⊠연구논문

중도 절단된 자료에 대한 적응 로버스트 회귀*

김철기
이화여대 통계학과

Adaptive Robust Regression for Censored Data

Chul-Ki Kim
Dept. of Statistics, Ewha Womans University

Abstract

In a robust regression model, it is typically assumed that the errors are normally distributed. However, what if the error distribution is deviated from the normality and the response variables are not completely observable due to censoring? For complete data, Kim and Lai(1998) suggested a new adaptive M-estimator with an asymptotically efficient score function. The adaptive M-estimator is based on using B-splines to estimate the score function and a simple cross validation to determine the knots of the B-splines, which are a modified version of Kun(1992). We herein extend this method to right-censored data and study how well the adaptive M-estimator performs for various error distributions and censoring rates. Some impressive simulation results are shown.

Key words: Adaptive M-estimator; Asymptotically efficient score function; Right-censored data.

* This research was supported by the grant for promotion of scientific research in women's universities.

1. Introduction

Consider the linear regression model

$$y_j = \alpha + \beta^T x_j + \varepsilon_j \quad (j = 1, 2, \cdots, n), \tag{1.1}$$

where the ε_j are i.i.d. random variables and the x_j are independent $p \times 1$ random vectors independent of $\{\varepsilon_j\}$. Huber's M-estimators $\hat{\alpha}_H$, $\hat{\beta}_H$ of α, β based on $(x_1, y_1), \cdots, (x_n, y_n)$ are defined as a solution vector to the minimization problem

$$\sum_{j=1}^{n} \rho(y_j - a - b^T x_j) = \left(\int \rho(y - a) dF^*_{n,b}(y) \right) = \min!, \tag{1.2}$$

where $F^*_{n,b}$ is the empirical distribution constructed from $y_j(b) = y_j - b^T x_j$, $j = 1, \cdots, n$. When ρ is differentiable, the M-estimators $\hat{\alpha}_H$, $\hat{\beta}_H$ can also be defined as a solution of the system of estimating equations

$$\sum_{j=1}^{n} \rho'(y_j - a - b^T x_j) = 0, \quad \sum_{j=1}^{n} x_j \rho'(y_j - a - b^T x_j) = 0. \tag{1.3}$$

A well-known robust choice of ρ is Huber's score function

$$\rho'(u) = u \quad \text{if } |u| \le c \quad \text{and} \quad \rho'(u) = \pm c \text{ if } |u| > c, \tag{1.4}$$

where c represents some measure of dispersion of F. Using (1.4) in (1.2) is tantamount to applying the method of least-squares to "metrically Winsorized residuals" (cf. Huber(1981, p.180)).

In many applications, the responses y_j in (1.1) are not completely observable due to constraints on the experimental design. Thus, instead of (x_j, y_j), one observes

$$(x_j, \tilde{y}_j, \delta_j) \qquad j = 1, \cdots, n, \tag{1.5}$$

where $\tilde{y}_j = y_j \wedge c_j$, $\delta_j = I(y_j \le c_j)$ and we use $\wedge$ to denote minimum and maximum respectively, and the c_j represent censoring variables by $-\infty < c_j < \infty$. Unless otherwise stated, it will be assumed that (c_j, x_j) are independent of the sequence $\{\varepsilon_j\}$. The "censored regression model" is widely studied in quality engineering and economics. The system of estimating the equations (1.3) can be easily extended to censored data. Kim(1997) suggested an algorithm to compute the Huber's M-estimator for censored data. A modification of estimating equation (1.3) with $\rho'(u) = u$ results in the Buckley and James' estimator(cf. Buckley and James(1979)).

However, for non-normal error distributions, Huber's or B-J's (abbreviation of Buckley and James') estimator is not a pertinent estimator. Therefore, it behooves us to find a data-adaptive M-estimator. In Section 2, we first study a simple form of the estimate of the score function for the underlying density (cf. Kun (1992)) and learn two set cross-validation which is theoretically justified by Lai and Ying(1991b, 1992). In Section 3, some numerical examples show how well our new adaptive M-estimators perform for censored data with non-normal error distributions.

