Journal Title
XX(X):??–??
©The Author(s) 2013
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

A working likelihood approach for robust regression

Liya Fu[1,2], You-Gan Wang[2], Fengjing Cai[3]

Abstract

Robust approach is often desirable in presence of outliers for more efficient parameter estimation. However, the choice of the regularization parameter value impacts the efficiency of the parameter estimators. To maximize the estimation efficiency, we construct a likelihood function for simultaneously estimating the regression parameters and the tuning parameter. The 'working' likelihood function is deemed as a vehicle for efficient regression parameter estimation, because we do not assume the data are generated from this likelihood function. The proposed method can effectively find a value of the regularization parameter based on the extend of contamination in the data. We carry out extensive simulation studies in a variety of cases to investigate the performance of the proposed method. The simulation results show that the efficiency can be enhanced as much as 40% when the data follow a heavy-tailed distribution, and reaches as high as 468% for the heteroscedastic variance cases compared to the traditional Huber's method with a fixed regularization parameter. For illustration, we also analyzed two datasets: one from a diabetics study and the other from a mortality study.

Keywords

data driven, Huber's loss function, robust method, tuning parameter, working likelihood

[1]School of Mathematics and Statistics, Xi'an Jiaotong University, China
[2]chool of Mathematical Science, Queensland University of Technology, Brisbane, QLD 4001, Australia,
[3] College of Mathematics, Wenzhou University, China.

Corresponding author:
You-Gan Wang, School of Mathematical Science, Queensland University of Technology, Australia
Fengjing Cai, College of Mathematics, Wenzhou University, China
Email: you-gan.wang@qut.edu.au
caifj7704@wzu.edu.cn

1 Introduction

Suppose that observations $\{(x_i, y_i), i = 1, \cdots, n\}$ satisfy a robust linear regression model introduced by ?,

$$y_i = x_i^{\mathrm{T}}\beta + \sigma\epsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where β is a $p \times 1$ unknown parameter vector of interest, and $\sigma > 0$ is a scale parameter. We assume that the error terms $\epsilon_1, \cdots, \epsilon_n$ are independent and identically distributed with an unknown distribution F_ϵ that satisfies $F_\epsilon(0) = 1/2$.

The method of least squares can obtain an estimator of β, denoted as $\hat{\beta}_{LS} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$, where $X = (x_1, \cdots, x_n)^{\mathrm{T}}$ and $Y = (y_1, \cdots, y_n)^{\mathrm{T}}$. When F_ϵ is a normal distribution, $\hat{\beta}_{LS}$ is an uniformly optimal linear unbiased estimator. However, when F_ϵ deviates from the normal distribution and/or is contaminated by outliers, such as F_ϵ is a heavy-tailed distribution, the performance of the least square estimator is poor, and thus robust methods are more desirable in these cases[? ?] .

Several families of robust estimators have been developed. One of the robust methods is the M-estimation method, which can provide much better regression coefficient estimates when outliers are present in the data by minimizing the "maximum likelihood type" loss function. The most widely used loss function is the Huber's loss function given by

$$\rho_\tau(\epsilon) = \begin{cases} \dfrac{1}{2}\epsilon^2 & |\epsilon| \le \tau \\ |\epsilon|\tau - \dfrac{\tau^2}{2} & |\epsilon| > \tau \end{cases}, \tag{2}$$

where $\tau > 0$ is a regularization parameter and must be specified. The regularization parameter τ (also known as the tuning parameter) regulates the amount of robustness. When the data follow a normal distribution, the best value of τ is $+\infty$. When the data follow a Laplace distribution, the best value of τ should be very small. ? proposed selecting a value of τ between 1 and 2. The default value of τ is 1.345 in R package (rlm function), which can reach 95% efficiency when the data follow a normal distribution.

The corresponding subgradient function of $\rho(\cdot)$ is

$$\psi_\tau(\epsilon) = \begin{cases} \epsilon & |\epsilon| \le \tau \\ \mathrm{sign}(\epsilon)\tau & |\epsilon| > \tau \end{cases}.$$

The robust estimator of β can be obtained by solving the following equation

$$\sum_{i=1}^{n} x_i \psi_\tau\left(\frac{y_i - x_i^{\mathrm{T}}\beta}{\hat{\sigma}}\right) = \mathbf{0}, \tag{3}$$

where $\hat{\sigma}$ is a consistent estimator of σ. Let $\hat{\beta}_H$ be the estimator derived from equation (??), and then $\sqrt{n}(\hat{\beta}_H - \beta) \xrightarrow{d} N(\mathbf{0}, \Sigma)$, where

$$\Sigma = \sigma^2[E(X^{\mathrm{T}}X)]^{-1}\frac{E[\psi_\tau^2(\epsilon)]}{\{E[\psi_\tau'(\epsilon)]\}^2}.$$

It is worth noting that the asymptotic variance Σ depends on the regularization parameter τ via an efficiency factor $\zeta^{-1} = \frac{E[\psi_\tau^2(\epsilon)]}{\{E[\psi_\tau'(\epsilon)]\}^2}$. The value of τ has a great influence on the efficiency of regression parameter estimation. Hence, ? proposed maximizing $\hat{\zeta}$ to choose the regularization parameter, where $\hat{\zeta}$ is a moment estimator of ζ,

$$\hat{\zeta} = \frac{[\sum_{i=1}^n I(|\hat{e}_i| \leq \tau)]^2}{n \sum_{i=1}^n [I(|\hat{e}_i| \leq \tau)\psi^2(\hat{e}_i) + \tau^2 I(|e_i| > \tau)]}, \tag{4}$$

where $\hat{e}_i = (y_i - x_i^{\mathrm{T}} \hat{\beta}_H)/\hat{\sigma}$ and $\hat{\sigma} = 1.4826 * \mathrm{median}_i\{\hat{e}_i - \mathrm{median}_j(\hat{e}_j)\}$ is the median absolute deviation (MAD) estimator of σ. This data-driven method can automatically select a value of τ, and ? constructed an R package rlmDataDriven, which provides a rlmDD function to obtain $\hat{\beta}_H$ with an automatically chosen τ.

