FACULTY WORKING PAPER 91-0106

Momentary Lapses: Moment Expansions and the Robustness of Minimum Distance Estimation

Roger Koenker
Department of Economics

José A.F. Machado
Universidade Nova de Lisboa

Alan H. Welsh
Australian National University

Bureau of Economic and Business Research
College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 91-0106

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

January 1991

Momentary Lapses: Moment Expansions and the Robustness of Minimum Distance Estimation

Roger Koenker
Department of Economics
University of Illinois at Urbana-Champaign

José A.F. Machado
Universidade Nova de Lisboa

Alan H. Welsh
Australian National University

The authors wish to thank Joseph Altonji for stimulating their initial inquiry into this subject; he is, of course, absolved of any responsibility for the consequences. The research was initiated while Roger Koenker was on sabbatical leave visiting Universidade Nova de Lisboa and Australian National University, and complete while José Machado was on leave at University of Illinois, Urbana-Champaign. Roger Koenker's research was partially supported by NSF Grant SES-89-22472, and by research stipends at UNL and ANU.

The influential recent literature on generalized method of moments (GMM) minimum distance estimation (MDE) has found widespread econometric application.[1] In exponential family models where sample moments may be interpreted as sufficient statistics these methods are especially attractive, as for example under Gaussian error conditions. But, as we shall see, sample moments are quintessentially non-robust; slight departures from Gaussian conditions can provoke the complete collapse of classical GMM methods based on least-squares principles. In this paper we explore some simple modifications of minimum distance methods designed to insure against such statistically inclement, non-Gaussian weather. Our evaluation of estimator performance employs a second-order expansion of the asymptotic variance of GMM-type estimators enabling us to study the effect of estimating the covariance matrix of the initial estimator as well as the effect of the initial estimator itself. In this respect our approach is closely allied with recent work by Rothenberg (1984), Carroll, Wu and Ruppert (1988), and Koenker, Skeels and Welsh (1990) on generalized least squares estimation.

After a brief, somewhat polemical, introduction intended to motivate an inquiry into robustness of moment based methods, we describe in Section 2, a rather simple, stylized MDE setting and explore the performance of conventional least-squares-based preliminary estimation of the model. Our second-order variance expansions reveal some surprising consequences of Eicker-White estimation of the covariance matrix of the least-squares preliminary estimates. In Section 3 we turn to robust alternatives to least-squares preliminary estimation. Here the second-order variance expansion is somewhat more arduous, but repays the effort yielding interesting qualitative conclusions. Some further "stylization" of the design assumptions permits us to explore quantitatively the the interplay between model dimension, error assumptions, and the degree of robustness of the preliminary estimation method. A final section draws together some conclusions and suggests directions for subsequent research.

An exhaustive catalog of recent work on this subject is unrealistic here, but we should mention Hansen (1982), Gallant (1987), Chamberlain (1982, 1987), Manski (1988), Newey (1985, 1988) on theoretical aspects, and work by Hansen and Singleton (1983), Tauchen (1986) and others on applications. Our own interest in this topic was stimulated by Altonji, Martins, and Siow (1987).

1. Introduction

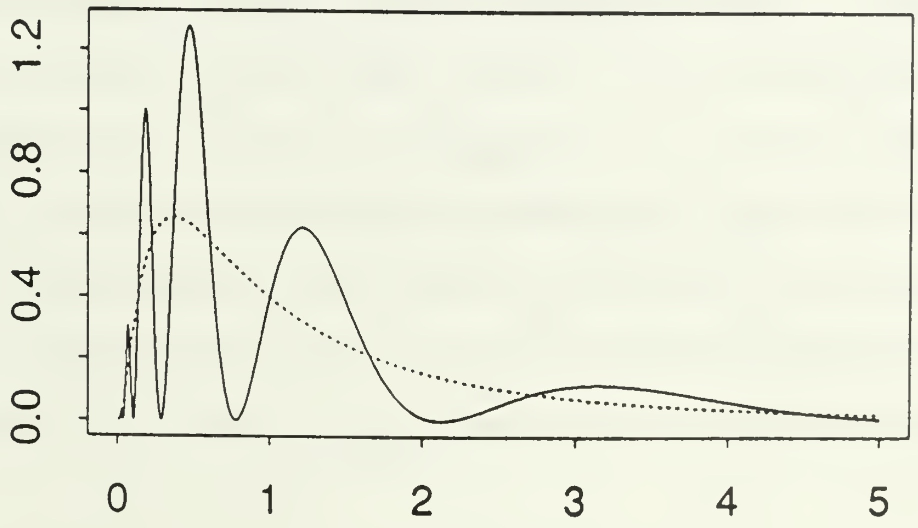
1.1. What's wrong with moments?

The question may appear heretical. It is an article of Gaussian faith that sample information inheres in sample moments. However, outside Gaussian theology this is not the case. Despite the sufficiency of sample moments in exponential family models like the Gaussian, slight departures from the exponential family play havoc with the sufficiency paradigm. Falling from their exalted status as "source of all knowledge", they become the tainted "fruit of forbidden tree."

1.2. Moments are difficult to interpret

The classical moment problem, e.g., Feller (1971) or Billingsley (1979, §30) asks: when is a distribution, μ, uniquely characterized by its moments? One answer is simply: If μ has a moment generating function which exists in a neighborhood of zero. In Figure 1.1 we illustrate two densities: one is the familiar lognormal density, the other bears little resemblance to the first, *but has the same sequence of integer moments.* Many such examples exist, even when, as here, the moment sequence is finite. When moments may be infinite the situation is much worse.

Figure 1.1. Two Densities with the same Moment Sequence



Take any density f and consider the Cauchy contaminated density $f_\epsilon(x) = (1-\epsilon)f(x) + \epsilon[\pi(1+x^2)]^{-1}$. *For any* $\epsilon > 0$, such a density has no moments of any order so the population moments convey no information about the form of the original f. It is tempting to dismiss such examples as "pathological", attributing them to the fiendishly long tail of the Cauchy density. "Infinite moments are impossible in the real world" - one sometimes is told. Fair enough. But one need only replace the Cauchy with something similar in shape, truncated to a bounded support, or even a normal density with large variance, to see that ϵ-contamination can exert a grotesque influence on the moments of f, particularly the higher order moments. In Figure 1.2-4 we illustrate three examples of (f, f_ϵ) pairs with f standard normal, standard lognormal and a bimodal example. In all three cases $\epsilon = .05$ and it is difficult to distinguish the contaminated density represented by the dotted line from the parent density represented by the solid line, however all three contaminated densities have indistinguishable moment sequences.

Figure 1.2.

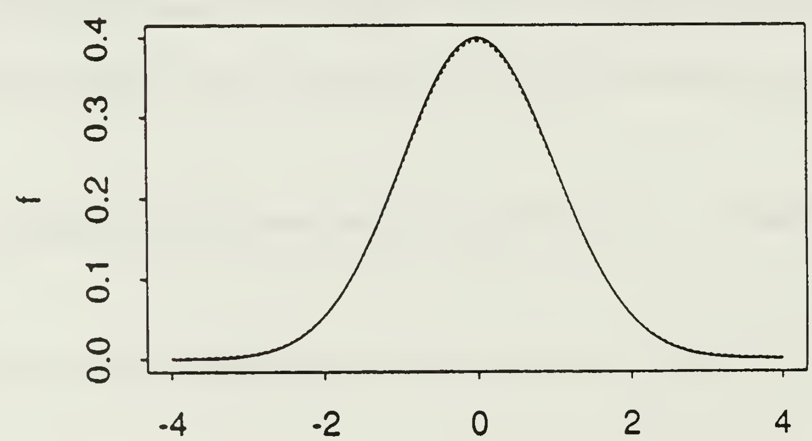


Figure 1.3.

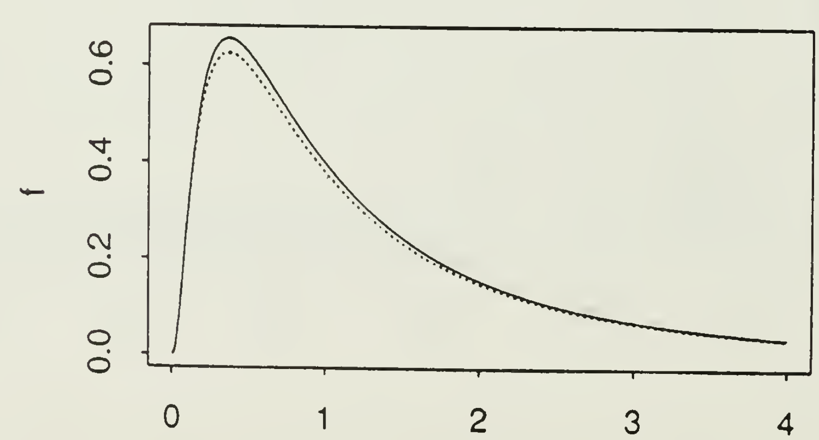
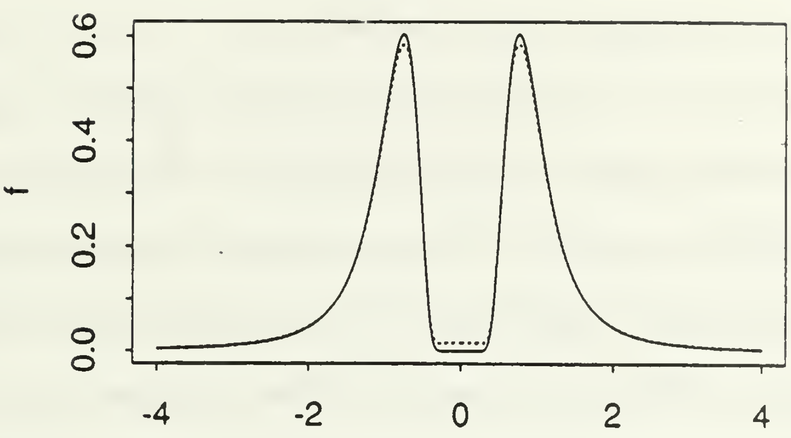


Figure 1.4.



