ECONOMETRIC REVIEWS
Vol. 22, No. 2, pp. 115–134, 2003

A Comparison of Partially Adaptive and Reweighted Least Squares Estimation

Brian H. Boyer,[1,*] James B. McDonald,[2] and Whitney K. Newey[3]

[1]University of Michigan, Ann Arbor, Michigan, USA
[2]Brigham Young University, Provo, Utah, USA
[3]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

ABSTRACT

The small sample performance of least median of squares, reweighted least squares, least squares, least absolute deviations, and three partially adaptive estimators are compared using Monte Carlo simulations. Two data problems are addressed in the paper: (1) data generated from non-normal error distributions and (2) contaminated data. Breakdown plots are used to investigate the sensitivity of partially adaptive estimators to data contamination relative to RLS. One partially adaptive estimator performs especially well when the errors are skewed, while another partially adaptive estimator and RLS perform particularly well when the errors are extremely leptokurtotic. In comparison with RLS, partially adaptive estimators are only moderately effective in resisting data contamination; however, they outperform least squares and least absolute deviation estimators.

Key Words: Least median of squares; Reweighted least squares; Partially adaptive estimation.

JEL Classification: C13; C20.

---

*Correspondence: Brian H. Boyer, University of Michigan Business School, 701 Tappan Street, Ann Arbor, MI 48109, USA; E-mail: boyerb@umich.edu.

DOI: 10.1081/ETC-120020459 0747-4938 (Print); 1532-4168 (Online)

1. INTRODUCTION

Regression analysis is a technique routinely used by researchers in many disciplines to fit mathematical models to observed data. The traditional estimation technique of least squares is efficient if the error terms are independent of the regressors and are identically and independently distributed (i.i.d.) as a normal. Otherwise, least squares need not be efficient. While the unobserved random disturbances in a regression model are often assumed to be normally distributed, real data are often replete with outliers which lie far from the pattern established by a majority of the data. Given that correct model assumptions have been made, outliers are generally the result of either mistakes in measuring and recording the data (data contamination) or data generating processes that are truly leptokurtotic or asymmetric. Hampel et al. (1986) review numerous studies that examine data for outliers caused by human or mechanical error and conclude that empirical data typically contains 1–10% gross inaccuracies. Ideally, such contaminated observations should be identified and corrected or discarded. However, outlier diagnostics often fail to properly identify such observations. On the other hand, many outliers are generated by genuinely thick tailed or asymmetric error distributions. For instance, the leptokurtotic nature of financial data is well documented (see Bollerslev et al., 1994 for a list of references). In this case, discarding outliers is inappropriate since they are representative of the true data generating process.

Rousseeuw's (1984) least median of squares (LMS) is a robust estimator that effectively identifies outliers. Outlying observations lie far from an LMS fit to the data, and hence, exhibit large standardized residuals. A weighted version of LMS, known as reweighted least squares (RLS), amounts to discarding all outliers and performing least squares on the remaining data. Alternatively, partially adaptive estimators (McDonald and Newey, 1988) are based on an underlying error distribution consistent with the entire sample. A general, nesting error distribution is assumed to approximate the underlying distribution and is used to obtain the regression estimates. Thus, RLS treats outliers as if they were the result of data contamination while partially adaptive estimators assume that outliers are the result of non-normal data generating processes.

In this paper, we compare the sensitivity of LMS, RLS, and partially adaptive estimators to both data contamination and outliers generated by non-normal error distributions. While there are several potential estimators we could have examined, we limited our attention to these estimators since these two classes of estimators are known to exhibit very different robust properties. Comparisons with other estimators are left for further research. First, the estimators are formally presented, and their robust properties discussed. We then use Monte Carlo methods to compare the small sample performance of LMS, RLS, and partially adaptive estimators to: (a) data contamination and (b) non-normal error distributions. Breakdown plots (Rousseeuw and Leroy, 1987) are employed in the analysis of the influence of data contamination. Two other estimators, least squares (OLS), and least absolute deviations (LAD), are also investigated as comparative benchmarks.

2. THE MODEL AND ESTIMATORS

The model we consider is a regression model of the form

$$y_i = \alpha_0 + x_i\beta_0 + u_i \quad (i = 1, \ldots, n) \tag{1}$$

where y is a dependent variable, x is a $1 \times k$ vector of regressors, u is an error term which is independent of the regressors, and β_0 denotes a $k \times 1$ vector of unknown coefficients. Assuming that x has a nonsingular variance matrix, β will be identified if u and x are independent. In that case, $x\beta$ is an additive shifter for the conditional distribution of y given x. An additional assumption is needed to identify the constant α. For example, if $E(u) = 0$ then $\alpha + x\beta$ is the conditional mean of y given x or if $\text{med}(u) = 0$ then $\alpha + x\beta$ is the conditional median of y given x. Generally, the interpretation of $\alpha + x\beta$ as a conditional location measure for y depends on the location restriction for u. Both α and β are also identified if u is symmetrically distributed around 0 conditional on x. In that case $\alpha + x\beta$ is both the conditional mean and the conditional median, as well as being equal to other conditional location measures.

Both α and β are also identified under a conditional location restriction, which is a weaker restriction than independence or conditional symmetry. For example, if u has conditional mean zero then $\alpha + x\beta$ is the conditional mean of y given x, while if the conditional median of u is zero then $\alpha + x\beta$ is the conditional median of y given x. We do not consider these weaker types of restrictions here. The reason for this is that associated with each conditional location restriction is a conditional moment restriction $E[\rho(u)|x] = 0$ for some $\rho(u)$ on which an estimator can be based. For example, $\text{med}(u|x) = 0$ corresponds to $E[\text{sgn}(u)|x] = 0$, which leads to LAD, while $E[u\,|\,x] = 0$ corresponds to OLS. Hence, we are interested in cases where u satisfies many conditional location restrictions, so that many alternative estimators are available.

This paper also considers the effects of outliers that do not satisfy this model, in the discussion of robustness and in the Monte Carlo results. The Monte Carlo simulations consider generalized autoregressive conditional heteroskedastic (GARCH) errors in addition to the i.i.d. data case for purposes of comparison.

2.1. Partially Adaptive Estimators

One type of estimators we consider are quasi (or pseudo) maximum likelihood estimators, each of which solves the following optimization problem:

$$(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \arg\max_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \ln f(y_i - \alpha - x_i\beta, \gamma) \tag{2}$$

where $f(u,\gamma)$ is a probability distribution function (pdf) with mean zero indexed by γ, a vector of distributional parameters. The least squares and LAD estimators are examples of this, as they can be obtained by taking the normal and Laplace distribution for the error distribution, respectively.

The normal and Laplace specifications do not allow for asymmetric error distributions or flexibility in the degree of leptokurtosis. The partially adaptive estimators circumvent this limitation by employing a more flexible distribution family for the error distribution. These estimators are constructed by maximizing (2) over both the regression and distributional parameters. We consider partially adaptive estimators associated with the generalized t distributions (GT) introduced by McDonald and Newey (1988), the power exponential distributions by Box–Tiao (1962) (BT), and exponential generalized beta distributions of the second kind (EGB2) proposed by McDonald and Xu (1995).