In this paper, we construct a robust density function that contains β, τ and a dispersion parameter θ. The proposed density function can fit the distribution of the data very well by automatically selecting the regularization parameter and the dispersion parameter. Thus, we propose selecting an "optimal" value for the regularization parameter τ by minimizing the negative log-working likelihood. The proposed method can automatically adjust the regularization parameter τ according to the data, and thus can improve the efficiency of regression parameter estimators. The working likelihood function is introduced in Section 2. A variety of simulation studies are carried out to evaluate the performance of the proposed method in Section 3. Two real data examples are used to illustrate the proposed method in Section 4. Finally, some conclusions are drawn in Section 5.

2 A working likelihood function

Considering the dispersion of the data, we rescale the Huber's loss function and define

$$\rho_{\tau,\theta}(r) = \begin{cases} \dfrac{r^2}{2\theta^2} & \dfrac{|r|}{\theta} \leq \tau \\ \dfrac{|r|\tau}{\theta} - \dfrac{\tau^2}{2} & \dfrac{|r|}{\theta} > \tau \end{cases},$$

where $r = y - x\beta$, and $\theta > 0$ is a parameter to control the dispersion of the data. We construct a density function based on $\rho_{\tau,\theta}(r)$

$$f(r; \beta, \tau, \theta) = C^{-1}(\tau, \theta) e^{-\rho_{\tau,\theta}(r)},$$

where $C(\tau, \theta)$ is a normalized constant and

$$\begin{aligned} C(\tau, \theta) &= \int_{-\infty}^{+\infty} e^{-\rho_{\tau,\theta}(r)} dr = \int_{|r| \leq \theta\tau} e^{-\frac{r^2}{2\theta^2}} dr + \int_{|r| > \theta\tau} e^{-|r|\tau/\theta + \frac{\tau^2}{2}} dr \\ &= \theta\sqrt{2\pi}[2\Phi(\tau) - 1] + \frac{2\theta}{\tau} e^{-\frac{\tau^2}{2}} = A(\tau)\theta. \end{aligned}$$

Therefore, the density function

$$f(r;\beta,\tau,\theta) \quad = \quad \frac{1}{A(\tau)\theta}e^{\frac{-r^2}{2\theta^2}}I(|r| \leq \theta\tau) + \frac{1}{A(\tau)\theta}e^{\tau(\frac{\tau}{2}-\frac{|r|}{\theta})}I(|r| > \theta\tau),$$

where $A(\tau) = \sqrt{2\pi}[2\Phi(\tau) - 1] + 2\tau^{-1}e^{-\frac{\tau^2}{2}}$. We refer to $f(r;\beta,\tau,\theta)$ as a robust working density function of r, which is a mixture distribution of a trimmed normal distribution $N(0,\theta^2)$ and a truncated Laplace distribution $\mathrm{LP}(0,\theta/\tau)$, and $2\tau^{-1}e^{\tau^2/2}/A(\tau)$ can be regarded as the proportion of outliers. Figure ?? shows the performance of the proposed density function fitting a normal distribution, a laplace distribution and two t-distributions with difference degrees of freedom.

Assume that $r_1, \cdots, r_n$ are independent and identically distributed random variables. Their joint working likelihood function is

$$L(\beta,\tau,\theta) = \prod_{i=1}^{n} f(r_i;\beta,\tau,\theta) = \frac{1}{C^n(\tau,\theta)}e^{-\sum_{i=1}^{n}\rho_{\tau,\theta}(r_i)}.$$

Therefore, $L(\beta,\tau,\theta)$ is a unified likelihood function of parameters β, τ and θ, and their estimators can be obtained by minimizing the negative log likelihood function,

$$\min_{\beta,\tau,\theta}\{-\log L(\beta,\tau,\theta)\} = \min_{\beta,\tau,\theta}\left\{\sum_{i=1}^{n}\rho_{\tau,\theta}(r_i) + n\log C(\tau,\theta)\right\}. \tag{5}$$

The corresponding estimating equations for β, τ and θ can be expressed as

$$\sum_{i=1}^{n} x_i\psi_{\tau,\theta}(r_i) = \mathbf{0}, \tag{6}$$

where $\psi_{\tau,\theta}(r)$ is the derivative of $\rho_{\tau,\theta}(r)$,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{|r_i|}{\theta} - \tau\right\}I(|r_i| > \theta\tau) = \frac{2e^{-\frac{\tau^2}{2}}}{A(\tau)\tau^2}, \tag{7}$$

and

$$\theta^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i^2 I(|r_i| \leq \theta\tau) + |r_i|\theta\tau I(|r_i| > \theta\tau)\right\}. \tag{8}$$

Therefore, the estimators of β, τ and θ can be obtained by minimizing equation (??) or solve equations (??), (??) and (??). If $f(r;\beta,\tau,\theta)$ is the true density function of $r_1, \cdots, r_n$, we can obtain

$$\frac{1}{n}\sum_{i=1}^{n}E\left\{\left[\frac{|r_i|}{\theta} - \tau\right]I(|r_i| > \theta\tau)\right\} = \left[\frac{2(\tau^2+1)e^{-\tau^2/2}}{\tau^2 A(\tau)} - \frac{2e^{-\tau^2/2}}{A(\tau)}\right] = \frac{2e^{-\frac{\tau^2}{2}}}{A(\tau)\tau^2};$$

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{r_i^2 I(|r_i| \le \theta\tau) + |r_i|\tau\theta I(|r_i| > \theta\tau)\right\} \quad = \quad \theta^2.$$

We now describe the working likelihood procedure as follows.
Step 1. Obtain an initial estimate of β using the median regression, denoted by $\hat\beta^{(k)}, \quad k = 0$.
Step 2. Obtain the residuals $\hat r_i = y_i - x_i^{\mathrm{T}}\hat\beta^{(k)}$.
Step 3. Obtain estimates of τ and θ via minimizing the negative log-likelihood (??).
Step 4. Obtain estimate $\hat\beta_{PL}^{(k+1)}$ of β by equation (??) via the iterative weighted least squares (IWLS) method with $\tau = \hat\tau^{(k)}$ and $\theta = \hat\theta^{(k)}$ obtained from Step 3.
Step 5. Repeat Steps 2-4 until the algorithm converges.

