Trimming the Least Squares Estimator in the Linear Model By Using a Preliminary Estimator

by

David Ruppert and Raymond J. Carroll

Institute of Statistics Mimeo Series #1220

April 1979

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

Trimming the Least Squares Estimator in the Linear Model By Using a Preliminary Estimator

by

David Ruppert* and Raymond J. Carroll**

Abstract

Let $\hat{\underline{\beta}}_0$ be an estimate of $\underline{\beta}$ in the linear model, $Y_i = \underline{x}_i'\underline{\beta} + e_i$. Define the residuals $Y_i - \underline{x}_i' \hat{\underline{\beta}}_0$, let $0 < \alpha < \frac{1}{2}$, and let $\hat{\underline{\beta}}_L$ be the least squares estimate of $\underline{\beta}$ calculated after removing the observations with the $[\alpha n]$ smallest and $[\alpha n]$ largest residuals. By use of an asymptotic expansion, the limit distribution of $\hat{\beta}_L$ is found under certain regularity conditions. This distribution depends heavily upon the choice of $\hat{\underline{\beta}}_0$. We discuss several choices of $\hat{\underline{\beta}}_0$, with special attention to the contaminated normal model. If $\hat{\underline{\beta}}_0$ is the median regression or least squares estimator then $\hat{\underline{\beta}}_L$ is rather inefficient at the normal model. If F is symmetric, then a particularly convenient, robust choice is to let $\hat{\underline{\beta}}_0$ equal the average of the αth and $(1-\alpha)$th regression quantiles (Koenker and Bassett, Econometrica (1978)). Then $\hat{\underline{\beta}}_L$ has a limit distribution analogous to the trimmed mean in the location model, and the covariance matrix of $\hat{\underline{\beta}}_L$ is easily estimated.

Key Words and Phrases: Linear model, trimmed least squares, robustness, regression quantiles, preliminary estimator, median regression

AMS 1970 Subject Classifications: Primary 62G35; Secondary 62J05, 62J10.

David Ruppert is an Assistant Professor and Raymond J. Carroll an Associate Professor, both at the Department of Statistics, the University of North Carolina at Chapel Hill. This research was supported by the National Science Foundation Grant NSF MCS78-01240 and the Air Force Office of Scientific Research under contract AFOSR-75-2796.

1. Introduction.

This paper is concerned with the linear model

$$\underline{y} = X\underline{\beta} + \underline{z}, \qquad (1.1)$$

where $\underline{y}' = (y_1,\ldots,y_N)$, X is a N×p matrix of known constant, $\underline{\beta}' = (\beta_1,\ldots,\beta_p)$ is a vector of unknown parameters, and $\underline{z}' = (z_1,\ldots,z_N)$ is a vector of i.i.d. random variables with distribution function F. The least squares estimator of $\underline{\beta}$ is said to be non-robust because it possesses two serious disadvantages, inefficiency when F has heavier tails than the Gaussian distribution and high sensitivity to spurious observations. These deficiencies are closely related and Huber (1977, p. 3) states that "for most practical purposes, 'distributional robust' and 'outlier resistant' are interchangeable". In the location model, three classes of estimators have been proposed to overcome these deficiencies: M, L, and R estimators; see Huber (1977) for an introduction. Among the L-estimates, the trimmed mean is particularly attractive because it is easy to compute, is rather efficient under a variety of circumstances, and can be used to form confidence intervals (Gross (1976) and Huber (1970)). Hogg (1974) favors trimmed means for the above reasons, and because they can serve as a basis for adaptive estimators. Stigler (1977) applied robust estimators to historical data and concluded that "the 10% trimmed mean (the smallest nonzero trimming percentage included in the study) emerges as the recommended estimator". It is therefore natural to seek a trimmed least squares estimator for the general linear model which possess these desirable properties of the trimmed mean.

For the linear model, Bickel (1973) has proposed a class of one-step L-estimators depending on a preliminary estimate of $\underline{\beta}$, but, while these have good asymptotic efficiencies, they are computationally complex and are generally not invariant to reparameterization.

Recently, Koenker and Bassett (1978) have extended the concept of quantiles to the general linear model. They suggest the following trimmed least squares estimator, $\hat{\underline{\beta}}_{KB}(\alpha)(=\hat{\underline{\beta}}_{KB})$: define the αth and $(1-\alpha)$th regression quantiles $\hat{\underline{\beta}}(\alpha)$ and $\hat{\underline{\beta}}(1-\alpha)$ (see their paper for a definition of regression quantile), remove from the sample any observation whose residual from $\hat{\underline{\beta}}(\alpha)$ is negative or whose residual from $\hat{\beta}(1-\alpha)$ is positive, and calculate the least squares estimator using the remaining observations. In the location model, this estimator reduces to the α-trimmed mean. Ruppert and Carroll (1978) studied the large sample behavior of $\hat{\underline{\beta}}_{KB}$ (p fixed and $N \to \infty$) and found that the variance of $N^{\frac{1}{2}}\,\hat{\beta}_{KB}(\alpha)$ is approximately $\sigma^2(\alpha,F)\ (N^{-1}\ X'X)^{-1}$, where in the location model $\sigma^2(\alpha,F)$ is the asymptotic variance under F of the α-trimmed mean (also normalized by $N^{\frac{1}{2}}$).

