M Estimation of Multivariate Regressions

ROGER KOENKER and STEPHEN PORTNOY*

Robust alternatives to the seemingly unrelated regression (SUR) estimator of Zellner (1962) are proposed for the classical multivariate regression model. These weighted *M* estimators achieve an asymptotic covariance matrix analogous to that of the SUR estimator. Comparisons for the l_1, least absolute deviation, case are made with the efficient estimator in the case of elliptically contoured distributions. An example reanalyzing the Grunfeld investment data using a smooth "l_1-like" *M* estimator is discussed in detail. In contrast to recent work of Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Rousseeuw (1987), and Oja (1983), the methods studied here are *not* affine equivariant; some remarks on the potential significance of this failing conclude the article.

KEY WORDS: l_1 estimation; Robustness; Seemingly unrelated regression.

1. INTRODUCTION

Consider the classical multivariate regression model

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_m \end{pmatrix} \qquad (1.1)$$

with m equations and n observations on each equation, which we will express more succinctly as

$$y = X\beta + u.$$

When $\operatorname{cov}(u) = \Omega \otimes I$ and β is an unknown $p = \sum_{i=1}^m p_i$ vector, it is well known that the ordinary least squares estimator $\hat{\beta} = (X'X)^{-1}X'y$ is inefficient relative to the (Gauss–Markov) generalized least squares estimator $\tilde{\beta} = (X'(\Omega^{-1} \otimes I)X)^{-1}X'(\Omega^{-1} \otimes I)y$. The former has covariance matrix

$$\hat{V} = V(\hat{\beta}) = (X'X)^{-1}X'(\Omega \otimes I)X(X'X)^{-1},$$

and the latter boasts

$$\tilde{V} = V(\tilde{\beta}) = (X'(\Omega^{-1} \otimes I)X)^{-1}.$$

The difference $\hat{V} - \tilde{V}$ is positive-semidefinite. Zellner (1962, 1963) contains the seminal analysis of this situation. See Srivastava and Giles (1987) for an exhaustive treatment of the recent literature on this subject.

Similarly, it is easy to show under analogous conditions that the ordinary least absolute deviation (l_1) estimator, $\hat{\beta}$, which minimizes

$$R(b) = \sum_{i=1}^m \sum_{j=1}^n |y_{ij} - x_{ij}b_i|,$$

has asymptotic covariance matrix of the form $V(\hat{\beta})$, but with Ω replaced by

$$\Omega = (\omega_{ij}) = \frac{\mathrm{E}\,\operatorname{sgn}(u_{ik})\operatorname{sgn}(u_{jl})}{4f_i(0)f_j(0)},$$

where f_k denotes the (marginal) density of the coordinate u_k. The numerator of ω_{ij} may be regarded as an l_1 correlation based on orthant probabilities between the errors in the ith and jth equation, and the terms in the denominator are the marginal densities of these error terms evaluated at their medians. Since the latter are inversely proportional to the scale of the marginal distributions, Ω may be regarded as an l_1 covariance matrix. The bivariate version of the numerator was considered in Blomqvist (1950); see also Devlin, Gnanadesikan, and Kettering (1975).

In light of the least squares results it is natural to ask: Can we construct a generalized l_1 estimator that has an asymptotic covariance matrix of the form $V(\tilde{\beta})$? In the next section we investigate a rather broad class of weighted *M* estimators that achieve a generalized version of this objective, and we shall see that a particular weighted l_1 estimator is an important special case. Since these estimators use one-dimensional kernels, Section 3 investigates their efficiency compared with the fully multivariate asymptotically optimal estimators. We consider elliptically contoured error distributions and specialize specifically to multivariate t distributions. The basic conclusions are that, although the methods based on univariate kernels can have arbitrarily small efficiency, this tends to occur only when the error coordinates are highly correlated (and hence when the asymptotic variance is small). Thus the simple one-dimensional methods (particularly, the appropriate l_1 estimator) will generally achieve quite reasonable asymptotic performance. Section 4 illustrates the methods by reestimating the well-known Grunfeld (1958) investment model. Section 5 concludes with some comments on the issue of affine equivariance.

2. *M* ESTIMATION OF MULTIVARIATE REGRESSION

Slight departures from Gaussian behavior of u can, of course, produce arbitrarily large disturbances in the behavior of the least squares estimators referred to in the previous section. To achieve some degree of robustness against such departures from normality we might consider estimators that minimize

$$R_0(b) = \sum_{i=1}^m \sum_{j=1}^n \rho(y_{ij} - x_{ij}b_i).$$

* Roger Koenker is Professor, Department of Economics, and Stephen Portnoy is Professor, Department of Statistics, both at the University of Illinois, Champaign, IL 61820. This research was supported in part by National Science Foundation Grants SES-8707169 and DMS-8802555.

© 1990 American Statistical Association
Journal of the American Statistical Association
December 1990, Vol. 85, No. 412, Theory and Methods

The ordinary l_1 estimator is an important special case. Estimation of the m-variate location and scatter model is also an important special case, where $X_i \equiv 1_n$, an n-vector of ones, and β is an m-vector of location parameters. Under mild conditions on ρ, minimizing $R_0(b)$ is equivalent to solving the equations

$$\sum_{j=1}^{n} \psi(y_{ij} - x_{ij}b_i)x_{ij} = 0, \qquad i = 1, \ldots, m,$$

for $\psi = \rho'$. We will refer to estimators that use such one-dimensional kernels as ordinary M estimators; in the location–scatter problem the terminology "coordinatewise M estimator" might be used. Like the ordinary least squares estimator, they can be computed one equation (coordinate) at a time.

It should also be remarked at this stage that most of the attractive choices for ρ involve some scale estimation to achieve scale invariance. For example, for the leading case of the Huber M estimator,

$$\begin{aligned} \rho(z) &= \tfrac{1}{2}z^2, & |z| &\le k, \\ &= k|z| - \tfrac{1}{2}k^2, & |z| &> k, \end{aligned}$$

we require some (scale-equivariant) scale estimators $s_i : i = 1, \ldots, m$, for example, the median absolute deviation from the l_1-fit, which can be used to rescale the objective function. In these cases we should presume that

$$\rho(y_{ij} - x_{ij}b_i) = \rho_0((y_{ij} - x_{ij}b_i)/s_i)$$

for some standardized ρ_0 and the rescaling by s_i is implicitly subsumed into the function ρ defined previously. Of course, in the case of the l_1 estimator, scale invariance requires no preliminary estimation of scale. The issue of scale estimation is treated in the illustrative data analysis of Section 5.

