Regression Mode Interval

SUMMARY

In this paper, we extend the concept of the mode type quantile interval of Huang (2003) for the linear regression model. The population type mode interval for some distributions are derived and their estimation in parametric and nonparametric ways are introduced. Simulations for nonparametric estimation and trimmed means based on this quantile interval are done in comparison with those techniques based on the traditional symmetric type intervals.

1. Introduction

In the linear regression model

$$y = x'\beta + \epsilon,$$

interval estimation is a very useful technique to monitor the picture of the response variable y or its distribution. Three types of this interval, popular in application and theoretical study, include (1) the confidence interval (C.I.) for conditional mean $E(y|x) = x'\beta$, (2) the C.I. for prediction variable y_0 given a future covariate vector x, and (3) estimation of a quantile interval $(F_{y|x}^{-1}(\alpha), F_{y|x}^{-1}(1-\alpha))$, where $F_{y|x}$ is the conditional distribution of y given a vector x. Basically, the first two are aiming mainly in construction of C.I. for a point, an unknown conditional mean or an unknown future variable. On the other hand, the third is trying to estimate an unknown parameter type interval.

The quantile interval has two interesting application aspects. The first is to use the width, denoted by $\tau(1-2\alpha) = F_{y|x}^{-1}(1-\alpha) - F_{y|x}^{-1}(\alpha)$, of this interval as a scale-like parameter. One example is the interquartile range $\tau(0.5)$, using a 50% quantile interval, playing as a robust type scale parameter. On the other hand, suppose that random variable y represents a quality characteristic for one product. Then the width of a 99.73% quantile interval, $\tau(0.9973)$, is used to represent a measure of the product's manufacturing process capability which is a very useful tool in improving product's quality.

The second application is to use the quantile interval to classify the data into sets of good and bad observations. This application means in two aspects. First, the set of good data is used to construct robust type location estimators such as the trimmed mean and the Winsorized mean and the scale estimator such as the trimmed variance,e.g., see Staudte and Sheather (1990) for details. The second aspect is that when variable y represents, again, a characteristic value for one product, a quantile interval serves a control chart to investigate if a manufacturing process is out of control. This is another important tool in engineering quality control.

When we need to use a quantile interval, why should we choose the symmetric one $(F_{y|x}^{-1}(\alpha), F_{y|x}^{-1}(1-\alpha))$? In fact, the following class

$$\{(F_{y|x}^{-1}(\alpha_0), F_{y|x}^{-1}(1-2\alpha+\alpha_0)) : 0 \leq \alpha_0 < 1-2\alpha\} \tag{1.1}$$

provides all possible interval with the same coverage probability $1-2\alpha$. There must have some reasons for us to choose one from this interval class. Two criteria may be applied to evaluate the suitability as reasons for a quantile interval. (a). A quantile interval actually serves an interval type location parameter. Then we may expect that it fullfills several desirable equivariant properties for any location parameter. Not every version in the class in (1.1) fullfills this expectation. However, the symmetric one does. The second criterion will encourage us to search an alternative one. (b). Among all choices in (1.1), is there one that has evidence of advantages from the point of statistical inferences? We may evaluate this through two aspects. (b.1). From the view of point estimation, can the trimmed mean induced from one quantile interval be with efficiency relatively higher than those induced from the other quantile intervals? (b.2). When the quantile interval serves a control chart, can we find one that its power in observing the fact that a process is out of control is larger than the symmetric one with the same coverage probability when the process is in control?

Section 2 introduces the population type regression mode interval and Section 3 provides it under several examples of distributions. In Section 4, a data analysis of some real examples is introduced and in Section 5, we introduce a technique for nonparametric estimation of the regression mode interval. Finally, in Section 6, we apply this mode interval to construct a robust trimmed mean for estimation of the regression parameters and a simulation for measuring its efficiency has been done.

2. Population Type Regression Mode Interval

Let's introduce a γ mode interval in the following definition.

Definition 2.1. The γ-mode interval for linear regression model is defined as $C_{mod}(\gamma) = (F_{y|x}^{-1}(\alpha^*), F_{y|x}^{-1}(\gamma+\alpha^*))$ with

$$\alpha^* = \arg \min_{0 \le \alpha \le 1-\gamma} [F_{y|x}^{-1}(\gamma+\alpha) - F_{y|x}^{-1}(\alpha)],$$

The γ-mode interval may be represented as $C_{mod}(\gamma) = x'\beta + (F_\epsilon^{-1}(\alpha^*), F_\epsilon^{-1}(\gamma+\alpha^*))$ with

$$\alpha^* = \arg \min_{0 \le \alpha \le 1-\gamma} (F_\epsilon^{-1}(\gamma+\alpha) - F_\epsilon^{-1}(\alpha)),$$

where we let F_ϵ be the distribution function of error variable ϵ. If we let x be a $p \times 1$ covariate vector with the first component 1, the γ-mode interval may also be represented as

$$C_{mod}(\gamma) = x'(\beta(\alpha^*), \beta(\gamma+\alpha^*)),$$

where p-vector $\beta(\delta) = \beta + \begin{pmatrix} F_\epsilon^{-1}(\delta) \\ 0_{p-1} \end{pmatrix}$ for $0 < \delta < 1$ and 0_{p-1} denotes the $(p-1) \times 1$ vector $(0, \dots, 0)'$.

The reason that we call it a mode interval is that a γ mode interval shrinks to a single set of mode point when γ approaches to zero. On the other hand, the symmetric interval $x'\beta + (F_\epsilon^{-1}(\alpha), F_\epsilon^{-1}(1-\alpha))$ will shrink to a set of single median point when α approaches to 0.5. We then call it the median interval and denote it by C_{med}. Moreover, when the regression model has an intercept term, the median interval may be represented as $C_{med} = x'(\beta(\alpha), \beta(1-\alpha))$.