2. Adaptive Score Function

We use the linear B-splines to estimate the score function $\rho' = \phi = (\log f)'$ and derive a simple form of the estimate $\widehat{\phi_k}$ of ϕ , where k, the number of knots, is a smoothing parameter of which the empirical selection rule is based on two set cross-validation. This rule is a modified version of Kun(1992), which is theoretically supported by Lai and Ying(1991b, 1992).

2.1 Notation

On an interval (b_l, b_r), for any integer k, let the knots $\{\xi\}_k$ be $b_l = \xi_{k(0)} < \xi_{k(1)} < \cdots < \xi_{k(k)} = b_r$ and let the linear B-spline basis be $B_{k(i)}(x)$, $i = 1, \cdots, k$. Also let $D_{k(i)}(x)$, $i = 1, \cdots, k$, be their piecewise derivatives.

Denote $B_k(x) = (B_{k(1)}(x), \cdots, B_{k(k)}(x))^t$, $D_k(x) = (D_{k(1)}(x), \cdots, D_{k(k)}(x))^t$, and $A_k(x) = B_k \cdot B_k^t(x)$. Define

$$B_k(F)=\left(\int B_{k(1)}(x)\,dF(x),\cdots,\int B_{k(k)}(x)\,dF(x)\right)^t. \tag{2.1}$$

Similarly define $D_k(F), A_k(F), B_k(F_n), D_k(F_n)$ and $A_k(F_n)$.

2.2 Spline Interpolation and Knot Placement

The interpolation of $\phi(x)$ is defined as $a_k^t(F)B_k(x)$, where $a_k(F)$ minimizes $\int_{b_l}^{b_r}(a_k^tB_k(x)-\phi(x))^2 f dx$ for all $a_k\in R^k$. By partial integration,

$$\begin{aligned}&\int_{b_l}^{b_r}(a_k^tB_k(x)-\phi(x))^2 f dx\\&=a_k^t\left(\int_{b_l}^{b_r}B_kB_k^t(x)\,fdx\right)a_k-2a_k^t\int_{b_l}^{b_r}B_k\phi f dx+\int_{b_l}^{b_r}\phi^2 f dx\\&=a_k^tA_k(F)a_k+2a_k^tD_k(F)+I_b(\phi).\end{aligned}\tag{2.2}$$

Minimizing (2.2) is equivalent to minimizing $a_k^tA_k(F)a_k+2a_k^tD_k(F)$. Therefore, the $a_k(F)$ exists and $a_k(F)=-A_k^{-1}(F)D_k(F)$. Then, ϕ_{b_l,b_r} is interpolated as $\phi_k(x)=a_k^t(F)B_k(x)$. Naturally, we take $\hat{a}_k=a_k(F_n)=-A_k^{-1}(F_n)D_k(F_n)$ as the esti-mate of $a_k(F)$, and we estimate ϕ_{b_l,b_r} by $\widehat{\phi_k}(x)=a^t{}_k(F_n)B_k(x)$. The partial integration in (2.2) was first used by Cox(1985) to interpolate ϕ by smoothing splines.

For a given integer k, let $\{e_{(i)}\}, i=1,\cdots,m(n)$, be the order statistics of the residuals $\{e_i\}$ in (b_l, b_r). Then, the knots are $\xi_{k(i)}=e_{(i\cdot m(n)/(k+1))}$, $i=1,\cdots,k$, and this is the equally spaced quantiles method approached by Faraway(1992).