Remark: In Step 1, any consistent estimator of β can be as an initial value, such as the MM estimator, the Huber's estimator with $\tau = 1.345$, and the finial estimator of the regression coefficient β is consistent.

3 Numerical studies

To investigate the performance of the proposed method, we calculate the relative efficiency (REF) of the estimator from the least-squares method (LS), the estimator (DD) by the Huber's method with τ selected by the data-driven method of ?, and the estimator (PL) by the Huber's method with τ selected by the proposed method, to the estimator (HF) via the Huber's method with fixed $\tau = 1.345$. The larger value of REF, the more efficient of the estimator relative to the estimator with $\tau = 1.345$. As one referee suggested, we also compare the proposed estimator with a fully efficient and enjoying high breakdown weighted-least estimator (RWLS) proposed by ?. The RWLS is with a hard rejection weight $\omega(u) = I(u < 1)$ and starting from the least median of squares estimator? . To consider the influence of the initial value on the estimates, we also use HF as an initial value to obtain the estimate in our procedures according to the reviewer's opinion. The corresponding estimate is denoted as PLH.

In our simulation studies, we consider a linear model

$$y_i = \beta_0 + z_i\beta_1 + \sigma\epsilon_i, \quad i = 1, \cdots, n,$$

with $\beta_0 = 1,\ \beta_1 = 2$ and sample size $n = 1000$. The covariates $z_1, \cdots, z_n$ are independently generated from $N(0, 1)$. A variety of error distribution types are considered:
Case 1. Normal errors, $N(0, 1)$, and σ takes a value of 1, 3 and 4.
Case 2. Normal errors, $N(0, 1)$, and the errors are contaminated by either t-distribution with three degrees of freedom (t(3)), $\chi^2(3) - 3$ or each value becomes 6 or -6 with a probability of γ. Different contaminated rates are considered, namely, $\gamma = 10\%$, 20% and 40%.
Case 3. Cauchy errors, $\epsilon \sim f(\epsilon) = [\pi s(1 + \epsilon/s)^2]^{-1}$ (denoted as Cauchy$(0, s)$), and $s = 1$, 3 and 5.
Case 4. Student's t-distribution, assume that ϵ follows a t-distribution with ν degrees of freedom, and $\nu = 2$, 5 and 10.
Case 5. Laplace distribution, $\epsilon \sim f(\epsilon) = (2\lambda)^{-1}e^{-|\epsilon|/\lambda}$ (denoted as LP$(0, \lambda)$), and $\lambda = 1$, 3 and 5.
Case 6. Heteroscedastic variance, $\sigma^2 = \mu_i^2 = (\beta_0 + \beta_1 z_i)^2$, or $\sigma = |z_i|$. The errors follow LP$(0, 1)$,

$N(0,1)$ and Cauchy$(0,1)$.

Case 1 and Case 2 are the normal errors without and with contaminations. Cases 3-5 are the errors with a heavy-tailed distribution. Case 6 is heteroscedastic variance. The scale parameter σ takes a value of 1 for Cases 2-5. We carry out 1000 independent realizations for all cases. The least squares estimates are obtained except for the Cauchy distribution. The simulation results are presented in Tables ??-??.

When the errors follow a normal distribution without contaminations, the proposed method is comparable with the least squares method which results in the optimal parameter estimation. Furthermore, the two data-driven methods (DD and PL) outperform the fixed regularization parameter method. The selected regularization parameters by the proposed method are quite large for the purely normal case. When the errors follow a normal distribution with contaminations, the efficiency of the two data-driven methods increases as the contaminated rate increases, and the performance of PL is similar to that of DD in most of the cases. The PL has a efficiency loss in the case of a normal distribution with 10% two-point contaminations. The WLS outperforms when there are 10% and 20% contaminations.

When the distribution of errors is heavy-tailed, the robust methods perform better than the least squares method. The least squares method cannot obtain consistent estimators when errors follow a Cauchy distribution. The proposed method obtains substantial efficiency, and the efficiency gain can be as high as 48% compared to that with the fixed regularization parameter (Cauchy(0,3)). The DD method has a 30% efficiency loss when the distribution of errors is Cauchy(0,1). For the heteroscedastic variance cases, the proposed method outperforms, and its efficiency can reach as high as 4.68 compared to the Huber's method with a fixed $\tau = 1.345$ when errors follow Cauchy(0,1) and $\sigma = |z_i|$. The estimators based on Huber's method perform better than REWLS for the heteroscedastic variance cases. For all the cases, the PL and PLH are very similar, which indicates that the estimator is not affected by the initial value.

4 Real data analysis

In this section, we illustrate the proposed method by two real datasets. The first data illustrate the relationship between diabetes and obesity measured through the body mass index and the waist/hip ratio of the participants[?, ?]. The second data were collected on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality[?, ?].

4.1 Diabetes data

The diabetes data consist of 19 variables on 403 subjects screened for diabetes in a study to understand the prevalence of cardiovascular risk factors such as obesity and diabetes in central Virginia for African Americans[?]. The dataset is available at http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets. The response variable is glycocholated hemoglobin (glyhb) which is usually taken as a positive diagnosis of diabetes when it exceeds the value of 7. We consider the following covariates: age, gender (male=1 and female=0), body mass index (bmi), waist/hip ratio (whip), body frame with three levels (small, medium and large), location (Louisa =1, Buckingham=0) and stabilized glucose (stabglu). We use two dummy variables 'framm' with one for medium frame and zero otherwise, and 'framl' with one for large frame and zero otherwise. There were missing values in the data, and we assume the data are missing completely at random. We use a regression model to fit the data,

$$\text{glyhb} = \beta_0 + \beta_1\text{age} + \beta_2\text{gender} + \beta_3\text{bmi} + \beta_4\text{whip} + \beta_5\text{framm} + \beta_6\text{framl} + \beta_7\text{location} + \beta_8\text{stabglu} + \sigma\epsilon.$$

Figure ?? indicates that there exists many underlying outliers in glycocholated hemoglobin, and the distribution of the residuals is heavy-tailed. The estimates of the regression parameters and their standard errors are presented in Table ??. The least squares estimates are quite different with those obtained by the robust methods. The two data-driven methods produce $\tau = 0.5$ and $\tau = 0.19$, which lead to smaller standard errors than the least squares method and the Huber's method with a fixed $\tau = 1.345$ for all the parameter estimates. With the least squares estimation, only the variables age and stabglu are significant at the 5% level, while with the robust estimation, the variable location is also significant. The proposed method has the smallest standard errors compared to other three methods. The computation time of the proposed method is about 15.04 seconds on Double Core class PC with 2.60 GHZ processors and 8.00 GB of RAM.