In this paper we investigate a class of estimators that represent a third possible method of defining a regression analogue of the trimmed mean. Specifically, let $\hat{\underline{\beta}}_0$ be a preliminary estimator. Form the residuals from $\hat{\underline{\beta}}_0$ and remove from the sample those observations corresponding to the $[N\alpha]$ smallest and $[N\alpha]$ largest residuals. Then the α-trimmed least squares estimator, $\hat{\underline{\beta}}_L(\alpha)\ (=\hat{\underline{\beta}}_L)$, is a least squares estimator using the remaining observations.

The definition of $\hat{\underline{\beta}}_L$ was motivated by the applied statisticians' practice of examining the residuals from a least squares fit, removing the points with large (absolute) residuals, and recalculating the least squares solution with the remaining observations. Generally, there is no formal rule for deciding which points to remove, but $\hat{\underline{\beta}}_L$ is at least similar to this practice. Furthermore, the authors do know of practitioners who have used $\hat{\underline{\beta}}_L$.

Theorems 1 and 2, which are a general results allowing a wide class of preliminary estimates, give asymptotic representations for $\hat{\underline{\beta}}_L$. These representation enables one to calculate the asymptotic bias (which is 0 if F is symmetric and $\hat{\underline{\beta}}_0$ is unbiased) and variance of $\hat{\underline{\beta}}_L$.

When the preliminary estimate is the least squares or median regression (L_1) estimate, two somewhat surprising conclusions emerge. First, for neither choice is $\hat{\beta}_L$ a multivariate analogue to a trimmed mean. Second, either choice causes $\hat{\beta}_L$ to be inefficient at the normal model, particularly when compared to the Koenker and Bassett estimate or the M-estimates. For symmetric F, we show that the "right" choice of a preliminary estimate is a regression analogue to averaging the α^{th} and $(1-\alpha)^{th}$ sample quantiles.

Hogg (1974, p. 917) mentions that adaptive estimators can be constructed from estimators similar or identical to $\underline{\hat{\beta}}_L(\alpha)$ with α a function of the residuals from $\underline{\hat{\beta}}_0$. The advantage of this class of adaptive estimators, he feels, is that they "would correspond more to the trimmed means for which we can find an error structure". However, from the above results, we can conclude, that even if α is non-stochastic, estimators of the type suggested by Hogg will not necessarily have error structures which correspond to the trimmed mean.

The methods of this paper can be applied to estimators similar to $\underline{\hat{\beta}}_L$. For example, let $\underline{\hat{\beta}}_A(\alpha)(=\underline{\hat{\beta}}_A)$ be the least squares estimate after the points with the $[2\alpha N]$ largest absolute residulas from $\underline{\hat{\beta}}_0$ are removed. In section 6 we state results for $\underline{\hat{\beta}}_A$. Their proofs are omitted, but are similar to the proofs of analogous results for $\underline{\hat{\beta}}_L$.

2. Notation and Assumptions. Although $\underline{y}$, X and $\underline{z}$ in (1.1) depend upon N, this will not be made explicit in the notation. Let $\underline{e}' = (1,0,\ldots,0)$ $(1\times p)$ and let I_p be the p×p identity matrix. For $0 < p < 1$, define $\xi_p = F^{-1}(p)$.

Throughout, we will make the following three assumptions.

C1. $N^{\frac{1}{2}}(\underline{\hat{\beta}}_0-\underline{\beta}) = 0_p(1)$

C2. Fix $0 < \alpha < 1$, and define $\xi_1 = \xi_\alpha$ and $\xi_2 = \xi_{1-\alpha}$.

Assume F has a continuous positive density f in neighborhoods of ξ_1 and ξ_2.

C3. Assume $X_{i1} = 1$ for $i = 1,\ldots,N$, (2.1) $$\lim_{N\to\infty} [N^{-\frac{1}{2}} \max_{i\le N, j\le p} |X_{ij}|] = 0 ,$$

(2.2) $$\sum_{i=1}^{N} X_{ij} = 0 \quad \text{for} \quad j = 2,\ldots,p,$$

i.e., the design is centered, and for Q some positive definite matrix

(2.3) $$\lim_{N\to\infty} N^{-1} X'X = Q.$$

Note taht the probability distribution of $\underline{Y}$ is unchanged if we replace $\underline{\beta}$ by $\underline{\beta} + \theta\underline{e}$ and F(.) by F(· + θ) where θ is any real number. Because of (2.2), many possible preliminary estimates, $\hat{\underline{\beta}}_0$, satisfy

$$N^{\frac{1}{2}}(\hat{\underline{\beta}}_0 - \underline{\beta} - \theta\underline{e}) = 0_p(1)$$

for some θ. In particular, the LAD (least absolute deviation or median regression) estimate has this property (Ruppert and Carroll (1978)). In this case, we can reparameterize so that C1 holds.

