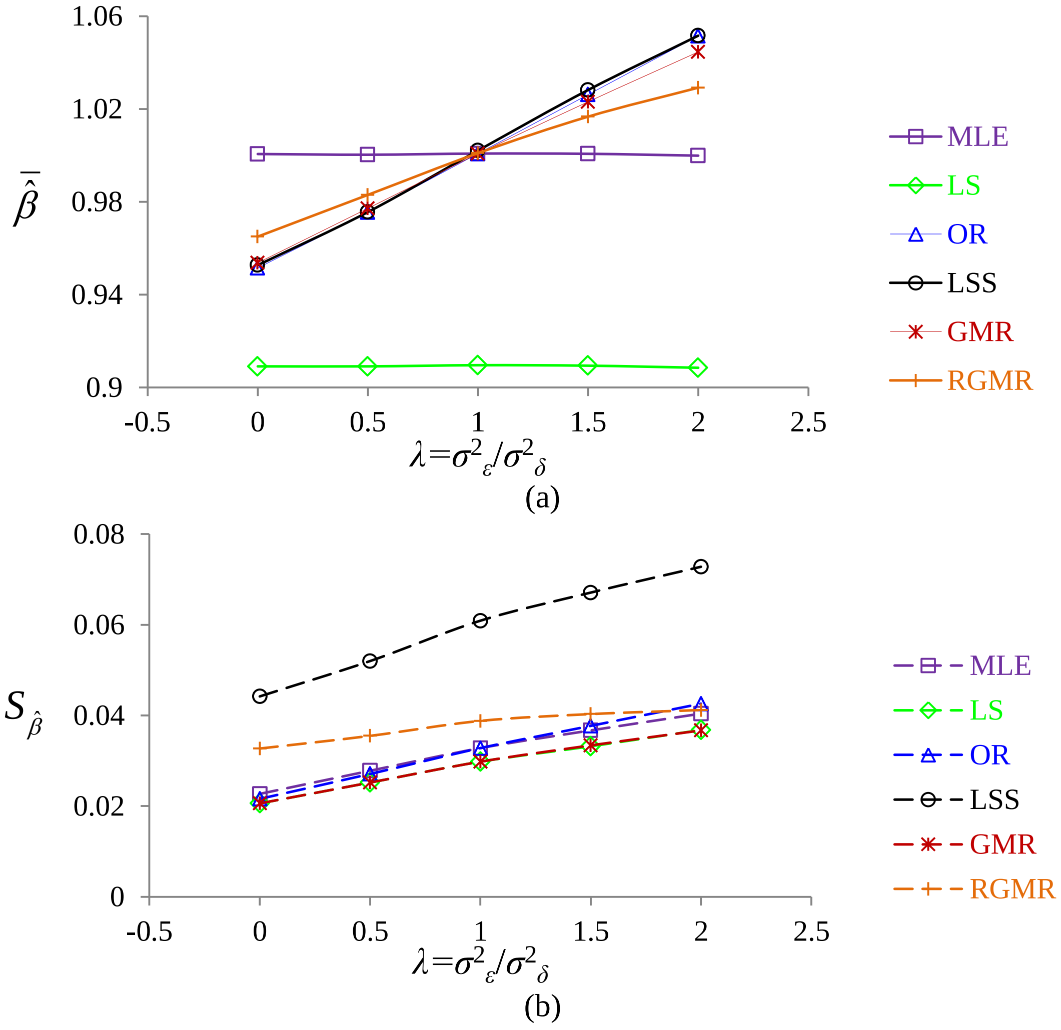
(c)

Figure 5.3 Comparison of five different estimators with the MLE when $0 \le \lambda \le 1$ is known, β = 1, n = 200 with 10,000 replications. We see that from (a) bias: (LS > OR > LSS > GMR > RGMR), the RGRM is relatively less biased compared to others while the LS is highly biased in

this case, and from (b) standard error: (LSS > RGMR > OR > LS > GMR), it seems one disadvantage of the LSS estimator is its standard error higher than the others. In terms of (c) RMSE of $\hat{\beta}$: (LS > LSS > RGMR > OR > GMR), the GMR in general is the optimal one in this simulation case.

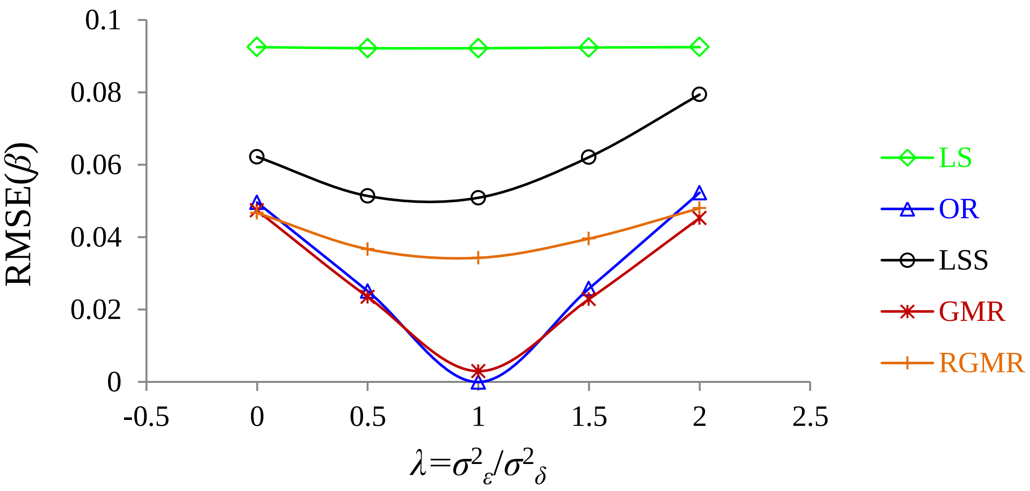
Table 5.5 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.4

λ	0	0.5	1	1.5	2
Mean(S_{YY}/S_{XX})	0.91	0.96	1	1.05	1.09
Std(S_{YY}/S_{XX})	0.04	0.05	0.06	0.07	0.08



(a)

(b)

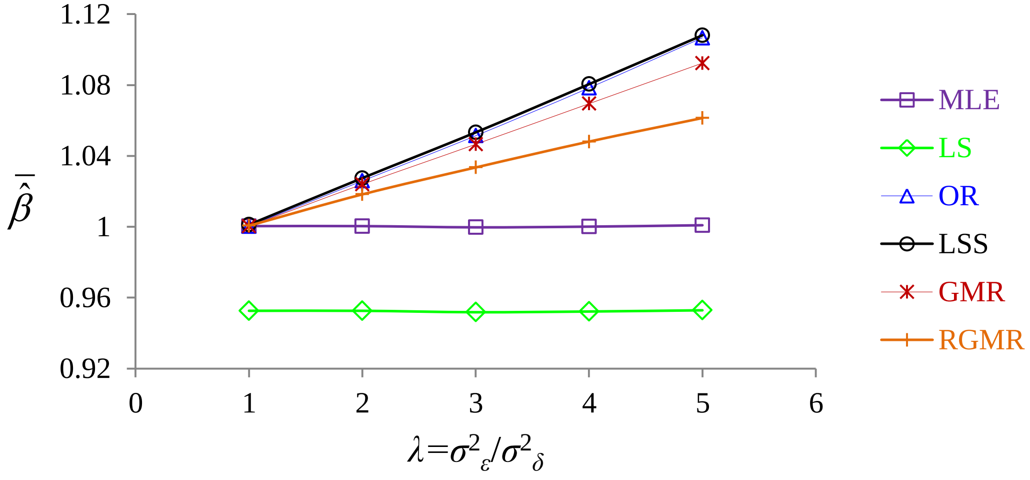


(c)

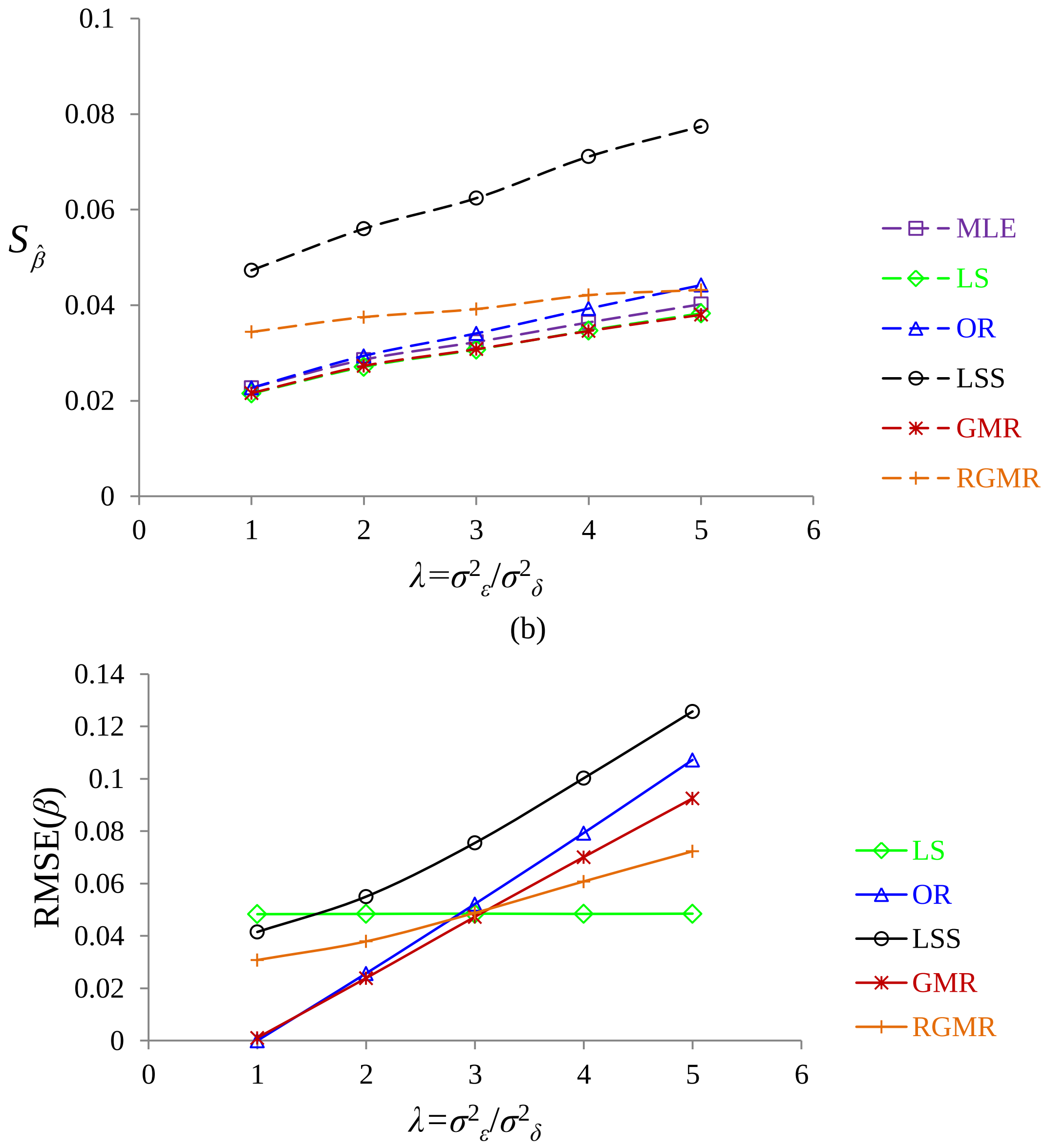
Figure 5.4 Comparison of five different estimators with the MLE when $0 \le \lambda \le 2$ is known, $\beta = 1$, $n = 200$ with 10000 replicates. We see that from (a) bias: (LS > OR > LSS > GMR > RGMR), (b) standard error: (LSS > RGMR > OR > LS > GMR), and (c) RMSE of $\hat{\beta}$: (LS > LSS > RGMR > OR > GMR), we can draw the similar conclusion as from Figure 5.3 that the GMR in general is the optimal one in this simulation case.

Table 5.6 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.5

λ	1	2	3	4	5
Mean(S_{YY}/S_{XX})	1.00	1.05	1.10	1.15	1.19
Std(S_{YY}/S_{XX})	0.04	0.05	0.06	0.07	0.08



(a)



(b)

(c)

Figure 5.5 Comparison of five different estimators with the MLE when $1 \leq \lambda \leq 5$ is known, $\beta = 1$, $n = 200$ with 10,000 replications. We see that from (a) bias: (LSS > OR > LS > GMR > RGMR), the RGRM relatively less biased compared to others, and from (b) standard error: (LSS > RGMR > OR > LS > GMR), it seems one disadvantage of the LSS estimator is its standard error higher than the others. In terms of (c) RMSE of $\hat{\beta}$: (LSS > OR > RGMR > LS > GMR), again the GMR is generally the optimal one in this simulation case.

As we have developed two new robust estimation approaches, it is more important to compare the robustness of different estimators in the presence of outliers. Although the well-known LMS robust estimator is specifically designed for the ordinary regressions, we ought to compare its performance with our new approaches for regressions with errors in variables. Due to the severe computational complexity of the LMS, we hereby only simulate 1000 replications.

There are basically two types of outlier contaminations in linear regression analysis. One is the outliers in the response variable direction, and the other is the leverage points outlying from the predictor variables domain. From the simulation point of view, one can view the random errors of the outlier contaminated data as data coming from a mixture of the assumed random error distributions. In terms of distribution functions this can be written as this mixture normal form

$$N_{\text{data}} = \alpha\, N_{\text{assumed}} + (1 - \alpha)\, N_{\text{contamination}}, \quad 0 < \alpha << 1$$

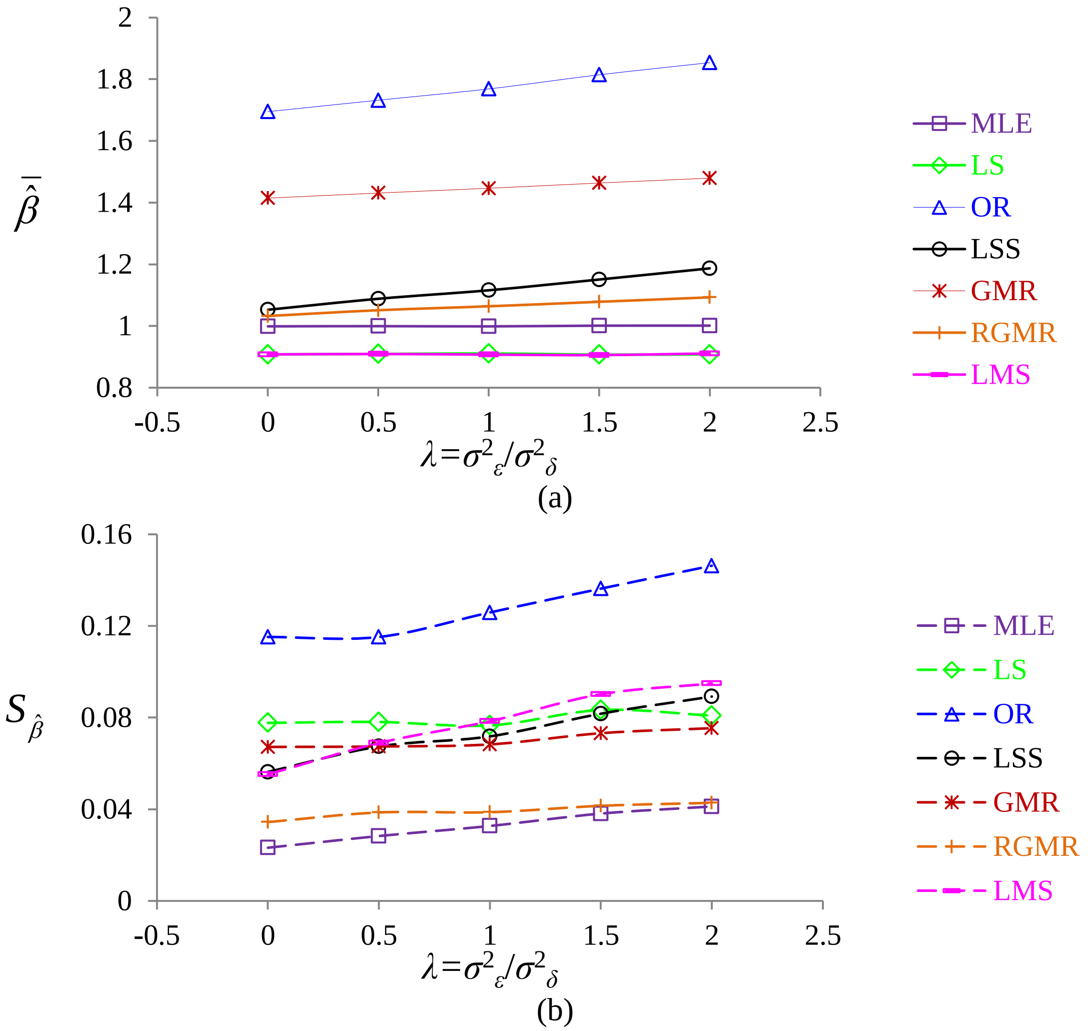
Case 4: Performance on EIV data with outliers in *Y* direction

We replace 5% of the ‘good’ points with outliers having contaminated $\varepsilon_{\text{c}} \sim N(50, \sigma^2_{\varepsilon})$ in the Y direction, and the noise-to-signal settings (b) in Table 5.1 is incorporated here. The mean $\bar{\hat{\beta}}$, standard error $S_{\hat{\beta}}$, and RMSE($\hat{\beta}$) over the 1000 runs from six different approaches are summarized through Figure 5.6. Of note, the MLE used for the calculation of RMSE of each method is obtained based on the uncontaminated part of data.

