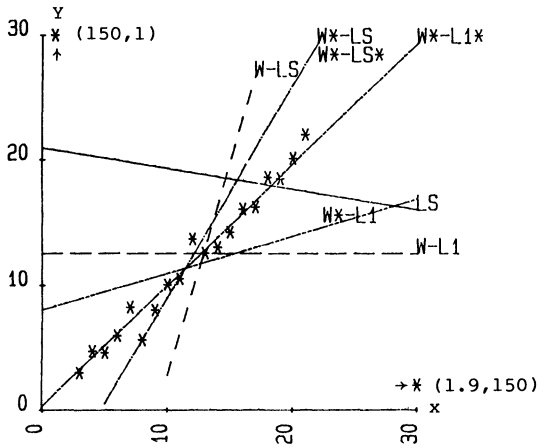

FIG. 3. *One y and one x outlier.*

Although our results are based on a limited number of data sets, we conclude that the Welsh trimmed mean should be used with caution in practice.

REFERENCES

BICKEL, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.* **1** 597–616.

DE JONGH, P. J. and DE WET, T. (1985). Trimmed means and bounded influence estimators for the parameters of the AR(1) process. *Comm. Statist. A—Theory Methods* **14** 1361–1375.

DENBY, L. and LARSEN, W. A. (1977). Robust regression estimators compared via Monte Carlo. *Comm. Statist. Theory Methods* **6** 335–362.

KOENKER, R. W. and BASSETT, G. W. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KRASKER, W. S. and WELSCH, R. E. (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604.

RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–838.

INSTITUTE FOR MARITIME TECHNOLOGY
P.O. BOX 181
SIMON'S TOWN 7995
REPUBLIC OF SOUTH AFRICA

ROGER KOENKER

University of Illinois, Urbana

1. Introduction. Alan Welsh has resolved an intriguing puzzle posed by Ruppert and Carroll (1980) in their influential study of analogues of the trimmed mean for the linear regression model. They showed that an estimator with "appropriate" asymptotic behavior could be constructed based on "regression quantiles," and they also showed that naive trimming based on residuals from a preliminary fit of the model had a considerably different, and far less satisfactory, asymptotic theory. Welsh has now shown that a less naive, but still

remarkably simple, form of "Winsorizing" preliminary residuals *can* succeed in achieving the desired asymptotic performance.

In these brief remarks I would like to offer an elementary interpretation of why Welsh's method works, mention some circumstances in which it doesn't perform quite so well, and finally, to sketch an idea for a more general theory of L estimators for the linear model.

2. Interpretation. Why does Welsh's method, which involves a "Winsorizing" transformation of the response vector, succeed while naive trimming fails? Ultimately, an answer lies in the intimate connection between the trimmed mean and the Huber M estimator and between Welsh's one-step and Bickel's (1975) Huber one-steps.

To emphasize these connections, we may write Welsh's estimator, τ_n, as

$$\tau_n = \theta_n + \Big(\sum x_j x_j' K_j\Big)^{-} \sum x_j \tilde{e}_j, \tag{1}$$

with $\tilde{e}_j$ denoting the "Winsorized" preliminary residuals,

$$\tilde{e}_j = \xi_{n\alpha}\big(J_j - \alpha\big) + e_j K_j + \xi_{n\beta}\big(L_j - (1-\beta)\big), \tag{2}$$

and the other notation as in Welsh, Section 1. By contrast, a (type-1) Bickel–Huber one-step is

$$\mu_n = \theta_n + \Big(\sum x_j x_j' K_j^*\Big)^{-} \sum x_j \tilde{e}_j^*, \tag{3}$$

where $\pm Ms_n(\theta_n)$ replaces $\xi_{n\alpha}$ and $\xi_{n\beta}$ in (2) and in the definitions of the revised indicators J_j^*, K_j^*, and L_j^*. If we adopt symmetric trimming and choose a symmetric (nominal) model for F, then Huber's constant (M in the present notation) can always be chosen to make these alternative quantile estimates asymptotically equivalent, reducing the difference between τ_n and μ_n to the quantity