The residuals from the preliminary estimate $\hat{\underline{\beta}}_0$ are

(2.4) $$r_i = y_i - \underline{x}_i'\hat{\underline{\beta}} = z_i - \underline{x}_i'(\hat{\underline{\beta}} - \underline{\beta}).$$

Let r_{1N} and r_{2N} be the [Nα]th and [N(1-α)]th ordered residuals, respectively. Then the estimate $\hat{\underline{\beta}}_L$ is a least squares (LS) estimate calculated after removing all observations satisfying

(2.5) $$r_i \le r_{1N} \quad \text{or} \quad r_i \ge r_{2N} .$$

Because of C2, asymptotic results are unaffected by requiring strict inequalities in (2.5). Let $a_i = 0$ or 1 according as i satisfies (2.1) or not, and let A be the N×N diagonal matrix with $A_{ii} = a_i$. Thus

$$\hat{\underline{\beta}}_L(\alpha) = (X'AX)^- X'A\underline{y},$$

where $(X'AX)^-$ is a generalized inverse for $X'AX$. (Later we show that $N^{-1}(X'AX) \overset{P}{\to} (1-2\alpha)Q$, whence $P(X'AX$ is invertible$) \to 1$.)

3. Main Results. The analysis of the asymptotic behavior of $\hat{\beta}_L(\alpha)$ relies heavily on techniques developed by Ruppert and Carroll (1978). The proofs are sketched in the appendix. Lemma 1, which may be of some interest in intself, is an asymptotic linearity result and is a generalization of work by Bahadur (1966) and Ghosh (1971) for the location model.

For $0 < \theta < 1$, define

$$\psi_\theta(x) = \theta - I(x < 0) \ . \tag{3.1}$$

Lemma 1. For $\theta = \alpha$ or $(1-\alpha)$, let $r_{\theta N}$ be the $[N\theta]$th ordered residual. Then,

$$N^{\frac{1}{2}}(r_{\theta N}-\xi_\theta) = f(\xi_\theta)^{-1}[N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi_\theta(z_i-\xi_\theta)] - \underline{e}_N' N^{\frac{1}{2}}(\hat{\underline{\beta}}_0-\underline{\beta}) + o_p(1) \ . \tag{3.2}$$

Theorem 1. Define $a = \xi_2 f(\xi_2) - \xi_1 f(\xi_1)$, $\underline{c}_i = (I-\underline{e}'\underline{e})x_i = (0,x_{i2},\ldots,x_{ip})'$, and

$$h(x) = xI(\xi_1 \le x \le \xi_2) + \xi_2(I(x > \xi_2) - \alpha) + \xi_1(I(x < \xi_1) - \alpha) \ . \tag{3.3}$$

Then,

$$(1-2\alpha)N^{\frac{1}{2}}(\hat{\underline{\beta}}_L-\underline{\beta}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} Q^{-1} \underline{c}_i \ z_i \ I(\xi_1 \le z_i \le \xi_2) + N^{-\frac{1}{2}} \sum_{i=1}^{N} \underline{e} \ h(z_i) + a \ N^{\frac{1}{2}} (I-\underline{e} \ \underline{e}')(\hat{\underline{\beta}}_0-\underline{\beta}) + o_p(1) \ . \tag{3.4}$$

For our next theorem we require another condition.

C4. For some function g,

$$N^{\frac{1}{2}} (\hat{\underline{\beta}}_0-\underline{\beta}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} Q^{-1} \underline{x}_i \ g(z_i) + o_p(1) \ .$$

As is well-known, C4 holds with $g(x) = x$ if $\hat{\underline{\beta}}_0$ is the LS estimate. By Ruppert and Carroll (1978), Theorem 2), C4 holds with $g(x) = (f(F^{-1}(0)))^{-1}(\frac{1}{2} - I(x < F^{-1}(0)))$ if $\hat{\underline{\beta}}_0$ is the LAD estimate. As a consequence of Theorem 1, we have our main result.

Theorem 2. Assume C4. Then

$$(3.5) \qquad (1-2\alpha)\, N^{\frac{1}{2}}(\hat{\underline{\beta}}_L - \underline{\beta}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} Q^{-1} \underset{\sim}{C}_i \{Z_i\, I(\xi_1 \leq Z_i \leq \xi_2) + a\, g(Z_i)\} + N^{-\frac{1}{2}} \sum_{i=1}^{N} \underline{e}\, h(Z_i) + o_p(1).$$

As a special case of corollary 1, we obtain a result of deWet and Venter (1974).

Corollary 1. In the location model ($p=1$ and $x_i=1$ for all i)

$$(1-2\alpha)\, N^{\frac{1}{2}}(\hat{\underline{\beta}}_L - \hat{\underline{\beta}}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} h(Z_i) + o_p(1).$$

4. Asymptotics. In this section we show the Theorem 2 leads to the basic conclusions:

1) The intercept estimate is asymptotically unbiased if F is symmetric.

2) The slope estimates are asymptotically unbiased even if F is asymmetric.

3) The asymptotic variance of the intercept, which does not depend upon the choice of $\hat{\beta}_0$, is that of the trimmed mean in the location model.

4) The asymptotic covariance matrix of the slopes depends upon $\hat{\beta}_0$ and, in general, will be difficult to estimate.