We are tempted to regard first moments as location parameters, and higher moments about the mean as dispersion, skewness, kurtosis, etc., but all of these concepts have considerable ambiguity in any general setting where arbitrarily small amounts of contamination by long-tailed distributions can distort these familiar quantities radically. Bickel and Lehmann (1975) address this problem in elegant generality and conclude that even in the case of location, the mean is a poor choice to characterize this fundamental descriptive aspect of distributions. Higher moments are similarly condemned.

1.3. Moments are difficult to estimate

If ϵ-contamination can wildly distort the population moments of an uncontaminated density, f, making population moments difficult to interpret, such contamination is ruinous to estimation based on sample moments. Suppose, for example, that we have a random sample from a density we believe is normal--mean θ, variance 1. We wish to estimate θ, and confidently, if perhaps naively, compute $\bar{X}_n = n^{-1}\sum X_i$. If we are correct, we have behaved "optimally" and $\bar{X}_n \sim N(\theta, 1/n)$; our estimator is consistent, fully efficient, "true-BLUE." But what if our ϵ Cauchy contamination comes slithering back and the X_i have density $f_\epsilon(x - \theta)$? Then $\bar{X} - \theta$ is a mixture of normal and Cauchy random variables and since the normal component degenerates at θ, $\bar{X}_n$ is asymptotically Cauchy, hence is inconsistent, has unbounded mean-square error, zero efficiency, in short is an unmitigated disaster. These same remarks apply for *any* fixed contamination proportions, ϵ, *however small.*

If we had instead chosen to estimate θ by the foolishly inefficient sample median $\hat{\theta} = X_{(n/2)}$, it is well-known that $\sqrt{n}\,(\hat{\theta} - \theta) \rightsquigarrow N(0, \omega^2)$ where $\omega^2 = [2f_\epsilon(F_\epsilon^{-1}(0)]^{-2}$ which in the present case is about 1.60, if we assume $\epsilon = .05$ and that the Cauchy contamination is centered at θ. For comparison, an asymptotically optimal estimator of θ, like the MLE, can approach the Cramer-Rao bound of 1.0538. So the median sacrifices considerable efficiency to the MLE, but nevertheless it clearly represents a huge improvement over the unbounded mean square error of the sample mean in this setting.

The extreme sensitivity of the performance of the sample mean to modest departures from strictly Gaussian conditions is the first lesson of robust statistics. Tukey's (1960) summary of extensive research during the late 40's and 50's is the seminal treatment. See Huber (1981), Hampel, et.al. (1986) and Rousseeuw and Leroy (1988) for recent treatments.

If the mean is difficult to estimate in contamination models, higher moments pose even greater challenges. The widely cited debate between R.A. Fisher and the physicist A.S. Eddington illustrates this point vividly. Eddington advocated the use of the mean absolute deviation $d_n = n^{-1}\sum|x_i - \bar{x}|$ as a measure of dispersion, while Fisher argued that $s_n = (n^{-1}\sum(x_i - \bar{x})^2)^{1/2}$ was 12% more efficient than d_n at the normal model. Huber (1981) computes the asymptotic relative efficiency of d_n to s_n at the contaminated normal model $F_\epsilon(x) = (1-\epsilon)\Phi(x) + \epsilon\Phi(x/3)$. He finds that $\epsilon \geq .002$ contamination is enough to reverse Fisher's efficiency claim for the standard deviation. At $\epsilon = .05$, d_n is roughly twice as efficient as s_n! Even d_n is not robust in the formal sense of Hampel (1968) and most robustniks would prefer the median absolute deviation from the median as a measure of dispersion. For symmetric F, this so-called MAD estimator estimates the interquartile range and has the virtue of being entirely insensitive to the tail behavior of the error density. See Welsh and Morrison (1990) for a recent discussion of robust methods for estimating dispersion.

What relevance does the foregoing discussion have to the performance of moment-based minimum distance methods? In practice these methods typically involve minimizing a quadratic form, $(m - \mu(\alpha))'\hat{V}^{-1}(m - \mu(\alpha))$ in sample moments, m, where V is an estimate of the covariance matrix of m. Thus V typically consists of higher order sample moments as we climb the "misty staircase" of Mosteller and Tukey (1977) on which less and less reliable higher moments are employed to assess the higher reliability of lower moments.

A critical practical question about GMM-MDE methods is the effect of estimating the covariance matrix V on the performance of the estimator $\hat{\alpha}_n$. Unfortunately, the familiar first-order asymptotics of $\hat{\alpha}_n$ are silent on this effect and we must turn to higher-order expansions.

2. A moment expansion for a least-squares-based minimum-distance estimator

In this section we wish to explore a simple, stylized version of minimum-distance estimation based on least-squares principles. We consider the extremely simple, yet representative model

$$Y_i = x_i'g(\alpha_0) + u_i \qquad i = 1, \dots, n \tag{2.1}$$

where $\{x_i\}$ is a sequence of known q-vectors, g is a smooth, known function from $\mathbf{R}^q$ to $\mathbf{R}^p$, $p << q$, and $\{u_i\}$ is a sequence of independent random variables. Estimation of the model (2.1) proceeds naturally in two steps. We begin with an estimator $\hat{\beta}_n$ of $g(\alpha_0)$ and solve either:

$$\min_a (\hat{\beta}_n - g(a))'\hat{V}_n^{-1}(\hat{\beta}_n - g(a)) \tag{2.2}$$

or, equivalently,

$$0 = -2\nabla g(a)'\hat{V}_n^{-1}(\hat{\beta}_n - g(a)) \tag{2.3}$$

where the $q \times q$ matrix $\hat{V}_n^{-1}$ is typically chosen as an estimate of the asymptotic covariance matrix of $\hat{\beta}_n$.

Our objective in this section is to study the asymptotic performance of the estimator $\hat{\alpha}_n$ solving (2.2) or (2.3) when $\hat{\beta}_n$ is the (ordinary) least-squares estimator and $\hat{V}_n$ has the (Eicker-White) form

$$\hat{V}_n = n(X'X)^{-1}\sum_{i=1}^{n} x_i x_i' \hat{u}_i^2 (X'X)^{-1}. \tag{2.4}$$

It is well known that the limiting distribution of $\sqrt{n}\,(\hat{\alpha}_n - \alpha_0)$ does not depend upon the specific form of $\hat{V}_n$, but only upon its limiting value. The silence of first-order asymptotics on this important issue has led us to second-order expansions of $V(\sqrt{n}\,(\hat{\alpha}_n - \alpha_0))$ incorporating a term of $O(n^{-1})$ which sheds some light on the effect of $\hat{V}_n$.

To further simplify matters we will assume that the function g takes the linear form

$$\beta = g(\alpha) = G\alpha \tag{2.5}$$

for some $q \times p$ matrix G of rank p. Nonlinear g could be easily incorporated into our framework, but specific assumptions about the curvature of g would then be required to interpret the resulting expansions. With g linear and $\hat{\beta}_n$ linear as well, an expansion for $\sqrt{n}\,(\hat{\alpha}_n - \alpha_0)$ is relatively direct. Solving (2.3) we have

$$\hat{\alpha}_n - \alpha_0 = (G'\hat{V}_n^{-1}G)^{-1}G'\hat{V}_n^{-1}(X'X)^{-1}X'u \tag{2.6}$$

which we would like to expand asymptotically. Before attacking this directly we describe a simple scalar example which illustrates the strategy of the expansion.

2.1. An Example

To illustrate the methods employed, we begin with an elementary problem where the scalar computations are sufficiently transparent to illuminate our general moment expansion strategy. Consider the problem of refining the usual "δ-method" scheme for computing the asymptotic variance of a nonlinear function of the sample mean. More explicitly let $h(x)$ be a smooth function $\mathbf{R} \rightarrow \mathbf{R}$ admitting the cubic expansion,

$$h(x) = h(\mu) + h'(\mu)(x-\mu) + \frac{1}{2}h''(\mu)(x-\mu)^2 + \frac{1}{6}h'''(\mu)(x-\mu)^3 + O(|x-\mu|^4)$$

For i.i.d. observations $X_1, \ldots, X_n$ with finite 4th moment and $\bar{X}_n = n^{-1}\sum X_i$,

$$Eh(\bar{X}_n) = h(\mu) + \frac{1}{2}h''(\mu)\sigma^2/n + \frac{1}{6}h'''(\mu)\mu_3/n + r_n$$

hence we may write

$$\Delta_n \equiv \sqrt{n}\,(h(\bar{X}) - Eh(\bar{X})) = Z_{1n} + \frac{1}{\sqrt{n}}Z_{2n} + \frac{1}{n}Z_{3n} + o_p(n^{-1})$$

where

$$Z_{1n} = \sqrt{n}\,(\bar{X} - \mu)h'(\mu)$$

$$Z_{2n} = n[(\bar{X} - \mu)^2 - \sigma^2/n](h''(\mu)/2)$$

$$Z_{3n} = n^{3/2}[(\bar{X} - \mu)^3 - \mu_3/n](h'''(\mu)/6)$$

are all terms of $O_p(1)$. The asymptotic variance of $h(\bar{X}_n)$ may be expressed as,

$$V(\Delta_n) = V(\sqrt{n}h(\bar{X}_n)) = V(Z_{1n}) + \frac{1}{n}[V(Z_{2n}) + 2(n^{1/2}Cov(Z_{1n}, Z_{2n}) + Cov(Z_{1n}, Z_{3n}))] + o(n^{-1}).$$