2.3 Empirical Selection of k

In order to choose the number of knots, we would like to pick $\hat{k}$ to minimize

$$\int_{b_l}^{b_r}(a_k(F_n)^tB_k(x)-\phi(x))^2 f dx. \tag{2.3}$$

This cannot be done since (2.3) depends on the unknown f. This is typical situation in which cross-validation can be applied. Minimizing (2.3) is equivalent to minimizing

$$L(k,F_n,F)= a'_k(F_n)A_k(F)a_k(F_n)+2\ a'_k(F_n)D_k(F). \tag{2.4}$$

Following procedure was suggested by Kun(1992):

1. Split the residuals into $e_1,\cdots,e_{n_1}$, $e_{n_1+1},\cdots,e_{n_1+n_2}$, where $n_1=\min\{3n^{1/2}, n/2\}$ and $n_2=n-n_1$.
2. Minimize $L(k,F_n,F)$ which is estimated by

$$L(k,F_{n_1},F_{n_2})= a'_k(F_{n_1})A_k(F_{n_2})a_k(F_{n_1})+2\ a'_k(F_{n_1})D_k(F_{n_2}),$$

   where F_{n_1} and F_{n_2} are the empirical distribution functions of $\{e_1,\cdots,e_{n_1}\}$ and $\{e_{n_1+1},\cdots,e_{n_1+n_2}\}$, respectively.
3. Select the first local minimizer $\hat{k}_{cv}$ as a cross-validatory estimate of k satisfying

$$L(1,F_{n_1},F_{n_2})\geq\cdots\geq L(\hat{k}_{cv},F_{n_1},F_{n_2})<L(\hat{k}_{cv}+1,F_{n_1},F_{n_2}).$$

4. Split the residuals again, this time into $e_1,\cdots,e_{n_2}$, $e_{n_2+1},\cdots,e_{n_2+n_1}$ and go to step 2 and 3, where another first local minimizer $\hat{k}'_{cv}$ is picked.
5. Define $ST(k,F_n)=(1/k)\sum_{j=0}^{k-1}\int_{b_l}^{b_r}(a_j^t(F_n)B_j(x)-a_k^t(F_n)B_k(x))^2dF_n$ and $\hat{k}_n$ is the first local minimizer of $ST(k,F_n)$ over $k\in I(n)$, where $I(n)=\{k\colon \hat{k}'_{cv}\leq k\leq \widehat{k^2}_{cv}\}$ if $\hat{k}'_{cv}\leq\hat{k}_{cv}$. Choose $\hat{k}_n$ within $I(n)$ satisfying

$$ST(\hat{k}'_{cv},F_n)\geq\cdots\geq ST(\hat{k}_n,F_n)<ST(\hat{k}_n+1,F_n).$$

If there is no such $\hat{k}_n$ within $I(n)$, choose $\hat{k}_n=\widehat{k^2}_{cv}$.

However, the step 5 called a stationary correction procedure is not theoretically proved rule but just empirical selection one. Thus, instead of the stationary correction procedure, we apply another cross-validatory method proposed by Lai and Ying(1991b, 1992), which starts from dividing the sample into two disjoint subsets. One might randomly split it into two sets of data. From the first subsample, define the residuals $e_i = \tilde{y}_i - bx_i - a$ ($i \leq n/2$) and let $n_1 = [n/2]$, i.e. the largest integer $\leq n/2$, where (a, b) is the B-J's estimators of (α, β) for the first subsample. And we do the step 2 and 3 to get the score function $\hat{\phi}_k(x) = a^t_k(F_{n_2})B_k(x)$ for the first subsample from the second one of $n_2 = n - n_1$ observations $(x_r, \tilde{y}_r, \delta_r)$, $n_1 < r \leq n$. Likewise, we define the residuals from the second subsample and do the same procedure after we switch F_{n_1} and F_{n_2}. Also we obtain the score function $\hat{\phi}_{k'}(x) = a^t_{k'}(F_{n_1})B_{k'}(x)$ for the second subsample from the first one.

Once we evaluate the score function at each data point, we plug all the values in the algorithm for computing the M-estimator for censored data proposed by Kim(1997) and get an adaptive M-estimator.