4.2 Air pollution data

The air pollution data include 16 variables: variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. The data is available at https://www3.nd.edu/ busiforc/handouts/Data%20and%20
Stories/regression/Air%20Pollution/airpullution.html. The response variable is mortality. We consider the following covariates: mean January temperature (JanTem, degrees Farenheit), mean July temperature (JulyTem, degrees Farenheit), relatively humidity (RelHum), annual rainfall (Rain, inches), median education (Education), population density(PopDensity), percentage of non whites (X.NonWhite), percentage of white collar workers (X.WC), population (Pop), population per house (Pop.house), median income (Income), hydrocarbon pollution potential (HCPot), nitrous oxide pollution potential (NOxPot), sulfur dioxide pollution potential (SO2Pot). Due to the skewness of the observations of the air pollution variables, we consider logarithm of them. Observation number 21 contains two missing values, and we assume they are missing completely at random. The response variable and the covariates are scaled to have mean equal to zero and variance equal to one. Such that all the variables are in a similar observation range.

We use the linear regression model to fit the data. The estimates of the regression parameters and their standard errors are presented in Table ??. The Q-Q plot in Figure ?? indicates the there may exist some extreme values in the data, and the kernel density plot indicates the error distribution is heavy-tailed. The parameter estimates are similar obtained by different methods. However, the standard errors of the parameter estimators obtained by the robust methods are smaller than those by the least squares method. The DD method obtains $\tau = 1.5$, and the proposed method obtains $\tau = 1.49$ and has the smallest standard errors. In addition to covariates JanTem, Rain, X.NonWhite and NOxPot, covariate X.WC is also significant at the 5% level using the proposed method. Furthermore, we use only these five significant covariates to fit the data (excluding other nonsignificant covariates), and find that X.WC is now also significant at the 5% level using others three methods, which indicates that the proposed method is more reliable than other three methods.

5 Conclusion and Discussion

The Huber's method has been widely used in robust analysis. ? proposed an approach for automatic selection of the regularization parameter by maximizing the efficiency factor. Their method based on the moment estimator of the efficiency factor, and hence the efficiency may be when the In this paper, we utilize the idea of ? to construct a working density function based on the Huber's loss function, and then obtain the regularization parameter by minimizing the negative log-likelihood. The proposed density function can approximate the true density function very well by automatically choosing the appropriate regularization parameter and dispersion parameter. The simulation results indicate that proposed method has appealing efficiency for the regression parameter estimation, especially for the Cauchy type distributions.

The Huber's method is robust against outliers in response variable. As we known, the Huber estimator is not fully efficient and does not enjoy high breakdown. Researchers have proposed some robust methods, which are fully efficient and with high breakdown, such as ?,?, ? and ?. When there exists outliers in covariates, or in both covariates and response, we can consider a weighted estimating function based on the Huber's score function, where the weight function can down-weight the influence of the extreme values in covariates[?] . The weighted equation is given as follows:

$$\sum_{i=1}^{n} w_{x_i} x_i \psi_{\tau,\theta}(r_i) = \mathbf{0}, \tag{9}$$

where $w_{x_i} = min\{1, c/d^2(x_i)\}$ is a function to down weight the influence of the outliers or/and leverage points in covariates. The tuning constant c in w_{x_i} takes a value of $\chi^2_{0.95}(p)$ and $d^2(x_i)$ denotes the squared Mahalanobis distance of x_i based on some robust measure of location and dispersion for x_i (see ?). The proposed procedures in Section 2 can be used to obtain the weighted estimates by replacing (??) with (??). We construct simulation studies to explore the performance of the weighted method. We consider the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \cdots, 1000,$$

where x_{i1} follows a contaminated normal distribution $(1-\alpha)N(0,1) + \alpha N(5,9)$ with $\alpha = 0.03$, and x_{i2} follows a uniform distribution $U(-1,1)$. Three distributions for ϵ_i are considered: the standard normal distribution $N(0,1)$, a contaminated distribution $0.8N(0,2) + 0.2t(5)$, a laplace distribution $f(\epsilon) = 1/4e^{-|\epsilon|/2}$ and a contaminated normal distribution $0.8N(0,4) + 0.2N(0,25)$. The results based on 1000 replications are presented in Table ??. When a covariate is contaminated, the weighted method performs much better than the unweihgted Huber's methods, especially for the coefficient estimator corresponding to the contaminated covariate. The robust full efficient estimator RWLS outperforms.

In this short article, we only consider the Huber's loss function for the linear regression model. The proposed method can be extended to generalized linear models or hypothesis testing. It is of interest to investigate the performance of the proposed method in these cases in future work.

Acknowledgments

This research was supported by the Australian Research Council Discovery Project (DP160104292), Natural Science Foundation of Zhejiang province under grant number (Y19A010054), Science Foundation of China (No.11871390), the Fundamental Research Funds for the Central Universities (No. xjj2017180), the Natural Science Basic Research Plan in Shaanxi Province of China (No.2018JQ1006).

References

. Huber PJ. Robust Statistics. New York: Wiely, 1981.

. Terpstra JT. and McKean JW. Rank-based analysis of linear models using R. Journal of Statistical Software 2005; 14: (7).

. Wang Y-G, Lin X, Zhu M. and Bai ZD. Robust estimation using the Huber function with a data-dependent tuning constant. Journal of Compuatational and Graphcial Statistics 2007; 16: (2) 468–481.