Let $\underline{0}$ be a $(p-1) \times 1$ vector of zeroes. By (2.2), there is a $\tilde{Q}$ such that

$$(4.1) \qquad Q = \begin{bmatrix} 1 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \tilde{Q} \end{bmatrix} \quad \text{and} \quad Q^{-1} = \begin{bmatrix} 1 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \tilde{Q} \end{bmatrix}.$$

Moreover,

(4.2) $$N^{-1} \sum_{i=1}^{N} \underline{C}_i \underline{C}_i' = \begin{bmatrix} 0 & \underset{\sim}{0}' \\ \underset{\sim}{0} & \tilde{Q} \end{bmatrix}$$

and

(4.3) $$Q\, N^{-1} \sum_{i=1}^{N} \underline{C}_i = \begin{bmatrix} 0 \\ \underset{\sim}{0} \end{bmatrix}.$$

We will call the first entry of $\underline{\beta}$ the intercept and the remaining entries will be call the slopes. If we estimate $\underline{\beta}$ with $\hat{\underline{\beta}}_L$, then the asymptotic bias of the intercept is

$$E\, h(Z_1) = (1-2\alpha)^{-1} \int_{\xi_1}^{\xi_2} x\, dF(x),$$

which is zero if F is symmetric about zero. By (3.4) and (4.3) the slope estimates are asymptotically unbiased, even if F is asymmetric. The asymptotic variance of the intercept, normalized by $N^{\frac{1}{2}}$, is

(4.4) $$\sigma^2(\alpha,F) = (1-2\alpha)^{-2} \text{ Var } h(Z_i)$$

the asymptotic variance of the normalized α-trimmed in the location model. The intercept is asymptotically uncorrelated with the slopes, and the asymptotic covariance matrix of the normalized slopes is $\tilde{Q}^{-1} \sigma^2(\alpha,g,F)$ where

(4.5) $$\sigma^2(\alpha,g,F) = (1-2\alpha)^{-2} \text{ Var}(Z_1\, I(\xi_1 \leq Z_1 \leq \xi_2) + a\, g(Z_1)) .$$

We see that the asymptotic distribution of the intercept estimate does not depend upon the choice of $\hat{\underline{\beta}}_0$ provided $(\hat{\underline{\beta}}_0 - \underline{\beta}) = O_p(N^{-\frac{1}{2}})$.

On the other hand, we see from (3.4) that the slope estimates depend upon $\hat{\underline{\beta}}_0$, since the unusual situation where a = 0 is ruled out by assumption C2. Using the Lindeberg central limit theorem and corollary 1, it is easy to show that under C4, $N^{\frac{1}{2}} (\hat{\underline{\beta}}_L - \underline{\beta} - \underline{e}\, (1-2\alpha)^{-1} Eh(Z_1))$ converges in distribution to a normal law.

In general, large sample statistical inference based on $\hat{\underline{\beta}}_L$ will be a challenging problem, because of the difficulties of estimating $a = (\xi_2 f(\xi_2) - \xi_1 f(\xi_1))$. Obtaining reasonably good estimates of the density f might take very large sample sizes.

5. A Close Analog to the Trimmed Mean. There is one choice of $\hat{\underline{\beta}}_0$ (the average of the α^{th} and $(1-\alpha)^{th}$ "regression quantiles") for which the asymptotic covariance matrix of $\hat{\underline{\beta}}_L$ is relatively simple to estimate when F is symmetric about 0. For $0 < \theta < 1$, let $\hat{\underline{\beta}}(\theta)$ be the θ the regression quantile (Koenker and Bassett (1978)). Let $\xi(\theta) = F^{-1}(\theta)$ and define $\psi_\theta(x) = \theta - I(x < 0)$. By theorem 2 of Ruppert and Carroll (1978), if F has a continuous positive density f in a neighborhood of $\xi(\theta)$, then

$$(5.1) \quad N^{\frac{1}{2}} (\hat{\underline{\beta}}(\theta) - \underline{\beta} - \xi(\theta)\underline{e}) = N^{-\frac{1}{2}} (f(\xi(\theta))^{-1} Q^{-1} \sum_{i=1}^{N} \underline{x}_i \psi_\theta(z_i - \xi(\theta)) + o_p(1) .$$

Let $\hat{\underline{\beta}}_L(RQ)$ equal $\hat{\underline{\beta}}_L$ when $\hat{\beta}_0 = (\hat{\underline{\beta}}(\alpha) + \hat{\underline{\beta}}(1-\alpha))/2$. By C2 and (5.1) this $\hat{\underline{\beta}}_0$ satisfies (C4) with

$$g(x) = (2 f(\xi_1))^{-1} \psi_\alpha(x-\xi_1) + (2 f(\xi_2))^{-1} \psi_{1-\alpha}(x-\xi_2) .$$

If F is symmetric, then $\xi_1 = -\xi_2$, $f(\xi_1) = f(\xi_2)$, and therefore

$$(5.2) \qquad a\, g(x) = \xi_1 I(x \leq \xi_1) + \xi_2 I(x \geq \xi) .$$

By (3.5) and (5.2),

$$(5.3) \qquad (1-2\alpha) N^{\frac{1}{2}} (\hat{\underline{\beta}}_L - \underline{\beta}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} Q^{-1} \underline{x}_i h(Z_i) + o_p(1) ,$$

and therefore by (4.4),

$$N^{\frac{1}{2}}(\hat{\underline{\beta}}_L - \underline{\beta}) \xrightarrow{L} N(0, Q^{-1} \sigma^2(\alpha,F)) .$$