3. Adaptive M-estimator

3.1 A Gauss-Newton-Type Algorithm

In Section 2, we studied the data-adaptive score function $\hat{\phi}(x)$. Now we fit the linear regression model with $\hat{\phi}(x)$. Throughout the sequel, we shall use the following notation for the right-censored data (1.5). Let $\tilde{y}_i(b) = \tilde{y}_i - b^T x_i$ and define

$$N(b, u) = \sum_{i=1}^{n} I(\tilde{y}_i(b) \geq u), \quad \triangle(b, u) = I(\tilde{y}_i(b) = u, \delta_i = 1),$$

$$\hat{F}_b(u \mid v) = 1 - \prod_{i:\, v < \tilde{y}_i(b) \leq u,\, \delta_i = 1} \{1 - \triangle(b, \tilde{y}_i(b)) / N(b, \tilde{y}_i(b))\}. \tag{3.1}$$

The notation $\hat{F}_b(u \mid v-)$ will be used to denote (3.1) in which " $v < \tilde{y}_i(b)$" is replaced by " $v \le \tilde{y}_i(b)$". The function $\hat{F}_b(u | -\infty)$ is the product-limit estimate of the common distribution function $F(u)$ of the $\epsilon_j + a$ in (1.1). Put $\phi = \rho'$ in (1.3). To extend (1.3) to right-censored data (1.5), Lai and Ying(1994) applied "missing information principle" which leads to replacing (1.3) by the estimating equations

$$\sum_{i=1}^{n} \phi_i^*(a, b) = 0, \quad \sum_{i=1}^{n} x_i \phi_i^*(a, b) = 0, \tag{3.2}$$

where

$$\phi_i^*(a, b) = \delta_i \phi(\tilde{y}_i(b) - a) + (1 - \delta_i) \int_{u > \tilde{y}_i(b)} \phi(u - a) d\hat{F}_b(u | \tilde{y}_i(b)) \tag{3.3}$$

(cf. (2.24) and (2.26) of Lai and Ying(1994)).

Ordering the $\tilde{y}_k(\tilde{\beta})$ as $\tilde{y}_{[1]}(\tilde{\beta}) \ge \cdots \ge \tilde{y}_{[n]}(\tilde{\beta})$ and let

$$y_{[r]} = \tilde{y}_{[r]}(\tilde{\beta}), \tag{3.4}$$

where r is the risk set size trimmed away for smoothness of the product-limit estimate at its tail (cf. Lai and Ying(1991a)). Let X denote the $n \times (p+1)$ matrix whose ith row is $I(\tilde{y}_i(\tilde{\beta}) \le y_{[r]})(1, x_i^T)$. We now describe an iterative algorithm for computing the M-estimator for censored data (cf. Kim(1997)). Let $\theta = (a, \beta^T)^T$ and let $\theta^{(k)} = (a^{(k)}, \beta^{(k)T})^T$ denote the result after the kth iteration to compute the M-estimator of θ, and we initialize $\theta^{(0)} = (\tilde{a}, \tilde{\beta}^T)^T$ by the B-J's estimators for the whole data. The algorithm consists of the following procedure.

1. For $k=0$, set $\theta^{(0)} = (\tilde{a}, \tilde{\beta}^T)^T$ and compute $y_{[r]}$.
2. Compute $\tilde{y}_i(\beta^{(k)}), i = 1, \cdots, n$.
3. Evaluate $\hat{F}_{\beta^{(k)}}(u | v)$ or $\hat{F}_{\beta^{(k)}}(u | v-)$ by (3.1) at $u \in \{\tilde{y}_i(\beta^{(k)}) : \delta_i = 1, i \le n\}$ and $v \in \{\tilde{y}_i(\beta^{(k)}) : i \le n\}$, $u \ge v$.

4. Compute the $n\times 1$ vector $\Phi^{(k)}$, whose ith component is $\phi^*{}_i(\alpha^{(k)}, \beta^{(k)}) \times I(\tilde{y}_i(\tilde{\beta})\leq y_{[r]})$.
5. Solve the linear equation $X^TXz = X^T\Phi^{(k)}$ to find $z = z^{(k)}$.
6. Put $\theta^{(k+1)} = \theta^{(k)} + z^{(k)}$
7. Increase counter from k to $k+1$ and go to step 2.

We will see in the next section that the data-adaptive M-estimator works well compared with other M-estimators.