The leading term is the familiar δ-method asymptotic variance

$$V(Z_{1n}) = (h'(\mu))^2 nV(\bar{X}_n - \mu) = (h'(\mu))^2\sigma^2,$$

while the components of the $1/n$ term are

$$V(Z_{2n}) = \frac{1}{4}(h''(\mu))^2 n^2 E[(\bar{X}_n - \mu)^4 - \sigma^4/n^2]$$

$$= \frac{1}{4}(h''(\mu))^2 n^2[\frac{\mu_4}{n^3} + \frac{3(n-1)\sigma^4}{n^3} - \frac{\sigma^4}{n^2}]$$

$$= \frac{1}{2}(h''(\mu))^2\sigma^4 + o(1)$$

$$Cov(Z_{1n}, Z_{2n}) = \frac{1}{2}h'(\mu)h''(\mu)n^{3/2}E(\bar{X}_n - \mu)^3$$

$$= \frac{1}{2} h'(\mu) h''(\mu) \frac{1}{\sqrt{n}} \mu_3$$

and

$$Cov(Z_{1n}, Z_{3n}) = \frac{1}{6} h'(\mu) h'''(\mu) n^2 E(\overline{X}_n - \mu)^4$$

$$= \frac{1}{2} h'(\mu) h''(\mu) \sigma^4.$$

The last line follows from the fact that $E(\overline{X} - \mu)^4 = 3\sigma^4/n^2 + o(n^{-2})$, as in $V(Z_{2n})$ above. So we have finally

$$V(\Delta_n) = (h'(\mu))^2 \sigma^2 + n^{-1} \sigma^4 [\frac{1}{2}(h''(\mu))^2 + h'(\mu)(h''(\mu)\mu_3/\sigma^4 + h'''(\mu))] + o(n^{-1}).$$

The crux of this example, and its relevance for our expansions, is the cubic expansion of $h(\cdot)$ required to compute the $O(n^{-1})$ "correction" term for the expansion of $V(\Delta_n)$. The facts that $Cov(Z_{1n}, Z_{2n}) = O(n^{-1/2})$ and $Cov(Z_{1n}, Z_{3n}) = O(1)$ are perhaps not immediately obvious. Indeed the latter term is inexplicably missing from the treatment of this example in Bickel and Doksum (1977). One may well ask what about $Cov(Z_{1n}, Z_{4n})$? However, note that this term may be relegated to the remainder since, if both Z_{1n} and Z_{4n} are $O_p(1)$, their covariance is, by Cauchy-Schwartz, $O_p(1)$ and Z_{4n} appears in the expansion multiplied by $n^{-3/2}$.

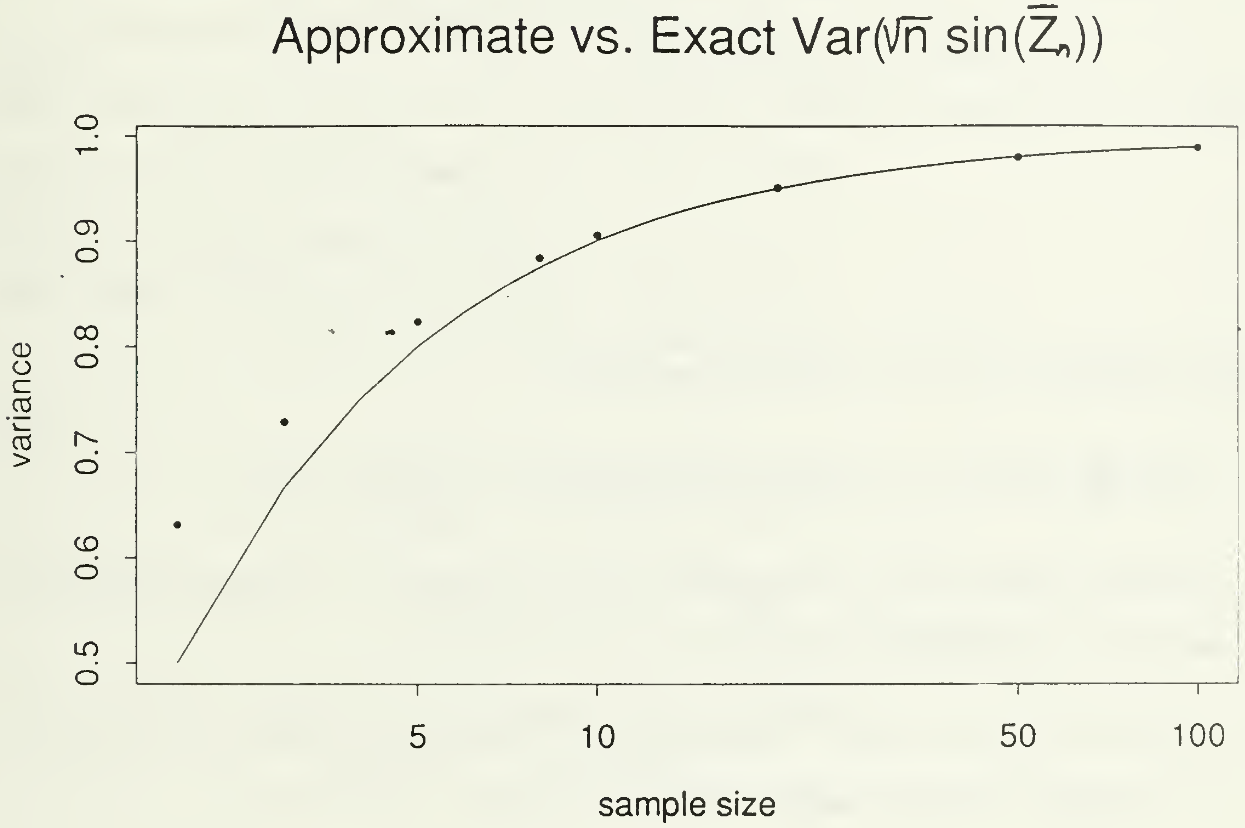
A simple application of this variance expansion for functions of the sample mean, and one that permits us to evaluate the quality of the $1/n$ correction term, is the following example. Suppose $\overline{Z}_n$ is the sample mean of n i.i.d. standard normal random variables, so $\overline{Z}_n \sim> N(0, n^{-1})$ and consider $Y_n = \sqrt{n}\sin(\overline{Z}_n)$. Our variance expansion for this case is simply,

$$V(Y_n) = 1 - \frac{1}{n} + o(n^{-1}).$$

The exact density of Y_n is "easily" computed to be

$$f_n(y) = \sum_{k=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{\frac{-(k\pi - \sin^{-1}(y/\sqrt{n}))^2}{2/n}\}/\sqrt{1 - y^2/n} \ ,$$

where only the 3 central terms $k = 1, 0, 1$ are needed to capture *virtually* all of the standard normal density in the interval $[-3\pi/2, 3\pi/2]$. Obviously, as $n \to \infty$, $f_n(y)$ tends to the standard normal density.



To evaluate the performance of the $O(n^{-1})$ variance expansion we plot this approximation versus the true variance as computed numerically, in *Mathematica*, Wolfram (1988), directly from $f_n(y)$. It is striking how well the simple two term variance expansion performs even at what might be regarded as rather small n. At $n = 5$, for example, the approximation gives $V(Y_n) \simeq .8$ while the exact computation yields $V(Y_n) \simeq .8242$. Obviously, the expansion is absurdly optimistic at $n = 1$, predicting a variance of zero when the correct value is about .5. A similar breakdown will be seen to occur in the more complicated settings considered below.

2.2. The expansion in the independent case

As the previous example suggests we require an asymptotic expansion of $\sqrt{n}\,(\hat{\alpha}_n - \alpha_0)$ of the form

$$\sqrt{n}\,(\hat{\alpha}_n\ \alpha_0) = A_{0n} + n^{-1/2}A_{1n} + n^{-1}A_{2n} + o_p(n^{-1}) \tag{2.7}$$

with $A_{in} = O_p(1)$ for $i = 0, 1, 2$. Starting from (2.6) and using the fact that

$$\hat{V}^{-1} - V^{-1} = \hat{V}^{-1}(V - \hat{V})V^{-1} \tag{2.8}$$

$$= V^{-1}(V - \hat{V})V^{-1} + (\hat{V}^{-1} - V^{-1})(\hat{V} - V)V^{-1}$$

$$= -V^{-1}(\hat{V} - V)V + V^{-1}(\hat{V} - V)V^{-1}(\hat{V} - V)V^{-1} + O_p(\|\hat{V} - V\|^3).$$

we may establish (2.7) under the following conditions

D. The design sequence $\{X_n\}$ satisfies the condition that $n^{-1}\sum_{i=1}^{n}\|x_i\|^{12} = O(1)$.

F. The error sequence $\{u_i\}$ is independent, each component has a symmetric distribution about zero, and finite *sixth* moment.

G. The matrix G has rank p, as has $Q_n = (G'H_nJ_n^{-1}H_nG)$ where $H_n = n^{-1}\sum x_ix_i'$, $J_n = n^{-1}\sum\sigma_i^2x_ix_i'$, and $\sigma_i^2 = Eu_i^2$.