. Wang Y-G and Wang N. rlmDataDriven: Robust Regression with Data Driven Tuning Parameter. https://CRAN.R-project.org/package=rlmDataDriven, 2018.

. Gervini, D., Yohai, V. J. A class of robust and fully efficient regression estimators. The Annals of Statistics 2002; 30: 583–616.

. Rousseeuw, P. J. Least median of squares regression. Journal of the American Statistical Association 1984; 79: 871–880.

. Harrell F. Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer New York, 2001.

. Heritier S. Robust Methods in Biostatistics. John Wiley & Sons, 2009.

. Gijbels I and Vrinssen I. Robust nonegative garrote variable selection in linear regression. Computational Statistics and Data Analysis 2015; 85: 1–22.

. Smucler E and Yohai V J. Robust and sparse estimators for linear regression models. Computational Statistics & Data Analysis 2017; 111: 116–130.

. White H. Maximum likelihood estimation of misspecified models. Econometrica 1982; 50: 1–26.

. Bondell, H. D. and Stefanski, L. A. Efficient robust regression via two-stage generalized empirical likelihood. Journal of the American Statistical Association 2013; 108: 644–655.

. Kong, D., Bondell, H., and Shen, W. Outlier detection and robust estimation in nonparametric regression. International Conference on Artificial Intelligence and Statistics (2018a); 84: 208–216.

. Kong, D., Bondell, H., and Wu, Y. Fully efficient robust estimation, outlier detection and variable selection via penalized regression. Statistica Sinica 2018b; 28: 1031–1052.

Table 1. Relative efficiency of the least quare estimator (LS), and the estimator (DD) via the Huber's method with τ selected by Wang et al. (2007), the estimator (PL) via the Huber's method with τ selected by the proposed working likelihood method, the proposed estimator (PLH) using the Huber's method with $\tau = 1.345$ an initial value,, and the weighted-least estimator (RWLS) proposed by Gervini and Yohai to the estimator (HF) via the Huber's method with $\tau = 1.345$ for β_1 and β_2. $\hat{\tau}$ is the mean value of τ using the DD and PL methods based on 1000 simulations.

	$\epsilon \sim N(0,1)$								
	$\sigma = 1$			$\sigma = 3$			$\sigma = 4$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	1.0294	1.0582	——	1.0430	1.0820	——	1.0417	1.0655	——
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	1.0046	1.0395	2.434	1.0327	1.0488	2.437	1.0393	1.0483	2.448
PL	1.0263	1.0546	21.44	1.0424	1.0771	18.28	1.0429	1.0638	16.34
PLH	1.0263	1.0546	21.44	1.0424	1.0772	18.28	1.0431	1.0639	16.34
REWLS	0.9920	0.9769	——	1.0013	1.0081	——	0.9980	0.9749	——
	$N(0,1)$ with $t(3)$ contamination								
	$\gamma = 10\%$			$\gamma = 20\%$			$\gamma = 40\%$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	0.8925	0.8467	——	0.7650	0.7872	——	0.6540	0.7003	—-
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	1.0065	0.9775	1.755	0.9702	0.9835	1.424	0.9771	0.9562	1.145
PL	1.0022	1.0023	1.759	1.0016	0.9925	1.435	1.0019	0.9847	1.208
PLH	1.0022	1.0023	1.760	1.0016	0.9920	1.436	1.0014	0.9843	1.210
REWLS	0.9273	0.9772	——	0.9608	0.9041	——	0.8981	0.8957	——
	$N(0,1)$ with $\chi^2(3) - 3$ contamination								
	$\gamma = 10\%$			$\gamma = 20\%$			$\gamma = 40\%$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	1.2457	0.7501	——	1.8176	0.6706	——	4.0990	0.6463	——
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	0.9789	0.9813	1.350	1.0177	0.9874	1.041	1.0337	0.9717	0.738
PL	0.9959	0.9881	1.322	1.0316	0.9919	1.076	1.0367	0.9821	0.883
PLH	0.9966	0.9882	1.322	1.0329	0.9930	1.078	1.0359	0.9848	0.887
REWLS	0.6585	0.9267	——	0.4932	0.9193	——	0.4232	0.8387	——
	$N(0,1)$ with two-point contamination								
	$\gamma = 10\%$			$\gamma = 20\%$			$\gamma = 40\%$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	0.3455	0.3311	——	0.2959	0.3026	——	0.7233	0.7279	——
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	0.9768	0.9947	0.813	1.1259	1.1570	0.523	2.4853	2.4744	0.244
PL	0.8351	0.8626	0.371	1.0035	1.0513	0.201	2.2853	2.3224	0.635
PLH	0.8402	0.8698	0.375	1.0104	1.0570	0.205	2.2342	2.2990	0.714
REWLS	1.3280	1.4191	——	2.1010	1.9394	—-	0.4695	0.2434	—-

Table 2. Relative efficiency of the least quare estimator (LS), and the estimator (DD) via the Huber's method with τ selected by Wang et al. (2007), the estimator (PL) via the Huber's method with τ selected by the proposed working likelihood method, the proposed estimator (PLH) using the Huber's method with $\tau = 1.345$ as an initial value, and the weighted-least estimator (RWLS) proposed by Gervini and Yohai to the estimator (HF) by the Huber's method with $\tau = 1.345$ for β_1 and β_2. $\hat{\tau}$ is the mean value of τ using the DD and PL methods based on 1000 simulations.