The results show that both the OR and the GMR estimates are highly biased and badly influenced by the outliers, but the LS estimator performs much better than them. From my humble opinion, the reason for the unexpectedly fair performance of the non-robust LS estimator is that, under this simulation experiment, the outliers with high Y values happen to be balanced with the usually attenuated LS slope estimates. The RGMR not only has the smallest bias but the smallest standard error among all the estimators concerned here. In terms of RMSE, the RGMR is the most robust in this simulation.

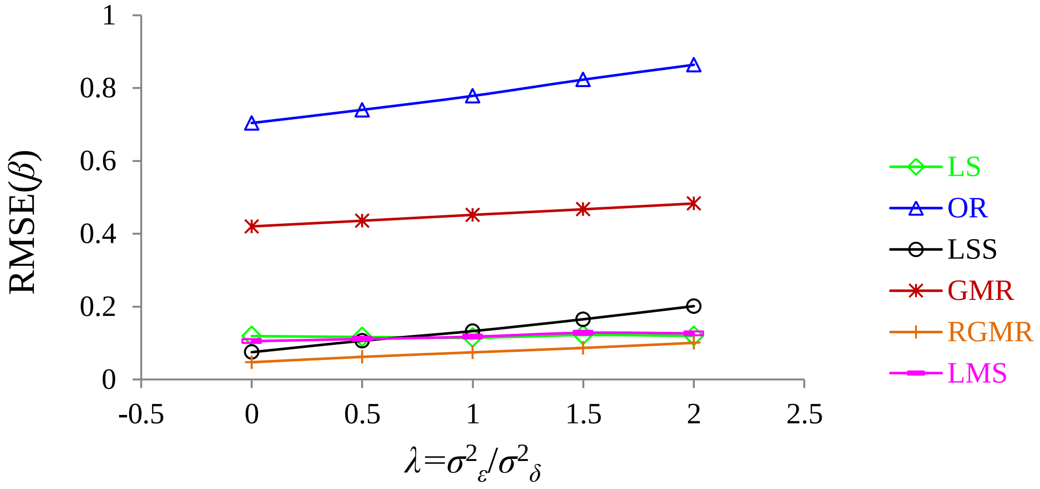
Table 5.7 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.6

λ	0	0.5	1	1.5	2
Mean(S_{YY}/S_{XX})	2.01	2.05	2.10	2.15	2.18
Std(S_{YY}/S_{XX})	0.19	0.20	0.20	0.22	0.22



(a)

(b)



(c)

Figure 5.6 Comparison of six different estimators with the MLE when $\lambda = \sigma^2_\varepsilon/\sigma^2_\delta$ is known, on contaminated data with 5% outliers $\varepsilon_c \sim N(50, \sigma^2_\varepsilon)$: $\beta = 1$, $n = 200$ with 10,000 replications. We can see that from (a) bias: (OR > GMR > LSS > LMS > LS > RGMR), the OR, GMR estimators are highly biased but the LS estimator performs much better than them, and from (b) standard error: (OR > LS > LMS > LSS > GMR > RGMR), the RGMR has the smallest bias and standard error among all the estimators concerned here. In terms of RMSE: (OR > GMR > LSS > LS > LMS > RGMR), hence the RGMR is the most robust estimator.

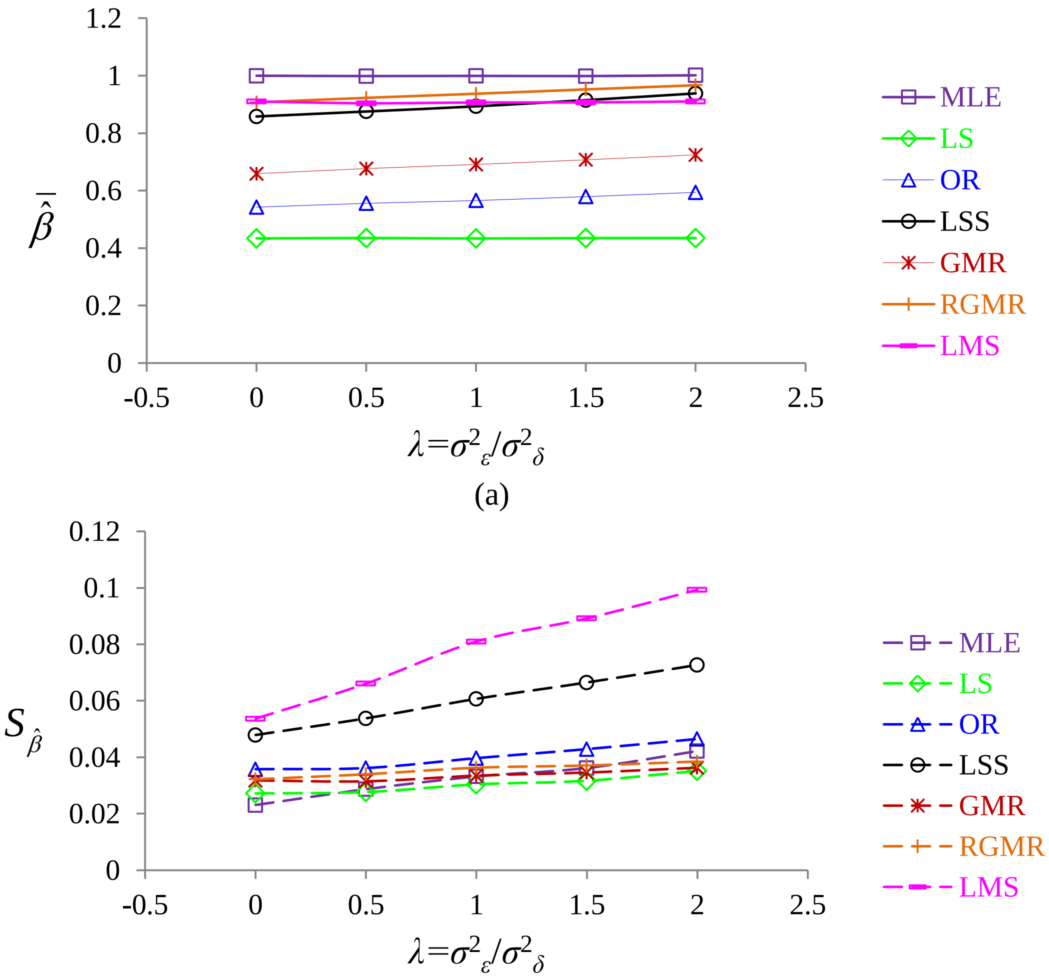
Case 5: Performance on EIV data with leverage points in *X* direction

Similarly, we replace 5% of the 'good' points with leverage points having contaminated $\delta_c \sim N(50, \sigma^2_\delta)$ in the X direction, and the noise-to-signal settings (b) in Table 5.1 is again utilized. The mean $\bar{\hat{\beta}}$, standard error $S_{\hat{\beta}}$, and RMSE($\hat{\beta}$) over the 1000 runs from six different approaches are summarized through Figure 5.7.

The RGMR is still the most robust, and the near optimal LSS outperforms the classic robust LMS approach, which implies that the LSS performs particularly well in the presence of leverage points. By contrast, the OR and GMR estimators are badly influenced by the outlier, and not to mention the worst estimation from the LS approach.

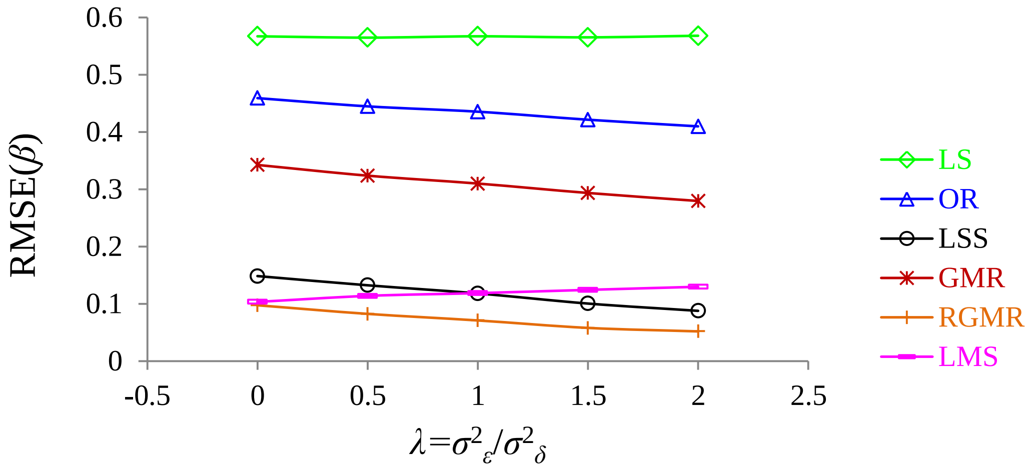
Table 5.8 Cross reference of the true λ and the sample S_{YY}/S_{XX} for Figure 5.7

λ	0	0.5	1	1.5	2
Mean(S_{YY}/S_{XX})	0.44	0.46	0.48	0.50	0.52
Std(S_{YY}/S_{XX})	0.04	0.05	0.05	0.05	0.06



(a)

(b)



(c)

Figure 5.7 Comparison of six different estimators with the MLE when $\lambda = \sigma^2_\varepsilon/\sigma^2_\delta$ is known, on contaminated data with 5% leverage points $\delta_c \sim N(50, \sigma^2_\delta)$: $\beta = 1$, $n = 200$ with 10,000 replications. We view that from (a) bias: (LS > OR > GMR > LSS > LMS > RGMR), the traditional nonrobust LS, OR, GMR estimators are highly biased and badly influenced by the outliers, from (b) standard error: (LMS > LSS > OR > RGMR > GMR > LS), the RGMR has the smallest standard error among all the robust estimators concerned here. In terms of RMSE: (LS > OR > GMR > LMS > LSS > RGMR), it confirms that the RGMR is the most robust estimator in this simulation.

5.2 Calibration of RCR through Efficiency Plot

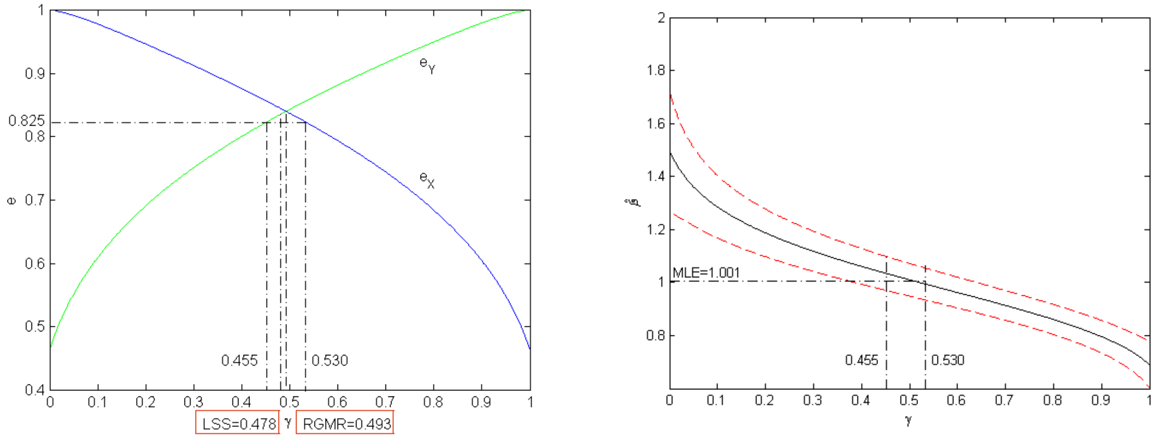
In this section, we did some simulation study to test how well our model estimates the underlying true model. The data are generated in the same way, using the structural EIV model assumptions, as described at the beginning of this chapter except we only have one single dataset in each of the following simulation. In addition, since there are so many scenarios of combinations of λ ratios and different kinds of contaminations, for the purpose of a clear illustration, we will fix $\lambda = \sigma^2_\varepsilon/\sigma^2_\delta = 1$ with a moderate ($\sigma^2_\varepsilon/\sigma^2_\eta = 10\%$) noise-to-signal level of Y and a moderate ($\sigma^2_\delta/\sigma^2_\xi = 10\%$) noise-to-signal level of X.

Case 1: Performance on uncontaminated EIV data

Model fitting:

We fit the data by our RCR approach and we have the following efficiency plot (Horizontal-axis: the compound parameter γ, vertical-axis: RCR efficiency). In this simulated dataset, we eyeball that the γ range from 0.45 to 0.55 is presumably the optimal solution we desire (in this situation, both e_Y and e_X is above 0.825). From the efficiency plot, we can get the rough idea of obtaining the corresponding γ to make sure the efficiency for both variables would be enough high. The constrained RCR gives us the resulting $\gamma \in [0.455, 0.530]$ in which range both efficiencies would be at least 0.825.

The RCR approach gives us the alternative selection method when we do not know the error variances ratio. Since the structural approach assumptions are satisfied in this case, the OR estimate of $\hat{\beta}$ = 1.001 (the MLE when λ = 1) should be suitable here, and the GMR estimate gives $\hat{\beta}$ = 1.001 as well. This means our model with selected $\hat{\beta}$ varying from 0.997 to 1.034 is close to the existing suitable model, even when we have no information on the λ ratio. In addition, the LSS estimate $\hat{\beta}$ = 1.022 with corresponding γ = 0.478 falls inside the selected interval, while the LMS estimate $\hat{\beta}$ = 0.899 is quite biased.