As that we treat a quantile interval as an interval type location parameter, we may expect that it satisfies some usual desired properties of equivariance. Let's now define these conditions that are extended from those for location parameter in Staudte and Sheather (1990). Let A be a set of real numbers. We denote addition and multiplication of set A with a real number b by $A + b = \{x + b : x \in A\}$ and $bA = \{bx : x \in A\}$. We also say that $A \ge b$ if $x \ge b$ for $x \in A$.

Definition 2.2. Let X be a random variable. A measure of coverage set, with confidence coefficient γ, is a set $D(X)$ with $P(X \in D(X)) = \gamma$ that satisfies the following conditions:

(1). $D(X+b) = D(X) + b$ for $b \in R$.

(2). $D(aX) = aD(X)$ for $a \in R$.

(3). $X \geq 0$ implies $D(X) \geq 0$.

We now show that the mode interval does satisfy the conditions of a measure of coverage set.

Theorem 2.3. The γ mode interval is a measure of coverage set with confidence coefficient γ.

Proof. Redenote the γ mode interval by $C(\gamma, y)$ and α^* by $\alpha^*(y)$. We note that the population quantile $F^{-1}_{\epsilon|x}$ satisfies $F^{-1}_{\epsilon+b|x}(\alpha) = F^{-1}_{\epsilon|x}(\alpha) + b$ for $b \in R$ and $F^{-1}_{a\epsilon|x}(\alpha) = aF^{-1}_{\epsilon|x}(\alpha)$ if $a > 0$ and $= aF^{-1}_{\epsilon|x}(1-\alpha)$ if $a < 0$. With these and the fact that $y + b = x'\beta + (\epsilon + b)$ and $ay = a(x'\beta) + a\epsilon$, we see that $F^{-1}_{y+b|x}(\alpha) = F^{-1}_{y|x}(\alpha) + b$ and $F^{-1}_{ay|x}(\alpha) = aF^{-1}_{y|x}(\alpha)$ if $a > 0$ and $= aF^{-1}_{y|x}(1-\alpha)$ if $a < 0$.

Consider condition (1). Since $F^{-1}_{y+b|x}(\alpha) = F^{-1}_{y|x}(\alpha) + b$,

$$\begin{aligned}
&\alpha^*(y+b) = \arg \min_{0\leq\alpha\leq1-\gamma} [F^{-1}_{y+b|x}(\alpha+\gamma) - F^{-1}_{y+b|x}(\alpha)] \\
&= \arg \min_{0\leq\alpha\leq1-\gamma} [F^{-1}_{y|x}(\alpha+\gamma) - F^{-1}_{y|x}(\alpha)] \\
&= \alpha^*(y).
\end{aligned}$$

Then

$$\begin{aligned}
&C(\gamma, y+b) = (F^{-1}_{y+b|x}(\alpha^*(y+b)), F^{-1}_{y+b|x}(\gamma + \alpha^*(y+b))) \\
&= (F^{-1}_{y+b|x}(\alpha^*(y)), F^{-1}_{y+b|x}(\gamma + \alpha^*(y))) \\
&= (F^{-1}_{y|x}(\alpha^*(y)), F^{-1}_{y|x}(\gamma + \alpha^*(y))) + b \\
&= C(\gamma, y) + b.
\end{aligned}$$

Next, consider condition (2) with $a > 0$. Since $F^{-1}_{ay|x}(\alpha) = aF^{-1}_{y|x}(\alpha)$,

$$\begin{aligned}
&\alpha^*(ay) = \arg \min_{0\leq\alpha\leq1-\gamma} [F^{-1}_{ay|x}(\alpha+\gamma) - F^{-1}_{ay|x}(\alpha)] \\
&= \arg \min_{0\leq\alpha\leq1-\gamma} a[F^{-1}_{y|x}(\alpha+\gamma) - F^{-1}_{y|x}(\alpha)] \\
&= \alpha^*(y).
\end{aligned}$$

Then

$$\begin{aligned}
&C(\gamma, ay) = (F^{-1}_{ay|x}(\alpha^*(ay)), F^{-1}_{ay|x}(\gamma + \alpha^*(ay))) \\
&= (F^{-1}_{ay|x}(\alpha^*(y)), F^{-1}_{ay|x}(\gamma + \alpha^*(y))) = a(F^{-1}_{y|x}(\alpha^*(y)), F^{-1}_{y|x}(\gamma + \alpha^*(y))) \\
&= aC(\gamma, y).
\end{aligned}$$

Now, consider condition (2) with $a < 0$.

$$\begin{aligned}\alpha^*(ay) &= \arg \min_{0\le\alpha\le 1-\gamma} [F^{-1}_{ay|x}(\alpha+\gamma) - F^{-1}_{ay|x}(\alpha)] \\ &= \arg \min_{0\le\alpha\le 1-\gamma} [F^{-1}_{a\epsilon|x}(\alpha+\gamma) - F^{-1}_{a\epsilon|x}(\alpha)] \\ &= \arg \min_{0\le\alpha\le 1-\gamma} a[F^{-1}_{\epsilon|x}(1-(\alpha+\gamma)) - F^{-1}_{\epsilon|x}(1-\alpha)] \\ &= \arg \min_{0\le\alpha\le 1-\gamma} [F^{-1}_{\epsilon|x}(1-\alpha) - F^{-1}_{\epsilon|x}(1-(\alpha+\gamma))] \\ &= \arg \min_{0\le\delta\le 1-\gamma} [F^{-1}_{\epsilon|x}(\delta+\gamma) - F^{-1}_{\epsilon|x}(\delta)] \\ &= \arg \min_{0\le\delta\le 1-\gamma} [F^{-1}_{y|x}(\delta+\gamma) - F^{-1}_{y|x}(\delta)]\end{aligned}$$