To relax the implausible and potentially dangerous Gaussian hypothesis on u in Section 1 we will assume the following.

Condition A1. The m-vectors $u_j = (u_{1j}, u_{2j}, \ldots, u_{mj})'$ for $j = 1, \ldots, n$ are iid with joint distribution function F.

Following Ruppert and Carroll (1980) and Jurečková (1977), we also require the following.

Condition P1. The function $\psi(u)$ is bounded and monotonically nondecreasing.

Condition P2. The matrix

$$R \otimes I = (\mathrm{E}\ \psi(u_{ik})\psi(u_{jl})) = (\rho_{ij}\delta_{kl})$$

is positive-definite. Either ψ or the marginal densities f_i for the marginal distributions F_i $(i = 1, \ldots, m)$ are absolutely continuous and satisfy

$$\phi_i \equiv \int_{-\infty}^{\infty} \psi'(u)\, dF_i(u) \quad \text{or} \quad \equiv -\int_{-\infty}^{\infty} \psi(u) f_i'(u)\, du$$

for constants $0 < \phi_i < \infty$, $i = 1, \ldots, m$, and $\mathrm{E}\ \psi(u_{ik}) = 0$ for $i = 1, \ldots, m$; $j = 1, \ldots, n$.

Condition X1. Each design matrix X_i has first column equal to a vector of ones.

Condition X2. $n^{-1/2} \max|x_{ij}| = o(1)$ as $n \to \infty$.

Condition X3. For each $i = 1, \ldots, m$, $n^{-1}X_i'X_i \to D_{ii}$, where D_{ii} is a positive-definite matrix.

Note that in the least squares case $\rho(u) = \frac{1}{2}u^2$, so R is simply the usual covariance matrix of the u_i, and $\phi_i \equiv 1$. In the l_1 case, R is the "orthant probabilities correlation matrix" of covariances of the signs of the errors, and $\phi_i = 2f_i(0)$.

The asymptotic theory of the ordinary M estimator is immediately obtained from the asymptotically linear representation of the M estimator for each equation,

$$\hat{\beta}_i - \beta_i = n^{-1}(\phi_i D_{ii})^{-1}X_i'\psi_i + o_p(n^{-1/2}), \qquad i = 1, \ldots, m, \quad (2.1)$$

where $D_{ii} = \lim n^{-1}X_i'X_i$ and $\psi_i = (\psi(u_{ij}))$, $i = 1, \ldots, m$. The joint asymptotic normality of these vectors follows immediately as in single equation context. A typical block of the covariance matrix is

$$\begin{aligned} &\mathrm{cov}((\hat{\beta}_i - \beta_i), (\hat{\beta}_j - \beta_j)) \\ &\qquad = n^{-2}\phi_i^{-1}\phi_j^{-1}\rho_{ij}D_{ii}^{-1}X_i'X_jD_{jj}^{-1} + o_p(n^{-1}). \end{aligned}$$

Thus the covariance matrix for the entire vector $(\hat{\beta} - \beta) = ((\hat{\beta}_i - \beta_i))$ may be written as

$$\hat{V} = (X'X)^{-1}X'(\Delta \otimes I)X(X'X)^{-1},$$

where $\Delta = \Phi^{-1}R\Phi^{-1}$ with $\Phi = \mathrm{diag}(\phi_i)$. It might be noted that we can also write

$$\hat{V} = (X'PX)^{-1}X'(R \otimes I)X(X'PX)^{-1},$$

where $P = \Phi \otimes I$. Clearly the block diagonality of X as well as the Kronecker product form of P is essential to the preceding "simplification". The latter form for the asymptotic covariance matrix of the l_1 estimator has recently been derived by Kuester (1987).

As we observed previously, it is natural to ask whether we can improve on the asymptotic performance of this ordinary M estimator, designing a generalized M estimator that would achieve asymptotic covariance matrix,

$$\tilde{V} = (X'(\Delta^{-1} \otimes I)X)^{-1}.$$

This objective is easily achieved if we simply replace the "normal equations" of the unweighted objective function, which we may express in more compact form as

$$X'\psi(b) = 0,$$

with the *weighted* normal equations

$$X'P(R^{-1} \otimes I)\psi(b) = 0. \quad (2.2)$$

In cases where ψ is not continuous, Theorem 2.1 will apply to any estimator satisfying $X'P(R^{-1} \otimes I)\psi(b) = o_p(n^{-1/2})$. A natural question at this point is whether or not there is an optimization problem that implies (2.2), but differentiating (2.2) with respect to b and noting that the resulting matrix is *not* symmetric resolves the question negatively.

Our main result is the following asymptotic representation of $\tilde{\beta}_n$, the estimator solving (2.2).

Theorem 2.1. In the multivariate linear model (1.1), suppose that Conditions A, P, and X hold. Then

$$\tilde{\beta}_n - \beta = (X'P(R^{-1} \otimes I)PX)^{-1}X'P(R^{-1} \otimes I)\psi(0) + o_p(n^{-1/2}), \qquad (2.3)$$

where $\psi(0) = (\psi(u_{ij}))$.

Proof. Consider the normalized gradient,

$$g(\delta) = n^{-1/2}X'P(R^{-1} \otimes I)\psi(\delta),$$

where $\psi(\delta) = (\psi(u_{ij} + n^{-1/2}x_{ij}\delta_i))$, an mn-vector. Familiar arguments from Ruppert and Carroll (1980) and Bickel (1975) imply for fixed $L > 0$,

$$\sup_{\|\delta\|<L} \|g(\delta) - g(0) - \mathrm{E}(g(\delta) - g(0))\| = o_p(1). \qquad (2.4)$$

Further, $\tilde{\delta} = n^{1/2}(\tilde{\beta} - \beta) = O_p(1)$, $\mathrm{E}\, g(0) = 0$, and $g(\tilde{\delta}) = o_p(1)$. Finally, by expanding $\psi(\cdot)$, we have

$$\sup_{\|\delta\|<L} \|\mathrm{E}\ g(\delta) - n^{-1}X'P(R^{-1} \otimes I)PX\delta\| = o_p(1), \qquad (2.5)$$

so substituting $\tilde{\delta}$ in (2.5) and then in (2.4) completes the argument for $\|\tilde{\delta}\| \leq L$. As in Ruppert and Carroll (1980) or Jurečková (1977), monotonicity of ψ completes the argument.