The generalized t distribution GT(σ, p, q) whose pdf, $f_{GT}(\cdot, \sigma, p, q)$, is defined by

$$f_{GT}(u, \sigma, p, q) = \frac{p}{2\sigma q^{1/p} B\left(\frac{1}{p}, q\right)\left[1 + \frac{|u|^p}{q\sigma^p}\right]^{q+1/p}} \quad \text{for } -\infty < u < \infty \tag{3}$$

where $B(\cdot,\cdot)$ denotes the beta function. The GT pdf is symmetric about zero. The parameter σ is a scale factor and governs the variance, while p and q determine the kurtosis. Larger values of p and q correspond to thinner tails, and smaller values correspond to thicker tails. The GT distribution includes five other distributions as special or limiting cases: the Student t, Cauchy, BT, normal, and Laplace distributions. The Student t and Cauchy distributions correspond to (3) with $p = 2$ and $(p, q) = (2, 1/2)$, respectively. Further, the pdf of GT(σ, p, q) converges in distribution to that of the BT(σ, p) as $q \to \infty$ with pdf defined by

$$f_{BT}(u, \sigma, p) = \frac{pe^{-\left(\frac{|u|}{\sigma}\right)^p}}{2\sigma\Gamma\left(\frac{1}{p}\right)}. \tag{4}$$

The normal and Laplace represent special cases of the BT corresponding to $p = 2$ and $p = 1$, respectively. Figure 1 summarizes these interrelationships.

To accommodate potential asymmetric probability distributions, McDonald and Xu (1995) propose the exponential generalized beta distributions of the second kind [EGB2(δ, σ, p, q)] with pdf's defined by

$$f_{EGB2}(u, \delta, \sigma, p, q) = \frac{e^{p(u-\delta)/\sigma}}{\sigma B(p, q)[1 + e^{(u-\delta)/\sigma}]^{p+q}} \quad \text{for } -\infty < u < \infty. \tag{5}$$

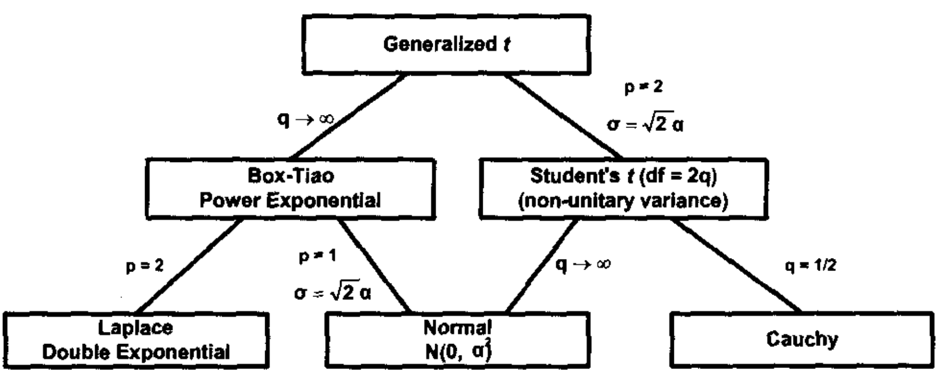

Figure 1. Distribution tree.

The second, third, and fourth moments about the mean of the EGB2 are flexible. Of the parameters in (5), δ governs the mean,[a] σ is a scale parameter, and p and q determine both the skewness and kurtosis. The distribution is asymmetric unless $p = q$. The pdf of f_{EGB2} converges in distribution to the pdf of the normal as $p = q$ grow indefinitely large.

The relationships between the GT, BT, and EGB2 make it convenient to perform statistical tests. For example, a test of normality can be performed by testing the hypothesis $p = 2$ in Eq. (4), using a likelihood ratio, Wald, or Lagrange multiplier test. Likewise, a test of symmetry can be performed by testing the hypothesis $p = q$ in (5).

McDonald and White (1993) investigate the small sample performance of partially adaptive estimators in comparison with that of LAD, adaptive maximum likelihood (AML), generalized method of moments (GMM), and two *M*-estimators suggested by Huber (1981). The authors' study closely parallels those of Hsieh and Manski (1987) and Newey (1988). The EGB2 partially adaptive estimator is found to be the most efficient for nearly all cases with non-normal errors. Further, the efficiency loss from estimating additional parameters via partially adaptive estimation with normal errors is found to be quite small, even though the sample size considered is as small as 50.

The performance of LAD is also investigated in this paper to provide some understanding of how a traditional robust regression procedure performs in comparison to the other techniques considered. Least absolute deviations estimation is maximum likelihood and yields unbiased and asymptotically efficient estimators when the actual data generating process is based on a Laplace error distribution. These and other statistical properties of LAD are summarized in Narula and Wellington (1982). Least absolute deviations has gained some respect in the literature because of its strong resistance to outliers in the y direction and its ability to efficiently estimate regression models with thick tails (for example, see Bloomfield and Steiger, 1983; Devroye and Gyorfi, 1984; and Koenker and Bassett, 1978). He et al. (1990) find LAD to perform very well according to a finite-sample measure of robustness based on tail behavior. However, LAD is sensitive to leverage points. As a result, just one outlying observation is sufficient to carry LAD across all bounds.

2.2. Least Median of Squares and Reweighted Least Squares

In an effort to make the estimators less sensitive to outliers, it has been suggested that the summation over the observations be replaced by the median in the construction of the estimators. For least squares this replacement leads to the LMS estimator, that minimizes the median of the squared residuals, Rousseeuw (1984). Formally, the LMS estimator is

$$(\tilde{\alpha}, \tilde{\beta}) = \arg\min{}_{\alpha,\beta} \text{med}_i (y_i - \alpha - x_i\beta)^2 \tag{6}$$

In essence, LMS determines the linear pattern established by the majority of the data. Rousseeuw (1984) and Rousseeuw and Leroy (1987) discuss computer algorithms that calculate the LMS estimate.

---

[a]The mean of an EGB2 random variable with pdf given by (5) is $\delta + \sigma[\psi(p) - \psi(q)]$ where $\psi(s) = d \ln \Gamma(s)/ds$; thus, δ is the mean if and only if $p = q$. The mean is zero if $\delta = \sigma[\psi(q) - \psi(p)]$, thus three parameters determine the third and fourth moments as well as the scale.

An alternative estimator that uses LMS for outlier detection but is based on least squares is the RLS estimator. This estimator is based on applying least squares to a subset of the data where the LMS residuals are not too large. Let

$$\tilde{u}_i = y_i - \tilde{\alpha} - x_i\tilde{\beta} \quad (i = 1, \ldots, n)$$

denote the LMS residuals. The RLS estimator is defined by

$$(\bar{\alpha}, \bar{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} 1\left(\left|\frac{\tilde{u}_i}{\hat{\gamma}}\right| \leq 2.5\right)(y_i - \alpha - x_i\beta)^2 \tag{7}$$

where $\hat{\gamma}$ is an estimator of scale based on the LMS residuals and $1(\cdot)$ denotes the indicator function. The threshold of 2.5 in Eq. (7) is arbitrary, but is probably a reasonable value because with a normal distribution few residuals will be deleted. This estimator was presented in Rousseeuw and Leroy (1987) and will be considered in our Monte Carlo work.