(a) (b)

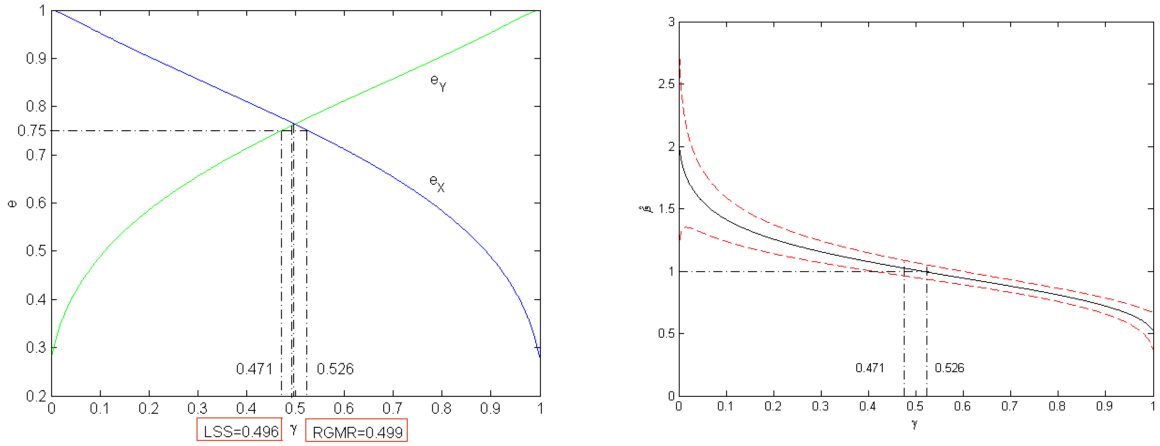
Figure 5.8 (a) RCR efficiency plot for simulation case 1 when λ = 1 for the uncontaminated data. If we set both e_Y and e_X to be no less than 0.825, the corresponding γ interval would be [0.455, 0.530], and the corresponding slope estimate $\hat{\beta}$ varies from 0.997 to 1.034. The cross point corresponds to the RGMR estimate of $\hat{\beta}$ = 1.015 with γ = 0.493. The LSS estimate is $\hat{\beta}$ = 1.022 with the γ = 0.478 included in the selected γ interval. (b) 95% C.I. of RCR slope coefficient estimator. The true β = 1 always lies in the 95% C.I. of the selected RCR estimates.

Resampling:

Since for our nonparametric RCR we cannot conduct theoretical inference on the slope estimate, we will use the bootstrap resampling (1000 replicates) to obtain the 95% confidence interval (C.I.) of $\hat{\beta}$. From panel (b) of Figure 5.8, we can see that the C.I. when $\gamma \in [0.455, 0.530]$ covers the MLE $\hat{\beta}$ = 1.001.

Case 2: Performance on EIV data with outliers in *Y* direction

Exactly the same as the introduce of outliers in Section 5.1, we replace 5% of the 'good' points with outliers having contaminated $\varepsilon_c \sim N(50, \sigma^2_\varepsilon)$ in the Y direction. From the following RCR efficiency plot, we can see that if we want both e_Y and e_X to be no less than 0.75, we can limit the range of γ inside the interval of $\gamma \in [0.471, 0.526]$. In this case, the MLE/OR $\hat{\beta}$=1.881 is highly biased due to the outlier contamination. The LSS estimate $\hat{\beta}$=1.004 with corresponding γ=0.496 falls inside the selected interval, while the LMS estimate $\hat{\beta}$=0.925 is biased. The true β=1 always lies in 95% C.I. of selected $\hat{\beta}$ as can be seen from panel (b) of Figure 5.9.



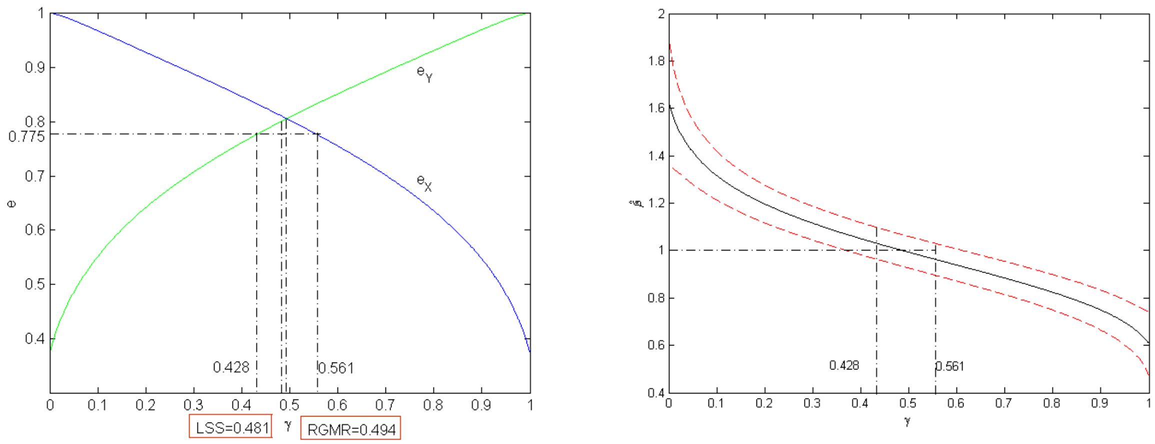
(a) (b)

Figure 5.9 (a) RCR efficiency plot for simulation case 2 when λ = 1 having 5% outliers. If we set both e_Y and e_X to be no less than 0.75, the corresponding γ interval would be [0.471, 0.526], and the corresponding slope estimate $\hat{\beta}$ varies from 0.98 to 1.05. The cross point corresponds to the RGMR estimate of $\hat{\beta}$ = 1.002 with γ = 0.499. The LSS estimate is $\hat{\beta}$ = 1.004 with the γ = 0.496 included in the selected γ interval. (b) 95% C.I. of RCR slope coefficient estimator. The true β = 1 is always covered by the 95% C.I. of the selected RCR estimates.

Case 3: Performance on EIV data with leverage points in *X* direction

Similarly, we now replace 5% of the 'good' points with leverage points having contaminated $\delta_c \sim N(50, \sigma^2_\delta)$ in the X direction. From the following efficiency plot, we can limit the range $\gamma \in [0.428, 0.561]$ to make both e_Y and e_X be at least 0.775. In this case, the MLE/OR $\hat{\beta}$ = 0.566 is highly biased due to the outlier contamination. The LSS estimate $\hat{\beta}$ = 1.020 with corresponding γ = 0.481 does fall inside the selected interval, while the LMS estimate $\hat{\beta}$ = 0.891

is biased. The true $\beta = 1$ always is always inside the 95% C.I. of selected $\hat{\beta}$ as can be seen from the following confidence interval plot.



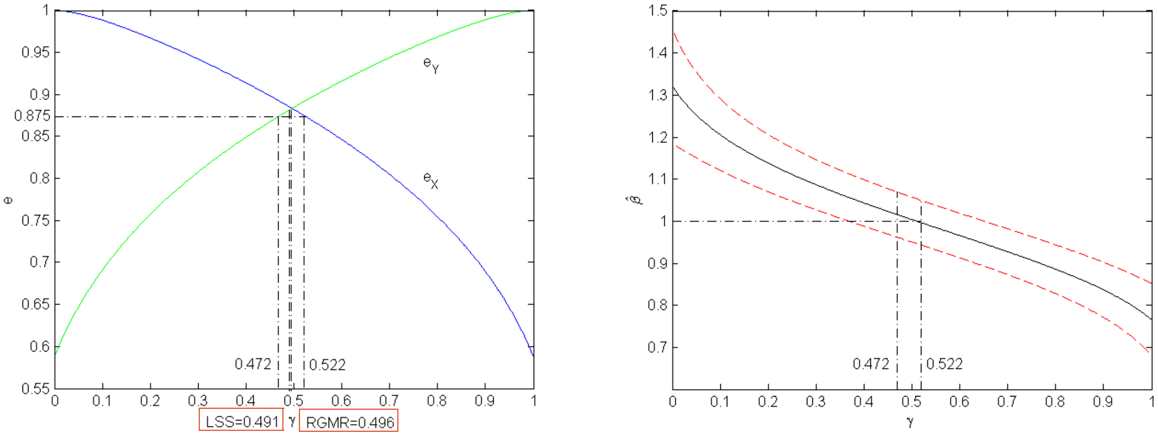
(a) (b)

Figure 5.10 (a) RCR efficiency plot for simulation case 3 when $\lambda = 1$ having 5% leverage points. If we set both e_Y and e_X to be no less than 0.775, the corresponding γ interval would be [0.428, 0.561], and the corresponding slope estimate $\hat{\beta}$ varies from 0.975 to 1.050. The cross point corresponds to the RGMR estimate of $\hat{\beta} = 1.012$ with $\gamma = 0.494$. The LSS estimate is $\hat{\beta} = 1.020$ with the $\gamma = 0.481$ included in the selected γ interval. (b) 95% C.I. of RCR slope coefficient estimator. The true $\beta = 1$ is always covered by the 95% C.I. of the selected RCR estimates.

An advantage of our RCR approach is that it is distribution-free, which means our approach should provide decent estimate even if we encounter data which do not follow normal distribution.

Case 4: Performance on uniformly distributed EIV data

Here we let random latent variable $\xi \sim U(0, 100)$ distribution, and random errors ε and δ also follow the uniform distribution with mean 0 such that $\sigma^2_\varepsilon/\sigma^2_\eta = \sigma^2_\delta/\sigma^2_\xi = 10\%$ i.e. $\lambda = 1$. From the following efficiency plot, we can limit the range $\gamma \in [0.472, 0.522]$ to make both e_Y and e_X be at least 0.875. In this case, the MLE as of $\hat{\beta} = 1.021$ is not the optimal any more due to the violation of the underlying normality assumption. The GMR estimate $\hat{\beta}$ = 1.019 is close to our RCR estimates with selected $\hat{\beta}$ varying from 0.997 to 1.016, however. The LSS estimate $\hat{\beta}$ = 1.009 with corresponding γ = 0.491 does fall inside the selected interval, while the LMS estimate $\hat{\beta}$ = 0.910 is quite biased. The true β = 1 always is always inside the 95% C.I. of selected $\hat{\beta}$ as can be seen from the following confidence interval plot.



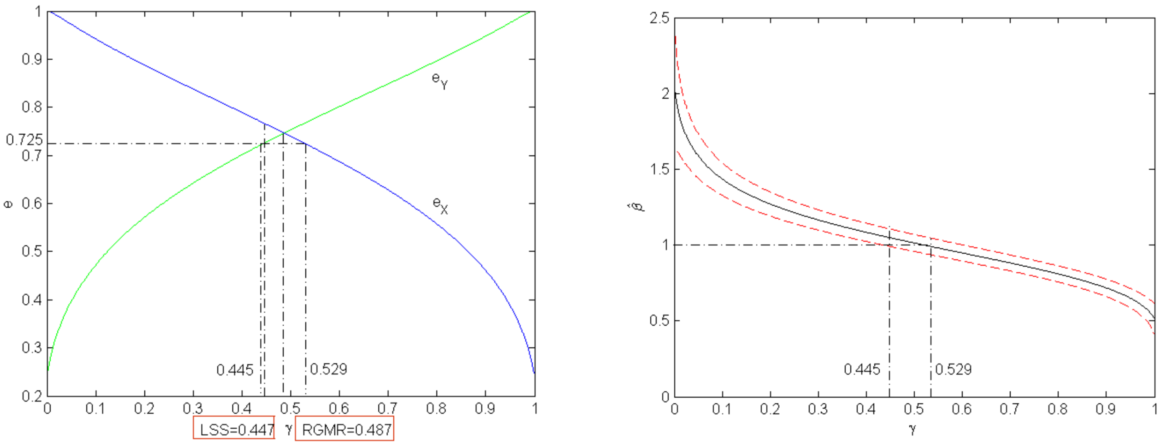
(a) (b)

Figure 5.11 (a) RCR efficiency plot for simulation case 4 when data follows uniform distribution. If we set both e_Y and e_X to be no less than 0.875, the corresponding γ interval would be [0.472, 0.522], and the corresponding slope estimate $\hat{\beta}$ varies from 0.997 to 1.016. The cross point corresponds to the RGMR estimate of $\hat{\beta}$ = 1.007 with γ = 0.496. The LSS estimate is $\hat{\beta}$ = 1.009 with the γ = 0.491 included in the selected γ interval. (b) 95% C.I. of RCR slope coefficient estimator. The true β = 1 always lies in the 95% C.I. of the selected RCR estimates.

Case 5: Performance on *t*(3) (heavy-tail) distributed EIV data

In this case, we set the random latent variable $\xi \sim \sqrt{5}t(3)$ distribution, and the random errors ε and δ both follow the student's *t*(3) distribution such that $\sigma^2_\varepsilon/\sigma^2_\eta = \sigma^2_\delta/\sigma^2_\xi$ = 20% i.e. λ = 1. From the following efficiency plot, we can limit the range $\gamma \in [0.445, 0.529]$ to make both e_Y and e_X to be no less than 0.725. Again, the MLE $\hat{\beta}$ = 1.033 cannot be treated as the optimal estimate here due to the violation of the underlying normality assumption. The GMR estimate $\hat{\beta}$

= 1.026 is included in our selected RCR estimates with $\hat{\beta}$ varying from 0.998 to 1.056, however. The LSS estimate $\hat{\beta}$ = 1.055 with corresponding γ = 0.447 barely falls inside the selected γ interval, while the LMS estimate $\hat{\beta}$ = 0.953 does not. The true β = 1 always is always inside the 95% C.I. of the selected RCR estimates $\hat{\beta}$ as can be seen from the following confidence interval plot.



(a) (b)

Figure 5.12 (a) RCR efficiency plot for simulation case 5 when data follows student's t distribution. If we set both e_Y and e_X to be no less than 0.725, the corresponding γ interval would be [0.445, 0.529], and the corresponding slope estimate $\hat{\beta}$ varies from 0.998 to 1.056. The cross point corresponds to the RGMR estimate of $\hat{\beta}$ = 1.027 with γ = 0.487. The LSS estimate is $\hat{\beta}$ = 1.055 with the γ = 0.447 included in the selected γ interval. (b) 95% C.I. of RCR slope coefficient estimator. The true β = 1 always lies in the 95% C.I. of the selected RCR estimates.

To sum up the calibration results of our RCR approach, we can conclude that, the selected RCR estimates are close to the suitable MLE solution in all situations including (1) when there is no violations of underlying model assumptions, (2) in the presence of outlier contamination, and (3) for the uniformly distributed data, and (4) for the heavy-tail t(3) distributed data.

Chapter 6

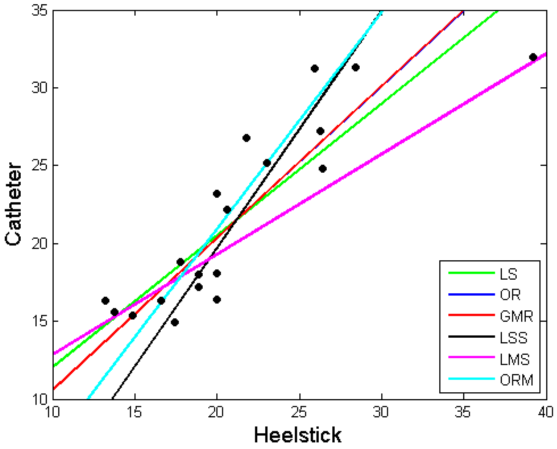
Real-life Examples

6.1 Data from Method Comparison Studies

Table 6.1 Serum kanamycin levels in blood samples drawn simultaneously from an umbilical catheter and a heel venapuncture in twenty babies

Baby	Heelstick (*X*)	Catheter (*Y*)	Baby	Heelstick (*X*)	Catheter (*Y*)
1	23.0	25.2	11	26.4	24.8
2	33.2	26.0	12	21.8	26.8
3	16.6	16.3	13	14.9	15.4
4	26.3	27.2	14	17.4	14.9
5	20.0	23.2	15	20.0	18.1
6	20.0	18.1	16	13.2	16.3
7	20.6	22.2	17	28.4	31.3
8	18.9	17.2	18	25.9	31.2
9	17.8	18.8	19	18.9	18.0
10	20.0	16.4	20	13.8	15.6

In order to illustrate our method, let us consider a simple example which is given by (Kelly 1984) and reanalyzed by (Zamar 1989) in the context of EIV models. The data in Table 6.1 on simultaneous pairs of measurements of serum kanamycin levels in blood samples drawn from twenty babies. One of the measurements was obtained by a heelstick method (*X*), the other by using an umbilical catheter (*Y*). The question was whether the two methods are systematically different and so that one could be substituted for the other after correction for bias. It seems reasonable to assume that both methods are subject to measurement errors with equal variances (Kelly 1984). To better illustrate the behavior of different approaches in the presence of outliers, we change the original value (33.2, 26.0) of case 2 to (39.2, 32.0) as in the numerical example given by (Zamar 1989). The case 2 is clearly separated from others in the upper right-hand corner of Figure 6.1.