If we examine deWet and Venter's (1974) representation of the trimmed mean (cf. corollary 1 of this paper), we see that (5.3) is a generalization of their result to the general linear model. Therefore, this $\hat{\underline{\beta}}_0$ appears to be the "correct" choice. Also by theorem 3 of Ruppert and Carroll (1978)

$$N^{\frac{1}{2}}(\hat{\underline{\beta}}_{KB}-\hat{\underline{\beta}}_L) \xrightarrow{P} 0 \,, \tag{5.4}$$

so that asymptotically there is no difference between trimmed with this preliminary estimate and using Koenker and Bassett's (1978) proposal. (However, (5.4) does not necessarily hold if F is asymmetric.)

Let $\hat{\underline{\beta}}_L$(LS) and $\hat{\underline{\beta}}_L$(LAD) be $\hat{\underline{\beta}}_L$ when $\hat{\underline{\beta}}_0$ is the LAD and LS estimate, respectively. Table 1 displays $\sigma^2(\alpha,g,F)$ for several choices of α, ε, and b, and for g corresponding to $\hat{\underline{\beta}}_L$(LS), $\hat{\underline{\beta}}_L$(LAD), and $\hat{\underline{\beta}}_L$(RQ). For comparison purposes, we include the asymptotic variance of the LS estimate, Huber's proposal 2 M-estimate, and a one-step Hampel estimate using Huber's proposal 2 as a preliminary estimate (Huber's (1973), (1977)). (By asymptotic variance, we mean σ^2 where the asymptotic covariance is $\sigma^2 Q^{-1}$). For discussion of the last two estimates see Carroll and Ruppert (1979). Several conclusions emerge from Table 1.

1) $\hat{\underline{\beta}}_L$(LS) and $\hat{\underline{\beta}}$(LAD) are rather inefficient at the normal distribution.
2) $\hat{\underline{\beta}}_L$(RQ) is quite efficient at the normal model.
3) Under heavy contamination (b large or ε large) $\hat{\underline{\beta}}_L$(LS), $\hat{\underline{\beta}}$(LAD), and $\hat{\underline{\beta}}_L$(RQ) are relatively efficeint compared with LS. Also $\hat{\underline{\beta}}_L$(RQ) and $\hat{\underline{\beta}}_L$(LAD) compare well against the M-estimates, but $\hat{\underline{\beta}}_L$(LS) does poorly compared to the M-estimates if $\varepsilon = .25$, $b = 10$, and $\alpha = .25$. (Intuitively, one can expect that when $\alpha = .25$, $\hat{\underline{\beta}}_L$(LS) will be heavily influenced by its preliminary estimate, which estimates $\underline{\beta}$ poorly for these b and ε.)

Because of 1) and 3), the practice of fitting by least squares or LAD, removing points corresponding to extreme residuals, and computing the least squares estimate from the trimmed sample, is not an adequate substitute for robust methods of estimation.

If, instead of removing those observations with the [Nα] smallest and [Nα] largest residuals from $\hat{\underline{\beta}}_0$, we remove those observations with the [2Nα] largest absolute residuals, then the asymptotic variance of the intercept is the same as that of the slopes. Specially, let $\hat{\underline{\beta}}_A(\alpha)(=\hat{\underline{\beta}}_A)$ be the estimate formed in this manner. Then, if F is symmetric,

$$(6.1)\qquad (1-2\alpha)N^{\frac{1}{2}}\,(\hat{\underline{\beta}}_A-\beta) = N^{-\frac{1}{2}}\sum_{i=1}^{N} Q^{-1}\,\underline{x}_i\{Z_i\ I(\xi_1 \le Z_i \le \xi_2) + a(\hat{\underline{\beta}}_0-\underline{\beta})\}$$

and if C4 holds, then

$$(6.2)\qquad (1-2\alpha)N^{\frac{1}{2}}\,(\hat{\underline{\beta}}_A-\underline{\beta}) = N^{-\frac{1}{2}}\sum_{i=1}^{N} Q^{-1}\,\underline{x}_i\{Z_i\ I(\xi_1 \le Z_i \le \xi_2) + a\ g(Z_i)\}$$

which in the location case reduces to

$$(6.3)\qquad (1-2\alpha)N^{\frac{1}{2}}\,(\hat{\underline{\beta}}_A-\underline{\beta}) = N^{-\frac{1}{2}}\sum_{i=1}^{N} \{Z_i\ I\{\xi_1 \le Z_i \le \xi_2\} + a\ g(Z_i)\}.$$

The proofs are similar to those of theorems 1 and 2 and are omitted.