Theorem 2.1: Under conditions D, F, G, the expansion (2.7) holds with,

$$A_{0n} = Q_n^{-1}G'H_nJ_n^{-1}n^{-1/2}\sum x_iu_i$$

$$A_{1n} = -Q_n^{-1}G'H_nJ_n^{-1}[n^{-1}\sum_i\sum_j x_ix_i'K_nx_j(u_i^2 - \sigma_i^2)u_j]$$

$$A_{2n} = Q_n^{-1}G'H_nJ_n^{-1}n^{-3/2}[2\sum_i\sum_j\sum_k x_jx_j'x_j'H_n^{-1}x_iK_nx_ku_iu_ju_k$$

$$- n^{-1}\sum_i\sum_j\sum_k\sum_l x_lx_l'H_n^{-1}x_ix_j'H_n^{-1}x_lx_l'K_nx_ku_iu_ju_k$$

$$+ \sum_i \sum_j \sum_k x_i x_i' K_n x_j x_j' K_n x_k (u_i^2 - \sigma_i^2)(u_j^2 - \sigma_j^2) u_k]$$

and K_n takes the form

$$K_n = J_n^{-1} - J_n^{-1} H_n G Q_n^{-1} G' H_n J_n^{-1}.$$

Proof: (See Appendix A)

A more general expansion like (2.7) but based on an initial M-estimator and corresponding $\hat{V}_n$ matrix will be developed in the next section. Computing moments as in the scalar example we have the following result.

Theorem 2.2: Under conditions D, F, G,

$$V(\sqrt{n}(\hat{\alpha}_n - \alpha_0)) = Q_n^{-1} + n^{-1} Q_n^{-1} G' R_n G Q_n^{-1} + o(n^{-1})$$

where

$$R_n = n^{-1} \sum \sigma_i^2 (1 - k_i) H_n J_n^{-1} x_i x_i' K_n x_i x_i' J_n^{-1} H_n$$

$$+ 4n^{-1} \sum H_n J_n^{-1} x_i x_i' K_n (2\sigma_i^2 I - J_n H_n^{-1}) x_i x_i'$$

and $k_i = Eu_i^4/\sigma_i^4$. When the errors $\{u_i\}$ are identically distributed

$$V(\sqrt{n}(\hat{\alpha}_n - \alpha_0)) = \sigma^2 \Omega_n + n^{-1} \sigma^2 (5 - k) \Omega_{1n} + o(n^{-1})$$

with $\Omega_n = (G' H_n G)^{-1}$ and $\Omega_{1n} = \Omega_n G' n^{-1} \sum x_i x_i' [H_n^{-1} - G(G' H_n G)^{-1} G'] x_i x_i' G \Omega_n$.

Proof: See Appendix A.

The form of the leading $O(1)$ term in this variance expansion is quite familiar. The form of the second $O(1/n)$ term is perhaps less so. Consider the iid error context. The matrix Ω_{1n} is clearly positive definite. Thus for $k(F) < 5$ the effect of using $\hat{V}_n$ in lieu of the appropriate and simpler H_n, is to *increase* the variance of $\hat{\alpha}_n$, up to the order of the expansion. Recalling that in the Gaussian case $k(F) = 3$, this finding provides a quantitative assessment of the efficiency loss due to $\hat{V}_n$. See Rothenberg (1984) for a discussion, from a sufficiency standpoint of a similar result for generalized least squares (GLS) estimation under Gaussian conditions, and see Carroll, Wu and Ruppert (1988) and Koenker, Skeels and Welsh (1990) for closely related expansions again in the (GLS) context.

Most curious is the effect of the $O(1/n)$ term when $k(F) > 5$, that is in long tailed circumstances. Here the expansion seems to suggest that estimating $\hat{V}_n$ is advantageous, actually decreasing the variance of $\hat{\alpha}_n$ below what it would be if the true V_n were used. (Had we known that the errors were iid and used H_n instead of $\hat{V}_n$, the expansion (2.7) simplifies to just the leading term and the $O(1/n)$ term vanishes entirely in the variance expansion.) This may seem paradoxical at first. How can ignorance of the true V_n lead us to an improved estimator of α_0? The answer lies in the robustifying effect of $\hat{V}_n$: Since $\sqrt{n}\,(\hat{\beta}_n - \beta_0) = O_p(1)$, estimating $\hat{\alpha}_n$ by (2.6), thereby downweighting observations with big residuals from the initial least-squares fit, has the effect of downweighting observation with large error realizations. And this is exactly what robust M-estimation would accomplish under long-tailed error conditions. The effect is very much like the effect of a one-step (robust) M-estimator. Of course, since the $O(1/n)$ variance term is linear in $k(F)$ we have the ultimately absurd conclusion that, to the order of the variance expansion, one can drive the variance of $\sqrt{n}\,\hat{\alpha}_n$ to zero and beyond by increasing the kurtosis of the $\{u_i\}$. The same effect may be observed in the $O(1/n)$ variance expansion for GLS estimators in Carroll, Wu and Ruppert (1988) and Koenker, Skeels and Welsh (1990).

To explore the design contribution to the $O(n^{-1})$ term of the variance expansion we rewrite the expansion in the iid error case as

$$V(\sqrt{n}\,(\hat{\alpha}_n - \alpha_0)) = \sigma^2\Omega_n[I + n^{-1}(5-k)\Delta_n] + o(n^{-1}), \tag{2.9}$$

where $\Delta_n = G'\tilde{M}_n G'\Omega_n$ so

$$n^{-1}\Delta_n = G'[\textstyle\sum x_i x_i'(X'X)^{-1}x_i x_i']G'(G'X'XG)^{-1}$$

$$+ G'[\textstyle\sum x_i x_i' G(G'X'XG)^{-1}G'x_i x_i]G'(G'X'XG)^{-1}$$

Set $Z = XG$, so $z_i' = x_i'G$, and taking the trace, we have,

$$\text{tr}\,(n^{-1}\Delta_n) = \sum_{i=1}^{n} P_{ii}(X)P_{ii}(Z) - \sum_{i=1}^{n} P_{ii}^2(Z)$$

where $P_{ii}(Z) = z_i'(Z'Z)^{-1}z_i$, the diagonal elements of the "hat" matrix of Z.

Following Box and Watson (1962) we will make some (heroic) simplifying assumptions in the effort to facilitate the interpretation of the design effect. In particular, we show in Appendix D that when the design elements $\{x_{ij}\}$ are iid

$$\sum P_{ii}(X)P_{ii}(Z) = \frac{p}{n}[k(x) + q - 1]$$

where $k(x)$ is the kurtosis of the design elements. Similarly, we show that

$$\sum (z_i'(Z'Z)^{-1}z_i)^2 = p(p+2) + (k(x) - 3)\left[\sum_{k=1}^{p} \frac{S_4^k}{(S_2^k)^2} + \sum_k \sum_{l \neq k} \frac{S_2^{kl}}{S_2^k S_2^l}\right] + 2\sum_k \sum_{l \neq k} \frac{(S_1^{kl})^2}{S_2^k S_2^l}$$

where, in the notation of Box and Watson (1962), $S_m^{kl} = \sum_{j=1}^{q} g_{kj}^m g_{lj}^m$ and $S_m^k = \sum_{j=1}^{q} g_{kj}^m$. Finally, noting that if the $\{x_{ij}\}$ are Gaussian so $k(x) = 3$, and if the columns of G are orthogonal the last two terms of the previous expression vanish and we have

$$\text{tr}\,(\Delta_n) = n^{-1}[3p + p(q-1) - p(p+2)] = \frac{p}{n}(q-p).$$

This dramatically simplified expression sheds some light on the choice of parametric dimension of the original model, and the quality of the usual first-order asymptotic approximation for the variance of $\hat{\alpha}_n$. Even under the highly favorable design conditions imposed above, we must have $2p(q-p)/n << 1$ under ideal Gaussian error conditions if the usual first order asymptotics are to be reliable. It is interesting to compare this result to those of Huber (1973) who (Proposition 2.2) that a necessary and sufficient condition for any linear function of the least-squares estimator to be asymptotically normal (with natural parameters) is that $\max_i P_{ii}(X) \to 0$ as $n \to \infty$. This condition implies that $p/n \to 0$ as $n \to \infty$. In our least-squares minimum distance framework the analogous condition is $(q-p)/n \to 0$. Note that p in (2.9) may be factored out because we are summing over the asymptotic variances of p

parameters of vector $\hat{\alpha}_n$ in taking the trace. We will return to this problem of "large-p, q asymptotics" in Section 3, when we consider M-estimators.

When $k(x) > 3$ the situation is more difficult to characterize; now the kurtosis of g_{ij}'s is also relevant. Roughly speaking, high kurtosis in $\{x_{ij}\}$ and $\{g_{ij}\}$ amplify the $O(1/n)$ effect relative to Gaussian design conditions -- acting, in effect, like an increase in the parametric dimension of the model.

3. A moment expansion for M-MDE estimators

The estimator $\hat{\alpha}_n$, and the asymptotic expansions for $\hat{\alpha}_n$, of the previous section break-down completely in the face of long-tailed errors. Despite the salutory robustifying effect of the covariance matrix estimate $\hat{V}_n$, when moments are finite this effect is eventually dominated by the (inflated) effect of the leading term of the variance expansion. When our stringent $Eu_i^6 < \infty$ condition (F) fails, the outlook is bleak indeed. In this section we consider robustified versions of the MDE procedure studied above, in which the initial least-squares estimator $\hat{\beta}_n$ is replaced by an M-estimator $\tilde{\beta}_n$ defined as the solution to either

$$\min_{b \in \mathbb{R}^q} \sum_{i=1}^{n} \rho(y_i - x_i b) \tag{3.1}$$

or

$$0 = \sum_{i=1}^{n} \psi(y_i - x_i b) x_i \tag{3.2}$$

where $\psi = \rho'$. When the $\{u_i\}$ are iid with a known, common density f, it is natural to choose $\rho = \log f$. When f is unknown, one would like to choose (ρ, ψ) to achieve high efficiency over a broad class of error distributions. The *sine qua non* for this objective is the boundedness of ψ. The fatal flaw of the least-squares choice $\psi(u) = u$ is its unboundedness. See Huber (1981) and Hampel, et.al. (1986) for an elaboration of these ideas.