3. THEORETICAL PROPERTIES

Theory for the estimators is helpful for understanding the Monte Carlo results. Two kinds of properties are important. Asymptotic distribution theory helps to explain the efficiency results. In particular, the differing convergence rates for some of the estimators should lead to some of them being more efficient than others. Also, the robustness properties of some of the estimators help to explain the Monte Carlo results on breakdown points. In this section we give a brief account of these properties.

3.1. Asymptotic Distribution Theory

Each of the estimators is known to be consistent under certain conditions. When the disturbance is independent of the regressors, the sample variance matrix of the regressors is bounded away from singularity, and other regularity conditions are satisfied, then the partially adaptive estimators are consistent estimators of the true slope coefficients β_0, as shown in McDonald and Newey (1988) for the GT distribution, and can be shown for EGB2 by analogous results. Under independence the estimator of the constant α is not generally consistent, because GT and EGB2 impose implicit location restrictions on u that vary with the shape parameters. Thus, the limit of the estimator of the constant term varies as the shape parameters vary. Also for GT, consistency of slope and intercept estimators can be shown when the disturbance is distributed symmetrically around zero conditional on the regressors. The EGB2 estimator need not be consistent under conditional symmetry (without independence), because the log-likelihood need not be symmetric around zero as a function of u. Under independence or conditional symmetry the partially adaptive estimators of the constant α and the distributional parameters γ will be consistent estimators of

$$(\alpha_*, \gamma_*) = \arg\max_{\alpha,\gamma} E[\ln f(y_i - x_i\beta_0 - \alpha, \gamma)]. \tag{8}$$

The consistency of the LMS estimator holds under several conditions, including independence of the disturbance and regressors, the disturbance being symmetrically distributed around zero and unimodal, and a nonsingularity condition on the regressors. For the location parameter case (where x is not present) this result was shown in Rousseeuw and Leroy (1987). For the general regression case this result was shown by Kim and Pollard (1990). Consistency of RLS can also be shown under similar conditions.

The partially adaptive estimators will be root-n consistent. Due to the boundary restrictions on the distributional parameters (among which the intercept is essentially a location parameter) the asymptotic distribution of the intercept is quite complicated. For that reason we will focus on the slope coefficients. Let $s(u) = d\ln f(u, \gamma_*)/du$ and $s'(u) = ds(u)/du$. Then we have

$$\hat{\beta} = \beta_0 + \sum_{i=1}^{n} \frac{\Psi(z_i)}{n} + O_p(n^{1/2}), \tag{9}$$

where $z = (y, x)$ and $\Psi(z)$ takes the following form, with $u_i^* = u_i + \alpha_0 - \alpha_*$,

$$\Psi(z) = \{E[s'(u_i^*)]\}^{-1}[\text{var}(x_i)]^{-1}(x - E[x_i])s(y - \alpha_* - x\beta_0). \tag{10}$$

It follows that the partially adaptive estimators are root-n consistent and asymptotically normal with asymptotic variance equal to var[$\Psi(z)$].

The LMS will also have a limiting distribution but converges at a slower rate than the partially adaptive estimators. Specifically, as shown by Kim and Pollard (1990), when the disturbance is independent of the regressors, symmetrically distributed around zero, and additional regularity conditions are satisfied, then there is a random vector W such that

$$n^{1/3}(\tilde{\beta} - \beta_0) \xrightarrow{d} W. \tag{11}$$

The nature of W is quite complicated, so we omit a full description. The RLS estimator behaves very similarly to the LMS estimator. In Appendix A we give a result showing that for the pdf $f(u)$ and cumulative distribution function $F(u)$ of the disturbance, and for the limit γ_0 of $\tilde{\gamma}$,

$$\bar{\beta} = \beta_0 + \frac{5\gamma_0 f(2.5\gamma_0)}{2F(2.5\gamma_0) - 1}(\tilde{\beta} - \beta_0) + o_p(n^{-1/3}). \tag{12}$$

Thus, the RLS estimator will also converge at a cube root rate, and its asymptotic distribution is a rescaling of that of LMS. In many cases the scaling constant will be quite small, suggesting that RLS will be much more efficient than LMS. For example, in the case of a normal distribution where γ is the standard deviation, the scaling constant is 0.11, so that the asymptotic dispersion of the RLS estimator is only about 1/9 of the LMS estimator.

The slower rate of convergence for LMS and RLS means that for large enough sample sizes the distribution of these estimators will be more "spread out" than the distribution of the partially adaptive estimators. In this sense LMS and RLS are asymptotically less efficient, in small samples though their dispersion need not exceed that of the partially adaptive estimators, even for the asymptotic approximation. For example, if the dispersion of W is less than that of the normal distribution then for small enough samples the asymptotic approximation could give smaller dispersion for the LMS and/or RLS estimators. Of course the small sample properties of the estimators may not be well approximated by the asymptotic distribution.

3.2. Robustness

Robustness properties quantify how sensitive the estimators are to certain kinds of misspecification in the regression model of Eq. (1). In particular, robustness concerns sensitivity to a proportion of the data being equal to some arbitrary point (often thought of as an "outlier"). We consider two different properties, the influence function and the breakdown point of the estimator. The influence function of an estimator is the derivative of the limit of the estimator with respect to a change in a single observation, Hampel (1974). To be precise, suppose that the probability limit of the estimator is $T(F)$ when F is the distribution of a single observation. Let $\Delta(z)$ denote the distribution that puts probability one on z. The influence function of the estimator is then

$$\Psi(z) = \lim_{\eta \to 0} \frac{T[(1-\eta)F_0 + \eta\Delta(z)] - T(F_0)}{\eta} \tag{13}$$

where F_0 denotes the true distribution and it is assumed that convex combinations with $\Delta(z)$ are included in the domain of $T(F)$. When the estimator also satisfies Eq. (9) the two forms of the influence function generally coincide with the $\Psi(z)$ of Eq. (13), giving it another interpretation, as the effect of a single observation on the estimator in large samples. Indeed, in the modern literature on semiparametric estimation (see Bickel et al., 1993) Eq. (13) is sometimes used as the definition of the influence function.

From Eq. (10), we see that the influence functions of the partially adaptive estimators are linear in the regressors but nonlinear in the disturbance u. Butler et al. (1990) show that the $s(u)$ for GT is a smooth function that descends to zero as the residual grows in magnitude. Thus, these estimators should not be very sensitive to outliers in the values of y (or u), although they may be to outliers in x. The influence function of LMS and RLS does not exist, essentially because they do not have a representation as in Eq. (9). However, it is possible to get some understanding of the influence of a single observation on LMS by plotting stylized sensitivity curves as in Rousseeuw and Leroy (1987). Also, they do have another robustness property, a large breakdown point.