with $\delta = 1-(\alpha+\gamma)$. Then $\alpha^*(y) = 1-[\alpha^*(ay)+\gamma]$ or $\alpha^*(ay) = 1-[\alpha^*(y)+\gamma]$. Furthermore,

$$\begin{aligned}C(\gamma, ay) &= (F^{-1}_{ay|x}(\alpha^*(ay)), F^{-1}_{ay|x}(\gamma+\alpha^*(ay))) \\ &= (F^{-1}_{ay|x}(1-[\alpha^*(y)+\gamma]), F^{-1}_{ay|x}(1-[\alpha^*(y)+\gamma]+\gamma)) \\ &= (aF^{-1}_{y|x}(\alpha^*(y)+\gamma), aF^{-1}_{y|x}(\alpha^*(y))) \\ &= a(F^{-1}_{y|x}(\alpha^*(y)), F^{-1}_{y|x}(\alpha^*(y)+\gamma)) \\ &= aC(\gamma, y). \quad \square\end{aligned}$$

How can we estimate the regression mode interval $(x'\beta + F^{-1}_{\epsilon|x}(\alpha^*), x'\beta + F^{-1}_{\epsilon|x}(\gamma + \alpha^*))$? Basically, there are two directions we may consider. The first one is a two steps method. It consists of estimating the regression parameters β, denoted by $\hat{\beta}$, for the first step and then using the residuals $e_i = y_i - x_i'\hat{\beta}$ to estimate error quantiles $F^{-1}_{\epsilon|x}(\alpha^*)$ and $F^{-1}_{\epsilon|x}(\gamma+\alpha^*)$ for the second step. The second one is, when the model has the intercept term, that we may estimate $\beta(\alpha^*)$ and $\beta(\gamma+\alpha^*)$ in one step. In this paper, we consider only through the first direction; however, parametric and nonparametric estimation techniques are also discussed. Computing $\beta(\alpha)$ with known α has been introduced by Koenker and d'Orey (1987) where this parameter vector was called the regression quantile and introduced by Koenker and Bassett (1978). However, one stage to estimate the mode type quantile vector with unknown α^* needs further investigation. In the following section, we consider the parametric formulation of the mode interval that may be estimated by parametric estimation.

3. Parametric Formulation of Regression Mode Interval

The mode interval may be more explicitly formulated when the distribution function F is known, although it may involve unknown parameter θ in its ends. In this section, we will consider this parametric type mode interval and also display its point estimation whereas the case of nonparametric study will be introduced in subsequent sections. One interesting question for the parametric distribution is that whether or not there is a parametric family of distributions for r.v. X so that the mode interval has an explicit formula which makes the statistical inference easier to perform. The following theorem indicates that the location-scale family is the interesting one.

Theorem 3.1. If the linear regression model has error variable with distribution in the family of continuous location-scale distributions with p.d.f. of the form $f(\epsilon;\theta_1,\theta_2) = \frac{1}{\theta_2} f_0(\frac{\epsilon - a(\theta_1,\theta_2)}{\theta_2})$ for $\theta_1 \in R$ and $\theta_2 > 0$, then the regression median and mode type intervals are, respectively,

$$x'\beta + a(\theta_1,\theta_2) + \theta_2(F_0^{-1}(\alpha), F_0^{-1}(1-\alpha))$$

and

$$x'\beta + a(\theta_1,\theta_2) + \theta_2(F_0^{-1}(\alpha^*), F_0^{-1}(\gamma+\alpha^*)),$$

where

$$\alpha^* = \arg \min_{0 \le \alpha \le 1-\gamma} [F_0^{-1}(\alpha+\gamma) - F_0^{-1}(\alpha)]$$

and F_0 is the distribution function of p.d.f. f_0.
Proof. The proof is obvious from the fact that $F^{-1}(\alpha) = a(\theta_1,\theta_2) + \theta_2 F_0^{-1}(\alpha)$.

The benefit of the location-scale family is that the mode interval is explicitly displayed in terms of α^* and parameter θ and then we may easily develop the estimator of the coverage interval through the existing theorems for the statistical inference of parameter θ. The picture of median and mode type intervals are displayed in Figure 1 in the appendix.

Normal error: Suppose that we have the linear regression model

$$y = x'\beta + \epsilon,$$

where ϵ is independent of x and has the normal distribution $N(0,\sigma^2)$ for some $\sigma > 0$. The symmetric error distribution indicates the identity of the median type and mode type intervals, i.e.,

$$C_{med}(1-2\alpha) = C_{mod}(1-2\alpha) = x'\beta + \sigma(-z_\alpha, z_\alpha),$$

where z_α satisfies $P(Z > z_\alpha) = \alpha$ and $Z \sim N(0,1)$.

Exponential error: Suppose that we have the linear regression model

$$y = x'\beta + \epsilon,$$

where ϵ is independent of x and has the exponential distribution with p.d.f. $f(\epsilon) = \frac{1}{\theta}e^{-\frac{\epsilon+\theta}{\theta}}, \epsilon \geq -\theta$, for some $\theta > 0$. We see that the conditional quantile is $F^{-1}_{y|x}(\alpha) = x'\beta - \theta[1 + \ln(1-\alpha)]$. Then we further have

$$C_{med}(1-2\alpha) = x'\beta - \theta - \theta(\ln(\alpha), \ln(1-\alpha))$$

and

$$C_{mod}(1-2\alpha) = x'\beta - \theta - \theta(\ln(2\alpha), 0).$$

On the other hand, one distribution highly asymmetric and skewed to the left has the form

$$f(\epsilon) = \frac{1}{\theta}e^{\frac{\epsilon-\theta}{\theta}}I(\epsilon < \theta).$$