Methods	$\hat{\alpha}$	$\hat{\beta}$
LS	3.55	0.85
OR	0.86	0.97
GMR	0.79	0.98
LSS	-10.78	1.52
LMS	6.41	0.64
ORM*	-6.91	1.39

* Orthogonal Regression M-estimators (Zamar, 1989)

Figure 6.1 Different fitted regression lines for Example 1. The ratio of sample variances is S_{YY}/S_{XX} = 0.954. It seems that the LSS and the ORM estimators have the best fit to the main bulk of data, and the slope estimates from other approaches are greatly attenuated by the outlier in the upper right-hand corner. Of note, we also observe that the OR and the GMR gives almost same slope estimates that are close to 1, while the robust LMS estimate is badly influenced by the outlier in this situation.

The different regression estimates of the simple linear model $\hat{Y} = \hat{\alpha} + \hat{\beta}\, X$ are presented in the form of fitted regression lines in Figure 6.1. We observe that the LSS and the ORM estimators have the best fit to the main bulk of data, and the slope estimates from other approaches are greatly attenuated by the outlier in the upper right-hand corner. Of note, we also observe that the OR and the GMR gives almost same slope estimates that are close to 1, while the robust LMS estimate is badly influenced by the outlier in this situation. To illustrate our new methods, we will perform the RCR approach on this dataset as well. The selected robust compound regression results are tabulated in Table 6.2, and we can see that RGMR has the largest summation of regression efficiencies which indicates it is the best among all RCR estimates.

Table 6.2 Selected RCR results for Example 1

γ	α	β	$\sum_i \frac{(Y_i-\hat{Y}_i)^2}{R_i^2}$	$\sum_i \frac{(X_i-\hat{X}_i)^2}{R_i^2}$	e$_Y$	e$_X$	e$_Y$ + e$_X$
0	-20.79	2.00	16.04	4.02	0.443	1.000	1.443
0.10	-12.68	1.61	10.94	4.20	0.650	0.957	1.607
0.16 (LSS)	-10.77	1.52	10.05	4.33	0.708	0.928	1.636
0.20	-9.66	1.47	9.59	4.43	0.742	0.908	1.650
0.30	-7.67	1.38	8.85	4.67	0.803	0.861	1.664
0.36 (RGMR)	-6.68	1.33	8.54	4.83	0.833	0.833	1.666
0.40	-6.10	1.30	8.37	4.93	0.850	0.815	1.665
0.50	-4.75	1.24	8.01	5.22	0.887	0.770	1.657
0.60	-3.50	1.18	7.74	5.56	0.919	0.723	1.642
0.70	-2.27	1.12	7.51	5.98	0.947	0.673	1.620
0.80	-0.97	1.06	7.33	6.53	0.970	0.616	1.586
0.90	0.55	0.99	7.19	7.36	0.989	0.546	1.535
1.00	2.72	0.89	7.11	9.07	1.000	0.443	1.443

One advantage of the RCR approach is that it gives users the flexibility to choose their desired regression line from the entire class of RCR lines from the RCR efficiency plot in Figure 6.2. Suppose we want the desired line to be at least 95% efficient for the estimation of Y, we can clearly see that when e$_Y$ = 0.95, we have the compound parameter γ = 0.714 and e$_X$ = 0.67. In most situations, we do not have a preference on either variable of the analyzed data. For example, if we set both e$_Y$ and e$_X$ be no less than 0.825, from the RCR efficiency plot, the satisfied γ interval would be [0.344, 0.379], and the corresponding $\hat{\beta}$ varies from 1.32 to 1.34.

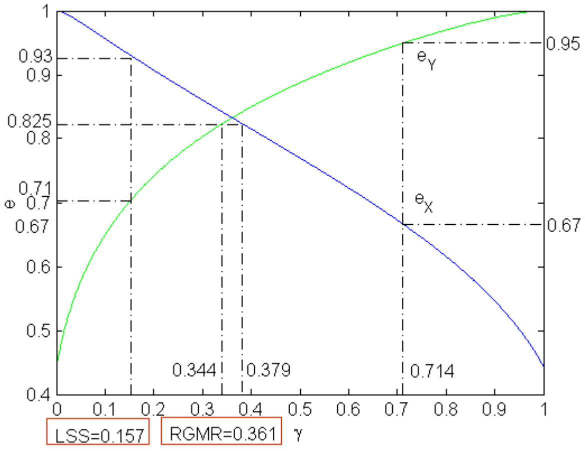


Figure 6.2 RCR efficiency plot for Example 1. Suppose we want the desired line to be at least 95% efficient for the estimation of Y, that is, e$_Y$≥ 0.95, the corresponding compound parameter γ = 0.714 with the corresponding efficiency for estimating X to be e$_X$ = 0.67. If we set both e$_Y$ and e$_X$ to be no less than 0.825, the corresponding γ interval would be [0.344, 0.379], and the corresponding slope estimate $\hat{\beta}$ varies from 1.32 to 1.34. The cross point corresponds to the RGMR estimate of $\hat{\beta}$ = 1.330 with γ = 0.361. The LSS estimate is $\hat{\beta}$ = 1.524 with γ = 0.157.

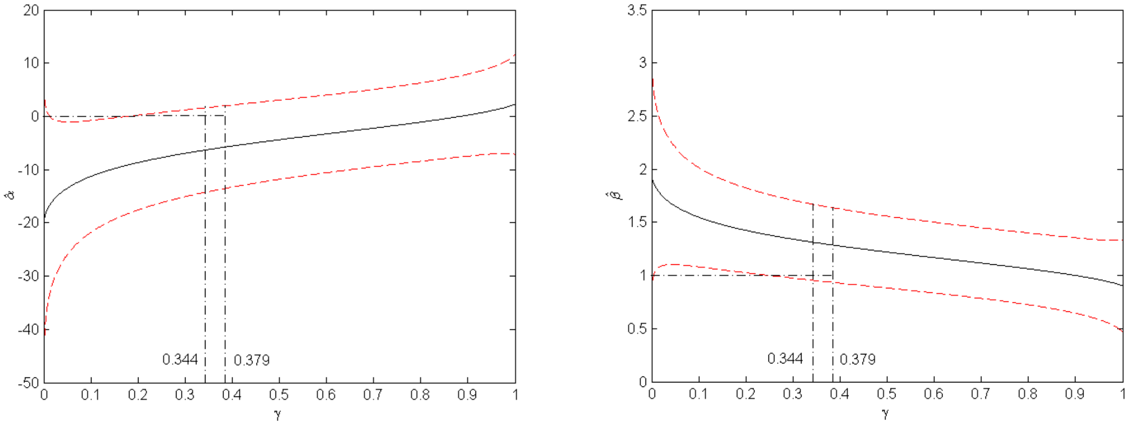
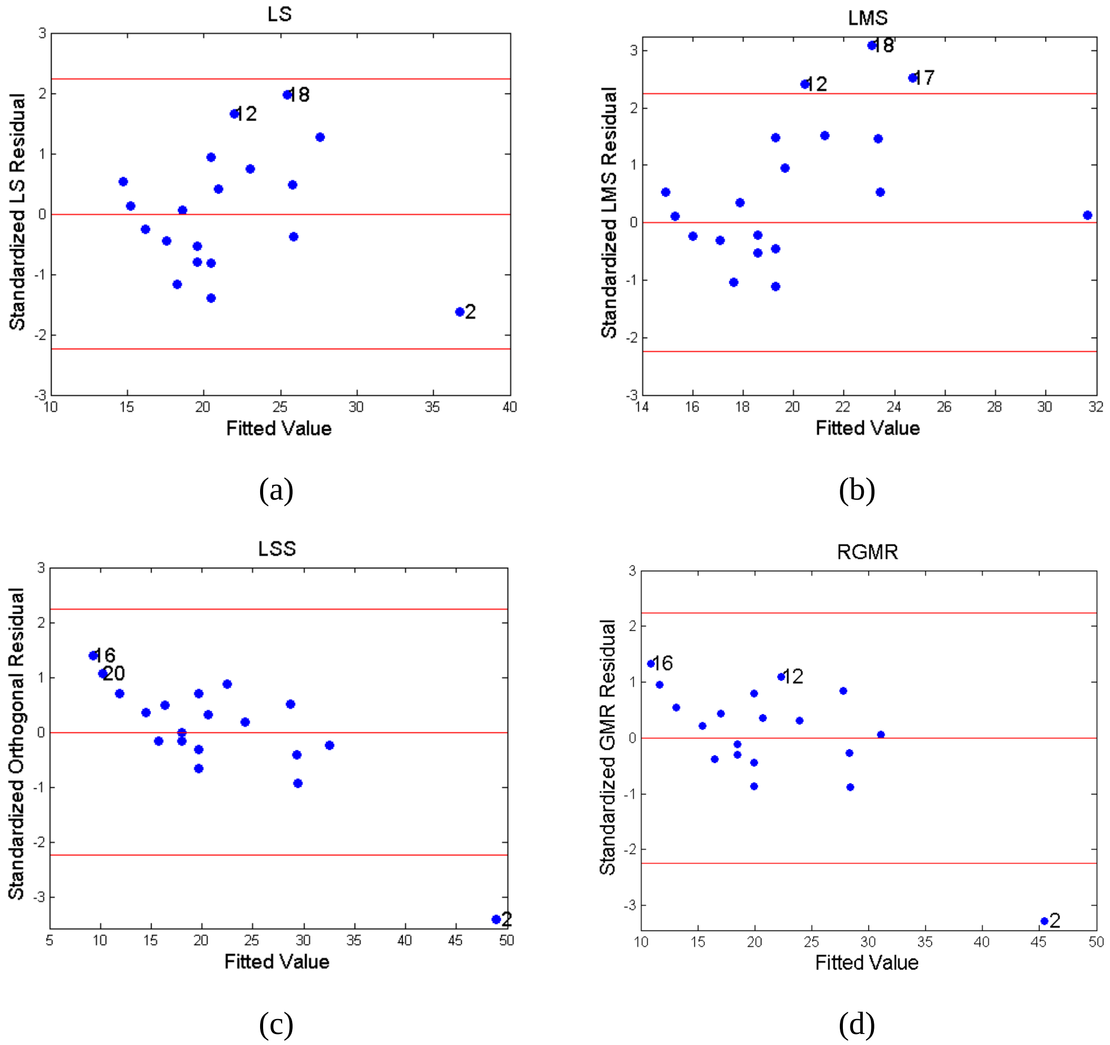


Figure 6.3 95% C.I.s of RCR slope coefficient estimator for Example 1. The null hypothesis of α = 0 and β = 1, that is, the two method are equivalent, are always covered by the 95% C.I. of the selected RCR estimates under the criterion that both efficiencies e$_Y$ and e$_X$ are no less than 0.825.

To detect individual cases that may differ from the bulk of the data, we performed the diagnostics in Figure 6.4 based on the residuals from the usual LS, the robust LMS, and our new robust approaches - LSS and RGMR respectively. The LS diagnostics firstly detects cases 18 and

12 as potential outliers before finding the gross outlier case 2, which is misleading. Of note, the robust LMS approach performs even worse. By contrast, both the new high breakdown LSS and RGMR diagnostics successfully detect case 2 as a gross outlier.



(a)

(b)

(c)

(d)

Figure 6.4 Outlier diagnostics for Example 1. (a) The usual LS diagnostics fails to detect the outlier - case 2; (b) the robust LMS diagnostics treat 'good' points as outliers; (c) the new LSS diagnostics and (d) the new RGMR diagnostics both show the case 2 as a gross outlier.

As we confirmed the presence of such a gross outlier, case 2 should be deleted. Assume the sample variations come from the random errors only, i.e. $\hat{\lambda} = \frac{s_{YY}}{s_{XX}} = 1.582$, the GMR would be the MLE solution here. We now apply the GMR estimator (i.e. the MLE) to the remainder, and the GMR gives the fitted line in Figure 6.5 with slope 1.26 and intercept -4.52. Moreover, the bootstrap 95% confidence intervals (0.97, 1.57) for $\hat{\beta}_{GMR}$ and (-10.93, 1.35) for $\hat{\alpha}_{GMR}$ support the null hypotheses that α = 0 and β = 1. As can be seen, compared to the optimal GMR estimates, the RCR estimate is the most robust one, better than Zamar's ORM estimate, when exposed to the contamination of case 2. Furthermore, the values α = 0 and β = 1 are always covered by the 95% C.I. of the selected optimal RCR estimates as shown in Figure 6.3, which also concludes that the two methods of measurement are not significantly different.

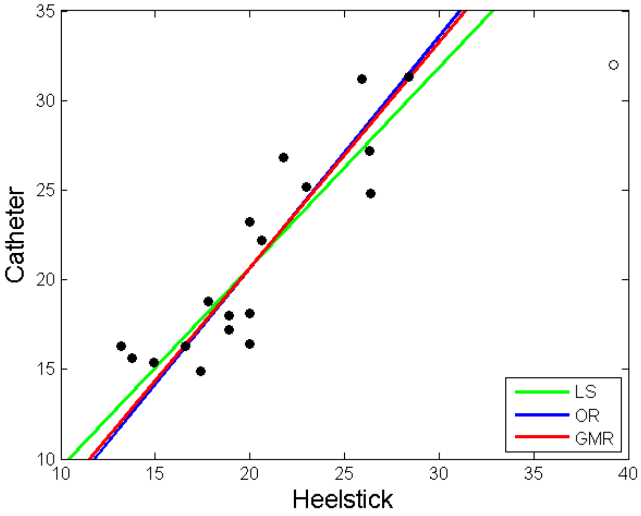


Figure 6.5 Reanalysis of Example 1 after outlier removal. The case 2 in the upper right-hand corner is deleted from the analysis, then we have the fitted LS line: $\hat{Y} = -1.92 + 1.11\,X$; the OR line: $\hat{Y} = -5.26 + 1.29\,X$; the GMR line: $\hat{Y} = -4.52 + 1.26\,X$.