$$\Delta_n = \Big(\sum x_j x_j' K_j\Big)^{-1} \sum x_i \big[\xi_{n\alpha}\alpha + \xi_{n\beta}(1-\beta)\big],$$

which can be seen to be, asymptotically, an adjustment solely to the intercept of the model, i.e., since $(1, \ldots, 1)'$ is in the column space of X,

$$\Delta_n = (\beta - \alpha)^{-1}\big[\xi_{n\alpha}\alpha + \xi_{n\beta}(1-\beta)\big] e_1 + o_p(1),$$

where $e_1 = (1, 0, \ldots, 0)$. Of course, $E\Delta_n \to 0$ when F is symmetric and trimming is symmetric, so Δ_n serves the role of restoring the asymptotic variance of the intercept component of τ_n to its proper form. One may recall that in Ruppert and Carroll's theory of the naively trimmed preliminary residuals estimator, different asymptotic behavior was exhibited by the intercept and slope parameters. Welsh's results clarify this and show that, while trimming is appropriate for the intercept parameter, "Winsorization" is required for the slope parameters. This is consistent with the results of Ruppert and Carroll (1980), who find that the intercept parameter after naive trimming has the desired asymptotic behavior, but the slope parameters do not. Jurečková and Sen (1984) study a similar scale-equivariant Huber M estimator that employs the inner-αth quantile range based on Ruppert and Carroll's TLS residuals to compute Huber's constant;

they show that the resulting M estimator is asymptotically equivalent to the TLS estimator to order $O_p(n^{-3/4})$ employing Bahadur representations of the regression quantile statistics.

3. Some experimental evidence. One theme, among many, of the Princeton robustness study (Andrews et al. (1972)) was that one-steps, and iterative estimators generally, inherited the virtues and defects of their initial estimates. This is especially true in regression, and becomes, of course, glaringly obvious when the start fails to satisfy the fundamental asymptotic requirement of $\sqrt{n}$ consistency. In Table 1, we report some Monte Carlo experience with two variants of Welsh's estimators, one starting from ordinary least squares (1SWL2) and the other starting from the l_1 estimate (1SWL1). We include, for reference, performance of the ordinary least-squares estimator (OLS) as well as the trimmed least-squares estimator (TLS) proposed by Koenker and Bassett (1978) and studied intensively by Ruppert and Carroll (1980). We have also included the performance of a fifth analogue of the trimmed mean (TRQ), which we will describe in detail shortly.

The entries in Table 1 are based on 1000 Monte Carlo replications. Each configuration had 50 observations, with $p = 3, 5, 10$ parameters as indicated by the column of the table. The design matrices, X, were generated as a column of ones with remaining entries drawn as iid realizations from the distribution specified in column one. The design matrix is *fixed* for each configuration. The estimators investigated are the ordinary least-squares estimator (OLS) and four analogues of the trimmed mean: the trimmed regression quantile estimator (TRQ), the trimmed least-squares estimator (TLS), the one-step Welsh estimator starting from the least absolute error estimate (1SWL1), and the one-step Welsh estimator starting from the least-squares estimate (1SWL2). All four of the latter had trimming proportion 0.1 and have the same asymptotic behavior to order $1/\sqrt{n}$. The entries are sample averages of the Monte Carlo efficiencies over the p coefficients of the model. All efficiencies were computed relative to the optimal weighted least-squares estimator using the discrepancy between each estimator and the optimal estimator as a variance reduction technique. A portable implementation (Fox (1976)) of the well-known Marsaglia random number generator was used to generate random uniform numbers, and the algorithms for normal and "student" variates were taken from the Princeton robustness study (Andrews et al. (1972)).