An immediate application of this result is the asymptotic normality of $n^{1/2}(\tilde{\beta} - \beta)$, which has mean 0 [since $\mathrm{E}\ \psi(0) = 0$]. The asymptotic covariance matrix of $(\tilde{\beta} - \beta)$ is

$$\tilde{V} = (X'P(R^{-1} \otimes I)PX)^{-1} = (X'(\Delta^{-1} \otimes I)X)^{-1}. \qquad (2.6)$$

Note that each component $(\tilde{\beta}_i - \beta_i)$ is expressed in Theorem 2.1 as a weighted sum of n independent components. Our design conditions ensure that these summands satisfy the Lindeberg condition; compare Koenker and Bassett (1978).

It may be noted that, as in the classical case, if the design matrix is the same in all m equations then there is no efficiency gain in solving (2.2). Indeed, it is easy to see that any solution to the equation-by-equation M estimation problem will also solve (2.2) in this case.

As in the classical least squares case it is important to consider the consequences of replacing P and R in (2.2) by estimates. Arguments similar to those in the classical context, however, yield an identical asymptotic theory provided that $\hat{\Delta} \rightarrow \Delta$ in probability. In subsequent work we hope to explore the practical consequences of various estimation schemes for Δ.

3. COMPARISONS WITH OPTIMAL ESTIMATORS IN THE ELLIPTICALLY CONTOURED CASE

Although solving (2.2) provides an asymptotic improvement over the naive M estimator, this method still depends on a one-dimensional kernel. Since the problem is inherently multidimensional, this poses the question of how much one is sacrificing for the sake of simplicity. Two comments can be made here.

First, the results of Portnoy [see Portnoy (1977) and, especially, Portnoy (1979, sec. 1)] suggest that if there is only small dependence between the equations, a one-dimensional kernel with a small amount of redescent provides the first-order correction to the optimal estimator. Thus there is little sacrifice of efficiency if the dependence is small. If the dependence is large, however, improvements can be made by using fully multivariate estimators; for example, the maximum likelihood estimator for model (1.1). Comparisons are somewhat difficult to make in the completely general case, but the elliptically contoured case provides relatively clear and simple comparisons. Consider $u = (u_1, \ldots, u_n)$ as a matrix of a sample of size n from a multivariate density, f, on $\mathbf{R}^m$, which is elliptically contoured with parameter Λ. That is, Λ^{-1} is the "precision matrix," or, equivalently, $\Lambda^{-1/2}u_j$ is spherically symmetric. The matrix Λ is not uniquely defined but is only determined up to a positive multiplicative constant. Thus, when variances exist, we will generally specify the constant by taking $\Lambda = \mathrm{cov}(u_j)$. Clearly, the results do not depend on having a finite variance, but this specification will permit direct comparisons to be made. The specific examples considered here will take u_j to have a multivariate t distribution (with covariance Λ) and will emphasize the case in which the dimension $m = 2$.

The results may be summarized as follows. The optimal asymptotic covariance matrix is the inverse Fisher information matrix, which Theorem A.1 in the Appendix shows to be

$$V^* = c^*(X'(\Lambda^{-1} \otimes I)X)^{-1}, \qquad (3.1)$$

where c^* is defined by (A.1). Since the asymptotic covariance for the solution to (2.2) (the weighted M estimator) is of rather different form, we can simplify the comparisons by considering two stages. First, consider the case where we transform by $\Lambda^{-1/2}$ to obtain spherical symmetry. Theorem A.2 shows that the asymptotic covariance for the weighted M estimator applied to the transformed data is

$$V_{\mathrm{tr}} = c_{\mathrm{tr}}(X'(\Lambda^{-1} \otimes I)X)^{-1}, \qquad (3.2)$$

where c_{tr} is given by (A.2). Thus efficiencies of weighted M estimators applied to the transformed data can be readily computed by comparing c_{tr} with c^*. As a specific example, consider the multivariate t distribution with q df (for $q > 0$) and dimensions $m = 2$, 5, 10, and scaled so that each coordinate has variance 1. In this case, values for c^* and c_{tr} are calculated in Proposition A.1 of the Appendix [Eq. (A.3)]; efficiencies for the weighted l_1 estimator, c^*/c_{tr}, are plotted in Figure 1, along with efficiencies for the least squares estimator (where the constant is $c = 1$). Note that although the efficiency of the l_1 estimator can tend to 0, it.does so only for extreme error distributions where the asymptotic covariance is already quite small.

Finally, we compare the asymptotic covariances for the weighted l_1 estimator applied to the original data with those of the same estimator applied to the transformed data in the case in which $m = 2$. Proposition A.2 computes the covariance matrix given in (2.6) under a bivariate t

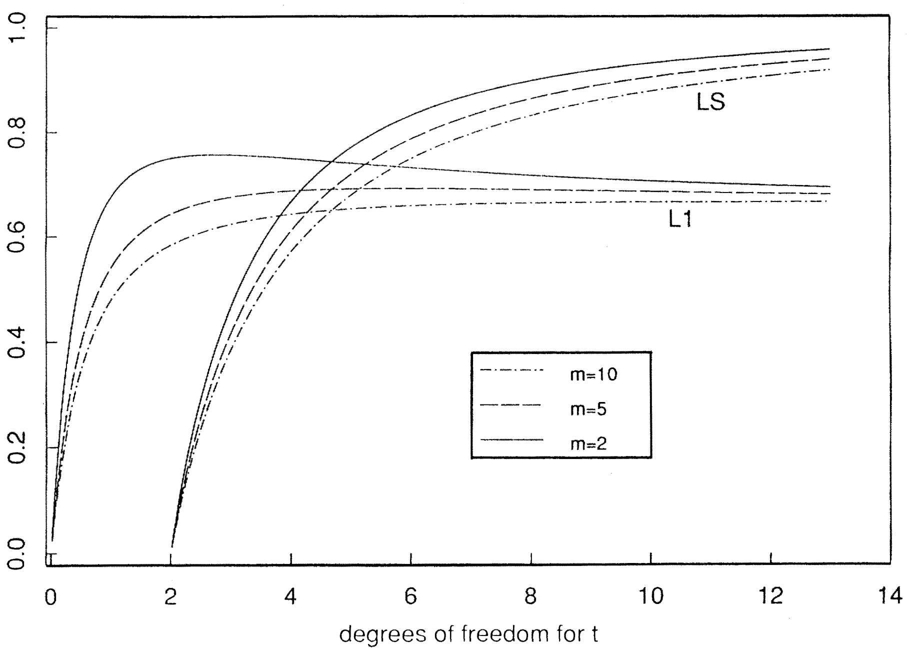

Figure 1. Efficiencies for L1 and LS.