Then

$$C_{med}(1-2\alpha) = x'\beta + \theta + \theta(\ln(\alpha), \ln(1-\alpha))$$

and

$$C_{mod}(1-2\alpha) = x'\beta + \theta + \theta(\ln(2\alpha), 0).$$

Gamma error: Suppose that the linear regression error is independent of x and follows the Gamma distribution with p.d.f. $f(\epsilon) = \frac{1}{\Gamma(\frac{k}{2})\theta^{\frac{k}{2}}}(\epsilon + \frac{k}{2}\theta)^{\frac{k}{2}-1}e^{-\frac{\epsilon+\frac{k}{2}\theta}{\theta}}, \epsilon \geq -\frac{k}{2}\theta$, for some $k \in N$ and $\theta > 0$. The conditional quantile is $F^{-1}_{y|x}(\alpha) = x'\beta - \frac{k}{2}\theta + \frac{\theta}{2}F_0^{-1}(\alpha)$. Then we have

$$C_{med}(1-2\alpha) = x'\beta - \frac{k}{2}\theta + \frac{\theta}{2}(F_0^{-1}(\alpha), F_0^{-1}(1-\alpha))$$

and

$$C_{mod}(\gamma) = x'\beta - \frac{k}{2}\theta + \frac{\theta}{2}(F_0^{-1}(\alpha^*), F_0^{-1}(\gamma + \alpha^*)),$$

where $\alpha^* = \arg\min_{0\leq\alpha\leq 1-\gamma}[F_0^{-1}(\alpha+\gamma) - F_0^{-1}(\alpha)]$ and F_0 denotes the distribution function of $\chi^2(k)$.

Basically, we may choose the regression mode and median intervals, for example, $C_{mod}(1-2\alpha)$ and $C_{med}(1-2\alpha)$, to have the same coverage probability for random variable y based on conditional distribution. Then why should we choose the mode type one?

In some statistical problems such as the quality control, we have a historical record of observations that we can compute the mode intervals for the ideal process for random variable y. For simplicity of interpretation, let's restrict this to the problem of the statistical process control. The computed quantile interval is then used as a control chart with two ends of the interval as the control limits. A new observation of variable y falling outside the limits may induce the conclusion that the process is out of control. Since the median type interval is the traditional way as the control chart, we then consider if a control chart based on mode type interval may improve the process control in some way. Consider the case where the two quantile intervals have the same coverage probabilities under the ideal process. Then these two have the same probability of making the wrong conclusion that the process is out of control when the process is still in control. A reasonable comparison is to see which one has larger power of concluding that the process is out of control when the process is out of control.

As an example, consider the case where y is the control variable and the process may be changed only through the error variable which has right skewed exponential distribution with $\theta = 1$ as the standard process. We compute the powers, for the two quantile intervals, of observation y falling outside the interval when the true parameter value is $\theta = \theta_1$:

$$\pi_{med}(\theta_1) = P_{\theta_1}(\{y \notin C_{med}(1-2\alpha)\}),$$
$$\pi_{mod}(\theta_1) = P_{\theta_1}(\{y \notin C_{mod}(1-2\alpha)\}).$$

Table 1. Powers for quantile intervals as control charts

θ_1	π_{med}	π_{mod}	π_{med}	π_{mod}
	$2\alpha = 0.1$		$2\alpha = 0.05$	
2	0.544	0.584	0.496	0.528
3	0.684	0.724	0.641	0.675
4	0.757	0.792	0.718	0.750
5	0.802	0.833	0.767	0.797
6	0.832	0.861	0.802	0.828
7	0.855	0.880	0.827	0.851
8	0.872	0.895	0.847	0.869
9	0.885	0.907	0.862	0.883
10	0.896	0.916	0.875	0.894

We have several conclusions drawn from the table above:

(a). For both quantile intervals, the power is increasing in θ_1 when the process is out of control with $\theta = \theta_1 > 1$. This means that both quantile intervals are appropriate to be used to construct control charts since larger value of $\theta_1(> 1)$ indicates that the process is out of control in a more serious situation.

(b). The powers of mode type intervals are uniformly larger than the corresponding ones of median type intervals. This suggests us to use the mode interval as the control chart when the underlying distribution is asymmetric.

4. Data Analysis

A company that sells and repairs small computers concerns the number of service engineers that will be required to serve the customers over the next few years. An important element to forecast this number is an analysis of the length of service calls which depends on the number of electronic components in the computer that must be repaired or replaced. Chatterjee and Price (1991) provided a data set of size $n = 24$ observations and studied this relationship by the following linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where y and x represent the minutes served and the number of units, respectively.

They found that the least squares estimate of regression parameters are $\hat{\beta}_0 = 37.213$ and $\hat{\beta}_1 = 9.969$, respectively, and the coefficient of determination is $R^2 = 0.900$. The high value of R^2 indicates a strong linear relationship between servicing time and the number of units repaired during a service call. However, they further observed that the residuals computed from the least squares estimate are not randomly distributed about zero and in fact these residuals move systematically from negative to positive and move back to negative as x increases. This problem happening in regression analysis was usually conquered by either variables transformation or selecting a new regression model such as polynomial regression or nonlinear regression models. It has been very scarcely developing a model with error variable which has distribution other than normality.

Based on the residuals computed from the least squares estimate, we consider the χ^2 goodness-of-fit test if the error variable follows an exponential distribution or Gamma distribution. In partition of the residuals into groups of numbers 4 and 5, the p-values are 0.2231 and 0.0534 respectively that both accept the null hypothesis of exponential distribution. We further found that the p-value for the Kolmogorov-Smirnov (K-S) test is 0.076 which also accepts the null hypothesis of choosing the exponential distribution.

Let's denote $a_0(x) = x' \begin{pmatrix} 37.21 \\ 9.969 \end{pmatrix}$. The following table display the estimated regression median and mode type intervals.