The second robustness property we consider is the breakdown point, Rousseeuw (1984). In comparison with the influence function, the breakdown point is a more global measure of robustness. To describe it we follow Rousseeuw and Leroy (1987). Let $Z = (z_1, \ldots, z_n)$ denote the data and $S(Z)$ an estimator. Let Z^* denote a "corrupted" sample obtained by replacing any m of the original data points by arbitrary values. Rousseeuw and Leroy (1987) define

$$\text{bias}(m; S, Z) = \sup_{Z^*} |S(Z^*) - S(Z)|.^{\text{b}} \tag{14}$$

[b]The right-hand side of Eq. (14) is not a bias in the sense of involving an expected value, but it does provide an indication of how far contamination can take an estimator from the uncontaminated value.

with the breakdown point then being defined as

$$\varepsilon_n^*(T, Z) = \min\left[\frac{m}{n};\ \text{bias}(m; S, Z) \text{ is infinite}\right]. \tag{15}$$

The breakdown point, $\varepsilon_n^*(T, Z)$, is the minimum fraction of outlying observations, m/n, that can cause T to become infinitely biased. The breakdown point of OLS is $1/n$, or rather, one outlying observation is sufficient to carry the least squares estimate over all bounds. Since the least squares breakdown point approaches zero as n grows indefinitely large, least squares is said to have a breakdown point of 0%. For further discussion of the breakdown point of least squares and other statistical estimators the reader is referred to Hampel et al. (1986).

Rousseeuw (1984) proves the breakdown point of LMS to be 50%. This implies it is impossible for any fraction of outliers less than 50% to have a boundless impact on the LMS estimate, no matter how distant they are from the true regression plane. As shown by Donoho and Huber (1983), LMS obtains the maximum finite sample breakdown point. Intuitively, if more than 50% of the data are outliers, it is impossible to distinguish them from "good" observations. In exploring the attributes of a number of robust estimators, Rouseeuw et al. (2001) demonstrate the breakdown point of RLS is 50%. Since the breakdown points have yet to be derived for partially adaptive estimators, this paper represents an initial attempt to understand the breakdown properties of these estimators. This is done by constructing breakdown plots and comparing the relative ability of these estimators to resist the influence of several "outliers" grouped together. While a complete understanding of breakdown properties would include an analytical derivation of the breakdown point, this is beyond the scope of this paper, and we leave this for further research.

Because of its high breakdown point, LMS is very effective in exposing outliers. Outliers are far from an LMS fit to the data, and hence, exhibit large residuals. In comparison, many other outlier diagnostics, based on OLS residuals, often *mask* outliers and/or are difficult to use. Cook's squared distance (Cook, 1977), amounts to examining the change in the OLS estimate after deleting one or more observations from the sample. As Rousseeuw and Leroy (1987) point out, the proper subset of observations to delete is not always obvious and with large data sets the computations involved in considering all possible subsets become impractical. Rousseeuw and van Zomeren (1990) also find that LMS is more robust in exposing masked outliers in multiple linear regression models than the classical method of computing Mahalanobis distances. In addition Chatterjee and Jacques (1994) demonstrate that LMS is more robust in exposing outliers than OLS when estimating Sharpe's (1964) market model.

4. METHODS AND RESULTS

4.1. Non-normal Data Generating Processes

Comparisons of small sample performance are based on Monte Carlo simulations. To provide the basis for some further comparisons with results obtained by other authors, the data generating process used by Hsieh and Manski (1987), McDonald and White (1993), and Newey (1988) was adopted. The true regression model is

$$y_i = -1.0 + 1.0x_i + u_i^{c} \tag{16}$$

where x is i.i.d. binomial such that $P(x_i = 0) = P(x_i = 1) = 0.5$, and the sample size is 50. Error terms were generated from the standard normal distribution, one leptokurtotic distribution (Laplace with unit variance), one extremely leptokurtotic distribution (Cauchy), and a skewed error distribution (transformed lognormal with zero mean and unit variance). Five hundred replications of random samples of size 50 were generated from each error distribution along with the given x_i to obtain y_i in (16). Using OLS, LAD, three partially adaptive estimation techniques, LMS, and RLS, seven sets of parameter estimates[d] were obtained for each sample. For samples with normal, Laplace, and lognormal distributed errors, efficiency was measured as the root mean square error (RMSE) of the slope estimator. Since the variance of the Cauchy is not defined, the inter-quartile ranges (Q3–Q1) are reported as for slope estimators in this case.

The "spread," RMSE or inter-quartile ranges, of the slope estimators are reported in Table 1. The least squares and LAD are included in the table to provide a comparative benchmark of how traditional regression procedures perform. As expected, OLS maintains the smallest RMSE when the errors are normally distributed, and LAD maintains the smallest RMSE when the errors are Laplace distributed. However, the improvement of OLS over EGB2 in the normal case is a mere 4%, and the improvement of LAD over EGB2 in the Laplace case is only 4%. On the other hand, EGB2 exhibits a 14% improvement over OLS in the Laplace case, and 17% improve-

Table 1. Spread of alternative slope estimators.

	Normal (RMSE)	Laplace (RMSE)	Cauchy Q3–Q1	Transformed LN (RMSE)
OLS	0.28	0.28	3.37	0.28
LAD	0.35	0.23	0.61	0.17
Partially adaptive				
BT	0.30	0.24	0.65	0.19
GT	0.32	0.25	0.62	0.12
EGB2	0.29	0.24	0.77	0.05
Robust estimators				
LMS	0.65	0.40	0.68	0.15
RLS	0.33	0.26	0.63	0.11

[c]The x_i used in this study are the same x_i used by McDonald and White (1993). Additional details about the data generating process can be obtained from the previously cited papers.

[d]Least median of squares and RLS were estimated using the algorithm outlined in Rousseeuw and Leroy (1987). The other estimation was performed using programs based on GQOPT.

ment over LAD in the normal case. Further, the relative performance of EGB2 in the case of lognormal errors is impressive: 71% less than that of LAD and 82% less than that of OLS. Hence, when the true error distribution is unknown, regression via OLS or LAD may be quite risky in terms of efficiency compared to EGB2. If the true error distribution is incorrectly assumed to be normal or Laplace, the loss in efficiency can be quite large.

The RMSE results for OLS and LAD agree very well with the results obtained by Hsieh and Manski (1987) and Newey (1988).

In comparison with the results obtained by McDonald and White (1993), RLS seems to perform similarly to the GMM, and outperform AML[e] in the case of lognormal errors. In the case of normal errors, however, both AML and GMM appear to be more efficient than RLS.

The nested relationships between the BT and GT distribution families help explain many of the results reported in Table 1. For example, since the BT nests the normal and Laplace distributions and the BT partially adaptive estimator estimates fewer parameters than the GT, the BT partially adaptive estimator may be expected to be more efficient than the GT when the true data generating process is based on either a normal or Laplace error distribution. The results of Table 1 agree with these expectations. The excellent performance of the EGB2 partially adaptive estimator, however, is not as well explained by nested relationships. Of all the error distributions considered in this paper, the only distribution nested in the EGB2 is the normal as a limiting case. Further, the EGB2 partially adaptive estimator is required to estimate more parameters than the BT and the GT. Hence, the low RMSE results for the EGB2 are impressive. In comparison, RLS also appears to be quite efficient when the errors are extremely leptokurtotic, but does not perform as well as the EGB2 in the case of normal, moderately leptokurtotic, or skewed errors.