6.2 Brain vs. Body Weights Data

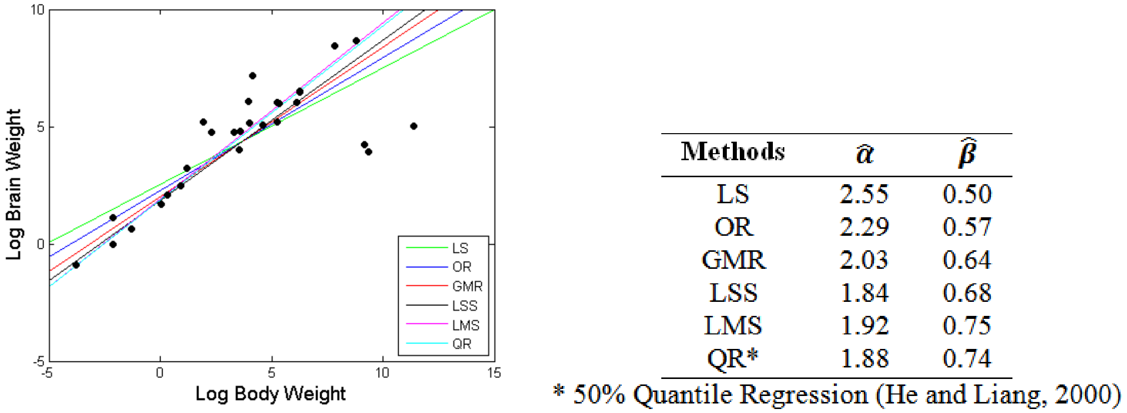
The data in Table 6.3 on brain and body weights of 28 animals is given by Rousseeuw and Leroy (1987) and reanalyzed by He and Liang (2000) in the context of EIV models. Here the predictor X is the body weight (in kilograms), and the response Y is the brain weight (in grams).

The logarithmic transformation was necessary to make the data look more linear and less heteroscedastic (Rousseeuw and Leroy 1987). We also take the view that both weights are assumed to be measured with error (He and Liang 2000). The question we are interested in is whether a larger brain is required to govern a heavier body, or from another perspective, whether the brain weight increases linearly as the body weight increases.

Table 6.3 Body and Brain Weight for 28 Animals

Index (*i*)	**Species**	**Body Weight (X_i)**	**Brain Weight (Y_i)**	**Index (*i*)**	**Species**	**Body Weight (X_i)**	**Brain Weight (Y_i)**
1	Mountain beaver	1.350	8.100	15	African elephant	6654.00	5712.00
2	Cow	465.000	423.000	16	Triceratops	9400.00	70.00
3	Gray wolf	36.330	119.500	17	Rhesus monkey	6.800	179.000
4	Goat	27.660	115.000	18	Kangaroo	35.000	56.000
5	Guinea pig	1.040	5.500	19	Hamster	0.120	1.000
6	Diplodocus	11700.0	50.0	20	Mouse	0.023	0.400
7	Asian elephant	2547.00	4603.00	21	Rabbit	2.500	12.100
8	Donkey	187.100	419.000	22	Sheep	55.500	175.000
9	Horse	521.000	655.000	23	Jaguar	100.000	157.000
10	Potar monkey	10.000	115.000	24	Chimpanzee	52.160	440.000
11	Cat	3.300	25.600	25	Brachiosaurus	87000.0	154.5
12	Giraffe	529.000	680.000	26	Rat	0.280	1.900
13	Gorilla	207.000	406.000	27	Mole	0.122	3.000
14	Human	62.00	1320.00	28	Pig	192.000	180.000

The scatter plot of the data along with different fitted regression lines are shown in Figure 6.6. The three animals clearly identified on the right side break the whole dataset into two parts. The three animals are categorized as dinosaurs, whereas the 25 remaining animals are all mammals. By visual inspection of the fitted lines in Figure 6.6, we see that the LMS and the QR estimators have the best fit to the main bulk of data, and the LSS having a relatively good fit as well compared to the attenuated LS, OR, and GMR estimates. We also want to point out that, obviously from the data in Figure 6.6, a curve that will flat out when the body weight passes beyond a certain threshold may be most reasonable.



Methods	$\hat{\alpha}$	$\hat{\beta}$
LS	2.55	0.50
OR	2.29	0.57
GMR	2.03	0.64
LSS	1.84	0.68
LMS	1.92	0.75
QR*	1.88	0.74

* 50% Quantile Regression (He and Liang, 2000)

Figure 6.6 Different fitted regression lines for Example 2. The ratio of sample variances is S_{YY}/S_{XX} = 0.405. It seems that the LMS and the QR estimators have the best fit to the main bulk of data, and the LSS having a relatively good fit as well compared to the attenuated LS, OR, and GMR estimates. However, obviously from the data, a curve that will flat out when the body weight passes beyond a certain threshold may be most reasonable.

Table 6.4 Selected RCR results for Example 2

γ	α	β	$\sum_i \frac{(Y_i-\hat{Y}_i)^2}{R_i^2}$	$\sum_i \frac{(X_i-\hat{X}_i)^2}{R_i^2}$	e_Y	e_X	$e_Y + e_X$
0	-1.20	1.49	26.44	11.90	0.296	1.000	1.296
0.10	-0.03	1.18	17.10	12.24	0.457	0.972	1.429
0.20	0.41	1.06	14.39	12.70	0.544	0.937	1.481
0.30	0.71	0.99	12.82	13.21	0.610	0.900	1.510
0.40	0.95	0.92	11.72	13.80	0.667	0.862	1.529
0.50	1.16	0.87	10.87	14.50	0.720	0.820	1.540
0.60 (RGMR)	1.36	0.81	10.15	15.38	0.770	0.774	1.544
0.70	1.57	0.76	9.52	16.56	0.821	0.718	1.539
0.80	1.79	0.70	8.94	18.34	0.875	0.649	1.524
0.82 (LSS)	1.84	0.68	8.82	18.83	0.886	0.632	1.518
0.90	2.09	0.62	8.36	21.75	0.935	0.547	1.482
1.00	2.76	0.44	7.82	40.22	1.000	0.296	1.296

The selected robust compound regression results are tabulated in the above table, and the RGMR is again verified having the largest sum of regression efficiencies which indicates it is the best among all the RCR estimates. Figure 6.7 is the corresponding RCR efficiency plot. Suppose we want the desired line to be at least 95% efficient for the estimation of X, we can clearly see that when $e_X = 0.95$, we have the compound parameter $\gamma = 0.163$ and $e_Y = 0.515$. If both variables are treated as equally important, we set both e_Y and e_X be no less than 0.75, the satisfied γ interval would be [0.560, 0.645], and the corresponding $\hat{\beta}$ varies from 0.79 to 0.83.

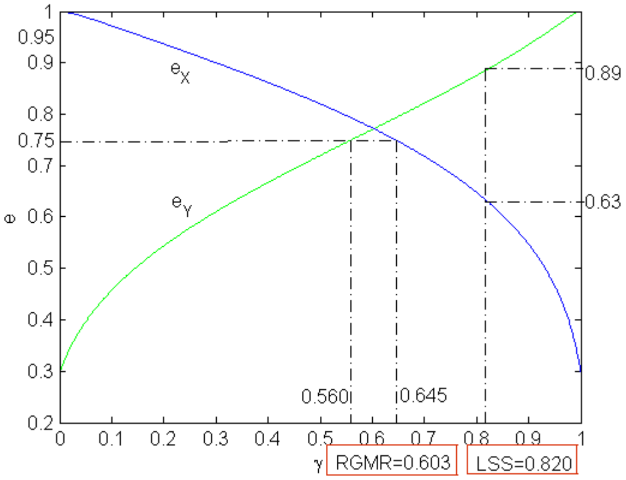
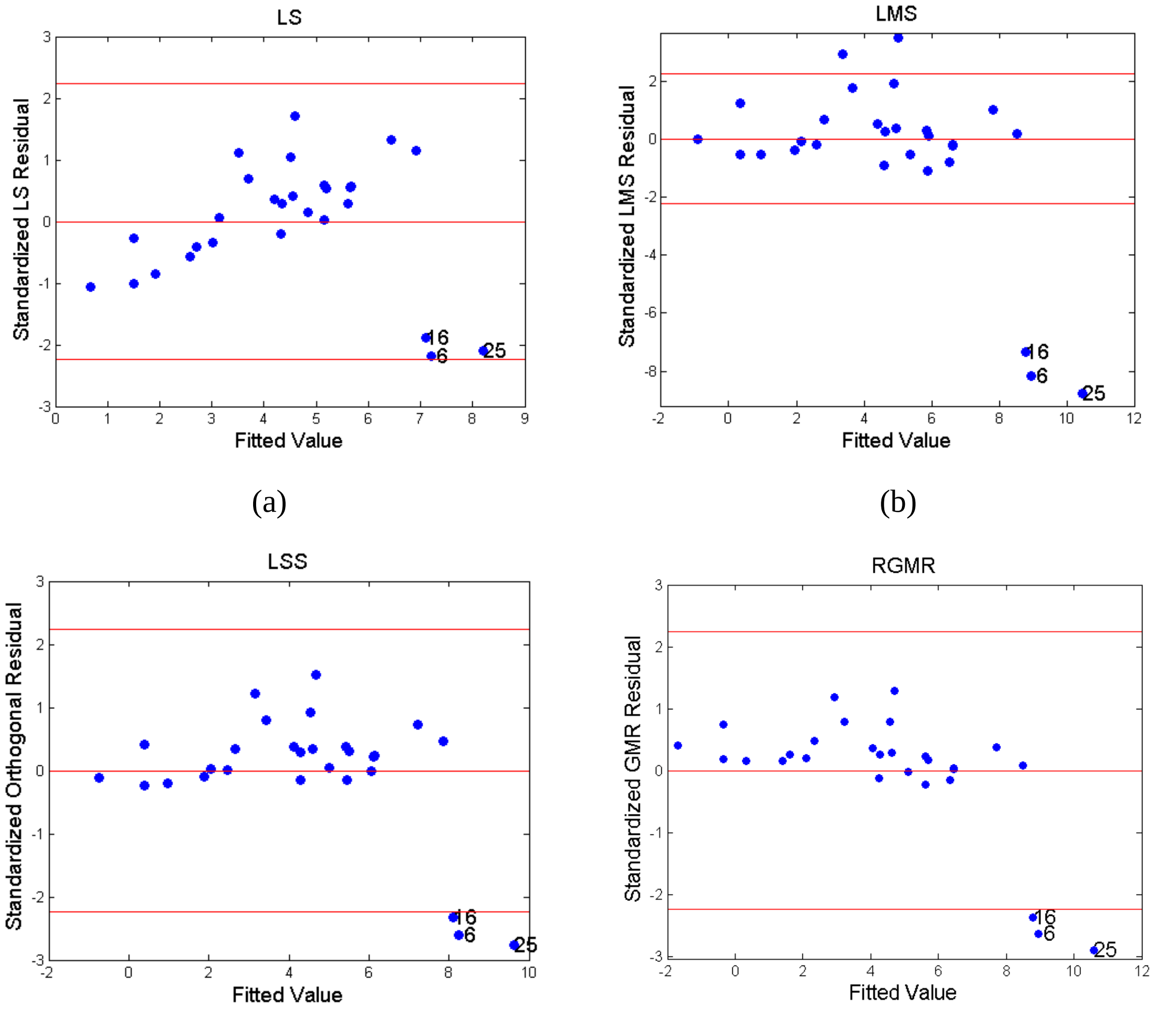


Figure 6.7 RCR efficiency plot for Example 2. If we set both e_Y and e_X to be no less than 0.75, the corresponding γ interval would be [0.560, 0.645], and the corresponding slope estimate $\hat{\beta}$ varies from 0.79 to 0.83. The cross point corresponds to the RGMR estimate of $\hat{\beta} = 0.811$ with $\gamma = 0.603$. The LSS estimate is $\hat{\beta} = 0.684$ with $\gamma = 0.820$.

Although the outliers are obvious in this example, we still perform the outlier diagnostics to test the power of each approach in detecting the three gross outliers. While the LSS barely detects the outliers near the 95% cut-off, the high breakdown diagnostics – the robust LMS, and the new LSS and RGMR all detect the cases 6, 16, and 25 as gross outliers.



(a) (b)

(c) (d)

Figure 6.8 Outlier diagnostics for Example 2. (a) The usual LS diagnostics barely detects the three outliers near the 95% cut-off; (b) the robust LMS diagnostics (c) the new LSS diagnostics and (d) the new RGMR diagnostics all detect the cases 6, 16, and 25 as gross outliers.

After the three dinosaurs are excluded from the analysis of the rest of the mammals, the traditional LS, OR, and GMR analysis methods all perform well as shown in Figure 6.9. Assume the sample variations come from the random errors only, i.e. $\hat{\lambda} = \frac{s_{YY}}{s_{XX}} = 0.614$, the GMR is the MLE solution here. As can be seen, the RCR estimates $\hat{\beta} \in (0.79, 0.83)$ is close to the GMR $\hat{\beta} =$ 0.78, and is totally covered by the 95% bootstrap C.I. (0.72, 0.85) of the GMR slope estimate.

Since a significant positive slope is confirmed, we can draw the conclusion that a larger brain is required to govern a heavier body.

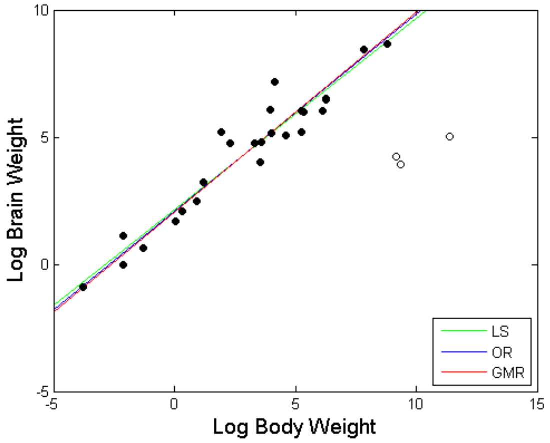


Figure 6.9 Reanalysis of Example 2 after outlier removal. The cases 2, 16, and 25 are deleted from the analysis, then we have the fitted LS line: $\hat{Y} = 2.15 + 0.75\,X$; the OR line: $\hat{Y} = 2.08 + 0.78\,X$; the GMR line: $\hat{Y} = 2.06 + 0.78\,X$.

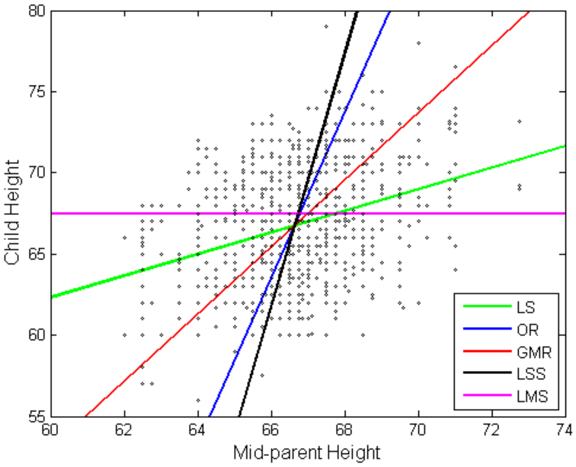
6.3 Galton's Family Heights Data

Galton's family heights data have been a preeminent historical dataset in regression analysis, and the original model and basic results on this dataset have survived the close scrutiny of statisticians for 125 years. Using Galton's data as a benchmark for different regression approaches including our newly developed robust approaches – LSS and RCR, we elucidated that the ordinary least squares regression has a strong bias leading to otherwise alternative conclusions on the true relationships between the heights of the child and his or her parents.

The statistical terminology of 'regression' was coined by Sir Francis Galton beyond dispute, while the family heights data was formally introduced in his study on *Regression towards Mediocrity in Hereditary Stature* (Galton 1886, 1889). Fortunately, the researchers were still able to retrieve Galton's family heights data from his firsthand notebook reserved at University College London despite the elapsing of more than a century (Hanley 2004). It consists of the records from 205 families with 962 adult children in total, among which 486 are sons and 476 are daughters. However, after excluding the non-numerical entries (tall, medium, short, etc.), the preprocessed data in our article is eventually formed from the records of 481 sons and 453 daughters as well as their parents.