Table 2. Estimated regression median and mode type intervals for the computer repairing data

$1-2\alpha$	$\hat{C}_{med}(1-2\alpha)$	$\hat{C}_{mod}(1-2\alpha)$
0.6	$a_0(x) + (-24.85, 19.50)$	$a_0(x) + (-32.00, -2.678)$
0.7	$a_0(x) + (-26.79, 28.70)$	$a_0(x) + (-32.00, 6.527)$
0.8	$a_0(x) + (-28.62, 41.68)$	$a_0(x) + (-32.00, 19.50)$
0.9	$a_0(x) + (-30.35, 63.86)$	$a_0(x) + (-32.00, 41.68)$
0.95	$a_0(x) + (-31.18, 86.04)$	$a_0(x) + (-32.00, 63.86)$

We draw a graph of $1-2\alpha = 0.7$ in Figure 2.

How can we apply the results in this table in helping make decision for this company? We now let $1-2\alpha = 0.9$, considering the 90% estimates of these two intervals and list the corresponding interval estimates for y given x

Table 3. Estimated regression median and mode type intervals of 90% for the computer repairing data

# of units x	$\hat{C}_{med}(0.9)$	$\hat{C}_{mod}(0.9)$
2	$(26.79, 121.0)$	$(25.14, 98.82)$
4	$(46.73, 140.9)$	$(45.08, 118.7)$
6	$(66.67, 160.8)$	$(65.02, 138.7)$
8	$(86.61, 180.8)$	$(84.96, 158.64)$
10	$(106.5, 200.7)$	$(104.9, 178.5)$
12	$(126.4, 220.6)$	$(124.8, 198.5)$
14	$(146.4, 240.6)$	$(144.7, 218.4)$
16	$(166.3, 260.5)$	$(164.7, 238.3)$

Consider the example of $x = 4$ to explain. Median interval and mode interval techniqes estimated that with 90% confidence the computer engineers may spend 46.73 to 140.9 minutes and 45.08 to 118.7 minutes, respectively. The wider range of interval estimate by the median interval technique makes the company more difficult in predicting the number of engineers for servicing the customers. This result reveals the fact very significant in the contribution of mode interval in estimating an interval for conditional random variable y given covariate x with some fixed confidence coefficient $1 - 2\alpha$.

As the second example, we consider the analysis of a cloud point of a liquid data. The cloud point is a measure of the degree of crystallization in a stock and can be measured by the refractive index and the percentage of I-8 in the base stock can be used as a predictor for the cloud point. There is a data of sample size 19 containing variables of percentage of I-8 and cloud point in Draper and Smith (1981) and has been analyzed by linear regression by Rousseuw and Leroy (1987).

Rousseuw and Leroy considered the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where x and y represent, respectively, the variables of percentage of I-8 and cloud point and the least squares estimate is $(\hat{\beta}_0, \hat{\beta}_1) = (23.35, 1.05)$ and R^2 is 0.955. However, they also observed that the residuals based on least squares estimate revealing systematically in the way that larger-valued residuals are relatively more densely spread whereas those with smaller-valued residuals are not. They argued the nonrandom display of residuals and further studied the data with multiple linear regression model by adding this linear regression model a quadratic term of x.

High value of R^2 already revealed that the simple linear regression model is satisfied in sense of fitting this data set and then adding an extra term to enlarge a bit for the R^2 may not be a correct direction to deal with this problem. We still treat this a simple linear regression model, however, we want to observe if there is more suitable distribution for error variable ϵ.

Based on the least squares residuals and considering the exponential distribution for the null hypothesis, we compute the p-values for groups of $3, 4$ and 5 that are, respectively, $0.3907, 0.2613$ and 0.7379. We also compute the K-S goodness-of-fit test that is 0.6034. All are completely not significant to reject the null hypothesis. Moreover, we also test several hypotheses with some Gamma distributions , however, the p-values revealed significant to reject the null hypotheses.

Let's denote $a_1(x) = x' \begin{pmatrix} 23.35 \\ 1.05 \end{pmatrix}$. The following table display the estimated regression median and mode type intervals.

Table 4. Estimated regression median and mode type intervals for the liquid data

$1-2\alpha$	$\hat{C}_{med}(1-2\alpha)$	$\hat{C}_{mod}(1-2\alpha)$
0.6	$a_1(x) + (-0.609, 0.777)$	$a_1(x) + (0.083, 1.000)$
0.7	$a_1(x) + (-0.897, 0.837)$	$a_1(x) + (-0.204, 1.000)$
0.8	$a_1(x) + (-1.303, 0.895)$	$a_1(x) + (-0.609, 1.000)$
0.9	$a_1(x) + (-1.996, 0.949)$	$a_1(x) + (-1.303, 1.000)$
0.95	$a_1(x) + (-2.690, 0.975)$	$a_1(x) + (-1.996, 1.000)$

A graph of these conditional quantile estimate of $1 - 2\alpha = 0.7$ is displayed in Figure 3.

Suppose that we let $1 - 2\alpha = 0.95$ and $x = 8$. We find that the median interval estimate is $(29.06, 32.72)$ and the mode interval technique estimate is $(29.75, 32.75)$. We may say that with 95% confidence when the percentage of I-8 is 8 the cloud point of a liquid is between 29.75 and 32.75 when we use the mode interval technique.

5. Nonparametric Study and Monte Carlo Simulation

In the previous work in this paper, the observations were assumed to come from some underlying distribution, whose general form is assumed known. If these assumptions about the shape of the distribution are not made, then a nonparametric method to estimate the mode interval $C_{mod} = (x'\beta + F^{-1}(\alpha^*), x'\beta + F^{-1}(\gamma + \alpha^*))$ must be used.