distribution with q df (again scaled to give variance 1): $\tilde{V} = c_{\text{tr}}(X'(\Delta^{-1}(u) \otimes I)X)^{-1}$,

$$\text{where } \Delta(u) = \begin{bmatrix} 1 & \frac{4}{\pi}\sin^{-1}\rho \\ \frac{4}{\pi}\sin^{-1}\rho & 4 \end{bmatrix} \qquad (3.3)$$

and where ρ is the correlation parameter in the specific example defined by (A.4). It turns out that Λ and Δ have the same diagonal elements (when $m = 2$), so $V_{\text{tr}} < \tilde{V}$ (in the sense of having a positive-definite difference) iff $\det(\Delta) < \det(\Lambda)$. In fact, the ratio of these determinants is just the ratio of generalized variances, $\det(V_{\text{tr}})/\det(\tilde{V})$. Thus $e \equiv \{\det(\Lambda)/\det(\Delta)\}^{1/2}$ is a measure of efficiency that is scaled as a ratio of variances. Direct computation shows that e monotonically decreases to 0 as $|\rho| \to 1$. Furthermore, e is moderately large unless there is substantial correlation among the equations, in which case the actual variance $\det(\tilde{V})$ is already small. In particular, $e \geq .82$ for $|\rho| \leq .7$ and $e \geq .62$ for $|\rho| \leq .9$.

As a final consequence, therefore, we can expect the weighted l_1 estimator to be reasonably efficient unless $\tilde{V}$ is already quite small. That is, inference based on the solution to (2.2) should be fairly good even though it does not take full account of the multivariate nature of the problem.

4. AN EXAMPLE

To illustrate the methods described previously, we now reconsider the well-known Grunfeld (1958) investment model. Grunfeld proposed and estimated a simple model in which a firm's investment in period $t + 1$ was linear in the firm's capital stock in period t and in the market value of the firm in period t. Grunfeld's data, which consist of annual observations on these quantities for several major U.S. corporations, 1935–1954, has been subsequently reanalyzed many times. See, for example, Boot and De Wit (1960) and the textbook treatment by Theil (1971) for the data and further details on the model.

We will consider, like Theil, only two firms: General Electric (GE) and Westinghouse (WH). Thus we have a model of the form (1.1) with $m = 2$, $n = 20$, $p_1 = p_2 = 3$. For numerical stability we have rescaled the data so market values are in billions of dollars, and the investment and capital stock variables are in hundreds of million dollars. In Table 1 we report ordinary least squares (OLS) and normal theory seemingly unrelated regression (SUR) estimation of the Grunfeld model. The estimated covariance matrix for the SUR estimates is

$$\begin{pmatrix} .066 & .018 \\ .018 & .009 \end{pmatrix},$$

which implies an estimated correlation between the errors of the two equations of .73.

We choose to illustrate our methods with a smooth l_1-like M estimator. This avoids some difficult computational problems in solving (2.2) when ψ is discontinuous and facilitates the computation of standard errors for reported estimates by avoiding the problem of sparsity estimation [e.g., see Welsh (1987)]. As in Amemiya (1982), we con-

Table 1. Classical Estimation of the Grunfeld Investment Model

	Intercept	*Market value*	*Capital stock*
OLS			
GE	−.100	.266	.152
	(.313)	(.156)	(.026)
WH	−.005	.529	.092
	(.080)	(.157)	(.056)
SUR			
GE	−.277	.383	.139
	(.289)	(.142)	(.025)
WH	−.012	.576	.064
	(.074)	(.143)	(.052)

NOTE: Standard errors appear in parentheses.

sider a logistic approximation to the l_1 ψ-function $\psi(u) = \operatorname{sgn}(u)$ as

$$\psi_\lambda(u) = -(1 - 2/(1 + e^{-\lambda u})),$$

where λ is a scale factor that controls the l_1-ness of the approximation. As with any such M-estimation method, some concomitant scale estimation is required to achieve scale equivariance. We adopt the prevalent device of starting our iterations as the coordinatewise l_1 estimate and using the mad scale estimate, that is,

$$s = 2c \operatorname{median}\{|\hat{u}_i - \operatorname{median}\{\hat{u}_i\}|\},$$

where $c = .7413$ is chosen to achieve (approximate) Fisher consistency at the Gaussian model.

In Table 2 we present single-equation estimates as well as the starting values provided by the l_1 estimates. The M estimates solve the equation

$$\sum_{j=1}^{n} x_j \psi_\lambda((y_j - x_j b)/s) = 0.$$

Since the Jacobian of this equation is easily computed analytically we employ the algorithm DZONEJ from the Port3 library (Fox 1984). To estimate standard errors we adopt a slight variation on one of the proposals of Huber (1981, sec. 7.6) for which we estimate the asymptotic covariance matrix of the M estimate $\hat{\beta}$ by $V_n = H_n^{-1} G_n H_n^{-1}$, where

$$G_n = \sum x_i x_i' \psi_\lambda^2((y_i - x_i\hat{\beta})/s)$$

and

$$H_n = \sum x_i x_i' \psi_\lambda'((y_i - x_i\hat{\beta})/s)(\lambda/s).$$

The scale factor λ is analogous to the Huber k; we have chosen it in such a way that under Gaussian conditions 20% of the observations would have $|\psi_\lambda(u)| < .99$. So the resulting M estimator behaves, roughly, like a 40% trimmed mean. In general, we may write

$$\lambda = \frac{\log((a + 1)/(1 - a))}{\Phi^{-1}(1 - b)},$$

where a is a bound on the ψ-function and b is a desired level of trimming. Here we have set $a = .99$ and $b = .40$.

Estimating the parameters of Ψ and R as

$$\hat{R}_{ij} = n^{-1} \sum_{k=1}^{n} \psi_\lambda(\hat{u}_{ik}/s_i)\psi_\lambda(\hat{u}_{jk}/s_j) \tag{4.1}$$

and

$$\hat{\Psi}_i = n^{-1} \sum_{k=1}^{n} \psi_\lambda'(\hat{u}_{ik}/s_i)(\lambda/s_i), \tag{4.2}$$

we obtain

$$\hat{R} = \begin{pmatrix} .854 & .518 \\ .518 & .865 \end{pmatrix}, \qquad \hat{\Psi} = \begin{pmatrix} 5.39 & 0 \\ 0 & 13.11 \end{pmatrix}.$$

The final M estimation of the two equations, obtained by solving (2.2), is reported in Table 3, where we have computed standard errors in accordance with the expression (2.6). Estimated standard errors are reported both by evaluating (2.6) at the initial estimates $\hat{R}$ and $\hat{\Psi}$ and reestimating R and Ψ using residuals from the multivariate fit.