	$\epsilon \sim f(\epsilon) = [\pi s(1+\epsilon/s)^2]^{-1}$								
	$s = 1$			$s = 3$			$s = 5$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	0.7261	0.7642	0.249	1.2370	1.1825	0.261	0.9227	0.8188	0.255
PL	1.4713	1.4444	0.144	1.4890	1.4772	0.102	1.4412	1.3962	0.088
PLH	1.4720	1.4456	0.144	1.4872	1.4777	0.102	1.4389	1.3974	0.088
REWLS	1.1385	1.0713	—-	1.0999	1.1410	—-	1.0532	1.0646	—-
	$\epsilon \sim t(\nu)$								
	$\nu = 2$			$\nu = 5$			$\nu = 10$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	0.1067	0.1172	—-	0.8279	0.7977	—-	0.9405	0.9588	—-
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	1.0934	1.0928	0.442	0.9620	0.9880	0.908	0.9762	0.9800	1.394
PL	1.0481	1.0413	0.329	0.9951	1.0020	1.185	1.0059	1.0051	1.763
PLH	1.0590	1.0471	0.336	0.9960	1.0026	1.187	1.0058	1.0051	1.764
REWLS	0.9057	0.9244	—-	0.8973	0.8700	—-	0.9018	0.9141	—-
	$\epsilon \sim f(\epsilon) = (2\lambda)^{-1}e^{-\|\epsilon\|/\lambda}$								
	$\lambda = 1$			$\lambda = 3$			$\lambda = 5$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	0.7632	0.7013	—-	0.7042	0.7363	—-	0.7305	0.7558	—-
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	1.2867	1.2526	0.209	1.3356	1.2892	0.220	1.2971	1.2774	0.214
PL	1.3026	1.2203	0.249	1.3486	1.3082	0.413	1.3104	1.2731	0.399
PLH	1.3082	1.2464	0.287	1.3299	1.3025	0.420	1.3121	1.2771	0.409
REWLS	0.7983	0.8121	—-	0.7808	0.8254	—-	0.8046	0.8502	—-
	Heteroscedastic Variance: $\sigma^2 = \mu_i^2$								
	$\epsilon \sim$ LP$(0,1)$			$\epsilon \sim N(0,1)$			$\epsilon \sim$ Cauchy$(0,1)$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	0.3511	0.3566	—-	0.7133	0.7712	—-	0.0000	0.0000	—-
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	1.9043	1.4063	0.103	1.2358	1.0028	0.103	0.7593	0.6304	0.103
PL	2.0315	1.4340	0.166	1.2568	0.9987	0.233	2.0173	1.2898	0.107
PLH	2.0275	1.4317	0.167	1.2578	0.9998	0.233	2.0149	1.2886	0.107
REWLS	0.6492	0.5252	—-	0.3161	0.2384	—-	0.9177	0.7755	—-
	Heteroscedastic Variance: $\sigma = \|z_i\|$								
	$\epsilon \sim$ LP$(0,1)$			$\epsilon \sim N(0,1)$			$\epsilon \sim$ Cauchy$(0,1)$		
	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$	REF$_{\beta_0}$	REF$_{\beta_1}$	$\hat{\tau}$
LS	0.2818	0.3847	—-	0.4837	0.7440	—-	0.0000	0.0000	—-
HF	1.0000	1.0000	1.345	1.0000	1.0000	1.345	1.0000	1.0000	1.345
DD	4.0536	1.3150	0.103	3.2880	0.9173	0.103	3.0925	0.9179	0.103
PL	3.0515	1.2908	0.373	3.0302	0.9190	0.333	4.6823	1.2701	0.142
PLH	3.0115	1.2879	0.374	2.9570	0.9220	0.335	4.5961	1.2691	0.143
REWLS	0.9938	0.5368	—-	0.8759	0.2116	—-	1.1382	0.7711	—-

Table 3. Parameter estimates for analysis of the diabetes data, and standard errors are in parentheses. LS is the least-squares method. HF is the Huber's method with a fixed regularization parameter $\tau = 1.345$. DD is the method proposed by Wang et al. (2007). PL is the proposed working likelihood method. The boldface indicates that p-value is less than 0.05.

	LS		HF		DD		PL	
Intercept	0.3116	(1.0221)	0.6866	(0.6649)	0.9373	(0.6119)	1.1342	(0.5535)
age	0.0183	(0.0052)	0.0132	(0.0034)	0.0119	(0.0031)	0.0127	(0.0028)
gender	-0.1122	(0.1827)	-0.0810	(0.1188)	-0.0933	(0.1094)	-0.0932	(0.0989)
bmi	0.0110	(0.0141)	0.0085	(0.0092)	0.0073	(0.0085)	0.0093	(0.0077)
whip	1.2462	(1.2098)	0.8547	(0.7870)	0.7419	(0.7243)	0.3723	(0.6551)
framm	0.2014	(0.1964)	0.1320	(0.1277)	0.1766	(0.1176)	0.2188	(0.1063)
framl	-0.0843	(0.2605)	0.0475	(0.1694)	0.1268	(0.1560)	0.1682	(0.1411)
location	-0.2092	(0.1565)	-0.2321	(0.1018)	-0.2924	(0.0937)	-0.2708	(0.0848)
stabglu	0.0288	(0.0015)	0.0303	(0.0010)	0.0295	(0.0009)	0.0296	(0.0008)
$\hat{\tau}$	$+\infty$		1.345		0.50		0.19	

Table 4. Parameter estimates for analysis of the air pollution data, and standard errors are in parentheses. LS is the least-squares method. HF is the Huber's method with a fixed regularization parameter $\tau = 1.345$. DD is the method proposed by Wang et al. (2007). PL is the proposed working likelihood method. The boldface indicates that p-value is less than 0.05.

	LS		HF		DD		PL	
JanTem	-0.3852	(0.1440)	-0.2959	(0.1350)	-0.3090	(0.1398)	-0.2801	(0.1152)
JulyTem	-0.1291	(0.1497)	-0.0941	(0.1404)	-0.0976	(0.1454)	-0.0898	(0.1198)
RelHum	0.0295	(0.0912)	0.0345	(0.0856)	0.0348	(0.0886)	0.0335	(0.0730)
Rain	0.2769	(0.1094)	0.2607	(0.1026)	0.2614	(0.1062)	0.2603	(0.0875)
Education	-0.1363	(0.1238)	-0.0731	(0.1162)	-0.0811	(0.1203)	-0.0642	(0.0991)
PopDensity	0.1045	(0.0975)	0.1088	(0.0915)	0.1096	(0.0947)	0.1068	(0.0780)
X.NonWhite	0.7426	(0.1444)	0.6979	(0.1355)	0.6993	(0.1402)	0.6973	(0.1155)
X.WC	-0.1529	(0.0973)	-0.1771	(0.0913)	-0.1735	(0.0945)	-0.1822	(0.0778)
log(Pop)	0.0575	(0.1011)	0.0279	(0.0948)	0.0336	(0.0981)	0.0208	(0.0808)
Pop.house	-0.1340	(0.1154)	-0.0823	(0.1083)	-0.0898	(0.1121)	-0.0740	(0.0923)
Income	-0.0494	(0.0956)	-0.0359	(0.0897)	-0.0365	(0.0929)	-0.0345	(0.0765)
log(HCPot)	-0.3984	(0.2753)	-0.3752	(0.2583)	-0.3775	(0.2674)	-0.3712	(0.2203)
log(NOxPot)	0.6466	(0.2712)	0.5172	(0.2544)	0.5308	(0.2633)	0.5005	(0.2169)
log(SO2Pot)	-0.0856	(0.1709)	0.0870	(0.1604)	0.0649	(0.1660)	0.1135	(0.1367)
$\hat{\tau}$	$+\infty$		1.345		1.5		1.49	