6.3.1 Regression of Child on Mid-parent

The regression of child on mid-parent was of greater scientific interest (Hanley 2005). All the heights in Galton's data are assumed to be subject to random measurement errors, which is suitable for our study of EIV models. Assume there exists the simple linear relationship $Y_i = \alpha + \beta X_i + \varepsilon_i$, $i = 1, 2, ..., 934$, where Y_i is the height of each child (son or daughter), and X_i is the mid-parent height (the average height of father and mother).



Methods	$\hat{\alpha}$	$\hat{\beta}$
LS	22.291	0.667
OR	-272.522	5.091
GMR	-70.882	2.065
LSS	-450.662	7.764
LMS	67.5	0.000

Figure 6.10 Different fitted regression lines for Example 3 part 1. From the large discrepancies between the regression estimates from different approaches, we view that the target dataset is

quite noisy. But when we eyeball the barely linear trend, we find the GMR among all others fairly represents the trend.

From the above scatter plot with different fitted regression lines, we can clearly see the target dataset is quite noisy and there is a very week linear relationship with a correlation coefficient as of r_{XY} = 0.323. However, our previous simulations give us the 'rule of thumb' to choose the GMR estimates as the most trustable result when there is no evidence of gross outliers, which is confirmed from the outlier diagnostics in Figure 6.11.

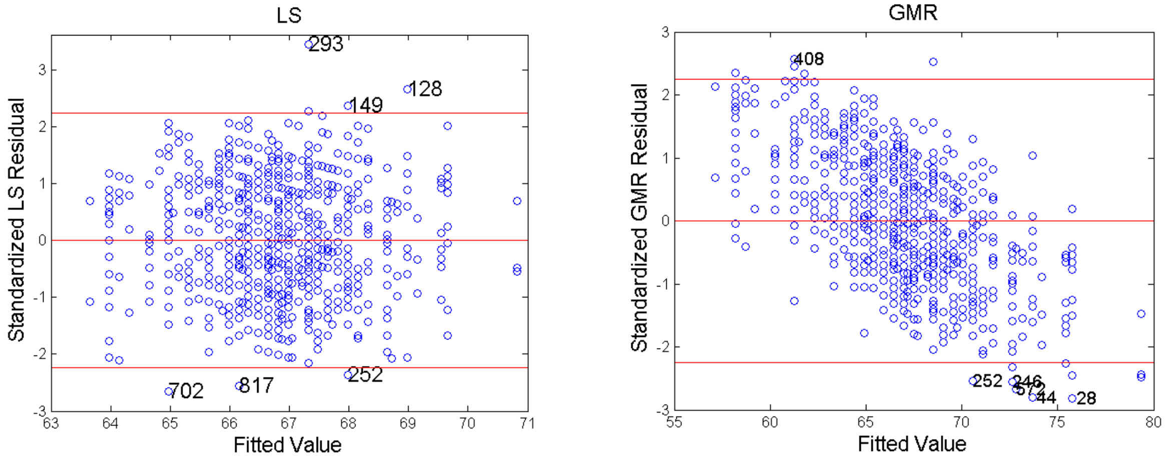


Figure 6.11 Outlier diagnostics for Example 3 part 1. There seems no appearance of gross outliers presented by the diagnostics from either the LS residuals or the GMR residuals.

From the EIV modeling point of view, if we classify the height problem as a pure functional EIV model where the variation should come from the errors only – then the GMR is the most suitable in the absence of outliers. To compare the MLE of the function EIV model with the GMR result, we need firstly estimate the error variances ratio λ from the data, which is feasible in the following sense. By the assumption of functional EIV models, the estimated λ can be obtained as $\hat{\lambda} = \frac{s_{YY}}{s_{XX}} = 4.265$. On the other hand, for general EIV problems, we are able to estimate λ when we find there are many identical values observed in both variables. In this dataset, the random error ${\sigma_\varepsilon}^2$ of child's height can be estimated from the 'repeated measurements' of child's height for the same mid-parent's height, and similarly we can estimate the random error ${\sigma_\delta}^2$ of mid-parent's height. Hence we have $\hat{\lambda} = {\hat{\sigma}_\varepsilon}^2 / {\hat{\sigma}_\delta}^2 = 4.346 \approx \frac{S_{YY}}{S_{XX}}$, which substantiates our claim that the sample variations are from the random errors only. The MLE is $\hat{\beta}_{MLE} = 2.065$ when we assume λ=4.265. We can see that it is consistent with the GMR estimate stated earlier. However, the normality assumption of the MLE approach is not attainable when we perform the residuals diagnostics (p = 0.016 under the Shapiro-Wilk test, and the data is light-tailed), which implies the parametric MLE approach is questionable but the distribution free GMR approach is still feasible.

As there are hardly any outliers here, we first perform the compound regression analysis on this dataset. From the CR efficiency plot of Figure 6.12, if we set both e$_Y$ and e$_X$ be no less than 0.65, the selected $\gamma \in [0.18, 0.20]$ with corresponding $\hat{\beta}$ varies from 2.039 to 2.093. The corss point corresponds to the GMR esimate γ = 0.19 & $\hat{\beta}$ = 2.065.

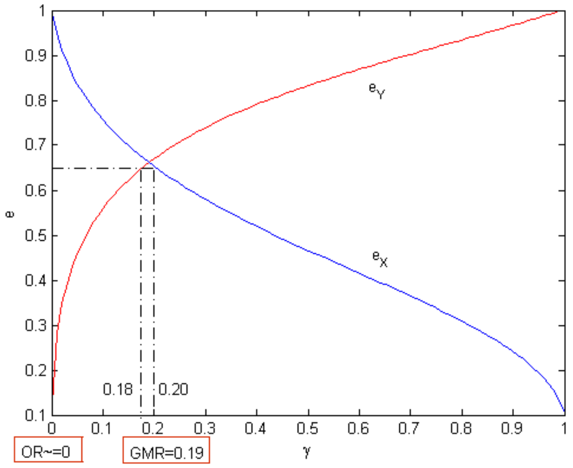


Figure 6.12 CR efficiency plot for Example 3 part 1. If we set both efficiencies e$_Y$ and e$_X$ to be no less than 0.65, the corresponding γ interval would be [0.18, 0.20], and the corresponding slope

estimate $\hat{\beta}$ varies from 2.039 to 2.093. The cross point corresponds to the GMR estimate of $\hat{\beta}$ = 2.065 with γ = 0.19. The OR estimate is $\hat{\beta}$ = 5.091 with $\gamma \approx 0$.

On the other hand, in order to illustrate our new RCR estimation approach, we will compare our robust RCR estimates with the reasonable GMR estimate. From the RCR efficiency plot of Figure 6.13, if we set both e_Y and e_X be no less than 0.55, the selected $\gamma \in [0.257, 0.302]$ with corresponding $\hat{\beta}$ varies from 1.566 to 1.650, which compared to other estimates is the most close to the GMR result. Especially, if $e_X > 0.70$ is desired, we find γ = 0.116 & $\hat{\beta}$ = 2.067 is surprisingly close to the GMR estimate.

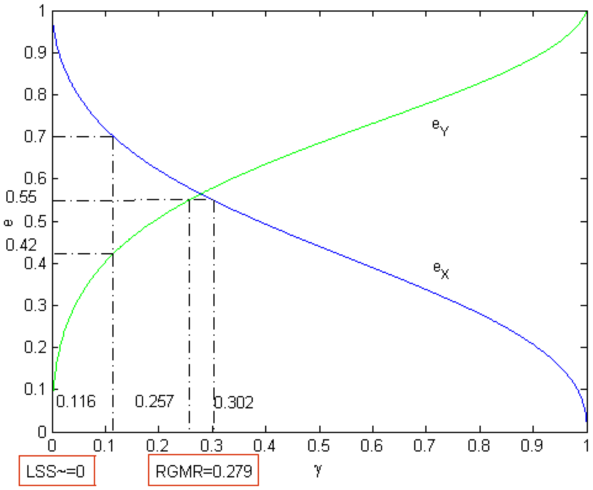


Figure 6.13 RCR efficiency plot for Example 3 part 1. If we set both e_Y and e_X to be no less than 0.55, the corresponding γ interval would be [0.257, 0.302], and the corresponding slope estimate $\hat{\beta}$ varies from 1.566 to 1.650. The cross point corresponds to the RGMR estimate of $\hat{\beta}$ = 1.608 with γ = 0.279. Suppose we want the desired efficiency for the estimation of X to be at least 70%, that is, $e_X \geq 0.70$, the corresponding compound parameter γ = 0.116, the regression efficiency of Y e_Y = 0.42, with the corresponding slope estimate $\hat{\beta}$ = 2.067 that is very close to the GMR estimate $\hat{\beta}$ = 2.065. While The LSS estimate gives $\hat{\beta}$ = 7.764 with $\gamma \approx 0$.

Upon comparing these two analysis approaches, we view that the CR approach is preferable when there is no evidence of gross outliers in the data, and the CR efficiency plot shows a higher efficiency for the GMR rather than the RCR efficiency for the RGMR. All in all, the reasonable slope estimate should be $\hat{\beta}_{GMR} = 2.065,$ which can be interpreted as an average one inch advantage in the mid-parent height will benefit their child about two inches taller than other children.

6.3.2 Estimation and Inference on Gender-specific Models

In reality, it is natural to raise the curiosity questioning whether the stature of the offspring inherits more from the father or the mother. In order to address this question, we proposed the pair of gender-specific multiple linear regression models (1) & (2) as we are interested in discriminating the model for sons from that for daughters.

$$Y_1 = \alpha_1 + \beta_{11}X_{11} + \beta_{12}X_{12} + \varepsilon_1 \quad (1)$$

$$Y_2 = \alpha_2 + \beta_{21}X_{21} + \beta_{22}X_{22} + \varepsilon_2 \quad (2)$$

The random variables of paternal height X_{11} and the maternal height X_{12} are bundled with the son's height Y_1 in model (1), while the daughter's height Y_2 together with the heights of her father X_{21} and her mother X_{22} are involved in model (2), where ε_1 and ε_2 are the corresponding error terms.

Table 6.5 above tabulates the regression analyses results of the gender-specific models on the 481 sons' and 483 daughters' datasets respectively. It clearly demonstrate that the LS slope estimates are always much smaller than that from the other regressions configured for EIV models, due to the nature that the LS will underestimate the regression slopes when the predictors are contaminated with measurement errors. Of note, referring to the estimated regression slopes for the model on daughters' heights, the LS slopes relative to the others are in a reverse pattern. To judge the goodness-of-fit of each method, I used the sum of regression efficiency as a general criterion for different methods. The reason is because we are trying to uncover a true linear relationship but not for the purposes of prediction of any variable, and thus all the regression variables are treated as equally important in the regression model. In terms of SRE, the OR is the most efficient method for the estimation of the first model, while the GMR estimate is most efficient for the second model.

Table 6.5 Results from different regressions for gender-specific models

Methods	$\boldsymbol{Y_1 = \alpha_1 + \beta_{11}X_{11} + \beta_{12}X_{12} + \varepsilon_1}$				$\boldsymbol{Y_2 = \alpha_2 + \beta_{21}X_{21} + \beta_{22}X_{22} + \varepsilon_2}$			
	$\hat{\alpha}_1$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	SRE$_1$	$\hat{\alpha}_2$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	SRE$_2$
LS	19.2182 (3.7420)	0.4187 (0.0431)	0.3290 (0.0454)	1.1978 (0.0313)	18.1594 (3.8485)	0.3753 (0.0381)	0.3109 (0.0447)	1.2341 (0.0373)
OR	-43.271 (10.940)	0.9811 (0.1522)	0.6976 (0.1488)	1.3053 (0.0454)	-21.889 (7.1581)	0.6142 (0.0700)	0.6773 (0.1067)	1.4142 (0.0496)
GMR	-53.162 (5.9810)	0.9471 (0.0534)	0.8888 (0.0502)	1.2974 (0.0470)	-41.522 (4.4929)	0.7674 (0.0396)	0.8179 (0.0441)	1.4393 (0.0512)
LSS	-49.569 (15.227)	1.0727 (0.1935)	0.6969 (0.2437)	1.3021 (0.0470)	-41.185 (11.137)	0.6924 (0.0917)	0.8936 (0.1413)	1.4301 (0.0534)
RGMR	-60.192 (4.7406)	0.9945 (0.0423)	0.9474 (0.0419)	1.2923 (0.0472)	-57.657 (4.4220)	0.8811 (0.0372)	0.9467 (0.0422)	1.4367 (0.0521)

Mean and standard error (in parentheses) of regression coefficients by B=5,000 bootstraps; SRE stands for the sum of robust regression efficiency with respect to each variable

Furthermore, we care about whether the identified patterns of unequal contributions from parents are statistically significant or merely occur by chance. Due to the violation of the normality assumption for the gender-specific models, our precedent insight of parametric hypotheses testing is questionable. Meanwhile, even if the underlying assumptions for parametric tests are fulfilled, it is still complicated for the inference based on the asymptotically estimated covariance matrix for regression coefficients of multivariate EIV models (Patefield 1981). Hence, we will take advantage of the prevailing non-parametric technique - the bootstrap (Efron 1979, 1982, Efron and Tibshirani 1993) to test the hypotheses.

Presumably, the parents are in equal roles to the stature of the offspring, our hypotheses would be for the model on the sons' heights, H_{01}: $\beta_{11}=\beta_{12}$ versus H_{a1}: $\beta_{11} > \beta_{12}$ i.e. the father rather than the mother has significantly larger influence on the height of their son; for the model on the daughters' heights, H_{02}: $\beta_{21}=\beta_{22}$ versus H_{a2}: $\beta_{21} < \beta_{22}$; i.e. the mother has significantly more contribution to their daughter than the father does. Under each null hypothesis, since both terms from father and mother in the regression model will be ultimately merged into one single term, it is therefore feasible to set up a permutation test by randomly swapping the father's and the mother's heights within each family.

If the observed positive differences of slopes in gender-specific models are denoted as $\hat{\Delta}_1 = \hat{\beta}_{11}$ - $\hat{\beta}_{12}$ and $\hat{\Delta}_2 = \hat{\beta}_{22}$ - $\hat{\beta}_{21}$ respectively, the corresponding resampled differences of $\hat{\Delta}_1^* = \hat{\beta}_{11}^*$ - $\hat{\beta}_{12}^*$ and $\hat{\Delta}_2^* = \hat{\beta}_{22}^*$ - $\hat{\beta}_{21}^*$ will then be hypothetically generated through the resampling procedures. Consequently, the permutation achieved significance level (ASL) of each hypothesis test is defined to be the permutation probability of observing at least that large a difference when the null hypothesis is true, that is the permutation probability that $\hat{\Delta}^*$ exceeds $\hat{\Delta}$. Hence,

$$\mathrm{ASL}_{Son} = P_{H_{01}}\left(\hat{\Delta}_1^* \geq \hat{\Delta}_1\right) = \#\{\hat{\Delta}_1^* \geq \hat{\Delta}_1\}/B$$

$$\mathrm{ASL}_{Daughter} = P_{H_{02}}\left(\hat{\Delta}_2^* \geq \hat{\Delta}_2\right) = \#\{\hat{\Delta}_2^* \geq \hat{\Delta}_2\}/B$$

where there are B = 1,000,000 permutation replications for each test. If the ASL is smaller than the specified significance level say α = 0.05, we reject the null hypothesis as there is a little chance the null hypothesis holds based on the data we already obtained.

Table 6.6 The ASL of hypotheses test on the unequal slope coefficients

Methods	$\mathrm{ASL}_{\mathrm{Son}}$	$\mathrm{ASL}_{\mathrm{Daughter}}$
LS	0.004697*	0.989513
OR	0.000000*	0.071841
GMR	0.063362	0.070548
LSS	0.000000*	0.000019*
RGMR	0.104749	0.024111*

B=1,000,000; * Significant at α=0.05

From the above table, we can conclude that at the significance level of 0.05, the LS, the OR, and the new LSS approaches from the first column show the father has significantly larger influence on the son's height rather than the mother does; to the daughter's height, both two new approaches - the LSS and the RGMR indicates the mother's contribution is significantly more than the father's.