We conclude from the table that the other estimators are clearly superior to TLS, especially in strictly Gaussian cases. Although the 1SWL2 does quite well in cases of mild kurtosis, it has unacceptable performance in long-tailed situations like the Cauchy where the start is poor. TLS performance deteriorates rapidly as p becomes large relative to n. It is interesting to note that this tendency is accentuated by a high degree of kurtosis in the design. The Welsh estimators starting from l_1 perform extremely well; like TRQ they show very little tendency to break down when p is moderately large. Further work on the theory of these estimators when p is large relative to n would be highly desirable.

TABLE 1
Monte Carlo efficiencies of least squares and four analogues of the trimmed mean. The double dagger indicates an entry greater than 10^3.

Sample Configuration	Estimator	Relative Efficiencies		
		$p = 3$	$p = 5$	$p = 10$
$Y \sim N$	OLS	1.000	1.000	1.000
$X \sim N$	TRQ	1.071	1.097	1.128
	TLS	1.166	1.250	1.604
	1SWL1	1.065	1.070	1.090
	1SWL2	1.058	1.065	1.074
$Y \sim N$	OLS	1.000	1.000	1.000
$X \sim t(3)$	TRQ	1.083	1.113	1.134
	TLS	1.205	1.379	1.757
	1SWL1	1.066	1.079	1.100
	1SWL2	1.062	1.063	1.071
$Y \sim N$	OLS	1.000	1.000	1.000
$X \sim t(1)$	TRQ	1.115	1.134	1.135
	TLS	1.620	3.173	7.364
	1SWL1	1.039	1.083	1.159
	1SWL2	1.074	1.115	2.839
$Y \sim t(5)$	OLS	1.521	1.587	1.594
$X \sim N$	TRQ	1.368	1.397	1.415
	TLS	1.450	1.552	1.916
	1SWL1	1.354	1.351	1.386
	1SWL2	1.336	1.336	1.355
$Y \sim t(3)$	OLS	2.532	2.613	2.758
$X \sim N$	TRQ	1.675	1.677	1.698
	TLS	1.728	1.813	2.308
	1SWL1	1.675	1.654	1.690
	1SLW2	1.619	1.654	1.741
$Y \sim t(1)$	OLS	‡	‡	‡
$X \sim N$	TRQ	5.531	5.956	6.093
	TLS	5.226	5.329	8.958
	1SWL1	5.774	6.155	6.384
	1SWL2	‡	‡	‡

4. Yet another analogue of the trimmed mean for the linear model. To conclude, I would like to briefly describe a general approach to L estimators for the linear model based on regression quantiles. This approach may be viewed as an alternative to the general theory of Bickel (1973) based on preliminary estimation. It yields as a leading example the mysterious TRQ estimator of Table 1. Koenker and Bassett (1978) suggested that p-dimensional analogues of the order statistics could be constructed for the linear model by solving

$$\text{(P)} \qquad \min_{b \in \mathbf{R}^p} \sum_{i=1}^{n} \rho_\theta(Y_i - x_i b),$$

where $\rho_\theta(\ )$ denotes, for $\theta \in [0,1]$, the "check" function,

$$\rho_\theta(u) = \begin{cases} \theta u, & \text{if } u \geq 0, \\ (\theta - 1)u, & \text{if } u < 0. \end{cases}$$

An asymptotic theory of finite linear combinations of such "regression quantiles" was developed, and led to "simple" analogues of the "quick-and-dirty (inefficient) statistics" of Mosteller (1946) and others. Ruppert and Carroll (1980) subsequently showed that certain ad hoc methods based on regression quantiles, notably the trimmed least-squares estimators, could achieve asymptotic performance like that of the trimmed mean.

Recently, several developments have coalesced to produce a rudimentary general theory of L estimators for the linear model. The essential element of this theory is the extension of the theory for finite linear combinations to a theory for smooth weight functions. Let $\hat{\beta}$: $[0,1] \to \mathbf{R}^p$ denote the random function that solves problem P. Under mild conditions on F, $\hat{\beta}(\theta)$ takes a unique value, except at a finite number of points in $[0,1]$. This exceptional set may be ignored since we are interested in estimators of the form

$$\tilde{\beta}_i[J] = \int_0^1 J(\theta)\hat{\beta}_i(\theta)\, d\theta, \qquad i = 1, \ldots, p,$$

for smooth choices of the weight function J.