Table 2. Single-Equation M Estimation of the Grunfeld Investment Model

	Intercept	Market value	Capital stock
GE	−.110 −.119 (.072)	.252 .252 (.028)	.150 .156 (.020)
WH	.051 .036 (.060)	.397 .417 (.096)	.139 .134 (.041)

NOTE: Line 1 in each table section contains the (l_1) starting values, line 2 reports M estimates, and the numbers in parentheses are standard errors for the M estimates computed from V_n.

Since the matrix $\Delta^{-1} = (\Psi R^{-1} \Psi)^{-1}$ plays the role of the covariance matrix in our M estimation of multivariate models, it is worth noting that, after reestimating R and Ψ,

$$\hat{\Delta}^{-1} = \begin{bmatrix} .022 & .006 \\ .006 & .004 \end{bmatrix},$$

which, if viewed as a conventional covariance matrix, implies a correlation of .65, compared with the .73 for the corresponding classical SUR estimates.

Since we are not privileged to know the *true* values of the parameters for this example, it is difficult to draw definite conclusions from the foregoing results. Clearly, the M estimates are quite stable with respect to the initial l_1 single-equation results, but rather substantial differences exist between this group of estimates and the SUR results. One way to illustrate the robustness of the M-estimation approach is to study the effects of introducing artificial contamination into an existing data set, like the Grunfeld data.

We undertake two simple experiments of this type. In the first we select an arbitrary observation from the first equation and introduce additive contamination to it. More explicitly, we let $y_{1,12}^* = y_{1,12} + d$ and study the resulting perturbation in our estimates as a function of the scalar d. The consequences of this contamination are displayed in sensitivity curves (Figures 2 and 3) and are quite different in the two equations. In the first equation, the SUR

Table 3. Multivariate M Estimation of the Grunfeld Investment Model

	Intercept	Market value	Capital stock
GE	−.114 (.186) (.159)	.255 (.092) (.078)	.151 (.016) (.013)
WH	.051 (.054) (.049)	.392 (.104) (.094)	.109 (.038) (.034)

NOTE: Two sets of standard errors are reported. The first set of figures in parentheses is based on evaluating (2.6) at $\hat{R}$, $\hat{\Psi}$ given in (4.1) and (4.2), and the second row is based on reestimation of R, Ψ.

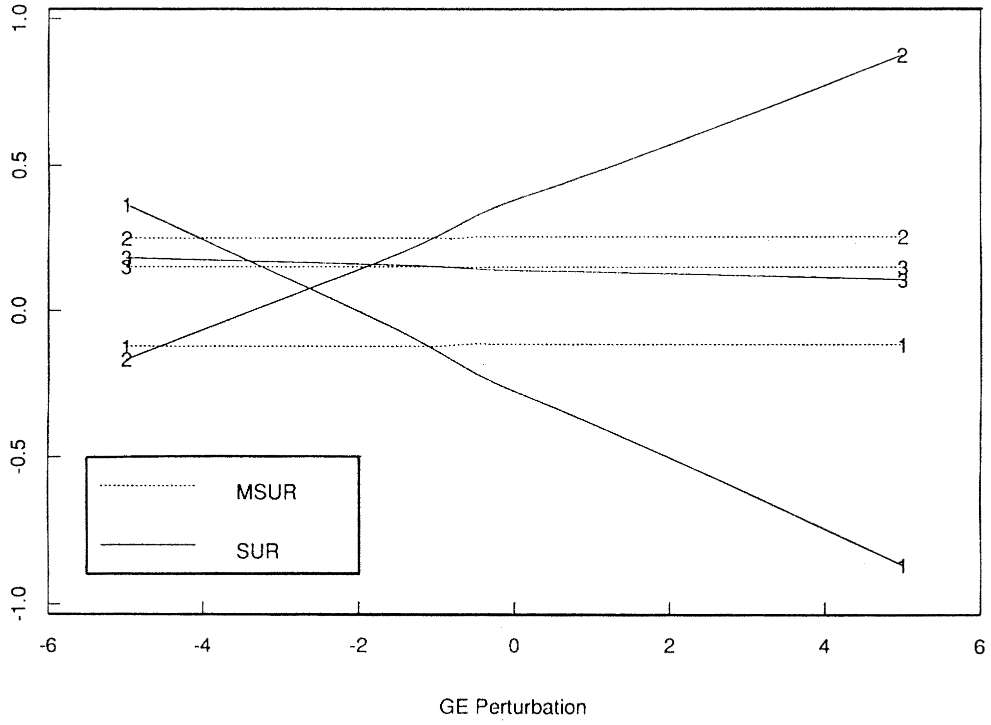

Figure 2. Sensitivity Curves for GE Parameters.

estimates appear essentially linear in d. So, as in OLS estimation, a single bad observation may create an arbitrarily large perturbation in the estimates. In the second equation the situation is somewhat more complicated. The contamination in the first equation has the effect of inflating the estimated variance of the first equation, thus decreasing its influence in the estimated parameters of the second equation. Correlation between the two equations diminishes but does not vanish. The net effect is a modest perturbation in the estimated parameters of the second equation, which gradually attenuates as the contamination becomes more extreme.

In contrast, the effect of the contamination on the M estimates is barely perceptible. A slight perturbation occurs as the contaminated observation crosses the plane determined by the initial fit, but further more extreme contamination has no further consequences.