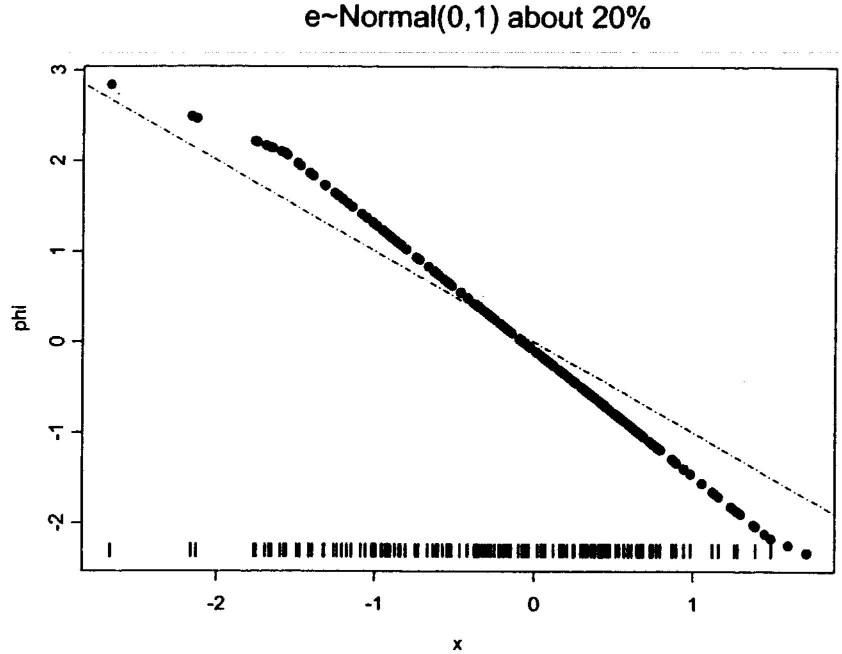
3.2 Numerical Examples

We consider a simple linear regression model $y_j = \alpha + \beta x_j + \varepsilon_j$, where the ε_j are i.i.d. random variables and we assume that the true values of α and β are equal to 0 and 1, respectively. For this simulation model, we compute B-J's, Huber's and the adaptive M-estimators, and then compare their mean square errors (MSE).

The simulation results show that the adaptive M-estimator performs well specially when the distribution of errors violates the normality assumption.

EXAMPLE 1. Generate the x_j from $U[-2,2]$ and the ε_j from $N(0,1)$. And the x_j are independent of the ε_j. The ε_j are subject to nearly 20% right censoring by the censoring variable c_j, where the c_j are i.i.d. $U[0,2]$. 100 censored data sets are generated from this model with sample size $n = 200$.

<Figure 1> shows the data-adaptive score function is quite close to the true one $\phi = f'/f$. Table 1 compares the mean square errors (MSEs) of the three estimators, where we mainly focus on the $\hat{\beta}$'s MSE because the location estimator $\hat{\alpha}$ in a censored regression model is meaningful only under some regularity condition of the error distribution (cf. Lai and Ying(1994)). As known, B-J's estimator is a good choice for a normal error distribution and we can check the fact in <Table 1>.