When the errors $\{u_i\}$ are independent but not identically distributed the analogue of $\hat{V}_n$, in (2.4) is

$$\tilde{V}_n \equiv \tilde{H}^{-1}\tilde{J}_n\tilde{H}_n^{-1} = (\sum\psi'(\tilde{u}_i)x_ix_i')^{-1}(\sum\psi^2(\tilde{u}_i)x_ix_i')(\sum\psi'(\tilde{u}_i)x_ix_i')^{-1} \tag{3.3}$$

with $\tilde{u}_i = y_i - x_i\tilde{\beta}_n$ which converges to $V_n = H_n^{-1}J_nH_n^{-1}$ where $H_n = n^{-1}\sum E\psi'(u_i)x_ix_i'$ and $H_n = n^{-1}\sum E\psi^2(u_i)x_ix_i'$. In the iid error context V_n obviously simplifies to

$$\frac{E\psi^2(u_1)}{(E\psi'(u_1))^2}n^{-1}\sum_{i=1}^{n}x_ix_i'.$$

Note that H_n and J_n as defined above specialize to the form taken in the previous section when $\psi(u) = u$ for the least-squares estimator. As in least-squares case we need an expansion of the form

$$\sqrt{n}\,(\tilde{\alpha}_n - \alpha_0) = A_{0n} + n^{-1/2}A_{1n} + n^{-1}A_{2n} + o_p(n^{-1}) \tag{3.4}$$

where as above

$$\tilde{\alpha}_n = (G'\tilde{V}_n^{-1}G)^{-1}G'\tilde{V}_n^{-1}\tilde{\beta}_n \tag{3.5}$$

and $\tilde{\beta}_n$ now solves (3.2), $\tilde{V}_n$ is defined as in (3.3) with $A_{in} = O_p(1)$ for $i = 1, 2, 3$. If $\tilde{Q}_n^{-1} = (G'\tilde{V}_n^{-1}G)^{-1}$, $\tilde{V}_n^{-1}$, and $\sqrt{n}\,(\tilde{\beta}_n - g(\alpha_0))$ possess expansions of the form (3.4), we may write

$$\sqrt{n}\,(\tilde{\alpha}_n-\alpha_0) = [Q_{0n}^{-1} + n^{-1/2}Q_{1n}^{-1} + n^{-1}Q_{2n}^{-1}]G'[V_{0n}^{-1} + n^{-1/2}V_{1n} + n^{-1}V_{2n}] \tag{3.6}$$

$$\cdot\,[B_{0n} + n^{-1/2}B_{1n} + n^{-1}B_{2n}] + o_p(n^{-1})$$

and we would have

$$A_{0n} = Q_{0n}^{-1}G'V_{0n}^{-1}B_{0n}$$

$$A_{1n} = Q_{0n}^{-1}G'V_{0n}^{-1}B_{1n} + Q_{0n}^{-1}G'V_{1n}^{-1}B_{0n} + Q_{1n}^{-1}G'V_{0n}^{-1}B_{0n}$$

$$A_{2n} = Q_{0n}^{-1}G'V_{0n}^{-1}B_{2n} + Q_{0n}^{-1}G'V_{1n}^{-1}B_{1n} + Q_{0n}^{-1}G'V_{2n}^{-1}B_{0n}$$

$$+ Q_{1n}^{-1}G'V_{0n}^{-1}B_{1n} + Q_{1n}^{-1}G'V_{1n}^{-1}B_{0n} + Q_{2n}^{-1}GV_{0n}^{-1}B_{0n}$$

As in the least squares case we can avoid expanding "inverses" using the "trick" (2.8) to express the expansions for $\tilde{Q}_n^{-1}$ and $\tilde{V}_n^{-1}$ in terms of Q_n and V_n.

The expansion for $\sqrt{n}\,(\tilde{\beta}_n - g(\alpha_0))$ however raises some new problems. In the least-squares case we had an exact linear representation for this quantity. Now, $\tilde{\beta}_n$ is defined by the nonlinear system of equations (3.2) and the following lemma is crucial to the sequel.

Inversion Lemma: If $h : \mathbf{R}^p \to \mathbf{R}^p$ admits the expansion, for nonsingular matrix A,

$$h(t) = h(0) + At + \begin{pmatrix} t'B_1t \\ \cdot \\ \cdot \\ \cdot \\ t'B_pt \end{pmatrix} + \begin{pmatrix} \sum_{ijk} t_it_jt_k c_{1ijk} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{ijk} t_it_jt_k c_{pijk} \end{pmatrix} + O(\|t\|^4)$$

and t solves $h(t) = 0$, then t admits the expansion in $h \equiv h(0)$,

$$t = \Gamma h + \begin{pmatrix} h'\Delta_1 h \\ \cdot \\ \cdot \\ \cdot \\ h'\Delta_p h \end{pmatrix} + \begin{pmatrix} \sum_{ijk} h_ih_jh_k \Lambda_{1ijk} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{ijk} h_ih_jh_k \Lambda_{pijk} \end{pmatrix} + O(\|h\|^4)$$

where $\Gamma = -A^{-1}$, $\Delta_i = -\sum_{j=1}^{P} A^{ij} A^{-1} B_j A^{-1}$, and

$$\Lambda_{ijk} = \begin{pmatrix} \Lambda_{1ijk} \\ \cdot \\ \cdot \\ \cdot \\ \Lambda_{pijk} \end{pmatrix} = 2A^{-1} \begin{pmatrix} (A^{-1}B_1)_{i\cdot}\,\Delta_{jk} \\ \cdot \\ \cdot \\ \cdot \\ (A^{-1}B_p)_{i\cdot}\,\Delta_{jk} \end{pmatrix} + A^{-1} \sum_{lmn} A^{li} A^{mj} A^{nk} c_{lmn}$$

Proof: The proof proceeds by equating coefficients. Substituting t in the original expansion we have

$$0 = (I + A\Gamma)h + A \begin{pmatrix} h'\Delta_1 h \\ \cdot \\ \cdot \\ \cdot \\ h'\Delta_p h \end{pmatrix} + A \sum_{ijk} h_ih_jh_k \Lambda_{ijk}$$

$$+ \begin{pmatrix} h'\Gamma' B_1 \Gamma h \\ \cdot \\ \cdot \\ \cdot \\ h'\Gamma B_p \Gamma h \end{pmatrix} + 2 \begin{pmatrix} h'\Gamma B_1 (h'\Delta_1 h \; ... \; h'\Delta_p h)' \\ \cdot \\ \cdot \\ \cdot \\ h'\Gamma B_p (h'\Delta_1 h \; ... \; h'\Delta_p h)' \end{pmatrix}$$

$$+ \sum_{ijk} \Gamma_{i.} h \Gamma_{j.} h \Gamma_{k.} h c_{ijk} + O(\|h\|^4).$$

Thus clearly $I + A\Gamma = 0$ implies $\Gamma = -A^{-1}$. Similarly equating the quadratic terms to zero gives,

$$0 = A \begin{pmatrix} h'\Delta_1 h \\ \cdot \\ \cdot \\ \cdot \\ h'\Delta_p h \end{pmatrix} + \begin{pmatrix} h'\Gamma B_1 \Gamma h \\ \cdot \\ \cdot \\ \cdot \\ h'\Gamma B_p \Gamma h \end{pmatrix} = \begin{pmatrix} h'\sum_i A_{1i} \Delta_i h \\ \cdot \\ \cdot \\ \cdot \\ h'\sum_i A_{pi} \Delta_i h \end{pmatrix} + \begin{pmatrix} h'A^{-1} B_1 A^{-1} h \\ \cdot \\ \cdot \\ \cdot \\ h'A^{-1} B_p A^{-1} h \end{pmatrix}$$

$$= (A \otimes h')\Delta h + (I \otimes h') \begin{pmatrix} A^{-1} B_1 A^{-1} \\ \cdot \\ \cdot \\ \cdot \\ A^{-1} B_p A^{-1} \end{pmatrix} h$$

so

$$(A \otimes I)\Delta = - \begin{pmatrix} A^{-1} B_1 A^{-1} \\ \cdot \\ \cdot \\ \cdot \\ A^{-1} B_p A^{-1} \end{pmatrix}$$

and the expression for Δ_i follows. Similarly, summing the cubic terms in h and equating to zero gives,

$$0 = \sum\sum\sum_{ijk} h_i h_j h_k A \Lambda_{ijk} - 2_{ijk} h_i h_j h_k \begin{pmatrix} (A^{-1}B_1)_{i\cdot} \Delta_{jk} \\ \cdot \\ \cdot \\ \cdot \\ (A^{-1}B_p)_{i\cdot} \Delta_{jk} \end{pmatrix} + \sum_{ijk} h_i h_j h_k \sum_{lmn} \Gamma_{li} \Gamma_{mj} \Gamma_{nk} c_{lmn}$$

and expression follows upon substituting for Γ and solving for Λ_{ijk}.

Remark: Note that Δ_{ijk} is the jk^{th} element of $\Delta_i = -\sum_{l=1}^{p} A^{il} A^{-1} B_l A^{-1}$ so $\Delta_{jik} = -\sum_l A^{il} (A^{-1} B_l A^{-1})_{jk} = \sum_l A^{il} (A_{j\cdot}^{-1} B_l A_{\cdot k}^{-1})$.

In the remainder of this section we will impose the following regularity conditions.

D. The design sequence $\{X_n\}$ satisfies the following conditions:

(i) $\max \|x_i\| = o(n^{1/2})$

(ii) $n^{-1} \sum \|x_i\|^{12} = O(1)$

(iii) The matrices $J_n = n^{-1} \sum x_i x_i' E\psi^2(u_i)$ and $H_n = n^{-1} \sum x_i x_i' E\psi'(u_i)$ converge to positive definite limits.