In the second experiment we contaminate both obser-

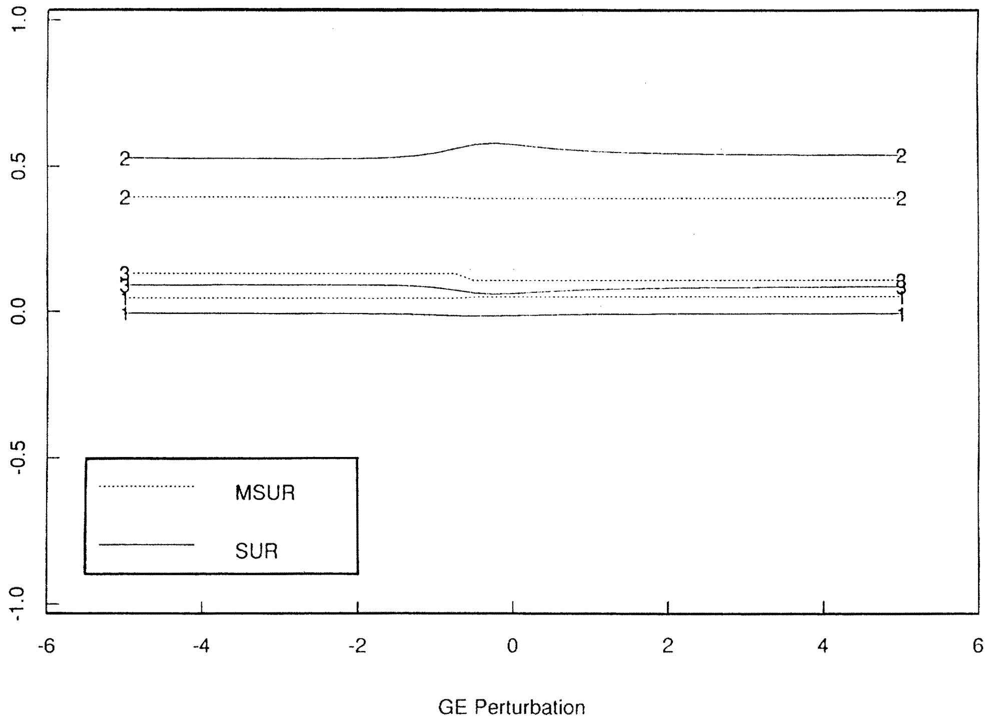

Figure 3. Sensitivity Curves for WH Parameters.

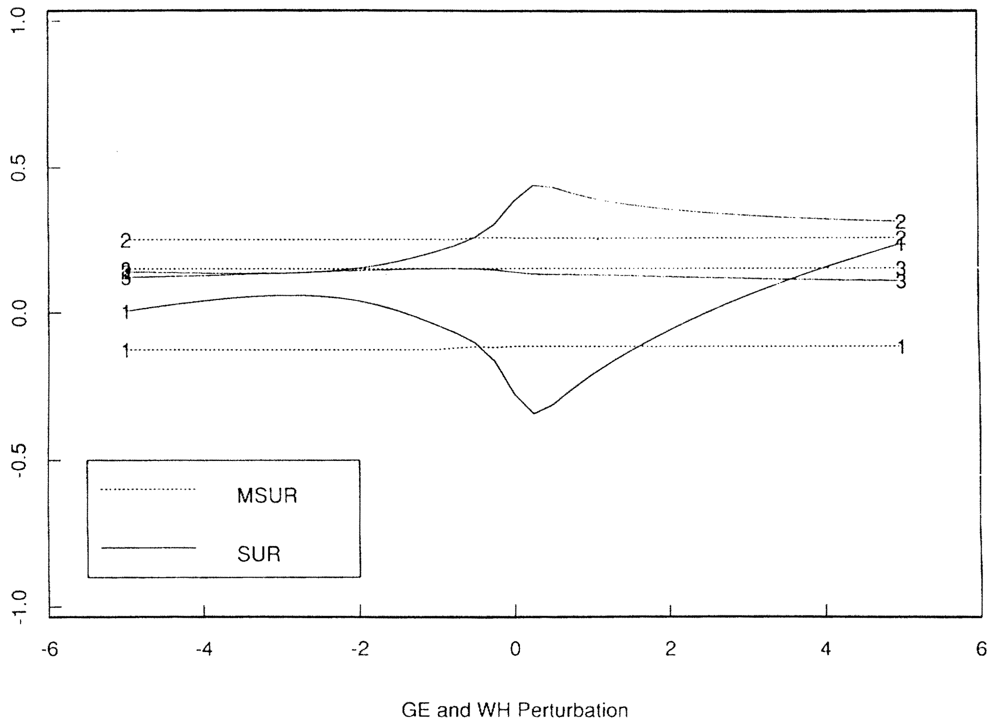

Figure 4. Sensitivity Curves for GE Parameters.

vations corresponding to a given year. Explicitly, $y^*_{1,12} = y_{1,12} + d$, $y^*_{2,12} = y_{2,12} + d$. The results appear in Figures 4 and 5. Now, the *pair* of contaminated observations gradually comes to dominate the correlation between the two equations, driving it to one. All of the SUR estimates behave linearly in d, for large values of $|d|$. In contrast, the MSUR estimates are completely insensitive to large values of the perturbation d.

5. ON AFFINE EQUIVARIANCE

To conclude, a brief apologia is required for the dereliction of affine equivariance. Most of the recent work on robust multivariate analysis [see Rousseeuw (1987) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986, chap. 5) and references cited there] has restricted attention to estimators that commute with affine transformations. Sup-

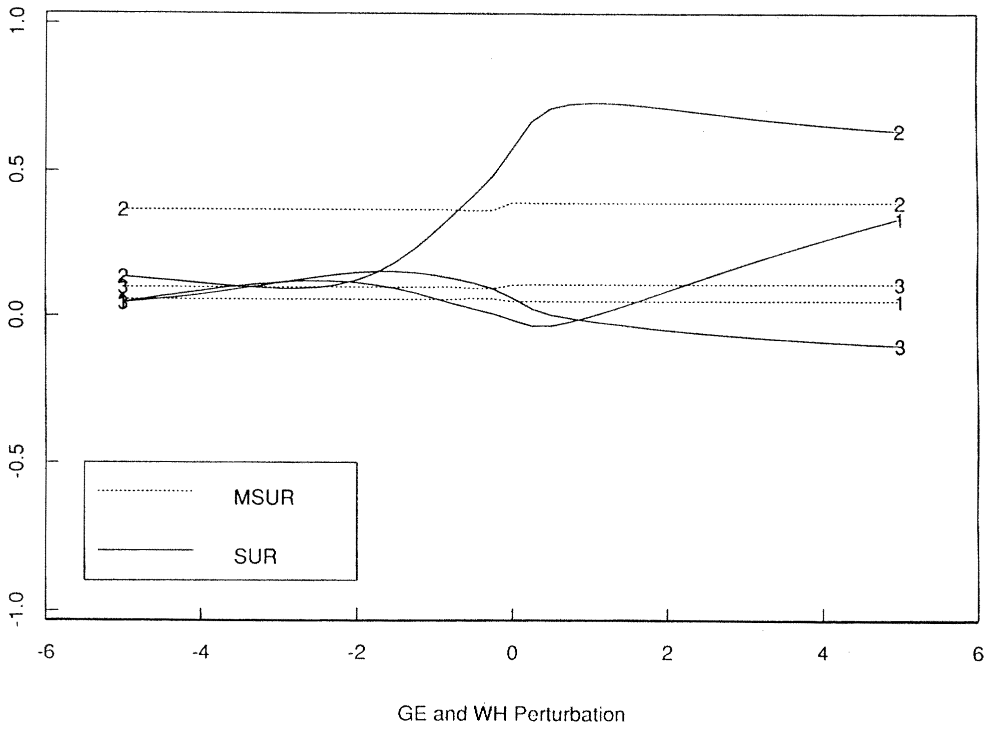

Figure 5. Sensitivity Curves for WH Parameters.