Table 5. Bias and relative efficiency of the least quare estimator (LS), and the estimator (DD) via the Huber's method with τ selected by Wang et al. (2007), the estimator (PL) via the Huber's method with τ selected by the proposed working likelihood method, the estimator (PLN) via the weighted Huber's method with τ selected by the proposed method to the estimator (HF) by the Huber's method with $\tau = 1.345$ for β_1, β_2 and β_3. $\hat{\tau}$ is the mean value of τ using the DD and PL methods based on 1000 simulations.

	$\epsilon \sim N(0,1)$						
	Bias$_{\beta_1}$	Bias$_{\beta_2}$	Bias$_{\beta_3}$	REF$_{\beta_1}$	REF$_{\beta_2}$	REF$_{\beta_3}$	τ
LS	-0.1454	-0.9975	-0.0002	0.1502	0.0781	0.4269	0.0000
HF	-0.0466	-0.2719	0.0003	1.0000	1.0000	1.0000	1.3450
DD	-0.0500	-0.2924	-0.0002	0.8965	0.8577	0.9799	1.4473
PL	-0.0368	-0.2125	0.0009	1.2361	1.6269	0.8934	0.7302
PLN	-0.0178	-0.0796	0.0006	2.0435	9.7143	0.9676	0.6637
REWLS	-0.0034	-0.0089	-0.0002	2.9938	60.968	1.2088	——
	$\epsilon \sim 0.8N(0,2) + 0.2t(5)$						
	Bias$_{\beta_1}$	Bias$_{\beta_2}$	Bias$_{\beta_3}$	REF$_{\beta_1}$	REF$_{\beta_2}$	REF$_{\beta_3}$	τ
LS	-0.2166	-1.5490	-0.0048	0.9552	0.8731	0.9669	0.0000
HF	-0.2114	-1.4456	-0.0032	1.0000	1.0000	1.0000	1.3450
DD	-0.2116	-1.4777	-0.0041	0.9954	0.9548	0.9819	1.8832
PL	-0.2134	-1.4900	-0.0041	0.9825	0.9406	0.9947	2.3976
PLN	-0.1514	-0.8070	-0.0008	1.7329	3.0954	1.0122	0.3342
REWLS	-0.0513	-0.1801	0.0017	6.7652	47.6912	1.4431	——
	$\epsilon \sim 1/4e^{-\lvert\epsilon\rvert/2}$						
	Bias$_{\beta_1}$	Bias$_{\beta_2}$	Bias$_{\beta_3}$	REF$_{\beta_1}$	REF$_{\beta_2}$	REF$_{\beta_3}$	τ
LS	-0.1471	-1.0076	0.0015	0.5308	0.3613	0.6899	0.0000
HF	-0.0974	-0.5915	0.0020	1.0000	1.0000	1.0000	1.3450
DD	-0.0702	-0.4167	0.0055	1.5128	1.9343	1.1809	0.3935
PL	-0.0637	-0.3791	0.0057	1.6874	2.3579	1.1954	0.4353
PLN	-0.0300	-0.1338	0.0034	2.8578	15.6746	1.3606	0.3703
REWLS	-0.0168	-0.0624	-0.0005	2.0412	27.7747	0.8710	——
	$\epsilon \sim 0.8N(0,4) + 0.2N(0,25)$						
	Bias$_{\beta_1}$	Bias$_{\beta_2}$	Bias$_{\beta_3}$	REF$_{\beta_1}$	REF$_{\beta_2}$	REF$_{\beta_3}$	τ
LS	-0.1465	-0.9960	-0.0031	0.5331	0.3719	0.6762	0.0000
HF	-0.0978	-0.5949	-0.0018	1.0000	1.0000	1.0000	1.3450
DD	-0.0885	-0.5351	-0.0015	1.1029	1.2171	0.9777	0.9330
PL	-0.0811	-0.4825	-0.0004	1.2206	1.4992	1.0010	0.8839
PLN	-0.0338	-0.1544	0.0002	1.8912	11.231	0.9115	0.7642
REWLS	-0.0246	-0.0895	-0.0015	2.0351	17.073	0.9178	——