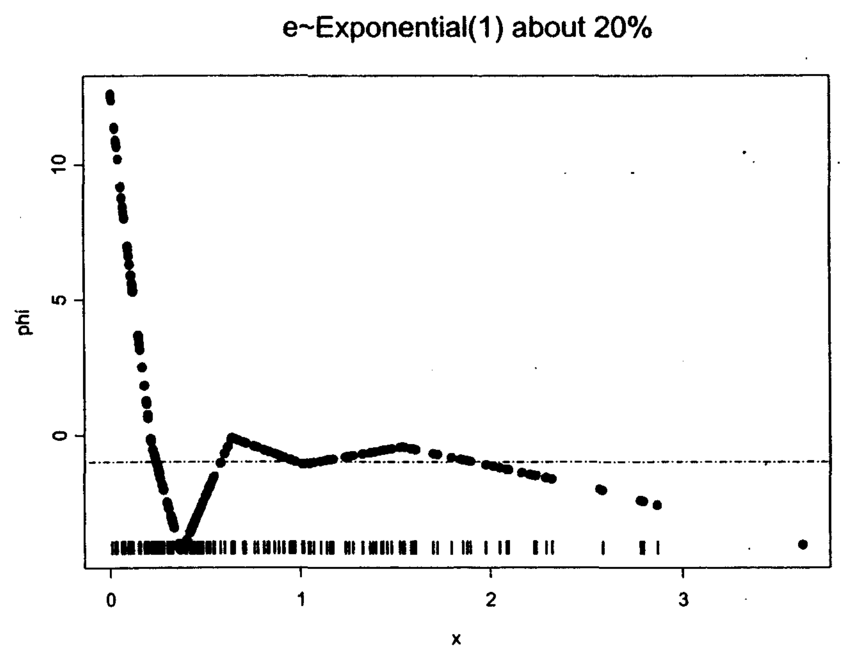


< Figure 1 > The broken line represents the true score function $\phi = f'/f = -x$ and the points stand for the estimated. The vertical lines show the density of uncensored residuals.

EXAMPLE 2. Consider the ε_j which are i.i.d. Exp(1), i.e. seriously deviated from normality, and the x_j uniformly distributed on [-2, 2] and independent of the ε_j. Also the ε_j are subject to right censoring by the c_j distributed as Exp(0.2) and the censoring rate is about 20% for all data sets. Like Example 1, 100 censored data sets are generated from this model with sample size $n=200$.

In <Figure 2>, we see the data-adaptive method pertinently estimate the true score function even under non-normality of the error distribution.

Although the error distribution is deviated from normality, the MSE of the adaptive M-estimator is smaller than those of others as shown in <Table 1>. It implies the adaptive M-estimator is more appropriate than others under the non-normal assumption. Notice that Huber's M-estimator which is resistant to outliers, is still as valid an estimator as the adaptive M-estimator, because the generated data have some outliers as shown in <Figure 2>.



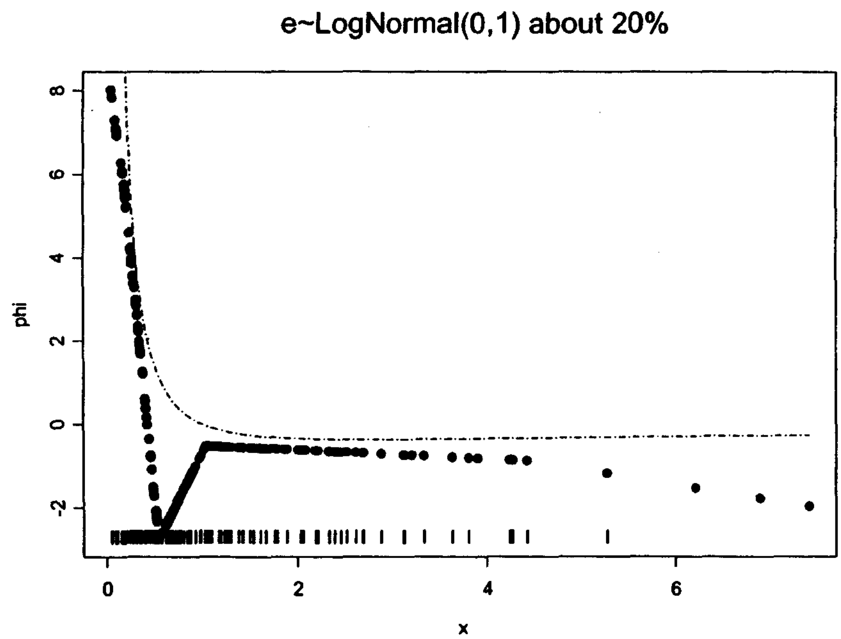
< Figure 2 > The broken line stands for $\phi = f'/f$, where $f = e^{-x}$, and the points represent the estimated score functions. The vertical lines show the density of uncensored residuals.