F. The error sequence $\{u_i\}$ is independent, and each component has a symmetric distribution about zero.

ψ The function $\psi : \mathbf{R} \to \mathbf{R}$ is odd, uniformly continuous, and possesses 3 uniformly continuous, bounded derivatives, and there exists $M < \infty$ such that $E\psi^6(u_i) < M$ for all $i = 1, 2, \ldots$

To apply the inversion lemma to $\sqrt{n}(\hat{\beta}_n - \beta)$ we first expand the implicit definition

$$0 = \sum_{i=1}^{n} \psi(u_i - x_i(\hat{\beta}_n - \beta_0)) x_i$$

writing ψ_i for $\psi(u_i)$, ψ_i' for $\psi'(u_i)$, and so on,

$$0 = n^{-1} \sum_i x_i \psi_i - n^{-1} \sum_i x_i x_i' \psi_i' (\hat{\beta}_n - \beta_0) \tag{3.7}$$

$$+ \frac{1}{2n}\{(\hat{\beta}_n - \beta_0)'[\sum x_i x_i' x_{ij} \psi_i''](\hat{\beta}_n - \beta_0)\}_{j=1,\ldots,1}$$

$$- \frac{1}{6n}\sum_{jkl}(\hat{\beta}_{jn} - \beta_{j0})(\hat{\beta}_{kn} - \beta_{k0})(\hat{\beta}_{ln} - \beta_{lo}) \cdot \sum_i x_{ij} x_{ik} x_{il} x_i E\psi_i'''$$

$$+ o_p(n^{-3/2}).$$

The order of the remainder may be established by noting that for $\|\beta_n^* - \beta_0\| < \|\tilde{\beta}_n - \beta_0\| = O_p(n^{-1/2})$,

$$n^{-1}\sum_{i=1}^{n} x_{ij} x_{ik} x_{il} x_i[\psi'''(u_i - x_i'(\beta^* - \beta_0)) - E\psi_i'''(u_i)] = o_p(1)$$

by the boundedness of ψ''', D(ii) , and the strong law of large numbers. Applying the lemma yields an expansion to order $O(n^{-1})$.

Theorem 3.1: Under conditions D, F, ψ

$$\sqrt{n}\,(\tilde{\beta}_n - \beta_0) = B_{0n} + n^{-1/2}B_{1n} + n^{-1}B_{2n} + o_p(n^{-1})$$

with,

$$B_{0n} = H_n^{-1} n^{-1/2} \sum_i x_i \psi_i$$

$$B_{1n} = -H_n^{-1} n^{-1} \sum_{ij} x_i x_i' H_n^{-1} x_j (\psi_i' - E\psi_i')\psi_j$$

$$B_{2n} = \frac{1}{2} n^{-3/2} \sum_{ijk} H_n^{-1} x_k x_i' H_n^{-1} x_k x_k' H_n^{-1} x_j \psi_i \psi_j \psi_{ik}''$$

$$+ n^{-3/2} \sum_{ijk} H_n^{-1} x_i x_i' H_n^{-1} x_j x_j' H_n^{-1} x_k (\psi_i' - E\psi_i')(\psi_j' - E\psi_j')\psi_k$$

$$- \frac{1}{6} n^{-5/2} \sum_{ijkl} H_n^{-1} \xi_{ijkl} \psi_i \psi_j \psi_k E\psi_l'''$$

and $\xi_{ijkl} = \sum_{uvw} \sum_{rst} H_n^{ru} H_n^{sv} H_n^{tw} x_{iu} x_{jv} x_{kw} x_l$.

An immediate application of this expansion is the following variance expansion for $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$ and hence for $\sqrt{n}\,(\tilde{\alpha}_n - \alpha_0)$.

Theorem 3.2: Under conditions D, F, ψ, with $\{u_i\}$ iid,

$$V(\sqrt{n}(\tilde{\beta}_n - \beta)) = v_0(\psi, F)D_n + n^{-1}v_1(\psi, F)D_{1n} + o(n^{-1})$$

where $v_0(\psi, F) = E\psi^2/(E\psi')^2$, $\quad D_n = n^{-1}\sum x_i x_i'$,

$$v_1(\psi, F) = \left[3\frac{V(\psi')V(\psi)}{(E\psi')^4} - 2\frac{Cov(\psi^2, \psi')}{(E\psi')^3} + 3\frac{V(\psi)Cov(\psi, \psi'')}{(E\psi')^4} - \frac{(E\psi^2)^2 E\psi'''}{(E\psi')^5}\right]$$

and $D_{1n} = D_n^{-1}n^{-1}\sum x_i x_i' D_n^{-1} x_i x_i' D_n^{-1}$ and consequently,

$$V(\sqrt{n}(\tilde{\alpha}_n(V_n) - \alpha_0)) = v_0(\psi, F)\Omega_n + n^{-1}v_1(\psi, F)\Omega_{1n} + o(n^{-1})$$

where as in Theorem 2.2, $\Omega_n = (G'D_nG)^{-1}$ and $\Omega_{1n} = \Omega_n G' D_n D_{1n} D_n G \Omega_n$.

Since Ω_{1n} is positive definite, the sign of $v_1(\psi, F)$ controls the effect of the $O(1/n)$ variance term. Factoring out Ω_n and noting as in the previous section that the design effect

$$tr\ (n^{-1}\sum x_i x_i' D_n^{-1} x_i x_i' D_n^{-1}) = \sum P_{ii}^2(X).$$

may be expressed as $n^{-1}p(k(x) + p + 2)$ following Box and Watson, we see that kurtosis of the design observations $\{x_{ij}\}$ accentuates the magnitude of the $v_1(\psi, F)$ effect on the $O(1/n)$ correction term, as in the least squares case.

Careful inspection of the argument leading to Theorem 3.2 reveals the critical role of the continuity of ψ. It is perhaps worth noting here that discontinuous ψ, exemplified by the $l_1 - \psi$, $\psi(u) = sgn(u)$, give rise to variance expansions in powers of $n^{-1/2}$ rather than the expansion in powers of n^{-1} seen here. This qualitative disparity in performance due to continuity of ψ has been previously noted by Jurečková (1985), Welsh (1989) and others.

To investigate the effect of estimating the matrix V_n by $\tilde{V}_n$ we must carry out the expansions of the inverse matrices as in the previous section. Expanding $\tilde{V}_n^{-1}$, we first note that

$$\tilde{V}_n^{-1} = \tilde{H}_n\tilde{J}_n^{-1}\tilde{H}_n = H_n\tilde{J}_n^{-1}H_n + n^{-1/2}2H_n\tilde{J}_n^{-1}[\sqrt{n}(\tilde{H}_n - H_n)] \tag{3.8}$$

$$+ n^{-1}[\sqrt{n}(\tilde{H}_n - H_n)]\tilde{J}_n^{-1}[\sqrt{n}(\tilde{H}_n - H_n)]$$

and using (2.8)

$$\tilde{J}_n^{-1} = J_n^{-1} - n^{-1/2} J_n^{-1}[\sqrt{n}\,(\tilde{J}_n - J_n)]J_n^{-1}$$

$$+ n^{-1} J_n^{-1}[\sqrt{n}\,(\tilde{J}_n - J_n)]J_n^{-1}[\sqrt{n}\,(\tilde{J}_n - J_n)]J_n^{-1}$$

$$+ n^{-3/2} O_p(\|\sqrt{n}\,(\tilde{J}_n - J_n)\|^3)$$

Thus, it is apparent that expansions of $\sqrt{n}\,(\tilde{J}_n - J_n)$ and $\sqrt{n}\,(\tilde{H}_n - H_n)$ to order $O_p(n^{-1/2})$ will suffice. Expanding $\sqrt{n}\,(\tilde{H}_n - H_n)$ we obtain,

$$\sqrt{n}\,(\tilde{H}_n - H_n) = n^{-1/2}\sum_i x_i x_i'[\psi'(u_i - x_i(\tilde{\beta}_n - \beta_n)) - E\psi_i']$$

$$= n^{-1/2}\sum x_i x_i'(\psi_i' - E\psi_i') - n^{-1/2}\sum x_i x_i' x_i'(\tilde{\beta}_n - \beta_0)\psi_i''$$

$$+ \frac{1}{2} n^{-1/2}\sum x_i x_i' x_i'(\tilde{\beta}_n - \beta_0)(\tilde{\beta}_n - \beta_0)' x_i E\psi_i'''$$

$$+ o_p(n^{-1})$$

and replacing $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$ by its *first order* approximation yields,

$$\sqrt{n}\,(\tilde{H}_n - H_n) = H_{0n} + n^{-1/2} H_{1n} + o_p(n^{-1/2})$$

with

$$H_{0n} = n^{-1/2}\sum x_i x_i'(\psi_i' - E\psi_i') = O_p(1)$$

$$H_{1n} = n^{-3/2}\frac{1}{2}\sum_{ijk} x_i x_i' x_i' H_n^{-1} x_j x_k' H_n^{-1} x_i \psi_j \psi_k E\psi_i'''$$

$$- n^{-1}\sum_{ij} x_i x_i' x_i' H_n^{-1} x_j \psi_j \psi_i''$$

where H_{0n} and H_{1n} are $O_p(1)$. Similarly,

$$\sqrt{n}\,(\tilde{J}_n - J_n) = n^{-1/2}\sum_i x_i x_i'(\psi^2(u_i - x_i'(\tilde{\beta}_n - \beta_0) - E\psi_i^2)$$

$$= n^{-1/2}\sum x_i x_i'(\psi_i^2 - E\psi_i^2) - n^{-1}2\sum x_i x_i' x_i(\tilde{\beta}_n - \beta_0)\psi_i\psi_i'$$