Consider again the linear regression model

$$y_i = x_i'\beta + \epsilon_i, i = 1, ..., n. \tag{5.1}$$

Let $\hat{\beta}$ be the least squares estimator for the linear regression model (5.1) and $e_{(1)}, e_{(2)}$, ..., $e_{(n)}$ be the order statistics of the residuals $e_i = y_i - x_i'\hat{\beta}, i = 1, ..., n$, based on the least squares estimator. By letting $k = [n(1-2\alpha)]$, where $[n(1-2\alpha)]$ denotes the largest integer less than or equal to $n(1-2\alpha)$, we define $\ell^* = \text{argmin}_i\{h_i = e_{(k+i-1)} - e_{(i)}, i = 1, ..., n-k+1\}$. This means that $h_{\ell^*} = e_{(k+\ell^*-1)} - e_{(\ell^*)}$ is the shortest width of k order statistics interval $[e_{(i)}, e_{(k+i-1)}]$. The nonparametric estimator of the regression mode interval is

$$\hat{C}_{mod} = x'\hat{\beta} + (e_{(\ell^*)}, e_{(k+\ell^*-1)}). \tag{5.2}$$

For comparison, we here also define the nonparametric estimator of the symmetric interval C_{med}. Let $k_1 = [n\alpha]$ and $k_2 = [n(1-\alpha)]$. We define

$$\hat{C}_{med} = x'\hat{\beta} + (e_{(k_1)}, e_{(k_2)}). \tag{5.3}$$

For a simulation study in comparison of the two nonparametric estimators, we consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, ..., n.$$

For all cases, we consider sample size $n = 30$ and the sample of x_i are constant i plus an error drawn from $N(0,1)$. Also, the replication is set to be $N = 10,000$. In each replication, we compute the estimates, denoted by $\hat{C}_{mod}(1-2\alpha) = (\hat{\beta}_0 + \hat{\beta}_1 x + e_{(\ell^*)}, \hat{\beta}_0 + \hat{\beta}_1 x + e_{(k+\ell^*-1)}))$ and $\hat{C}_{med}(1-2\alpha) = (\hat{\beta}_0 + \hat{\beta}_1 x + e_{(k_1)}, \hat{\beta}_0 + \hat{\beta}_1 x + e_{(k_2)})$, of mode type interval and median type interval. We define the following mean squares errors

$$MSE_{med} = \frac{1}{10,000}\sum_{j=1}^{10,000}\{\frac{1}{n}\sum_{i=1}^{n}([\hat{\beta}_0 + \hat{\beta}_1 x_i + e_{(k_1)} - (\beta_0 + \beta_1 x_i + F^{-1}(\alpha))]^2,$$
$$[\hat{\beta}_0 + \hat{\beta}_1 x_i + e_{(k_2)} - (\beta_0 + \beta_1 x_i + F^{-1}(1-\alpha))]^2)\}$$

and

$$MSE_{mod} = \frac{1}{10,000}\sum_{j=1}^{10,000}\{\frac{1}{n}\sum_{i=1}^{n}([\hat{\beta}_0 + \hat{\beta}_1 x_i + e_{(\ell^*)} - (\beta_0 + \beta_1 x_i + F^{-1}(\alpha^*))]^2,$$
$$[\hat{\beta}_0 + \hat{\beta}_1 x_i + e_{(k+\ell^*-1)} - (\beta_0 + \beta_1 x_i + F^{-1}(1-2\alpha+\alpha^*))]^2)\}.$$

With these nonparametric methods to estimate the two regression quantile intervals, it is then interesting to compare their performance in estimating their corresponding quantile intervals. Two factors may affect the performance of estimation of these two quantile intervals. First, the population mode interval is estimated by the nonparametric shortest width technique and the median quantile interval is estimated by the empirical distribution technique. Various techniques may induce different effects. Second, the mode interval itself is a parameter located at points with densities substantially high and the median interval is a parameter located at points one with high density value and one possibly relatively small density value. This fact also affect the performance of their estimation.

We consider several error distributions to compare their MSE's. In the first case, we consider that error variable ϵ follows the exponential distribution which is skewed to the right.

Table 5. MSE's with exponential error distributions

$1-2\alpha$	$\hat{C}_{med}(1-2\alpha)$	$\hat{C}_{mod}(1-2\alpha)$
$\lambda = 4$		
0.6	$0.7926, 3.1157$	$0.3489, 0.3264$
0.7	$0.8517, 2.6810$	$0.7544, 0.7216$
0.8	$0.6001, 3.7193$	$0.3858, 0.4919$
0.9	$0.2880.3.6290$	$0.2467, 0.7205$
$\lambda = 6$		
0.6	$0.3995, 1.8410$	$0.3626, 0.3056$
0.7	$0.5916, 2.5211$	$0.4664, 0.4386$
0.8	$0.8708, 4.2557$	$0.6127, 0.7312$
0.9	$0.7070, 5.3088$	$0.5824, 1.2413$
$\lambda = 8$		
0.6	$0.4968, 2.2954$	$0.3034, 0.2601$
0.7	$0.3500, 1.9597$	$0.3217, 0.2777$
0.8	$0.3493, 2.4464$	$0.3029, 0.3422$
0.9	$0.6109, 4.8872$	$0.5081, 1.1224$
$\lambda = 10$		
0.6	$0.5049, 2.4783$	$0.2319, 0.1988$
0.7	$0.4884, 2.6502$	$0.3045, 0.2815$
0.8	$0.4647, 3.2728$	$0.2989, 0.3800$
0.9	$0.4885, 3.3567$	$0.4762, 0.8637$

In this simulation result, the MSE's for mode type interval are relatively smaller than the corresponding one's for median type interval. We have two conclusions drawn from the results in Table 5:

(a). The location of the mode interval seems to be an important factor for efficiency of the interval estimation. Especially the second elements of the MSE's for mode interval are extremely smaller than those performed for median interval since the right ends of median intervals are too far from their corresponding mode points.