Since $\hat{\underline{\beta}}_A$ is particularly easy to compute in the location model, it is very suitable for Monte Carlo studies. It is hoped that such studies will indicate the degree of agreement between the asymptotic and finite sample variances of $\hat{\underline{\beta}}_L$ as well as $\hat{\underline{\beta}}_A$. Table 2 displays the variance of $\hat{\underline{\beta}}_A$(LS), i.e. $\hat{\underline{\beta}}_A$ with $\hat{\underline{\beta}}_0$ the LS estimate, for sample sizes of N = 50,100,200,300, and 400. The Monte-Carlo swindle (Gross (1973)) was employed as a variance reduction technique. One sees from this table that convergence of the variance to its asymptotic value can be extremely slow for some distributions, e.g. b = 10 and ε = .10 or .25.

7. Conclusions. Despite their intuitive appeal, trimmed regression estimates based on an arbitrary preliminary estimate will not be very satisfactory. However, provided the error distribution is symmetric, there is one such estimate that is closely analogous to the trimmed mean in the location model.

Appendix

Proposition A.1. For $\theta = \alpha$ or $(1-2\alpha)$ let μ_N be a sequence of solutions to

$$\sum_{i=1}^{N} (r_i-\mu_N)\ \psi_\theta(r_i-\mu_N) = \min.$$

Then,

$$N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi_\theta(r_i-\mu_N) = o_p(1).$$

Proof. The argument is very similar to that of theorem 1 of Ruppert and Carroll (1978) and will be omitted.

Proof of lemma 1. As pointed out by Koenker and Bassett (1978), $\mu = r_{\theta N}$ is a solution to

$$\sum_{i=1}^{N} (r_i-\mu)\ \psi_\theta(r_i-\mu) = \min\ ,$$

so that by Proposition A.1, for $\theta = \alpha$ or $(1-\alpha)$,

(A1) $$N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi_\theta(Z_i - \xi_\theta - \underset{\sim}{x}_i(N^{\frac{1}{2}}\ (\underline{\hat{\beta}}_0-\underline{\beta}) + \underset{\sim}{e}(r_{\theta N}-\xi_\theta)) = o_p(1)\ .$$

Here, we use the fact that $\underset{\sim}{x}_i'\underset{\sim}{e} = 1$. Define the processes

$$V_N(\Delta) = N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi_\theta(Z_i - \xi_\theta - \underline{x}_i'\ \Delta/N^{\frac{1}{2}})$$

and

$$W_N(\Delta) = V_N(\Delta) - V_N(0) - E(V_N(\Delta) - V_N(0)).$$

Following Bickel (1975) or as a special case of Lemma A2 of Ruppert and Carroll (1978), for all $M > 0$,

(A2) $$\sup_{0\le||\Delta||\le M} |W_N(\Delta)| = o_p(1)\ ,$$

and

(A3) $$\sup_{0\le||\Delta||\le M} |V_N(\Delta) - V_N(0) + f(\xi_\alpha)\; e'\Delta| = o_p(1) \;.$$

Further, following the method of Jurecková (1977) or Lemma A.3 of Ruppert and Carroll (1978), for all $\varepsilon > 0$ there exists η, K, and N_0 such that

(A4) $$P\{ \inf_{||\Delta||>K} |V_N(\Delta)| < \eta\} < \varepsilon \} \text{ for } N \geq N_0 \;.$$

By (A1) and (A4) we have that

(A5) $$N^{\frac{1}{2}} \{(\hat{\underline{\beta}}_0 - \underline{\beta}) + e(r_{\theta N} - \xi_\theta)\} = O_p(1) \;,$$

so by substituting the RHS of (A5) for Δ in (A3) we obtain by (A1) that

(A6) $$N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi_\theta(Z_i - \xi_\theta) - f(\xi_\theta)\; \underset{\sim}{e}'\; N^{\frac{1}{2}}\{(\hat{\underline{\beta}}_0 - \underline{\beta}) + \underline{e}(r_{\theta N} - \xi_\theta)\} = o_p(1). \qquad \square$$

Proposition A.2. (Lemma A.4 of Ruppert and Carroll (1978)). Let $D_{iN}(= D_i)$ be a r×c matrix whose (ℓ,k)th component is denoted by $D_{i\ell k}$. Suppose

$$\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} D_{i\ell k}^2 \text{ exists for all } \ell \text{ and } k.$$

Let h(x) be a function defined for all real x that is Lipschitz continuous on an open interval containing ξ_1 and ξ_2. For Δ_1, Δ_2, and Δ_3 in R^p and $\Delta = (\Delta_1, \Delta_2, \Delta_3)$, define

$$T(\Delta) = N^{-\frac{1}{2}} \sum_{i=1}^{N} D_i\; h(Z_i + \underline{X}_i'\; \Delta_3/N^{\frac{1}{2}})\; I\{\xi_1 + \underline{X}_i'\; \Delta_1/N^{\frac{1}{2}} \leq Z_i \leq \xi_2 + \underline{x}_i'\; \Delta_2/N^{\frac{1}{2}}\} \;.$$