$$+ n^{-1/2}\sum x_i x_i'(\tilde{\beta}_n - \beta_0)(\tilde{\beta}_n - \beta_0)' x_i E(\psi_i''\psi_i - \psi_i^2)$$

$$+ o_p(n^{-1/2})$$

so

$$\sqrt{n}\,(\tilde{J}_n - J_n) = J_{0n} + n^{-1/2}J_{1n} + o_p(n^{-1/2})$$

$$J_{0n} = n^{-1/2}\sum x_i x_i'(\psi_i^2 - E\psi_i^2)$$

$$J_{1n} = n^{-3/2}\sum_{ijk} x_i x_i' x_i H_n^{-1} x_j x_k' H_n^{-1} x_i x_i' E(\psi_i''\psi_i - (\psi_i')^2)\psi_j\psi_k$$

$$- n^{-1}2\sum_{ij} x_i x_i' x_i H_n^{-1} x_j \psi_i'\psi_j.$$

Thus, again using (2.8) we may write

$$\tilde{J}_n^{-1} - J_n^{-1} = n^{-1/2}J_n^{(0)} + n^{-1}J_n^{(1)} + o_p(n^{-1})$$

where $J_n^{(0)} = -J_n^{-1}J_{0n}J_n^{-1}$ and $J_n^{(1)} = -J_n^{-1}J_{1n}J_n^{-1} + J_n^{-1}J_{1n}J_n^{-1}J_{1n}J_n^{-1}$, and thus,

$$\tilde{V}_n^{-1} - V_n^{-1} = n^{-1/2}V_n^{(0)} + n^{-1}V_n^{(1)} + o_p(n^{-1})$$

where

$$V_n^{(0)} = H_n J_n^{(0)} H_n + 2H_n J_n^{-1} H_{0n}$$

and

$$V_n^{(1)} = H_n J_n^{(1)} H_n + 2H_n J_n^{-1} H_{1n} + 2H_n J_n^{(1)} H_n + H_{1n} J_n^{-1} H_{1n}.$$

Now substituting back into (3.8) we have established:

Theorem 3.3: Under D, F, ψ :

$$\sqrt{n}\,(\tilde{\alpha}_n - \alpha_0) = A_{0n} + n^{-1/2}A_{1n} + n^{-1}A_{2n} + o_p(n^{-1})$$

with

$$A_{0n} = Q_n^{-1}G'H_nJ_n^{-1}n^{-1/2}\sum x_i\psi_i$$

$$A_{1n} = Q_n^{-1}GH_nJ_n^{-1}n^{-1}\sum_{ij}[a_{ij}^{(1)}(\psi_i^2 - E\psi_i^2)\psi_j + a_{ij}^{(2)}(\psi_i' - E\psi_i')\psi_j]$$

$$A_{2n} = Q_n^{-1}GH_nJ_n^{-1}n^{3/2}\sum_{ijk}[a_{ijk}^{(1)}(\psi_i' - E\psi_i')(\psi_j' - E\psi_j')\psi_k$$

$$+ a_{ijk}^{(2)}(\psi_i^2 - E\psi_i^2)(\psi_j^2 - E\psi_j^2)\psi_k$$

$$+ a_{ijk}^{(3)}(\psi_i^2 - E\psi_i^2)(\psi_j' - E\psi_j')\psi_k$$

$$+ a_{ijk}^{(4)}\psi_i\psi_j\psi_k + a_{ijk}^{(5)}\psi_i\psi_j\psi_k'' + a_{ijk}^{(6)}\psi_i\psi_i'\psi_j\psi_k]$$

$$Q_n = G'V_n^{-1}G = G'H_nJ_n^{-1}H_nG$$

$$M_n = V_n - GQ_n^{-1}G'$$

$$K_n = J_n^{-1} - J_n^{-1}H_nGQ_n^{-1}G'H_nJ_n^{-1}$$

$$a_{ij}^{(1)} = -x_ix_i'J_n^{-1}H_nM_nH_nJ_n^{-1}x_j$$

$$a_{ij}^{(2)} = -x_ix_i'H_n^{-1}x_j + 2x_ix_i'M_nH_nJ_n^{-1}x_j$$

$$a_{ijk}^{(1)} = x_ix_i'H_n^{-1}x_jx_j'H_n^{-1}x_k - 2x_ix_i'M_nH_nJ_n^{-1}x_jx_j'H_n^{-1}x_k$$

$$+ x_ix_i'V_nx_jx_j'J_n^{-1}H_nM_nH_nJ_n^{-1}x_k$$

$$- 4x_ix_i'(V_n - M_n)H_nJ_n^{-1}x_jx_j'M_nH_nJ_n^{-1}x_k$$

$$a_{jik}^{(2)} = x_ix_i'K_nx_jx_j'K_nx_k$$

$$a_{ijk}^{(3)} = x_ix_i'J_n^{-1}H_nM_nH_nJ_n^{-1}x_jx_j'H_n^{-1}x_k - 2x_ix_i'J_n^{-1}x_jx_j'M_nH_nJ_n^{-1}x_k$$

$$+ 2x_ix_i'J_n^{-1}H_n(V_n - M_n)H_nJ_n^{-1}x_jx_j'M_nH_nJ_n^{-1}x_k$$

$$+ 2x_jx_j'(V_n - M_n)H_nJ_n^{-1}x_ix_i'J_n^{-1}H_nM_nH_nJ_n^{-1}x_k$$

$$a_{jik}^{(4)} = -\frac{1}{6}n^{-1}\sum_l b_{ijkl} E\psi_l''' + n^{-1}\sum_l x_l x_l' H_n^{-1} x_i x_j H_n^{-1} x_l x_l'$$

$$\cdot K_n \, x_k \, [E\psi_l''' - E(\psi_l''\psi_l + (\psi_l')^2)]$$

$$a_{ijk}^{(5)} = \frac{1}{2} x_k x_i' H_n^{-1} x_k x_k' H_n^{-1} x_j - 2 x_k x_k' x_k' H_n^{-1} x_i M_n H_n J_n^{-1} x_j$$

$$a_{ijk}^{(6)} = 2 x_i x_i' x_i' H_n^{-1} x_j J_n^{-1} H_n M_n H_n J_n^{-1} x_k .$$

Computing moments from this expansion yields the following two term expansion for the asymptotic variance of $\tilde{\alpha}_n$.

Theorem 3.4: Under conditions D, F, ψ with $\{u_i\}$ iid

$$V(\sqrt{n}\,(\tilde{\alpha}_n(\tilde{V}_n) - \alpha_0)) = v_0(\psi, F)\Omega_n + n^{-1}[v_1(\psi, F)\Omega_{1n} + v_2(\psi, F)\Omega_{2n}] + o_p(n^{-1}) \qquad (3.9)$$

where $v_0(\psi, F)$, $v_1(\psi, F)$, Ω_n, Ω_{1n} are as defined in Theorem 3.2, $\Omega_{2n} = \Omega_n G \tilde{M}_n G \Omega_n$, $\tilde{M}_n$ is as defined in Theorem 2.2, and

$$v_2(\psi, F) = [4\frac{(E\psi^2)^2 E\psi'''}{(E\psi')^5} - 4\frac{E\psi^2 E(\psi')^2}{(E\psi')^4} - 12\frac{E\psi^2 E\psi\psi''}{(E\psi')^4} + 8\frac{E\psi^2\psi'}{(E\psi')^3}$$

$$- 4\frac{E\psi^2 E(\psi' - E\psi')^2}{(E\psi')^4} + 4\frac{E(\psi' - E\psi')(\psi^2 - E\psi^2)}{(E\psi')^4} - \frac{(E\psi^2 - E\psi^2)^2}{E\psi^2 (E\psi')^2}].$$

Remark: Comparing Theorems 3.2 and 3.4 we see that the effect of estimating V_n is captured by the $O(1/n)$ term $v_2(\psi, F)\Omega_{2n}$ where Ω_{2n} is the same design matrix which appears in the $\hat{V}_n$-effect for the least-squares effect in Theorem 2.2. The consequence of using $\hat{V}_n$ rather than the true $V_n = v_0(\psi, F)\Omega_n$ is, up to $O(1/n)$, the effect of the $n^{-1}v_2(\psi, F)\Omega_{2n}$ term. Since the design contribution Ω_{2n} has already been discussed in the previous section we focus attention here on the $v_i(\psi, F)$ $i = 0, 1, 2$ terms.

In Table 3.1 we report computations of $v_1(\psi, F)$ $i = 0, 1, 2$ for some representative situations with Student t errors $\{u_i\}$ and logistic ψ of the form

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

which may be regarded as a smooth version of Huber's (1964) well-known minimax ψ. In Figure 3.1 we illustrate ψ_λ for several choices of λ. Relative to the Student t family $\lambda = 5$ corresponds to rather aggressive trimming, $\lambda = 1$ to light trimming, and $\lambda = 3$ to an intermediate position. The column labeled $\lambda = \infty$ gives corresponding coefficients for the least-squares estimator where appropriate.