pose that $T(y_1, \ldots, y_n)$ is an estimator of multivariate location based on observations $\{y_i \in \mathbf{R}^p : i = 1, \ldots, n\}$. Then T is said to be affine equivariant iff

$$T(y_1A + b, \ldots, y_nA + b) = T(y_1, \ldots, y_n)A + b \tag{5.1}$$

for any $b \in \mathbf{R}^p$ and nonsingular $(p \times p)$ matrix A. This property is particularly compelling in physical applications where, for example, the coordinate system for $\mathbf{R}^3$ is arbitrary. In many applications, however, the measured coordinates *are* meaningful—commodity bundles in economics, for example. Then, nondiagonal transformations A are difficult to interpret.

The methods suggested here satisfy (5.1) for diagonal A and, therefore, *are* affine equivariant coordinate by coordinate. They do not commute, however, with arbitrary nonsingular matrices A. Whether this failure is a mere peccadillo or a mortal sin seems debatable. Unless linear combinations of individual coordinates are meaningful quantities there appears to be little harm in restricting affine equivariance to be a coordinate-by-coordinate property. Unfortunately, the most appealing of the affine equivariant methods, due to Oja (1983) and Rousseeuw (1987), are extremely difficult to compute; this may offer another, at least temporary, rationale for the methods suggested here.

APPENDIX: THEORETICAL RESULTS FOR THE ELLIPTICALLY CONTOURED CASE

Theorem A.1. Consider the aforementioned elliptically contoured case. Define the function g on $\mathbf{R}^+$ by $g(u'\Lambda^{-1}u) = -\log f(u)$ for $u \in \mathbf{R}^m$. Assume appropriate regularity conditions for the maximum likelihood estimator to have an (optimal) asymptotic covariance matrix equal to the inverse of the Fisher information matrix. [For example, general conditions applicable to this SUR problem can be found in th. 4.2 (p. 194) of Ibragimov and Has'minskii (1981)]. Then this optimal covariance matrix is given by (3.1), where

$$\frac{1}{c^*} = \frac{4}{m} \mathrm{E}\|u_j\|^2(g'(\|u_j\|^2))^2. \tag{A.1}$$

Proof. First consider the spherically symmetric case ($\Lambda = I$). Using the coordinate notation of Section 1, the log-likelihood can be written

$$-L(\beta_1, \ldots, \beta_m) = \sum_{j=1}^{n} g\left(\sum_{i=1}^{m} (y_{ij} - x_{ij}\beta_i)^2\right).$$

For coordinates of β_{i_1} and β_{i_2} corresponding to different equations, we have

$$\mathrm{E}\,\frac{\partial^2 L}{\partial\beta_{i_1k_1}\partial\beta_{i_2k_2}} = 4\sum_{j=1}^{n} x_{i_1jk_1}x_{i_2jk_2}\,\mathrm{E}(y_{i_1j} - x_{i_1j}\beta_{i_1})(y_{i_2j} - x_{i_2j}\beta_{i_2})g''(\|u_j\|^2).$$

This equals 0 since the expectation equals 0 conditional on $\|u\|^2$. For coordinates of β_i in the same equation, we have

$$\mathrm{cov}\left(\frac{\partial L}{\partial\beta_{ik_1}}\right)\left(\frac{\partial L}{\partial\beta_{ik_2}}\right) = 4\sum_{j=1}^{n} x_{ijk_1}x_{ijk_2}\,\mathrm{E}(y_{ij} - x_{ij}\beta_j)^2(g'(\|u_j\|^2))^2.$$

This has the appropriate form (3.2). Since each coordinate of u_j has the same marginal distribution, the foregoing expectation is 4 times

$$\mathrm{E}\,u_{1j}^2(g'(\|u_j\|^2))^2 = \frac{1}{m}\mathrm{E}\|u_j\|^2(g'(\|u_j\|^2))^2,$$

and the result in the spherically symmetric case follows taking inverses. For general Λ, simply transform to symmetry by $\Lambda^{-1/2}$.

Theorem A.2. Consider the aforementioned elliptically contoured case and transform the problem so that the succinct form of model (1.1) becomes $\tilde{y} = \tilde{X}\beta + v$, where

$$\tilde{y} = (\Lambda^{-1/2} \therefore I)y,$$
$$\tilde{X} = (\Lambda^{-1/2} \otimes I)X,$$

and

$$v = (\Lambda^{-1/2} \therefore I)u.$$

Assume that Conditions A1 and A2 hold for the transformed problem. Assume, in addition, that the function ψ is antisymmetric. Then the solution to (2.2) with y and X replaced by $\tilde{y}$ and $\tilde{X}$ has asymptotic covariance matrix given by (3.2) with

$$x_{\mathrm{tr}} = \mathrm{E}\,\psi^2(v_{1j})/(\mathrm{E}\,\psi'(v_{1j}))^2. \tag{A.2}$$

Proof. It suffices to compute the matrices Φ and R given in Theorem 2.1 for the spherically symmetric random vector $v \in \mathbf{R}^m$. By spherical symmetry, for $i \neq j$, the coordinates $(-v_i, v_j)$ have the same distribution as (v_i, v_j). Hence, for $i \neq j$,

$$R_{ij} = \mathrm{E}\,\psi(v_i)\psi(v_j) = \mathrm{E}\,\psi(-v_i)\psi(-v_j) = -\mathrm{E}\,\psi(v_i)\psi(v_j).$$

Whence $R_{ij} = 0$. In addition, the coordinates of v have the same marginal distribution. Hence

$$R(v) = (\mathrm{E}\,\psi^2(v_1))I$$

and

$$\Phi(v) = (\mathrm{E}\,\psi'(v_1))I.$$

The result follows immediately from Theorem 2.1.