An algorithm for efficiently computing $\hat{\beta}$ is described in Koenker and D'Orey (1985); it is an application of standard methods of parametric linear programming. As θ traverses $[0,1]$, $\hat{\beta}(\theta)$ takes order n distinct values, each corresponding to (in linear programming parlance) a basic solution; here, a parameter estimate based on an exact fit to p distinct sample observations. Thus problem P serves to identify a small number of "interesting" basic solutions, roughly $O(n)$ in our empirical experience, out of the large $\binom{n}{p}$ number of possible basic solutions. Recent work by Wu (1986) and others has emphasized the fundamental role that these basic solutions play in the theory of least-squares estimation and diagnostics.

The asymptotic theory of $\tilde{\beta}[J]$ requires an invariance principle for the p-dimensional quantile process $\hat{\beta}(\theta)$. The finite-dimensional asymptotic distributions of this process are established in Koenker and Bassett (1978) and by rather different, simpler methods in Ruppert and Carroll (1980). Portnoy (1986) has recently established the tightness of the process on an interval $[\varepsilon, 1 - \varepsilon]$ for $0 < \varepsilon < \frac{1}{2}$, extending some of the results of Jurečkovà and Sen (1985). Thus, a reasonably broad theory of L estimators of the form $\tilde{\beta}[J]$ can be addressed by requiring $J(\theta)$ to be smooth on $[\varepsilon, 1 - \varepsilon]$ and vanish outside of it. The simplest interesting case is that of the trimmed mean of the regression quantile process,

$$\tilde{\beta}_i[J_\alpha] = \frac{1}{1 - 2\alpha}\int_\alpha^{1-\alpha} \hat{\beta}_i(\theta)\, d\theta, \qquad i = 1, \ldots, p.$$

This is the mystery estimator TRQ of Table 1 and one can see from the table that it performs quite well over the limited design-noise configurations treated by the experiment. In particular, it is much less sensitive to influential design

points than is the trimmed least-squares estimator and it is inherently insensitive to a preliminary estimator, which is a potentially serious problem with Welsh's estimator. Even when p, the number of parameters being estimated, is large relative to n, TRQ adheres fairly closely to the behavior predicted by its asymptotic theory. Like Welsh's estimator and trimmed least squares, it is scale- and reparameterization-of-design equivariant and therefore offers most of the attractions of the Huber M estimator without the difficulties created by the necessity of joint estimation of a scale parameter. This is also an advantage with respect to the estimators proposed by Bickel (1973).

As Welsh notes, L estimation plays an extremely useful role in the analysis of the one-sample problem; I believe that it could play a similarly constructive role in analyzing linear models. I hope others, like Welsh, will help to build a theory that would justify this belief.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location*. Princeton Univ. Press, Princeton, N.J.

BICKEL, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.* **1** 597–616.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434.

FOX, P. (1976). The PORT mathematical subroutine library. Bell Laboratories, Murray Hill, N.J.

JUREČKOVÁ, J. and SEN, P. K. (1984). Adaptive scale-equivariant m-estimators in linear models. *Statist. Decisions, Suppl.* **1** 31–46.

KOENKER, R. W. and BASSETT, G. W. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KOENKER, R. W. and D'OREY, V. (1985). Computing regression quantiles. To appear in *Appl. Statist.*

MOSTELLER, F. (1946). On some useful "inefficient" statistics. *Ann. Math. Statist.* **17** 377–408.

PORTNOY, S. (1986). Personal communication.

RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–838.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1295.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF ILLINOIS
CHAMPAIGN, ILLINOIS 61820

REJOINDER

A. H. WELSH

University of Chicago

The discussants have provided valuable insights into the nature of the one-step trimmed mean in the regression problem and made original proposals of their own. Their empirical results are both helpful and encouraging.

The choice of initial estimator for one-step estimators is important as both discussants note. In addition to the technical requirement that the initial