4.2. Sensitivity to Data Contamination

This section seeks to compare how well partially adaptive estimators resist data contamination relative to RLS and LMS. The least squares and LAD will also be examined to provide a relative understanding of how conventional regression procedures perform.

Breakdown plots are a convenient method of comparing the effect of data contamination on regression estimators. The specifications of the breakdown plots created in this paper follow those used by Rousseeuw and Leroy (1987). One hundred observations were generated according to the model

$$y_i = 2.0 + 1.0x_i + u_i \tag{17}$$

[e]The adaptive maximum likelihood estimator is based on a normal kernel. The fully iterative version of the transformed GMM estimators outlined by Newey (1988) was used. The slope RMSE for GMM is 0.30 and 0.11 for the normal and lognormal error distributions, respectively. The corresponding results for AML are 0.28 and 0.13.

where x is distributed uniformly on the interval $[1, 4]$, u_i is distributed with zero mean and standard deviation 0.2. The sample's R^2 was found to be approximately 0.95. Fifty "leverage points" were then generated according to a bivariate normal distribution with mean (7.2), standard deviation 0.5, and $\rho = 0$. The sample was then contaminated by replacing observations drawn at random from the sample with leverage points. The least squares, LAD, partially adaptive estimation, and RLS and LMS were used to estimate the model parameters between each step of the contamination process. Up to 55% of the data was contaminated. Breakdown plots, which map the estimated slope parameter of each estimator as a function of percent contamination, were then constructed and given in Fig. 2. The above experiment was repeated with standard deviation of u_i equal to 0.85 and R^2 approximately 0.50. Breakdown plots for this sample are given in Fig. 3.

In Fig. 2, the remarkable abilities of LMS and RLS to resist large proportions of data contamination are readily apparent. These estimators perform well until 50% of the data is contaminated. The least squares on the other hand, begins to collapse after only 1% contamination. In comparison, the partially adaptive estimators do not perform as well as LMS or RLS, but clearly prove to be much more resistant to the contamination process than OLS and LAD. Breakdown plots for BT, GT, and EGB2 break sharply at 15%, 10%, and 11% contamination respectively. While LAD also outperforms OLS, it is important to note that LAD does not maintain the true slope estimate even for small fractions of data contamination as well as the partially adaptive estimators. Although LAD does not break zero until 15% of the data has been contaminated, the estimator is attracted towards the leverage points very early in the contamination process. Rousseeuw and Leroy (1987) find that Huber's M-estimator with Mallows' weights (Mallows, 1975), and the repeated median estimator (Siegal, 1982), resist 6%, 30%, and 40% contamination respectively. In comparison then, RLS appears to be much more effective in resisting data contamination while the partially adaptive techniques seem to be only moderately effective.

In Fig. 3, which presents breakdown plots generated from data with R^2 approximately equal to 0.50, the effect of data contamination is more severe. Least median of squares and RLS sharply break at 30% contamination.[f] Least median of squares also begins to behave rather erratically around 13% contamination, increasing well above unity and then dropping to about 0.76 before collapsing below zero. The least squares estimator behaves almost exactly as it did in the previous experiment, falling towards the leverage points very early in the contamination process. Again, the partially adaptive estimators do not perform well as LMS or RLS, but are shown to outperform OLS. The GT estimator is able to soundly resist up to 7% contamination. The BT estimator on the other hand, is drawn to the leverage points early in the contamination process and breaks below zero at 6% contamination. The EGB2 estimator is also attracted to the contaminated data early on and breaks below zero at 9% contamination. Apparently, BT and EGB2 are not as effective in resisting small fractions of data contamination as GT. Again the LAD estimator out performs OLS, but does not maintain the true slope estimate even for small fractions of data contamination as well as any of the partially adaptive estimators.

[f]However, recall that the finite-sample breakdown point of RLS is still 50%. The finite-sample breakdown point represents the percentage of contamination required to carry the estimator over all bounds.

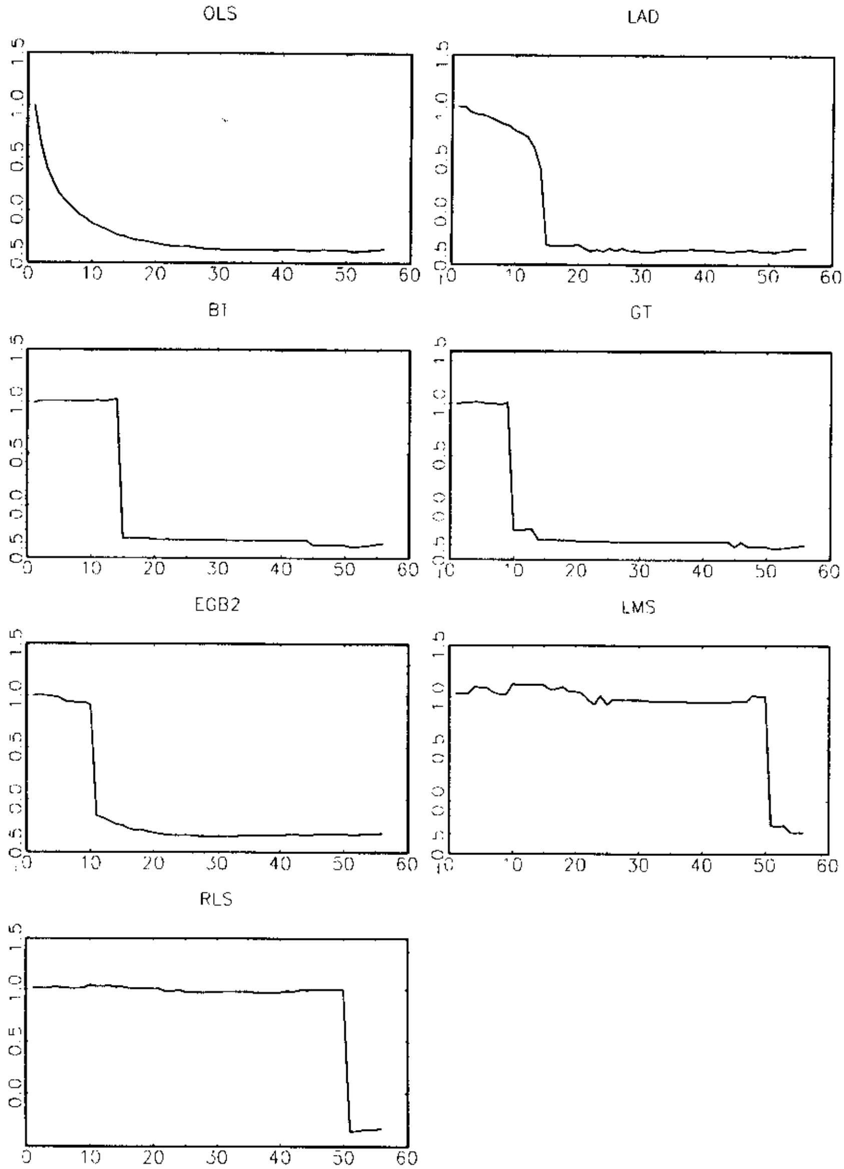

Figure 2. Breakdown plot with R^2 approximately equal to 0.95. The x-axis in each figure represents percent data contam-ination and the y-axis characterizes the estimated regression slope. The true regression slope is 1.0.