Proposition A.1. Consider the multivariate t distribution in m dimensions with q df and covariance Λ, scaled so that each coordinate has variance 1 [i.e., the distribution of $\mathbf{N}_m(0, \Lambda)/(\chi^2(q)/(q - 2))^{1/2}$]. Then c^* (A.1) and c_{tr} (A.2) are given by

$$c^* = \frac{(m + q + 2)(q - 2)}{q(m + q)}$$

$$\text{and } c_{\mathrm{tr}} = \frac{\pi(q - 2)\Gamma^2(q/2)}{4\Gamma^2((q + 1)/2)}. \tag{A.3}$$

Proof. First consider the optimal covariance. Let $w = \|v\|^2/(q - 2)$. Then the density of w is

$$f(w) = c(m, q)(1 + w)^{-(m+q)/2},$$

where

$$c(m, q) = \frac{\Gamma((m + q)/2)}{\Gamma(m/2)\Gamma(q/2)}$$

and c^* will be $(q - 2)$ times the value computed using this density. So the logarithmic derivative becomes

$$g'(w) = \frac{(m + q)}{2}\,\frac{1}{(1 + w)}.$$

Therefore, from (A.1),

$$\frac{1}{c^*(w)} = \frac{4}{m}\,\frac{(m + q)^2}{4}\int_0^\infty \frac{v}{(1 + v)^2}\,c(m, q)\,\frac{v^{(m-1)/2}}{(1 + v)^{(q+m)/2}}\,dv$$
$$= \frac{(m + q)^2}{m}\,\frac{c(m, q)}{c(m + 2, q + 2)} = \frac{(m + q)q}{(m + q + 2)},$$

from which (A.3) follows for c^*.

The result for c_{tr} follows easily from (A.2) and the calculations E $\psi^2(v) = 1$ and E $\psi'(v) = 2f_v(0)$, where f_v is just the density of a univariate t_q distribution times $(q - 2)/q$.

Last, we calculate $\tilde{V}$ (2.6) in a special case of a (scaled) bivariate t distribution. In particular, let $u_j \in \mathbf{R}^m$ be $((q - 2)/q)^{1/2}$ times an observation from a bivariate t distribution with q df and covariance matrix Λ given by

$$\Lambda = \begin{bmatrix} 1 & 2\rho \\ 2\rho & 4 \end{bmatrix} \qquad \text{(A.4)}$$

for $|\rho| \leq 1$.

Proposition A.2. Under the foregoing scaled multivariate t distribution, the asymptotic covariance (A.2) of the weighted l_1 estimator applied to the untransformed data is given by (3.3).

Proof. We only need to compute $R(u)$ and $\Phi(u)$ as given in Condition P2 for $\psi(u) = \mathrm{sgn}(u)$. Clearly, the diagonal entries of $R(u)$ are unity, and the off-diagonal entry is

$$R_{12}(u) = \mathrm{E}\ \mathrm{sgn}(u_1)\mathrm{sgn}(u_2) = \mathrm{E}\ \mathrm{sgn}(z_1)\mathrm{sgn}(z_2) = \frac{2 \sin^{-1}\rho}{\pi}$$

from (3.3), where formula 26.3.19 from Abramowitz and Stegun (1964) has been applied. In addition, since the marginal distribution of u_{1j} is the same t distribution as the marginal for v_{1j} above, and $u_{2j} \sim 2v_{2j}$, we have

$$\Phi(u) = c_{\mathrm{tr}}^{-1/2} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

where c_{tr} is exactly the same as in the expression for V_{tr}. Therefore, $\tilde{V}$ has the desired form with $\Delta(u) = \Phi^{-1}(u)R(u)\Phi^{-1}(u)$, from which (3.3) follows by direct calculation.

[*Received November 1988. Revised April 1990.*]

REFERENCES

Abramowitz, M., and Stegun, I. (1964), *Handbook of Mathematical Functions,* Washington, DC: National Bureau of Standards.

Amemiya, T. (1982), "Two-Stage Least Absolute Deviations Estimators," *Econometrica,* 50, 689–712.

Bickel, P. J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association,* 70, 428–433.

Blomqvist, N. (1950), "On a Measure of Dependence Between Two Random Variables," *Annals of Mathematical Statistics,* 21, 593–600.

Boot, J. C. G., and De Wit, G. M. (1960), "Investment Demand: An Empirical Contribution to the Aggregation Problem," *International Economic Review,* 1, 3–30.

Devlin, S. J., Gnanadesikan, R., and Kettering, J. R. (1975), "Robust Estimation and Outlier Detection With Correlation Coefficients," *Biometrika,* 62, 531–545.

Fox, P. A. (1984), The Port Mathematical Subroutine Library, Murray Hill, NJ: Bell Laboratories.

Grunfeld, Y. (1958), "The Determinants of Corporate Investment," unpublished Ph.D. thesis, University of Chicago, Dept. of Economics.

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

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

Ibragimov, I. A., and Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory,* (translated by S. Kotz), New York: Springer-Verlag.

Jurečková, Jana (1977), "Asymptotic Relations of *M*-Estimates and *R*-Estimates in Linear Regression Model," *The Annals of Statistics,* 5, 464–472.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica,* 46, 33–50.

Kuester, K. (1987), "Asymptotic Consistency and Normality of Least Absolute Deviations Applied to Seemingly Unrelated Regression Systems," technical report, Board of Governors of the Federal Reserve System.

Rousseeuw, P. J. (1987), "Identification of Multivariate Outliers and Leverage Points by Means of Robust Covariance Matrices," Report 87-15, Delft University of Technology, Faculty of Mathematics and Informatics.

Oja, H. (1983), "Descriptive Statistics for Multivariate Distributions," *Statistics and Probability Letters,* 1, 327–333.

Portnoy, S. (1977), "Robust Estimation in Dependent Situations," *The Annals of Statistics,* 22–43.

——— (1979), "Further Remarks on Robust Estimation in Dependent Situations," *The Annals of Statistics,* 7, 224–231.

Ruppert, D., and Carroll, R. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828–838.

Srivastava, U. K., and Giles, D. E. A. (1987), *Seemingly Unrelated Regression Equations Models: Estimation and Inference,* New York: Marcel Dekker.

Theil, H. (1971), *Principles of Econometrics,* New York: John Wiley.

Welsh, A. H. (1987), "Kernel Estimates of the Sparsity Function," in *Statistical Analysis Based on the L1 Norm,* ed. Y. Dodge, New York: North-Holland.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias, *Journal of the American Statistical Association,* 57, 348–368.

——— (1963), "Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results," *Journal of the American Statistical Association,* 58, 977–992.