< Table 1 > Comparison of the three $M-$estimators: BJ(Buckley and James' Estimator), H(Huber's M-Estimator) and AME(Adaptive M-Estimator)

	φ	$\hat{\alpha}$		$\hat{\beta}$	
		Mean	MSE	Mean	MSE
N(0, 1)	BJ	-0.00929	0.01305	1.00092	0.00245
	H	-0.06630	0.00803	1.00040	0.00260
	AME	0.08795	0.17485	0.99851	0.00299
-Exp(1)	BJ	-0.05167	0.00561	0.99766	0.00129
	H	-0.18811	0.03839	0.99328	0.00075
	AME	0.04351	0.02984	1.00132	0.00067
LN(0, 1)	BJ	-0.39414	0.19757	0.97926	0.00216
	H	-0.18111	0.03839	0.99328	0.00124
	AME	0.05685	0.17100	0.98088	0.00124

EXAMPLE 3. Consider the ε_i which are distributed as $LN(0, 1)$ (lognormal) deviated from normality. And we generate the x_i from $U[-2,2]$ independent of the ε_i. The ε_i are subject to about 22% right censoring by the c_i which are i.i.d. $LN(1,2)$. We sample 100 censored data sets from this model with size $n=200$.

<Figure 3> shows that the score function obtained by the adaptive method is approximate to $\phi=f'/f$ for the non-normal density function f.

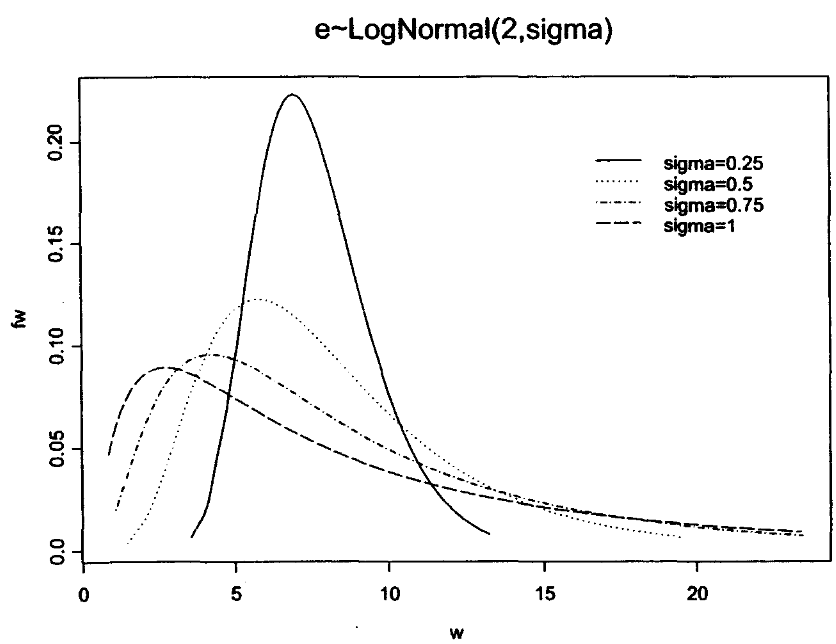


< Figure 3 > The broken line represents $\phi=f'/f$, and the points are the estimated score functions. The vertical lines show the density of uncensored residuals.

If the errors are not normally distributed unlike Example 1, B-J's estimator is not a good choice compared with others. We see that our adaptive M-estimator or Huber's is a better choice for the non-normal error distribution as shown in <Table 1>. The MSE of the adaptive M-estimator $\hat{\beta}$ is smaller than that of B-J's as in Example 2. The competitively small MSE of Huber's M-estimator probably results from the robustness to some outliers and the reason is the same as that in Example 2.

EXAMPLE 4. Generate the ε_j from four different lognormal distributions $LN(0, \sigma)$ with σ=0.25,0.5,0.75 and 1.0. Also the censoring variable c_j are generated from $LN(2.8,0.5)$, $LN(3,0.5)$, $LN(3,1)$ and $LN(3.2,1.25)$ with nearly 20% censoring rate, respectively.