4.3. An Empirically Based Monte Carlo Example with Generalized Autoregressive Conditional Heteroskedastic Effects

While the simulations discussed in Section 4.1 provide some information about the relative performance of the estimators, it is of interest to explore similar issues with an empirically based design.

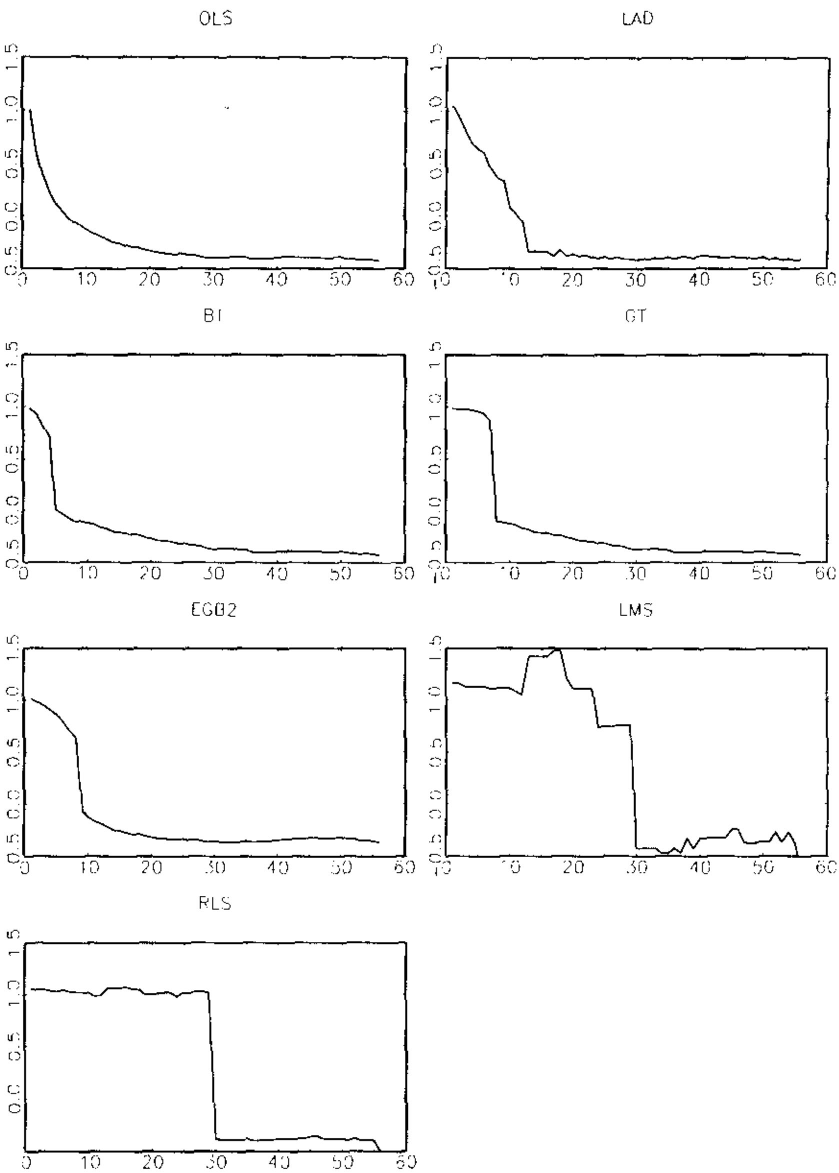

Figure 3. Breakdown plot with R^2 approximately equal to 0.50. The x-axis in each figure represents percent data contamination and the y-axis characterizes the estimated regression slope. The true regression slope is 1.0.

We consider a widely used model of excess stock returns

$$r_t^e = \alpha + \beta r_{mt}^e + u_t \tag{18}$$

where r_t^e is the excess return over the risk free rate realized on some stock, r_{mt}^e is the excess return realized on the market, u_t denotes the error term, and α and β are constants.

Many applications of this model are characterized by GARCH errors which can be modeled as $u_t = \sqrt{h_t} v_t$ where v_t is i.i.d. with zero mean and h_t evolves according to a GARCH(1, 1) process

$$h_t = \omega + \gamma u_{t-1}{}^2 + \lambda h_{t-1} \tag{19}$$

where ω, γ, and λ are all unknown parameters. Two background papers from the extensive and growing literature in this area are Bollerslev (1986; 1987).

The model defined by (18) and (19) was simulated using the CRSP monthly valued-weighted market return over the five year period from January 1997 through December 2001 for a sample size of 60 observations. Parameter values were arbitrarily set to the (rounded) estimated values using the monthly excess returns of Bird and Son Inc. obtained from the CRSP tapes.[g] The estimated parameter values were $\omega = 0.0011$, $\gamma = 0.087$, $\lambda = 0.85$, $\beta = 1.3$, and $\alpha = 0$. These estimated values are typical values for GARCH models.

The Capital Asset Pricing Model (CAPM) states that expected excess returns should be linear in regression betas,

$$E_{t-1}[r_t{}^e] = E_{t-1}[r_{\mathrm{mt}}{}^e]\beta. \tag{20}$$

Comparing the model in (20) and the expectation of the time-series regression in (18), it is clear that the model has one and only one implication for the data: the regression intercept should be zero.[h] Therefore, in estimating the regression model given in (18), we are not only interested in the estimated slope parameter as a measure of market risk but also the estimated intercept as a test of the CAPM itself. The CAPM holds in our simulations since $\alpha = 0$, and the innovation u_i represents an idiosyncratic or firm-specific shock to the stock return.

Our assumptions for the distribution of v_t include the standard normal, Laplace, a Student-t with 3 d.f, and lognormal. When necessary, the distributions were shifted to have zero mean and scaled to have unit variance. The model was simulated 500 times for each distribution.

Under the GARCH specification given in (19), the unconditional variance of u_i is given by

$$E[u_t{}^2] = \frac{\omega}{1 - \gamma - \lambda}. \tag{21}$$

To understand the impact of ARCH effects on estimator performance, we also simulate the model with constant h_t given by (21).

Table 2 reports the RMSE for the intercept and slope for the case of i.i.d. error terms as well as the GARCH error terms corresponding to the four error distributions and seven estimation procedures. The results with and without GARCH effects are similar. In the case of normally distributed errors OLS is associated with the smallest RMSE and performs the best. However, the partially adaptive estimators show little efficiency loss relative to OLS in this case. Reweighted least squares is more efficient than LMS, but less

---

[g]The data are available from the authors.