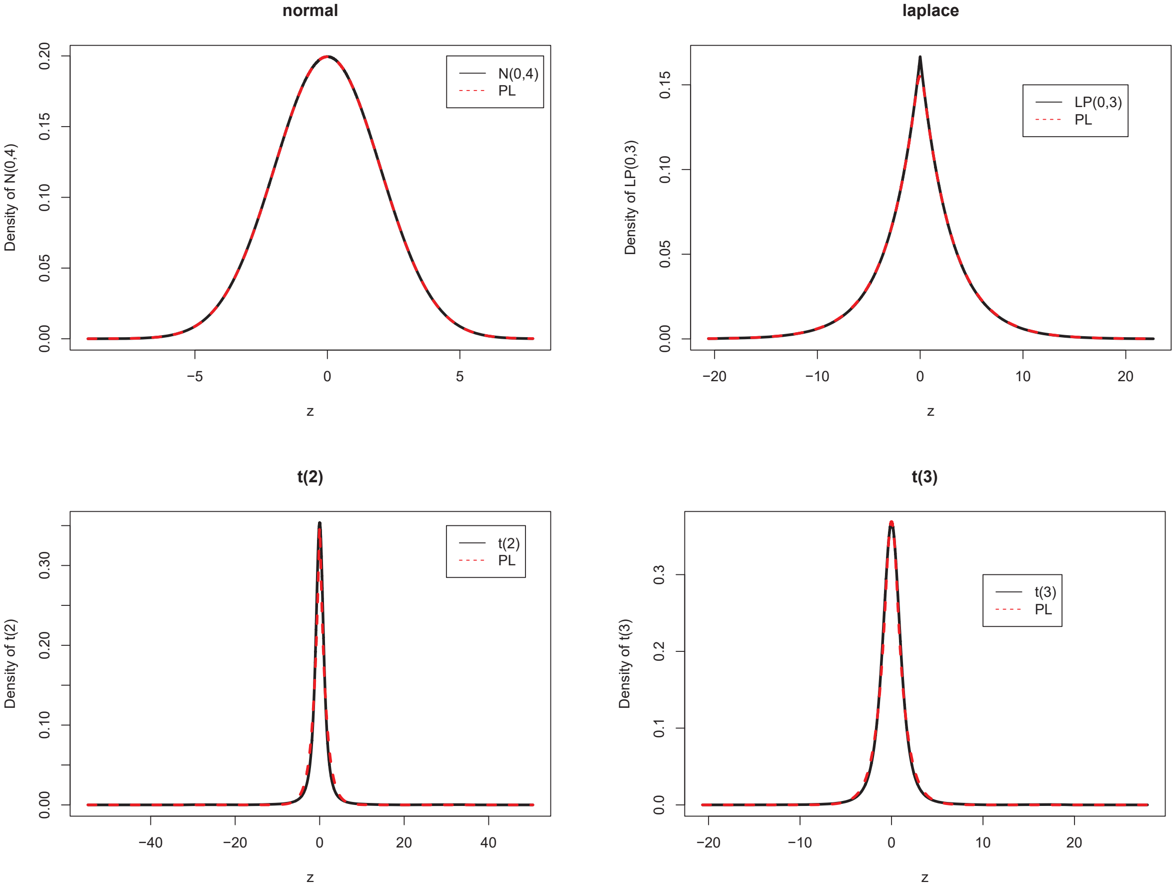

Figure 1. The density functions are fitted by the proposed density function (PL) with the selected regularization parameter and the dispersion parameter.

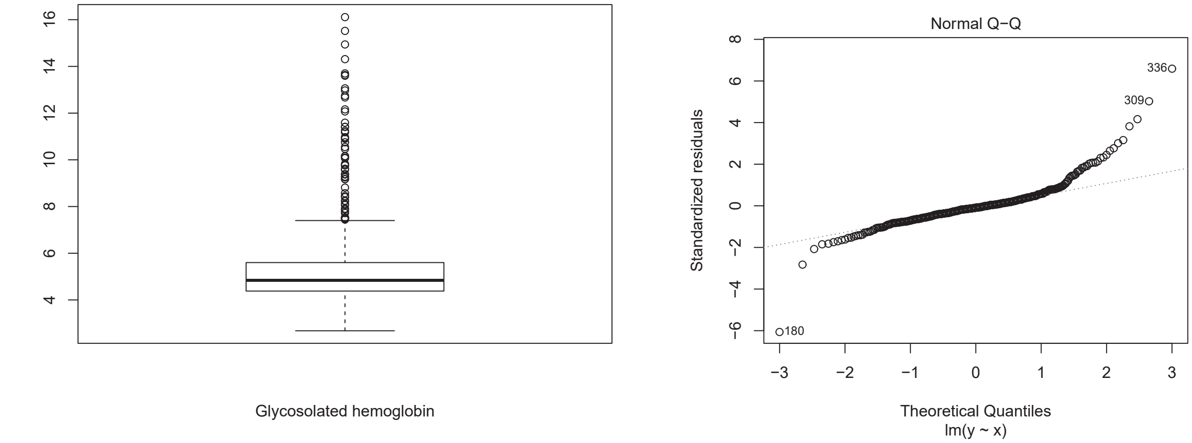

Figure 2. The left panel is a boxplot of glyhb, and the right panel is the Q-Q plot of the residuals from the least squares method.

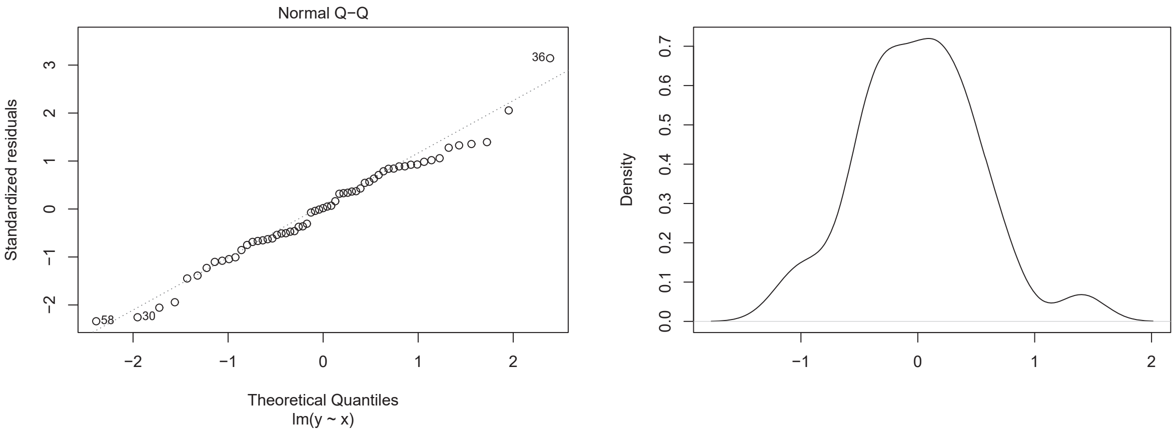

Figure 3. The left panel is the Q-Q plot of the residuals from the least squares method. The right panel is the kernel density function of the residuals obtained from the proposed method.