Define

$$S(\Delta) = T(\Delta) - T(0) - E(T(\Delta) - T(0)) \;.$$

Then, for all $M > 0$,

$$\sup_{0\le||\Delta||\le M} ||S(\Delta)|| = o_p(1) \;.$$

Proof of theorem 1. For Δ_1, Δ_2 in R^p and $\underset{\sim}{\Delta} = (\underset{\sim}{\Delta}_1, \underset{\sim}{\Delta}_2)$, define

$$U(\underset{\sim}{\Delta}) = N^{-1} \sum_{i=1}^{N} \underline{x}_i \, \underline{x}_i' \, I(\xi_1 + \underline{x}_i' \, \Delta_1/N^{\frac{1}{2}} \leq Z_i \leq \xi_2 + \underline{x}_i' \, \Delta_2/N^{\frac{1}{2}}\}$$

and

$$W(\underset{\sim}{\Delta}) = N^{-\frac{1}{2}} \sum_{i=1}^{N} \underline{x}_i \, z_i \, I\{\xi_1 + \underline{x}_i' \, \Delta_1/N^{\frac{1}{2}} \leq Z_i \leq \xi_2 + \underline{x}_i' \, \Delta_2/N^{\frac{1}{2}}\} .$$

Using Proposition A.2, it is easy to show (cf. Ruppert and Carroll (1978), proof of theorem 3) that for all $M > 0$,

(A7) $$\sup_{0 \leq ||\Delta|| \leq M} |U(\underset{\sim}{\Delta}) - (1-2\alpha)Q| = o_p(1)$$

and

(A8) $$\sup_{0 \leq ||\Delta|| \leq M} |W(\underset{\sim}{\Delta}) - W(\underset{\sim}{0}) - Q(\Delta_2 \, \xi_2 \, f(\xi_2) - \Delta_1 \, \xi_1 \, f(\xi_1))| = o_p(1) .$$

Then using the fact that $\underline{x}_i' \, \underline{e} = 1$, we have

$$I\{r_{1N} \leq r_i \leq r_{2N}\} = I\{\xi_1 + \underline{x}_i'((\underline{\beta}_0 - \underline{\beta}) + \underline{e}(r_{1N} - \xi_1)) \leq Z_i$$

$$\leq \xi_2 + ((\hat{\underline{\beta}}_0 - \underline{\beta}) + \underline{e}(r_{2N} - \xi_2)\}$$

and so replacing Δ_ℓ by $N^{\frac{1}{2}}((\hat{\underline{\beta}}_0 - \underline{\beta}) + \underline{e}(r_{\ell N} - \xi_\ell)$, for $\ell = 1, 2$, in (A7) and A8), we have

(A9) $$N^{-1} \, (X'AX) = (1-2\alpha)Q + o_p(1)$$

and

(A10) $$N^{-\frac{1}{2}} \, X'A(\underline{y} - AX \, \underline{\beta}) = W(\underline{0}) + Q\{\xi_2 \, f(\xi_2) \, N^{\frac{1}{2}} \, (\hat{\underline{\beta}}_0 - \underline{\beta} + \underline{e}(r_{2N} - \xi_2))$$

$$- \xi_1 \, f(\xi_1) \, N^{\frac{1}{2}} \, (\hat{\underline{\beta}}_0 - \underline{\beta} + \underline{e}(r_{1N} - \xi_1))\} + o_p(1) .$$

By (A9) and (A10),

(A11) $$N^{\frac{1}{2}}\,(X'A(y-AX\,\underline{\beta})) = (1-2\alpha)\,N^{\frac{1}{2}}Q(\hat{\underline{\beta}}_{LS}-\underline{\beta}) + o_p(1).$$

By (A10), (A11), and (3.2)

(A12) $$(1-2\alpha)\,N^{\frac{1}{2}}\,Q(\hat{\underline{\beta}}_{LS}-\underline{\beta}) = W(\underline{0})$$
$$+\,Q\{\xi_2\,\underline{e}\,N^{-\frac{1}{2}}\sum_{i=1}^{N}\psi_{1-\alpha}\,(Z_i-\xi_2) - \xi_1\,\underline{e}\,N^{-\frac{1}{2}}\sum_{i=1}^{N}\psi_{\alpha}(Z_i-\xi_1)$$
$$+\,N^{\frac{1}{2}}\,a(I-\underline{e}\,\underline{e}')(\hat{\underline{\beta}}_0-\underline{\beta})\} + o_p(1)\ .$$

Then (3.3) follows from (A12), (3.1), and the definition of $W(\underline{0})$.

<u>Proof of corollary 2</u>. By (2.2), the first row of Q is $\underline{e}'$. Therefore, the first row of Q^{-1} is also $\underline{e}$. Consequently, $(I-\underline{e}\,\underline{e}')\,Q^{-1}\,\underline{x}_i = Q^{-1}\,\underline{c}_i$. Thus, substituting (3.4) into (3.3) completes the proof.

Table 1 - Variances of the asymptotic distribution of slope estimators-(The asymptotic covariance matrix is Q_I^{-1} multiplied by the displayed quantity).