<Figure 4> shows the density functions of the four different distributions. In case of σ=0.25, it is quite close to a normal distribution, but the larger σ is, the more severely deviated from normality the distribution is. Therefore, we expect that the MSE of the adaptive M-estimator is constantly smaller than others, even though σ varies from 0.25 to 1.0. According to <Table 2>, the adaptive M-estimator be-haves relatively better than B-J's or Huber's M-estimator.



< Figure 4 > LN(2,σ). where σ=0.25, 0.5, 0.75, and 1.0

< Table 2 > Comparison of the three M-estimators for lognormal error distribution: BJ(Buckley and James' Estimator), H(Huber's M-Estimator) and AME (Adaptive M-Estimator)

LN(2, σ)	φ	$\hat{\alpha}$		$\hat{\beta}$	
		Mean	MSE	Mean	MSE
0.25	BJ	-0.05241	0.01449	1.04937	0.01529
	H	-0.18941	0.04667	1.04604	0.01529
	AME	0.39142	0.25456	1.03311	0.01270
0.5	BJ	0.36546	0.38391	0.99979	0.07853
	H	-0.41843	0.38568	1.11457	0.07378
	AME	0.58438	0.63096	1.00784	0.08930
0.75	BJ	-0.29137	0.20486	0.91587	0.14761
	H	-1.99602	4.49991	0.89200	0.17056
	AME	-0.07611	0.17601	0.94499	0.06487
1.0	BJ	-1.68275	4.07685	0.94329	0.37091
	H	-5.21241	4.06167	1.14572	0.23833
	AME	-1.52217	3.50419	0.98490	0.21207

4. Conclusion

In a linear regression model, there may exist censored data with non-normal errors which are often used in quality engineering and economics. Our simulation studies have shown that the adaptive M-estimator is better than other M-estimators such as Huber's or B-J's when the errors of a linear regression model are not normal. Moreover, even for censored data, this adaptive method is asymptotically optimal in the sense that the limiting distribution of the adaptive M-estimator is normal with a minimal variance (cf. Kim and Lai(1998)). In this study, we did the simulations only for the simple linear regression model, but the whole procedure can be easily extended to multiple linear regression models and we expect to get the similar results.

References

[1] Buckley, J. and James, I.(1979), "Linear regression with censored data," *Biometrica,* Vol. 66, pp. 429-436.

[2] Cox, D. D.(1985), "A penalty method for nonparametric estimation of the logarithmic derivative of a density function," *Annals of the Institute of Statistical Mathematics,* Vol. 37, pp. 271-288.

[3] Faraway, J. J.(1992), "Smoothing in adaptive estimation," *Annals of Statistics,* Vol. 20, pp. 414-427.

[4] Huber, P.(1981), *Robust Statistics,* Wiley, New York.

[5] Kim, C. K.(1997), "Robust regression for right-censored data," *Journal of the Korean Society for Quality Management,* Vol. 25, No. 2, pp. 47-59.

[6] Kim, C. K. and Lai, T. L.(1998), "Adaptive M-estimator with asymptotically efficient score function in a linear regression model," *Unpublished Manuscript.*

[7] Kun, J.(1992), "Empirical smoothing parameter selection in adaptive estimation," *Annals of Statistics,* Vol. 20, No. 4, pp. 1844-1884.

[8] Lai, T. L. and Ying, Z.(1991a), "Estimating a distribution function with truncated and censored data," *Annals of Statistics,* Vol. 19, pp. 417-442.

[9] Lai, T. L. and Ying, Z.(1991b), "Rank regression methods for left-truncated and right-censored data," *Annals of Statistics,* Vol. 19, pp. 531-556.

[10] Lai, T. L. and Ying, Z.(1992), "Asymptotically efficient estimation in censored and truncated regression models," *Statistica Sinica,* Vol. 2, pp. 17-46.

[11] Lai, T. L. and Ying, Z.(1994), "A missing information principle and M-estimators in regression analysis with censored and truncated data," *Annals of Statistics,* Vol. 22, pp. 1222-1255.