(b). The smaller values of the left elements of the MSE's for mode interval than those of median intervals indicates that the technique of shortest width does more efficient than the technique by empirical distribution.

Since the exponential distribution is a very skewed distribution, we here display the simulation results for a lightly skewed gamma distribution ($Gamma(a, b), a \neq 1$).

Table 6. MSE's with gamma error distributions $Gamma(a, b)$

$1-2\alpha$	$\hat{C}_{med}(1-2\alpha)$	$\hat{C}_{mod}(1-2\alpha)$
$(a, b) = (2, 5)$		
0.6	$0.2218, 0.8877$	$0.1979, 0.1349$
0.7	$0.2104, 1.0493$	$0.1503, 0.1419$
0.8	$0.7127, 2.1316$	$0.2846, 0.5398$
0.9	$0.3597, 2.1710$	$0.2195, 0.6256$
$(a, b) = (2, 9)$		
0.6	$0.3216, 1.0816$	$0.2297, 0.2123$
0.7	$0.4000, 1.3547$	$0.2522, 0.2986$
0.8	$0.1844, 1.2082$	$0.1286, 0.1842$
0.9	$0.6519, 3.0653$	$0.3305, 0.9976$
$(a, b) = (4, 2)$		
0.6	$0.1785, 0.4250$	$0.1754, 0.1678$
0.7	$0.1149, 0.4399$	$0.0882, 0.1183$
0.8	$0.1688, 0.4877$	$0.1870, 0.1925$
0.9	$0.2231, 1.1495$	$0.0973, 0.4836$
$(a, b) = (4, 8)$		
0.6	$0.4693, 0.9646$	$0.2842, 0.4598$
0.7	$0.0904, 0.1885$	$0.2063, 0.0759$
0.8	$0.1914, 0.7077$	$0.1119, 0.2554$
0.9	$0.1311, 0.6452$	$0.1503, 0.2487$

Two conclusions drawn from the Table 6:

(a). Since the second elements of MSE's for mode interval are relatively smaller than those for median interval, this does verify the effect of location in the way that the right ends of mode intervals are relatively closer than those of median intervals to mode points.

(b). For comparing the left elements of MSE's for $\hat{C}_{med}$ and $\hat{C}_{mod}$, the effect of technique seems does exist but not strong enough to cover the effect of location.

In case the error random variable obeys a symmetric distribution, the regression mode interval coincides the symmetric interval, i.e. $C_{mod} = C_{med}$. It is then interesting to see if the shortest width method in the nonparametric estimation can gain any advantage in estimation of the parameter. In the next table, we display the result induced from a simulation with a normal error variable.

Table 7. MSE's with normal error distributions

$1-2\alpha$	$\hat{C}_{med}(1-2\alpha)$	$\hat{C}_{mod}(1-2\alpha)$
0.6	$0.1496, 0.3328$	$0.1402, 0.3240$
0.7	$0.2017, 0.2758$	$0.1431, 0.2902$
0.8	$0.2330, 0.1799$	$0.1303, 0.1632$
0.9	$0.3231, 0.1138$	$0.2170, 0.1612$

In this symmetric distribution, there is no location effect for either one interval. Then the simulation results revealed that the shortest width to estimate a quantile interval does a bit more efficient than the empirical quantile technique in estimation of a quantile interval.

6. Trimmed Means based on Regression Mode Interval

The most popular technique in estimating the regression parameter vector β is the least squares estimation. Although this estimator has some interesting theoretical properties from both the parametric and nonparametric points of view. However, it is sensitive to departures from the normality and to the presence of outliers. Hence, we need to consider robust estimation.

Among many robust estimators proposed as alternatives to the least squares estimator, the trimmed mean has the advantages of simple computation and efficiency (see Ruppert and Carroll (1980) and Bickel (1973)). Basically, the trimmed mean in regression is the least squares estimator based on those observations lying in the sample median type interval $\hat{C}_{med} = (x'\hat{\beta}(\alpha), x'\hat{\beta}(1-\alpha))$. With the efficiency of a relatively smaller MSE of the mode type interval for nonparametric estimation than that of the median type interval, it is then interesting in seeing if the regression parameters may be more efficiently estimated by the trimmed mean for those observations lying in the sample mode type interval $\hat{C}_{mod}$.

For the linear regression model of (5.1), we inherit the notations of LSE ($\hat{\beta}$), mode interval estimate ($\hat{C}_{mod}$) and median interval estimate ($\hat{C}_{med}$),etc. Denote the matrices $y = (y_1, ..., y_n)'$ and $X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$. The mode type trimmed mean is defined as

$$\hat{\beta}_{mod} = (X'AX)^{-1}X'Ay,$$

where $A =$ diag $\{I_{\hat{C}_{mod}}(e_1), \ldots, I_{\hat{C}_{mod}}(e_n)\}$. On the other hand, the median type trimmed mean is defined as

$$\hat{\beta}_{med} = (X'BX)^{-1}X'By,$$

where $B =$ diag $\{I_{\hat{C}_{med}}(e_1), \ldots, I_{\hat{C}_{med}}(e_n)\}$.