ϵ*	b**	Least Squares	Huber Proposal 2	Hampel One Step	Trimmed Least Squares: $\hat{\underline{\beta}}_L$(LS) (Least Squares as Preliminary Estimate)			Trimmed Least Squares: $\hat{\underline{\beta}}_L$(LAD) (Least Absolute Deviation as Preliminary Estimate)			Trimmed Least Squares: $\hat{\underline{\beta}}_L$(RQ) (Average of αth and 1-αth Regression Quantiles as Preliminary Estimate)		
					α=.05	α=.10	α=.25	α=.05	α=.10	α=.25	α=.05	α=.10	α=.25
NORMAL		1.00	1.04	1.04	1.30	1.36	1.26	1.54	1.83	2.14	1.03	1.06	1.19
0.05	3.0	1.40	1.16	1.17	1.38	1.51	1.58	1.54	1.88	2.26	1.16	1.17	1.29
0.05	5.0	2.20	1.20	1.23	1.43	1.71	2.15	1.51	1.87	2.28	1.20	1.20	1.31
0.05	10.0	5.95	1.23	1.28	1.68	2.66	4.81	1.46	1.85	2.30	1.25	1.23	1.33
0.10	3.0	1.80	1.30	1.32	1.44	1.64	1.88	1.56	1.93	2.39	1.32	1.30	1.39
0.10	5.0	3.40	1.40	1.47	1.45	1.96	2.99	1.46	1.90	2.44	1.46	1.38	1.45
0.10	10.0	10.90	1.49	1.61	1.48	3.32	8.09	1.34	1.85	2.47	1.65	1.45	1.49
0.25	3.0	3.00	1.90	1.94	1.79	1.97	2.74	1.82	2.12	2.87	2.14	1.85	1.80
0.25	5.0	7.00	2.46	2.68	2.49	2.09	5.13	2.37	1.92	2.99	4.11	2.39	2.01
0.25	10.0	25.75	3.20	4.26	6.50	1.88	15.66	5.51	1.65	3.06	13.65	3.69	2.19

*Proportion of contamination

**Standard deviation of contamination

Table 2 - Finite and Asymptotic Variances of $N^{\frac{1}{2}}\ \hat{\underline{\beta}}_A(LS)$ in the Location Model

ϵ^+	b^{++}	N*=50 NI**=1000	N=100 NI=1000	N=200 NI=500	N=300 NI=500	N=400 NI=850	Asymptotic
NORMAL		1.31	1.36	1.37	1.32	1.35	1.36
.05	3	1.47	1.49	1.50	1.47	1.48	1.51
.05	5	1.57	1.65	1.70	1.66	1.65	1.71
.05	10	2.10	2.36	2.54	2.51	2.40	2.66
.10	3	1.58	1.58	1.65	1.63	1.60	1.64
.10	5	1.74	1.83	1.97	1.90	1.90	1.96
.10	10	2.24	2.51	2.92	2.99	3.03	3.32
.25	3	2.01	1.93	1.94	1.96	1.96	1.97
.25	5	2.12	2.05	2.08	2.11	2.07	2.09
.25	10	2.98	2.42	2.14	2.13	2.11	1.88

+Proportion of contamination

++Standard deviation of contamination

*Sample size

**Number of Monte-Carlo simulations

References

Anscombe, F.J. (1961). Examination of Residuals, *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, (J. Neyman, ed.), pp. 1-36, University of California Press, Berkeley and Los Angeles, California.

Bahadur, R.R. (1966). A note on quantiles in large samples, *Ann. Math. Statist. 37*, 577-580.

Bickel, Peter J. (1973). On Some Analogues to Linear Combinations of Order Statistics in the Linear Model, *Ann. Statist. 1*, 597-616.

Bickel, Peter J. (1975). One step Huber estimates in the linear model, *J. Amer. Statist. Assoc. 70*, 428-433.

Carroll, Raymond J. and Ruppert, David (1979). Almost Sure Properties of Robust Regression Estimates, Unpublished manuscript.

Ghosh, J.K. (1971). A new proof of the Bahadur representation of quantiles and an application, *Ann. Math. Statist. 42*, 1957-1961.

Gross, Alan M. (1976). Confidence Interval Robustness with Long-Tailed Symmetric Distributions, *J. Amer. Statist. Assoc. 71*, 409-416.

Gross, Alan M. (1973). A Monte-Carlo swindle for estimators of location, *App. Statist.*, *22*, 347-356.

Hogg, Robert V. (1974). Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Application and Theory, *J. Amer. Statist. Assoc. 69*, 909-927.

Huber, Peter J. (1970). Studentizing Robust Estimates. In *Nonparametric Techniques in Statistical Inference*, (M.L. Puri, ed.), pp. 453-463. Cambridge University Press.

Huber, Peter. J. (1977). *Robust Statistical Procedures*, SIAM, Philadelphia, PA.

Jurecková, Jana (1977). Asymptotic Relations of M-estimates and R-estimates in Linear Regression Model, *Ann. Statist.*, *5*, 464-472.

Koenker, Roger and Bassett, Gilbert, Jr. (1978). Regression Quantiles, *Econometrica, 46*, 33-50.

Ruppert, David and Carroll, Raymond J. (1978). Robust Regression by Trimmed Least-Squares Estimation. Institute of Statistics Mimeo Series #1186. Department of Statistics, Chapel Hill, North Carolina.

Stigler, Steven M. (1977). Do Robust Estimators Work with Real Data? *Ann. Statist. 5*, 1055-1098.