[h]This test was first proposed by Black et al. (1972).

Table 2. Root mean square error results for empirical simulation.

	Intercept				Slope			
	Normal	Laplace	Student t	Log normal	Normal	Laplace	Student t	Log normal
				Panel A: No GARCH effect				
OLS	0.016	0.016	0.018	0.017	0.330	0.326	0.316	0.320
LAD	0.020	0.013	0.014	0.039	0.404	0.268	0.252	0.189
BT	0.018	0.014	0.015	0.049	0.355	0.281	0.263	0.201
GT	0.018	0.014	0.015	0.042	0.361	0.289	0.250	0.136
EGB2	0.016	0.016	0.018	0.017	0.345	0.285	0.236	0.074
LMS	0.042	0.023	0.023	0.061	0.750	0.462	0.452	0.177
RLS	0.020	0.016	0.014	0.047	0.400	0.300	0.265	0.131
				Panel B: GARCH effect				
OLS	0.016	0.016	0.030	0.019	0.295	0.304	0.522	0.299
LAD	0.020	0.013	0.014	0.034	0.368	0.253	0.246	0.173
BT	0.017	0.013	0.015	0.041	0.324	0.266	0.257	0.180
GT	0.017	0.014	0.014	0.038	0.334	0.274	0.251	0.148
EGB2	0.016	0.016	0.028	0.019	0.302	0.269	0.249	0.135
LMS	0.040	0.022	0.022	0.050	0.724	0.425	0.415	0.164
RLS	0.018	0.015	0.017	0.039	0.347	0.278	0.263	0.129

than the partially adaptive estimators. For Laplace error terms, the partially adaptive estimators again show modest efficiency loss relative to LAD. This might be expected for the BT and GT estimators which "nest" the Laplace distribution; however, this does not explain the excellent performance of the EGB2 estimator. The Student t error distribution with three degrees of freedom has thick tails, relative to the normal and Laplace distributions and is a special case of the GT. In several cases, other estimators have the same or even marginally smaller RMSE than the GT estimator. In the case of the LN error distribution, the EGB2 has the smallest RMSE for the slope without GARCH effects and ties OLS in the case of the intercept with and without GARCH effects. The RLS estimator outperforms the partially adaptive estimators in the presence of GARCH effects with LN errors. A final observation, RLS consistently dominates LMS as an estimator for the slope and intercept.

5. SUMMARY AND RECOMMENDATIONS

From the results of this study, it is clear that outlier-resistant (LMS, RLS, and partially adaptive) estimators can provide much more efficient regression estimates than OLS if the errors do not follow the familiar normality assumption. The reduction in RMSE from using outlier-resistant methods over least squares was found to be as great as 82%. Moreover, even when the errors do follow the assumption of normality, the efficiency lost by using outlier-resistant estimation is small. While RLS does fairly well in estimating regression models that exhibit both extreme leptokurtosis and skewness, partially adaptive methods were typically found to be the most efficient outlier-resistant estimators. Further, OLS has been shown to be very sensitive to data contamination. Hence, OLS residual plots can be very misleading when trying to detect outlying observations. In comparison, RLS was shown to be very effective in resisting the attraction of data contamination, while partially adaptive estimators were shown to be only moderately effective. Of the partially adaptive estimators considered, the GT partially adaptive estimator appears to be the most effective in resisting data contamination.

Reweighted least squares and partially adaptive estimators then, both exhibit very desirable characteristics. If the error terms are truly generated by some thick tailed or skewed probability distribution, partially adaptive methods provided the best estimates. On the other hand, if the errors are truly normal, but the data is largely contaminated with gross errors, RLS will provide the best estimates, especially if the gross errors appear as leverage points.

In summary, when the error terms of a linear regression model deviate even slightly from normal, least squares becomes inefficient and masks outlying observations. Of the alternative estimation procedures considered in this paper, partially adaptive estimators appear to provide the most efficient estimates, and even perform well when the errors are extremely leptokurtotic or skewed. However, RLS was found to be considerably more effective in resisting data contamination and performed well with GARCH errors.

APPENDIX A

Asymptotic Representation of the Reweighted Least Squares Estimator

In this Appendix we give conditions for Eq. (12) in the text.

Theorem BI

If $(y_i x_i)$, $(i = 1, \ldots, n)$, are i.i.d., u_i and x_i are independent, u_i has an even (symmetric around zero) pdf $f(u)$ is continuous and bounded, $E[u_i^2]$ and $E[x_i' x_i]$ exist, var (x_i) is nonsingular, $(\tilde{\alpha}, \tilde{\beta}) = O_p(n^{-1/3})$, and $\tilde{\gamma} \xrightarrow{p} \gamma_0 > 0$, then for any $d > 0$ and for $\tilde{u}_i = y_i - \tilde{a} - x_i \tilde{\beta}$, the estimator

$$(\bar{\alpha}, \bar{\beta}) = \arg\min \sum_{i=1}^{n} 1\left(\left|\frac{\tilde{u}_i}{\tilde{\gamma}}\right| \leq d\right)(y_i - \alpha - x_i\beta)^2$$

satisfies

$$\bar{\beta} - \beta_0 = \frac{2d\gamma_0 f(d\gamma_0)}{2F(d\gamma_0) - 1}(\tilde{\beta} - \beta_0) + o_p(n^{-1/3}).$$

Proof

Let $\theta = (\alpha, \beta', \gamma)'$, $X = (1, x')'$, and $m(z, \theta) = 1(-d\gamma \leq y - \alpha - x\beta \leq d\gamma)Xu$. Note that $\|m(z, \theta)\| \leq \|X\||u|$. By an argument exactly like that of the second example in Pakes and Pollard (1989), $m(z, \theta)$ is Euclidean with envelope $\|X\||u|$. Also, $E[\|X\|^2 u^2] = E[\|X\|^2]E[u^2] < \infty$. Furthermore, by independence of u_i and x_i, and by continuity of the distribution if u_i, for all α, β, γ we have $\Pr(|y_i - \alpha - x_i\beta| = d\gamma) = 0$. Therefore, $m(z_i, \theta)$ is continuous at each θ probability one, so by the dominated convergence theorem, $E[\|m(z, \bar{\theta}) - m(z, \theta)\|^2] \to 0$ as $\bar{\theta} \to \theta$. Then by Lemma 2.17 of Pakes and Pollard and by $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}', \tilde{\gamma})' \xrightarrow{p} \theta_0 = (\alpha_0, \beta_0', \gamma_0)'$ it follows that for the true distribution F_z of z_i, for $\hat{m}(\theta) = \sum_{i=1}^{n} m(z_i, \theta)/n$ and

$$\bar{m}(\theta) = E[m(z, \theta)] = E\left[X \int_{-d\gamma + X'(b-b_0)}^{d\gamma + X'(b-b_0)} uf(u)\,du\right],$$

$$\hat{m}(\tilde{\theta}) - \hat{m}(\theta_0) - \bar{m}(\tilde{\theta}) = o_p(n^{-1/2}),$$

where $\bar{m}(\theta_0) = 0$ by symmetry. Let $b = (\alpha, \beta')'$. Note that $f(-d\gamma_0) = f(d\gamma_0)$ by symmetry. By the dominated convergence theorem and $f(u)$ continuous and bounded, $\bar{m}(\theta)$ is continuously differentiable in a neighborhood of θ_0 with

$$\frac{\partial \bar{m}(\theta_0)}{\partial b} = E[XX'(d\gamma_0 f(d\gamma_0)) - (-d\gamma_0)f(-d\gamma_0)] = 2d\gamma_0 f(d\gamma_0)Q_x, \quad Q_x = E[XX']$$

$$\frac{\partial \bar{m}(\theta_0)}{\partial \gamma} = E[X\{d(d\gamma_0 f(d\gamma_0)) - (-d)(-d\gamma_0)f(-d\gamma_0)\}] = 0.$$