For simplicity, we consider the simple linear regression and assume that the true regression parameter is $(1, 1)'$. With sample size $n = 30$, we randomly generate error variable ϵ from the following mixed gamma distribution

$$(1 - \delta)[\frac{1}{\sqrt{ab}}Gamma(a, b) - \sqrt{a}] + \delta[\frac{\sigma}{\sqrt{ab}}Gamma(a, b) - \sigma\sqrt{a}].$$

The covariates x_i are randomly generated from $i + N(0, 1), i = 1, ..., n$, and we let replication number $m = 10,000$. For the jth replication, we denote the corresponding estimates of mode and median types as, respectively, $\hat{\beta}^j_{mod}$ and $\hat{\beta}^j_{med}$. Finally we compute the following MSE's:

$$MSE_{mod} = \frac{1}{2m}\sum_{j=1}^{m}(\hat{\beta}^j_{mod} - \beta)'(\hat{\beta}^j_{mod} - \beta)$$

and

$$MSE_{med} = \frac{1}{2m}\sum_{j=1}^{m}(\hat{\beta}^j_{med} - \beta)'(\hat{\beta}^j_{med} - \beta).$$

In Tables 8 and 9, we, respectively, list the MSE's for the mixed gamma distribution with $(a, b) = (4, 1)$ and $(4, 5)$, where $(a, b) = (4, 1)$ represents a lightly skewed distribution and $(a, b) = (4, 5)$ represents a heavily skewed distribution.

Two conclusions may be drawn from Tables 8 and 9:

(a). Under these heavy-tailed distributions, trimmed means based on median and mode type intervals are with MSE's uniformly more smaller than those of the LSE's. This indicates that these two trimmed means are robust estimators.

(b). Besides cases of $1-2\alpha = 0.8, \delta = 0.1$, and $\sigma = 5$ for both $Gamma(4,1)$ and $Gamma(4,5)$, the trimmed means based on mode type intervals are with MSE's smaller than those of trimmed means based on median intervals. This shows that the mode interval is relatively more efficient than the median interval for constructing trimmed means when asymmetric errors exist.

References

Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *The Annals of Statistics* **1**, 597-616.

Chatterjee, S. and Price, B. (1991). *Regression Analysis by Example.* New York: Wiley.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis.* New York: Wiley.

Huang, J.-Y. (2003). Mode interval and its application to construct a new Shewhart control chart. Ph.D. dissertation, National Chiao Tung University.

Koenker, R. and Bassett, G. J. (1978). Regression quantile. *Econometrica*, **46**, 33-50.

Koenker, R. and d'Orey, V. (1987). Computing regression quantiles. *Applied Statistics.* **36**, 383-389.

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

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association* **75**, 828-838.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* New York: Wiley.

Table 8. MSE's for trimmed means and LSE under Gamma distribution $((a,b) = (4,1))$

$1-2\alpha$	LSE	$\hat{\beta}_{med}(1-2\alpha)$	$\hat{\beta}_{mod}(1-2\alpha)$
$(\delta,\sigma)=(0.1,5)$			
0.8	0.2348	0.1107	0.1240
0.9		0.1919	0.1069
0.95		0.1795	0.1218
$(\delta,\sigma)=(0.1,10)$			
0.8	0.7744	0.1593	0.1395
0.9		0.5355	0.1475
0.95		0.5067	0.2342
$(\delta,\sigma)=(0.1,25)$			
0.8	4.3547	0.5082	0.2077
0.9		2.9722	0.4557
0.95		2.7246	0.9244
$(\delta,\sigma)=(0.2,5)$			
0.8	0.4098	0.1865	0.1580
0.9		0.3334	0.2071
0.95		0.3135	0.2470
$(\delta,\sigma)=(0.2,10)$			
0.8	1.4608	0.4266	0.2225
0.9		1.0733	0.5269
0.95		1.0284	0.7300
$(\delta,\sigma)=(0.2,25)$			
0.8	8.7924	2.0073	0.6712
0.9		6.3253	2.7722
0.95		5.9835	4.1939
$(\delta,\sigma)=(0.3,5)$			
0.8	0.5790	0.3087	0.2497
0.9		0.4870	0.3814
0.95		0.4692	0.4238
$(\delta,\sigma)=(0.3,10)$			
0.8	2.0762	0.8989	0.5793
0.9		1.7003	1.2455
0.95		1.7221	1.4967
$(\delta,\sigma)=(0.3,25)$			
0.8	13.160	5.2867	3.0413
0.9		10.491	7.5763
0.95		10.169	9.2311

Table 9. MSE's for trimmed means and LSE under Gamma distribution ($(a,b) = (4,5)$)

$1-2\alpha$	LSE	$\hat{\beta}_{med}(1-2\alpha)$	$\hat{\beta}_{mod}(1-2\alpha)$
$(\delta,\sigma) = (0.1, 5)$			
0.8	0.2384	0.1155	0.1242
0.9		0.1926	0.1064
0.95		0.1786	0.1187
$(\delta,\sigma) = (0.1, 10)$			
0.8	0.7771	0.1659	0.1391
0.9		0.5557	0.1501
0.95		0.5069	0.2202
$(\delta,\sigma) = (0.1, 25)$			
0.8	4.4530	0.4475	0.1976
0.9		2.8180	0.4432
0.95		2.7492	1.0001
$(\delta,\sigma) = (0.2, 5)$			
0.8	0.4075	0.1848	0.1552
0.9		0.3439	0.2056
0.95		0.3148	0.2511
$(\delta,\sigma) = (0.2, 10)$			
0.8	1.4219	0.4203	0.2194
0.9		1.0999	0.5257
0.95		1.0345	0.7351
$(\delta,\sigma) = (0.2, 25)$			
0.8	8.6749	2.1516	0.6554
0.9		6.5789	2.8726
0.95		6.3160	4.3402
$(\delta,\sigma) = (0.3, 5)$			
0.8	0.5612	0.2995	0.2504
0.9		0.4745	0.3766
0.95		0.4477	0.4145
$(\delta,\sigma) = (0.3, 10)$			
0.8	2.1589	0.8969	0.5820
0.9		1.7102	1.2248
0.95		1.7013	1.5330
$(\delta,\sigma) = (0.3, 25)$			
0.8	13.079	5.0684	2.8441
0.9		10.239	7.3694
0.95		10.154	9.4117