Chapter 7

Discussion

In this thesis, we proposed two novel nonparametric approaches – the least sine squares (LSS) regression and the robust compound regression (RCR) analysis methods for the robust estimation of errors-in-variables (EIV) models. The RCR including the LSS as a special case provides the robust counterpart for every EIV regression line in a 1-1 mapping. We not only verified the robust least squares (RLS), the robust orthogonal regression (alias of the LSS), and the robust geometric mean regression (RGMR) are special members of the robust compound regression (RCR) family, but also proved the optimality of the RGMR in the respect of maximizing the sum of regression efficiencies in simple linear regression. Moreover, we provided the generalized versions of both new approaches for the further investigation, and proposed to use the sum of regression efficiencies as a general goodness-of-fit criterion to compare the estimates from different regression approaches in real data analysis.

The first advantage of both new approaches lies in their intuitive geometric interpretations, by minimizing the sum of squares of the projected orthogonal distances for the LSS, and the sum of weighted average of squares of the projected vertical and horizontal distances for the RCR. Meanwhile, both methods are distribution free, being direction generalizations of the nonparametric orthogonal regression analysis method and the nonparametric compound regression analysis method respectively. Moreover, both estimation approaches are independent to the ratio of the error variances, which is not the case for the usual

MLE approach of EIV models. Furthermore, their estimators are robust to outliers and other departures from the underlying assumptions. Although the LSS and RCR approaches are mainly designed for the robust estimation of EIV models, they are also good for robust estimation of the degenerated EIV model where only the response variable is random while the predictor variables are fixed. Of note, one particular merit of the LSS approach is because its estimator is analytically tractable by using either the principle component analysis (PCA) or the singular value decomposition (SVD), which makes the runtime of the LSS for high-dimensional and large datasets be at the same order of the runtime of the ordinary LS and OR estimation approaches, about 1000 times less than the runtime of the robust LMS approach.

Nevertheless, the common disadvantage for both methods is that their regression lines must pass through the center of the target dataset which will to some extent restrict their robustness performance. Additionally, another downside of the LSS is that it is only a special case of the RCR framework, and hence the LSS is not always the optimal choice as each situation has its unique optimal solution. In contrary, the RCR approach is advantageous in that it can provide a class of optimal robust estimators for the entire class of EIV model in a 1-1 mapping. The regression efficiency concept and the efficiency plots will aid us in searching for the optimal RCR for each data set analyzed.

In the future, we will consider extending more theoretical properties of the compound and constrained regression analysis methods (Leng and Zhu 2009) to the new RCR analysis approach in the high dimensional case. Secondly, we will search for a robust location (‘center of mass’) estimator from the data depth point of view (Liu 1990, Liu et al. 1999) to further enhance the robustness of our methods. Thirdly, it is worthwhile for us to explore the generalization of RCR by making a win-win partnership between the compound regression analysis method and any existing robust regression technique including the notable least median of squares method, which represents a promising direction for the development of other systematic classes of robust estimation methods like the RCR paradigm for the EIV models. Finally, we can develop a system of robust MLE estimators for the EIV models based on our RCR concept as well.

References:

Barker, F., Soh, Y. C., and Evans, R. J. (1988), "Properties of the Geometric Mean Functional Relationship," *Biometrics*, 44, 279-281.

Brown, M. L. (1982), "Robust Line Estimation with Errors in Both Variables," *Journal of the American Statistical Association*, 77, 71-79.

Carroll, R. J., and Gallo, P. P. (1982), "Some Aspects of Robustness in the Functional Errors-in-Variables Regression-Model," *Communications in Statistics Part a-Theory and Methods*, 11, 2573-2585.

Casella, G., and Berger, R. L. (2002), *Statistical Inference*, Duxbury Press.

Cheng, C. L., and Vanness, J. W. (1992), "Generalized M-Estimators for Errors-in-Variables Regression," *Annals of Statistics*, 20, 385-397.

Creasy, M. A. (1956), "Confidence Limits for the Gradient in the Linear Functional Relationship," *Journal of the Royal Statistical Society*, *Series B*, 18, 65-69.

Draper, N. R. (1992) "Straight Line Regression when Both Variables are Subject to Error," *Proceedings of the 1991 Kansas State University Conference on Applied Statistics in Agriculture*, 1-18.

Draper, N. R., and Smith, H. (1998), *Applied Regression Analysis*, Wiley.

Draper, N. R., and Yang, Y. H. (1997), "Generalization of the Geometric Mean Functional Relationship," *Computational Statistics & Data Analysis*, 23, 355-372.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Erickson, J., Har-Peled, S., and Mount, D. M. (2006), "On the Least Median Square Problem," *Discrete & Computational Geometry*, 36, 593-607.

Fekri, M., and Ruiz-Gazen, A. (2004), "Robust Weighted Orthogonal Regression in the Errors-in-Variables Model," *Journal of Multivariate Analysis*, 88, 89-108.

Forsythe, A. B. (1972), "Robust Estimation of Straight Line Regression Coefficients by Minimizing Pth Power Deviations," *Technometrics*, 14, 159-&.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Galton, F. (1886), "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246-263.

Galton, F. (1889), *Natural Inheritance*, London,: Macmillan.

Golub, G. H., and Van Loan, C. F. v. (1996), *Matrix Computations* (3rd ed.), The Johns Hopkins University Press.

Hanley, J. A. (2004), "Galton's Family Data on Human Stature," http://www.medicine.mcgill.ca/epidemiology/hanley/galton/.

Hanley, J. A. (2005), "Reply to Comment of 'Transmuting' Women into Men: Galton's Family Data on Human Stature, by Wilkinson, Wachsmuth, and Dallal," *The American Statistician*, 59, 1.

Hartmann, C., Vankeerberghen, P., SmeyersVerbeke, J., and Massart, D. L. (1997), "Robust Orthogonal Regression for the Outlier Detection When Comparing Two Series of Measurement Results," *Analytica Chimica Acta*, 344, 17-28.

Harvey, P. H., and Mace, G. M. (1982), "Comparison between Taxa and Adaptive Trends: Problems of Methodology," *Current Problems in Sociobiology*, Cambridge University Press, New York.

He, X. M., and Liang, H. (2000), "Quantile Regression Estimates for a Class of Linear and Partially Linear Errors-in-Variables Models," *Statistica Sinica*, 10, 129-140.

Jackson, J. D., and Dunlevy, J. A. (1988), "Orthogonal Least Squares and the Interchangeability of Alternative Proxy Variables in the Social Sciences," *The Statistician*, 37, 7-14.

Jung, K. M. (2007), "Least Trimmed Squares Estimator in the Errors-in-Variables Model," *Journal of Applied Statistics*, 34, 331-338.

Kelly, G. (1984), "The Influence Function in the Errors in Variables Problem," *THe Annals of Statistics*, 12, 87-100.

Ketellapper, R., and Ronner, A. (1984), "Are Robust Estimation Methods Useful in the Structural Errors-in-Variables Model?," *Metrika*, 31, 33-41.

Ketellapper, R. H., and Weisbeek, J. A. (1983), "Further Evidence on the Appropriateness of a Robust Estimation Procedure for the Structural Errors-in-Variables Model," *Communications in Statistics-Theory and Methods*, 12, 1511-1522.

Leng, L., Zhang, T., Kleinman, L., and Zhu, W. (2007), "Ordinary Least Square Regression, Orthogonal Regression, Geometric Mean Regression, and Their Applications in Aerosol Science," *Journal of Physics: Conference Series*, 78

Leng, L., and Zhu, W. (2009), "Compound and Constrained Regression Analyses," *PhD Dissertation at Stony Brook University*.

Lindley, D. B. (1947), "Regression Lines and the Functional Relationship," *Journal of the Royal Statistical Society Series B-Methodological*, 9, 219-244.

Liu, R. Y. (1990), "On a Notion of Data Depth Based on Random Simplices," *Annals of Statistics*, 18, 405-414.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference," *Annals of Statistics*, 27, 783-840.

Maronna, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *Annals of Statistics*, 4, 51-67.

Martin, R. D. (2002), *Robust Statistics with the S-Plus Robust Library and Financial Applications* (Vol. 1 and 2), New York, NY: Insightful Corp Presentation.

Morrison, D. F. (1976), *Multivariate Statistical Methods,* New York: Mcgraw-Hill.

Patefield, W. M. (1981), "Multivariate Linear Relationships - Maximum-Likelihood Estimation and Regression Bounds," *Journal of the Royal Statistical Society Series B-Methodological*, 43, 342-352.

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

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Sarabia, L. A., Ortiz, M. C., and Tomas, X. (1997), "Performance of the Orthogonal Least Median Squares Regression," *Analytica Chimica Acta*, 348, 11-18.

Sprent, P. (1969), *Models in Regression and Related Topics*, London: Methuen.

Sprent, P., and Dolby, G. R. (1980), "Query: The Geometric Mean Functional Relationship," *Biometrics*, 36, 547-550.

Van Huffel, S., and Lemmerling, P. (2002), *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, Dordrecht: Kluwer Academic Publishers.

Wong, M. Y. (1989), "Likelihood Estimation of a Simple Linear Regression Model When Both Variables Have Error," *Biometrica*, 76.

Zamar, R. H. (1989), "Robust Estimation in the Errors-in-Variables Model," *Biometrika*, 76, 149-160.

Appendix: Proof of Results

Theorem 1: The ordinary least squares (LS) estimator is also regression-, scale-, and affine-equivariant; the OR does not show any of these three equivariances; and the GMR is scale-equivariant only.

Proof.

For the LS, this follows from

$\sum_i [(Y_i + X_i V) - X_i(\beta + V)]^2 = \sum_i (Y_i - X_i\beta)^2$,

$\sum_i [cY_i - X_i(c\beta)]^2 = c^2 \sum_i (Y_i - X_i\beta)^2$,

and $\sum_i [Y_i - (X_i A)(A^{-1}\beta)]^2 = \sum_i (Y_i - X_i\beta)^2$, respectively.

For the OR, for any column vector *V*, any constant *c*, and any nonsingular square matrix *A*, in generally we do not see any equivariance as

$$\frac{\sum_i [(Y_i + X_i V) - X_i(\beta + V)]^2}{1 + (\beta + V)'(\beta + V)} \neq \frac{\sum_i (Y_i - X_i\beta)^2}{1 + \beta'\beta}$$

$$\frac{\sum_i [cY_i - X_i(c\beta)]^2}{1 + (c\beta)'(c\beta)} \neq f(c)\frac{\sum_i (Y_i - X_i\beta)^2}{1 + \beta'\beta}$$

$$\frac{\sum_i [Y_i - (X_i A^{-1})(A^{-1}\beta)]^2}{1 + (A^{-1}\beta)'(A^{-1}\beta)} \neq \frac{\sum_i (Y_i - X_i\beta)^2}{1 + \beta'\beta}$$

For the GMR, for any column vector *V*, and any nonsingular square matrix *A*, in generally we do not view the regression-, and affine- equivariance as

$$\frac{\sum_i [(Y_i + X_i V) - X_i(\beta + V)]^2}{[\prod_{j=1}^{p} (\beta_j + V_j)^2]^{\frac{1}{p+1}}} \neq \frac{\sum_i (Y_i - X_i\beta)^2}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p+1}}}$$

$$\frac{\sum_i [Y_i - (X_i A)(A^{-1}\beta)]^2}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p+1}}} \neq \frac{\sum_i (Y_i - X_i\beta)^2}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p+1}}}$$

But for any constant *c*, we always have the scale- equivariance for the GMR

$$\frac{\sum_i [cY_i - X_i(c\beta)]^2}{[\prod_{j=1}^{p} (c\beta_j)^2]^{\frac{1}{p+1}}} = c^{\frac{2}{p+1}} \frac{\sum_i (Y_i - X_i\beta)^2}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p+1}}}$$

Theorem 2: The least sine squares (LSS) slope estimate is the eigenvector corresponding to the smallest eigenvalue of the robust sample covariance matrix, and the LSS is a special case of the robust compound regression (RCR).

Proof. (1)

Without loss of generality, we form the multivariate regression model as $\sum_{j=1}^{p} \beta_j X_j = 0$ or in matrix form $\boldsymbol{X\beta}=\mathbf{0}$ for the centered data, where $\boldsymbol{X}=[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p]$ is a *n* by *p* matrix of observations, and $\boldsymbol{\beta}$ is a *p* by 1 column vector of regression coefficients. This linear relationship is uniquely specified by imposing the constraint $\boldsymbol{\beta}'\boldsymbol{\beta}=1$.

As we know, there is a close relationship between the principle component analysis (PCA) and the orthogonal regression (OR) (Jackson and Dunlevy 1988). Since the LSS is the robust analogy of the OR, we want to prove the LSS slope estimate can also be obtained through the principle component analysis on the robust sample covariance matrix.

We first define $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \ldots, \widetilde{\boldsymbol{X}}_p] = \begin{bmatrix} \frac{X_{11}}{R_1} & \cdots & \frac{X_{p1}}{R_1} \\ \vdots & \ddots & \vdots \\ \frac{X_{1n}}{R_n} & \cdots & \frac{X_{pn}}{R_n} \end{bmatrix}_{nxp}$ as the *n* by *p* transformed matrix of observations, where $R_i = \sqrt{\sum_{j=1}^{p} X_{ji}^2}$ is the distance from the *i*th observation to the origin. Then the LSS is defined to minimize

$$SS_{LSS} = (\widetilde{\boldsymbol{X}}\boldsymbol{\beta})'(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}) = \boldsymbol{\beta}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{\beta} = \boldsymbol{\beta}'(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})\boldsymbol{\beta} = \boldsymbol{\beta}'\widetilde{\boldsymbol{S}}\boldsymbol{\beta}$$

where $\widetilde{\boldsymbol{S}}$ is the *p* by *p* robust sample covariance matrix

$$\tilde{S} = \begin{bmatrix} \widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1 & \cdots & \widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_p \\ \vdots & \ddots & \vdots \\ \widetilde{\boldsymbol{X}}_p'\widetilde{\boldsymbol{X}}_1 & \cdots & \widetilde{\boldsymbol{X}}_p'\widetilde{\boldsymbol{X}}_p \end{bmatrix}_{pxp}$$

We define the eigenvectors of $\widetilde{\boldsymbol{S}}$ as ($\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, …, $\boldsymbol{\alpha}_p$) in the order of descending eigenvalue (λ_1, λ_2,…, λ_p). Assume $\widetilde{\boldsymbol{S}}$ is non-singular, the eigenvectors can expand the *p*-dimensional space, and then the slope estimate $\boldsymbol{\beta}$ can be expressed as a linear combination $l_1\boldsymbol{\alpha}_1$+ $l_2\boldsymbol{\alpha}_2$+…+ $l_p\boldsymbol{\alpha}_p$ subject to $\sum_{j=1}^{p} l_j = 1$. Hence, the minimization of SS_{LSS} is equivalent to the minimization of

$$(l_1\boldsymbol{\alpha}_1 + l_2\boldsymbol{\alpha}_2 + \ldots + l_p\boldsymbol{\alpha}_p)'\, \widetilde{\boldsymbol{S}}\, (l_1\boldsymbol{\alpha}_1 + l_2\boldsymbol{\alpha}_2 + \ldots + l_p\boldsymbol{\alpha}_p)$$

Since as we know that the eigenvectors are orthogonal and $\boldsymbol{\alpha}_j'\widetilde{\boldsymbol{S}}\boldsymbol{\alpha}_j = \lambda_j$, the problem becomes the minimization of $\sum_{j=1}^{p} \lambda_j l_j$. Under the constraints $\sum_{j=1}^{p} l_j = 1$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq$

λ_p, the minimum is achieved when we set l_p=1 and thus $\boldsymbol{\beta}$=$\boldsymbol{\alpha}_p$. That is the eigenvector corresponding to λ_p - the smallest eigenvalue of $\tilde{\boldsymbol{S}}$.