It follows by a mean-value expansion that

$$\bar{m}(\tilde{\theta}) = 2d\gamma_0 f(d\gamma_0)Q_x(\tilde{b} - b_0) + o_p(n^{-1/3}).$$

Also, by similar arguments and Lemma 2.4 of Newey and McFadden (1994), it follows that

$$\hat{Q} = \sum_{i=1}^{n} 1(|\tilde{u}_i/\tilde{\gamma}| \leq d)\frac{X_i X_i'}{n} \xrightarrow{p} Z = [F(d\gamma_0) - F(-d\gamma_0)]Q_x = [2F(d\gamma_0) - 1]Q_x.$$

By var(x_i) positive definite, we have Q_x positive definite, so that $\hat{Q}$ is positive definite with the probability approaching one. Also, the central limit theorem, $\hat{m}(\theta_0) = O_p(n^{-1/2})$. Then by $\tilde{b} = b_0 + O_p(n^{-1/3})$ we have

$$\begin{aligned}\hat{Q}^{-1}\hat{m}(\tilde{\theta}) &= \hat{Q}^{-1}[\hat{m}(\theta_0) + \bar{m}(\tilde{\theta})] + o_p(n^{-1/2}) \\ &= \hat{Q}^{-1}\bar{m}(\tilde{\theta}) + O_p(n^{-1/2}) = \hat{Q}^{-}2d\gamma_0 f(d\gamma_0)Q_x(\tilde{b} - b_0) + o_p(n^{-1/3}) \\ &= (Q_x{}^{-1}Q_x)\frac{2d\gamma_0(d\gamma_0)}{2F(d\gamma_0) - 1}(\tilde{b} - b_0) + o_p(n^{-1/3}).\end{aligned}$$

The conclusion then follows by the usual least squares algebra $\hat{b} = b_0 + \hat{Q}^{-1}\hat{m}(\tilde{\theta})$. ■

ACKNOWLEDGMENTS

The authors would like to thank Peter Rousseeuw for helpful suggestions, and for providing a copy of his software program *Progress*. Suggestions from the associate editor and referees are gratefully acknowledged. Any remaining errors are our own.

REFERENCES

Bickel, P. J., Klaasen, C. A., Ritov, Y., Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press.

Black, F., Jensen, M., Scholes, M. (1972). The capital asset pricing model: some empirical tests. In: Jensen, M., ed. *Studies in the Theory of Capital Markets.* New York: Praeger.

Bloomfield, P., Steiger, W. L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms.* Boston: Birkhauser Verlag.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* 31:307–327.

Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return. *Rev. Econ. Stat.* 69:542–547.

Bollerslev, T., Engle, R. F., Nelson, D. B. (1994). Arch Models. In: Engle, R. F., McFadden, D. L., eds. *Handbook of Econometrics.* Vol. IV. Amsterdam: Elsevier Science.

Box, G. E. P., Tiao, G. C. (1962). A further look at robustness via Bayes theorem. *Biometrika* 49:419–432.

Butler, R. J., McDonald, J. B., Nelson, R. D., White, S. B. (1990). Robust and partially adaptive estimation of regression models. *Rev. Econ. Stat.* 72:321–326.

Chatterjee, S., Jacques, W. E. (1994). An outlier-resistant approach to risk estimation. *Financ. Anal. J.* Sept.–Oct.:69–75.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* 19:15–18.

Devroye, L., Gyorfi, L. (1984). *Nonparametric Density Estimation: The L_1 View.* New York: John Wiley and Sons.

Donoho, D. L., Huber, P. J. (1983). The notion of breakdown point. In: Bickel, P., Doksum, K., Hodges, J. L. Jr., eds. *A Festschrift for Erich Lehmann*. Belmont CA: Wadsworth.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* 69:383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.

He, X., Jureckova, J., Koenker, R., Portnoy, S. (1990). Tail behavior of regression estimators and their breakdown points. *Econometrica* 58(5):1195–1214.

Hsieh, D. A., Manski, C. F. (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Ann. Stat.* 15:541–551.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley and Sons.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Stat.* 1:799–821.

Kim, J., Pollard, D. (1990). Cube root asymptotics. *Ann. Stat.* 18:191–219.

Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica* 46:33–50.

Mallows, C. L. (1975). On Some Topics in Robustness. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

McDonald, J. B., Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized *t* distribution. *Econmet. Theory* 4:428–457.

McDonald, J. B., White, S. B. (1993). A comparison of some robust, adaptive, and partially adaptive estimators of regression models. *Econ. Rev.* 12(1):103–124.

McDonald, J. B., Xu, Y. J. (1995). A generalization of the beta distribution with applications. *J. Econometrics* 66:133–152. Errata (1995). 69:427–428.

Narula, S. C., Wellington, J. F. (1982). The minimum sum of absolute errors regression: a state of the art survey. *Int. Stat. Rev.* 50:317–326.

Newey, W. K. (1988). Adaptive estimation of regression models via moment restrictions. *J. Econometrics* 25:219–251.

Newey, W. K., McFadden, D. (1994). Large sample estimation and hypotheses testing. In: Engle, R. F., McFadden, D. L., eds. *Handbook of Econometrics*. Vol. IV. Elsevier: New York, pp. 2111–2245.

Pakes, A., Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* 57:1027–1057.

Rousseeuw, P. J. (1984). Least median squares regression. *J. Am. Stat. Assoc.* 79(1):871–880.

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

Rousseeuw, P. J., Van Aelst, S., Van Driessen, K., Agullo, H. (2001). Robust Multivariate Regression. Mimeographed manuscript, University of Antwerp, Department of Mathematics and Computer Science.

Rousseeuw, P. J., van Zomeren, B. C. (1990). Unmasking mutivariate outliers and leverage points. *J. Am. Stat. Assoc.* 85(411):633–639.

Sharpe, W. F. (1964). Capital asset prices: a theory of equilibrium under conditions of risk. *J. Financ.* 19(Sept.):280–286.

Siegal, A. F. (1982). Robust regression using repeated medians. *Biometrika* 69:242–244.