(2)

It was already shown that the OR in higher dimensions is a special case of the structural approach and thus a special case of the compound regression (CR) by the equivalence between the structural approach and CR under the normality assumption (Leng and Zhu 2009).

In the context of principle component analysis, as we also know that the OR estimate is the eigenvector associated with the smallest eigenvalue of the ordinary sample covariance matrix $\boldsymbol{S}$, the OR estimate can be written as $\hat{\beta}_{OR}$=g($\boldsymbol{S}$), where g corresponds to the standard matrix manipulations to find the eigenvector of the smallest eigenvalue of any non-singular matrix. Similarly, as already shown in part (1), the LSS estimate can be expressed as $\hat{\beta}_{LSS}$=g ($\tilde{\boldsymbol{S}}$) with the identical matrix manipulations g as well.

On the other hand, in the context of compound regression analysis, given a set of $\boldsymbol{\gamma}_{OR}$ satisfying $\sum_{j=0}^{p} \boldsymbol{\gamma}_j = 1$, we can obtain the OR estimate $\hat{\beta}_{OR} = f(\boldsymbol{S})|\gamma_{OR}$, where f is a function of the sample covariance matrix $\boldsymbol{S}$ after solving a system of equations. Meanwhile, from section 4.1, we have shown that the estimation of RCR approach is also simplified to solve a system of equations simultaneously as follows.

$$\frac{\partial SS_{\tilde{\gamma}}}{\partial \beta_l} = -\frac{2\gamma_l}{\beta_l^3}(\tilde{S}_{YY} + \sum_{j=1}^{p}\sum_{k=1}^{p}\beta_j\beta_k\tilde{S}_{X_jX_k} - 2\sum_{j=1}^{p}\beta_j\tilde{S}_{X_jY})$$
$$+(\gamma_0 + \sum_{j=1}^{p}\frac{\gamma_j}{\beta_j^2})(2\sum_{\substack{j=1\\ j\neq l}}^{p}\beta_j\tilde{S}_{X_lX_j} + 2\beta_l\tilde{S}_{X_lX_l} - 2\tilde{S}_{X_lY}) = 0, \quad l = 1,2,...,p$$

We can see that the system of equations is exactly the same as in the CR except the $\boldsymbol{S}$ is replaced by $\tilde{\boldsymbol{S}}$. Suppose we are using the same set of γ_{OR} for the estimation of RCR, we obtain $\hat{\beta} = f(\tilde{\boldsymbol{S}})|\gamma_{OR}$. Now we want to show that this RCR estimate is indeed the LSS estimate.

Since the OR estimate should be the same in the context of different analyses, we have $\hat{\beta}_{OR} = f(\boldsymbol{S})|\gamma_{OR} = g(\boldsymbol{S})$. Hence, by induction we can easily draw the conclusion that $\hat{\beta} = f(\tilde{\boldsymbol{S}})|\gamma_{OR} = g(\tilde{\boldsymbol{S}}) = \hat{\beta}_{LSS}$. This implies the LSS is a special case of the RCR with the γ_{LSS} exactly the same as the γ_{OR} in the compound regression.

Theorem 3: There is a monotonic relationship between γ and $\hat{\beta}$ in robust compound regression.

Proof.

The equation (b) in Section 4.1 can be rewritten as

$$\frac{\gamma}{1-\gamma} = \frac{\tilde{S}_{YY} - \hat{\beta}\tilde{S}_{XY}}{\hat{\beta}\tilde{S}_{XX} - \tilde{S}_{XY}}\frac{1}{\hat{\beta}^3} = [-\hat{\beta}_{RLS\_Y} + \frac{\tilde{S}_{XY}(\hat{\beta}_{RLS\_X} - \hat{\beta}_{RLS\_Y})}{\tilde{S}_{XX}(\hat{\beta} - \hat{\beta}_{RLS\_Y})}]\frac{1}{\hat{\beta}^3}$$

We first denote
$$f(\hat{\beta}) = -\hat{\beta}_{RLS\_Y} + \frac{\tilde{S}_{XY}(\hat{\beta}_{RLS\_X} - \hat{\beta}_{RLS\_Y})}{\tilde{S}_{XX}(\hat{\beta} - \hat{\beta}_{RLS\_Y})}$$

When $\tilde{S}_{XY} \geq 0$ and $0 \leq \hat{\beta}_{\mathrm{RLS\_}Y} \leq \hat{\beta} \leq \hat{\beta}_{\mathrm{RLS\_}X}$, we have $f(\hat{\beta}) \geq 0 \downarrow$ and $\frac{1}{\hat{\beta}^3} \geq 0 \downarrow$ as $\hat{\beta} \uparrow$, thus γ is a decreasing function of $\hat{\beta}$ and vice versa;

When $\tilde{S}_{XY} < 0$ and $0 > \hat{\beta}_{\mathrm{RLS\_}Y} \geq \hat{\beta} \geq \hat{\beta}_{\mathrm{RLS\_}X}$, we have $f(\hat{\beta}) < 0 \downarrow$ and $\frac{1}{\hat{\beta}^3} < 0 \downarrow$ as $\hat{\beta} \uparrow$, thus γ is an increasing function of $\hat{\beta}$ and vice versa.

Thus, we have proven the theorem.

Theorem 4: The robust geometric mean regression (RGMR) is a special case of the robust compound regression (RCR).

Proof.

We have already proved this for the simple linear regression analysis. In the higher dimension case, a multivariable criterion for estimating the geometric mean regression (GMR) is given by Draper and Yang (1997). Consider the multivariate linear regression model defined in Theorem 3, their criterion can be written as

$$SS_{GMR} = \frac{(\boldsymbol{X\beta})'(\boldsymbol{X\beta})}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p}}} = \frac{1}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p}}} \boldsymbol{\beta}'\boldsymbol{S}\boldsymbol{\beta} = \boldsymbol{\tau}'\boldsymbol{S}\boldsymbol{\tau} \quad (1)$$

where $\boldsymbol{S}$ is the sample covariance matrix, and $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_p)'$ with $\tau_j = \frac{\beta_j}{(\prod_{j=1}^{p} \beta_j)^{\frac{1}{p}}}$.

Lemma 1. If all of the *p* LS regression solutions which use in turn each of the *p* variables as a dependent variable lie in the same hyper-octant, then for the ratio of error variances matrix $\boldsymbol{\Lambda}$ non-singular, the maximum likelihood estimator (as described in Section 2.2) of multivariate structural EIV model is a convex combination of the *p* LS solutions (Patefield 1981, Fuller 1987).

Based on the general conclusion in lemma 1, the following lemma 2 was proposed for the multivariate GMR.

Lemma 2. Assume the sample covariance matrix $\boldsymbol{S}$ is non-singular, if all of the *p* LS solutions lie in the same hyper-octant, there is unique solution $\widehat{\boldsymbol{\beta}}_{GMR}$ to the quadratic problem described in (1), and it lies within the simplex defined by the *p* LS solutions (Draper and Yang 1997).

For the RGMR, we can its multivariate criterion as minimizing

$$SS_{RGMR} = \frac{(\tilde{\boldsymbol{X}}\boldsymbol{\beta})'(\tilde{\boldsymbol{X}}\boldsymbol{\beta})}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p}}} = \frac{1}{(\prod_{j=1}^{p} \beta_j^2)^{\frac{1}{p}}} \boldsymbol{\beta}'\tilde{\boldsymbol{S}}\boldsymbol{\beta} = \boldsymbol{\tau}'\tilde{\boldsymbol{S}}\boldsymbol{\tau} \quad (2)$$

where the robust sample covariance matrix $\widetilde{\boldsymbol{S}}$ was defined in the same way as in Theorem 3.

Before we propose our corollary as the natural analogies of Lemmas 1 and 2, it is necessary for us to firstly define the robust least squares (RLS) regression of X_l on (X_l, ..., X_{l-1}, X_{l+1}, ..., X_l) as to solve the problem $\widetilde{\boldsymbol{X}}_l^c \boldsymbol{\beta}_{\mathrm{RLS\_}X_l} = \widetilde{\boldsymbol{X}}_l$, where $\widetilde{\boldsymbol{X}}_l^c = [\widetilde{\boldsymbol{X}}_1, \dots, \widetilde{\boldsymbol{X}}_{l-1}, \widetilde{\boldsymbol{X}}_{l+1}, \dots, \widetilde{\boldsymbol{X}}_l]$ is defined to be the complementary matrix of column vector $\widetilde{\boldsymbol{X}}_l$. Then we can easily know that the corresponding RLS solution is $\widehat{\boldsymbol{\beta}}_{\mathrm{RLS\_}X_l} = (\widetilde{\boldsymbol{X}}_l^{c\prime} \widetilde{\boldsymbol{X}}_l^c)^{-1} \widetilde{\boldsymbol{X}}_l^{c\prime} \widetilde{\boldsymbol{X}}_l$.

Corollary. When all of the p RLS solutions $\boldsymbol{\beta}_{\mathrm{RLS\_}X_l}$, $l = 1, 2, \ldots, p$ lie in the same hyper-octant, the set of feasible values for the β_i, $i = 1, 2, \ldots, p$ is a convex combination of the p RLS solutions, and if the robust sample covariance matrix $\widetilde{\boldsymbol{S}}$ is non-singular, there is a unique RGMR solution $\widehat{\boldsymbol{\beta}}_{RGMR}$ to the quadratic problem defined in (2), which lies within the simplex defined by the p RLS solutions.

Hence, by the above corollary, we know that each RCR estimates is a convex combination of the p RLS solutions, and meanwhile since the compound parameter $\boldsymbol{\gamma}$=(γ_1, γ_2, …, γ_p) is continuously defined on the (p-1) dimension hyper-plane $\sum_{j=1}^{p} \gamma_p = 1$, the RCR estimates by a 1-1 mapping ought to continuously expand and form the feasible set for the β_i, $i = 1, 2, \ldots, p$., which include the unique RGMR solution. Therefore, the theorem has been proved.

Theorem 5: In simple linear regression, the robust geometric mean regression (RGMR) will always yield the equal e_Y and e_X, and the maximum sum of regression efficiencies $e_Y + e_X$.

Proof. (1)

As we know the RGMR line passes through the mean ($\overline{X}$, $\overline{Y}$), we can write

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \tilde{S}_{YY} + \hat{\beta}^2 \tilde{S}_{XX} - 2\hat{\beta}\tilde{S}_{XY} \qquad \sum_{i=1}^{n}(X_i - \hat{X}_i)^2 = \frac{1}{\hat{\beta}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{\hat{\beta}^2}\tilde{S}_{YY} + \tilde{S}_{XX} - 2\frac{1}{\hat{\beta}}\tilde{S}_{XY}$$

Based on the RGMR slope estimate $\hat{\beta} = sign(\tilde{S}_{XY})\sqrt{\tilde{S}_{YY} / \tilde{S}_{XX}}$, we obtain

$$e_Y = \frac{\min \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,|_{\hat{\beta}=\frac{\tilde{S}_{XY}}{\tilde{S}_{XX}}}}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,|_{\hat{\beta}=sign(\tilde{S}_{XY})\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}}}} = \frac{\tilde{S}_{YY} + \frac{\tilde{S}_{XY}^2}{\tilde{S}_{XX}} - 2\frac{\tilde{S}_{XY}^2}{\tilde{S}_{XX}}}{\tilde{S}_{YY} + \tilde{S}_{YY} - 2\left|\tilde{S}_{XY}\right|\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}}} = \frac{\tilde{S}_{XX}\tilde{S}_{YY} - \tilde{S}_{XY}^2}{2\tilde{S}_{XX}\tilde{S}_{YY} - 2\left|\tilde{S}_{XY}\right|\sqrt{\tilde{S}_{XX}\tilde{S}_{YY}}}$$

$$e_X = \frac{\min \sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2} = \frac{\frac{1}{\hat{\beta}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,|_{\hat{\beta}=\frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}}}{\frac{1}{\hat{\beta}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,|_{\hat{\beta}=sign(\tilde{S}_{XY})\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}}}}$$

$$= \frac{\frac{\tilde{S}_{XY}^2}{\tilde{S}_{XX}\tilde{S}_{YY}}(\tilde{S}_{YY} + \frac{\tilde{S}_{YY}^2}{\tilde{S}_{XY}^2}\tilde{S}_{XX} - 2\tilde{S}_{YY})}{\tilde{S}_{YY} + \tilde{S}_{YY} - 2\left|\tilde{S}_{XY}\right|\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}}} = \frac{\tilde{S}_{XX}\tilde{S}_{YY} - \tilde{S}_{XY}^2}{2\tilde{S}_{XX}\tilde{S}_{YY} - 2\left|\tilde{S}_{XY}\right|\sqrt{\tilde{S}_{XX}\tilde{S}_{YY}}}$$

Hence, we have shown that e$_Y$= e$_X$ for the RGMR.

(2)

For any regression estimate $\hat{\beta}$, we have

$$\Sigma = e_Y + e_X$$

$$= \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,|_{\hat{\beta}=\frac{\tilde{S}_{XY}}{\tilde{S}_{XX}}}}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} + \frac{\frac{1}{\hat{\beta}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,|_{\hat{\beta}=\frac{\tilde{S}_{YY}}{\tilde{S}_{XY}}}}{\frac{1}{\hat{\beta}^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \frac{\tilde{S}_{YY} + \frac{\tilde{S}_{XY}^2}{\tilde{S}_{XX}} - 2\frac{\tilde{S}_{XY}^2}{\tilde{S}_{XX}}}{\tilde{S}_{YY} + \hat{\beta}^2\tilde{S}_{XX} - 2\hat{\beta}\tilde{S}_{XY}} + \frac{\frac{\tilde{S}_{XY}^2}{\tilde{S}_{YY}} + \tilde{S}_{XX} - 2\frac{\tilde{S}_{XY}^2}{\tilde{S}_{YY}}}{\frac{1}{\hat{\beta}^2}\tilde{S}_{YY} + \tilde{S}_{XX} - 2\frac{1}{\hat{\beta}}\tilde{S}_{XY}}$$

$$= \frac{(\tilde{S}_{YY} - \frac{\tilde{S}_{XY}^2}{\tilde{S}_{XX}}) + \beta^2(\tilde{S}_{XX} - \frac{\tilde{S}_{XY}^2}{\tilde{S}_{YY}})}{\tilde{S}_{YY} + \hat{\beta}^2\tilde{S}_{XX} - 2\hat{\beta}\tilde{S}_{XY}}$$

The maximum of Σ is achieved when $\frac{\partial\Sigma}{\partial\hat{\beta}} = 0$, and by straight-forward derivations it can be simplified as $\hat{\beta}^2 = \frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}$

When $\tilde{S}_{XY} \geq 0$, Σ = e$_Y$ + e$_X$ is unimodal and maximized at $\hat{\beta} = \sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}} = sign(\tilde{S}_{XY})\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}} = \hat{\beta}_{RGMR}$; while when $\tilde{S}_{XY} < 0$, Σ is unimodal and maximized at $\hat{\beta} = -\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}} = sign(\tilde{S}_{XY})\sqrt{\frac{\tilde{S}_{YY}}{\tilde{S}_{XX}}} = \hat{\beta}_{RGMR}$.

Therefore, we have proven that the sum e$_Y$+ e$_X$ is maximized for the RGMR.