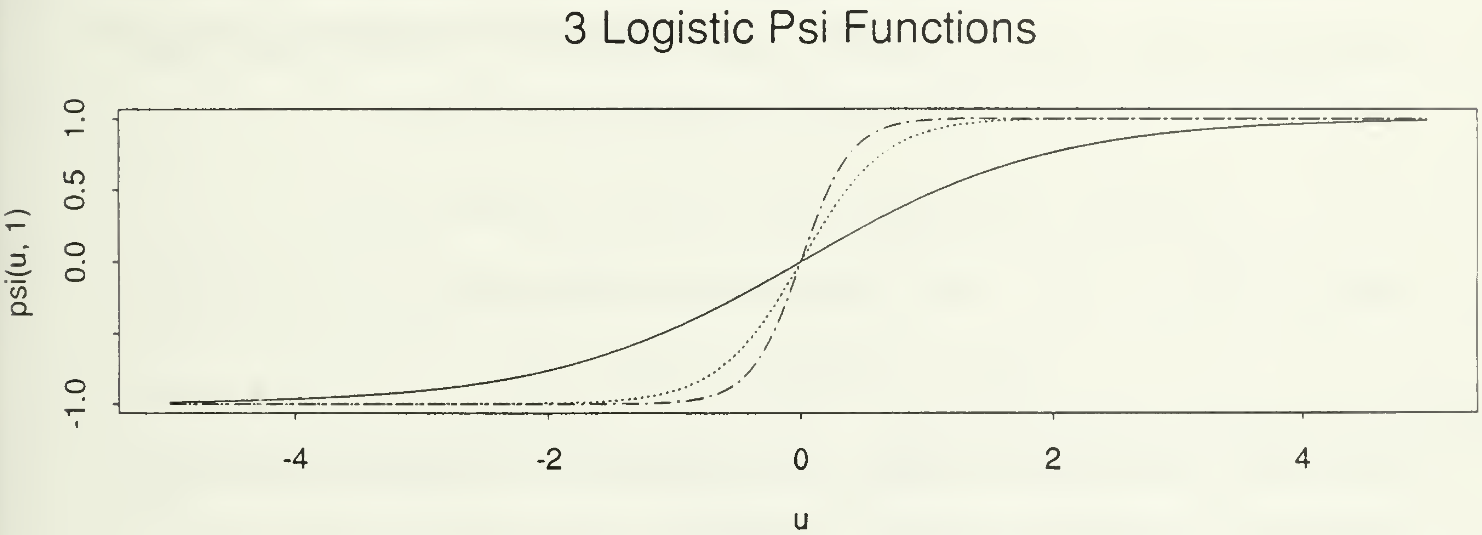


Table 3.1

Coefficients for the Moment Expansion for $\sqrt{n}\,(\hat{\alpha}_n - \alpha_0)^*$

Student DF	Scale of the Logistic ψ $\lambda = 0$	$\lambda = 1$	$\lambda = 3$	$\lambda = 5$
	∞	3.68	2.45	2.31
1	-	-16.79	-8.86	-3.38
	-	6.08	5.80	6.34
	∞	2.11	1.73	1.73
2	-	-2.81	1.04	5.70
	-	1.59	1.84	2.10
	1.67	1.38	1.36	1.42
5	-	1.12	3.85	8.04
	-	0.34	0.49	0.58
	1.25	1.18	1.25	1.33
10	1.00	1.82	4.39	8.43
	0	0.12	0.18	0.21

* Each "cell" of the table reports the three coefficients $v_i(\psi,F)$ $i = 0, 1, 2$, for the variance expansion of Theorem 3.4.

As in the least squares case, there appears to be some robustifying effect from the estimation of V_n when the errors are long-tailed and ψ is small, but this effect is confined to the Cauchy and t_2 situations. The order $O(1/n)$ effect when V_n is (correctly) treated as proportional to $D_n = n^{-1}X'X$ is to increase the asymptotic variance in all of our parametric examples, but $v_1(\psi, F)$ may also perhaps be negative.

The $\tilde{V}_n$-effect of the design is controlled by the matrix Ω_{1n} which also appears in the least-squares variance expansion for $\sqrt{n}\,(\hat{\alpha}_n - \alpha_0)$, in Theorem 2.2. When the design is Gaussian, $k(x) = 3$, a simple interpretation is again possible if we write (3.9) as

$$V(\sqrt{n}\,(\tilde{\alpha}_n - \alpha_0)) = \Omega_n[v_0(\psi, F)I_p + n^{-1}(v_1\Delta_{1n} + v_2\Delta_{2n})] + o_p(n^{-1})$$

where $\Delta_{in} = \Omega_n^{-1}\Omega_{in}$ $i = 1, 2$. Again taking traces, and dividing though by p, we may regard

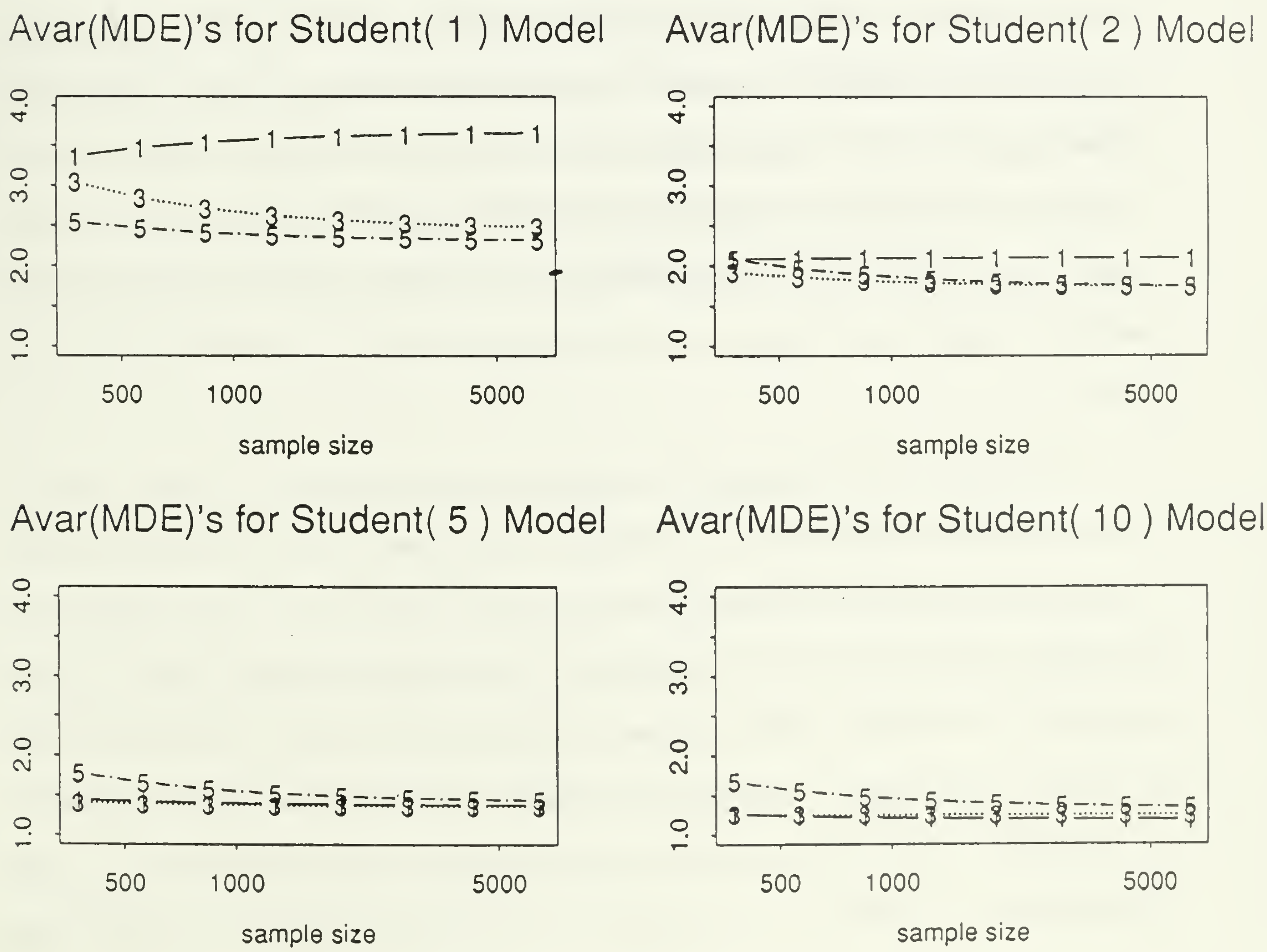
$$\varsigma(n) = v_0 + n^{-1}(v_1(1 - p) + v_2(q + 2))$$

as providing a $O(1/n)$ "correction term" for the standard first-order asymptotic variance, v_0, for each of the p parameters of the vector $\sqrt{n}\,\Omega_n^{-1/2}\tilde{\alpha}_n$.

Obviously, it is difficult to interpret this expression given the rather exotic form of v_1 and v_2. So we have to resort to plotting ς_n for a few representative situations. In Figure 3.2a we illustrate ς_n for Cauchy errors and the logistic ψ function with 3 different values of the scale parameter $\lambda \in \{1, 3, 5\}$. In all three panels of Figure 3.2 we take $p = 5$, $q = 20$. In the long-tailed t_1 situation severe trimming with $\lambda = 5$ is best, achieving $v_0 = 2.31$ as $n \to \infty$. The Cramer-Rao lower bounded is 2.0 in this case. The mild trimming of $\lambda = 1$ performs substantially worse-achieving $v_0 = 3.68$ asymptotically, but note that the robustifying effect of $\tilde{V}_n$ causes performance of the normalized estimator to approach this limit from below, while for the more successful severely trimmed estimators the $O(1/n)$ variance correction is positive, so $\varsigma(n)$ *decreases* toward v_0 as $n \to \infty$.

In Figure 3.2b where we illustrate the situation for t_2 errors the outcome is similar, however there is a slight preference for the moderate trimming of $\lambda = 3$ with n small. For t_5, and

t_{10} it is difficult to distinguish the performance of $\lambda = 3$ and $\lambda = 1$, but both are slightly preferred to the severe trimming of $\lambda = 5$. Reviewing the four cases, the fundamental observation of the robustness literature, e.g., Huber (1981), is apparent -- a small insurance premium in sacrificed efficiency in nearly Gaussian situations is justified by the protection afforded in long-tailed situations by robust methods. We should perhaps emphasize that even the mild trimming of $\lambda = 1$ is quite successful in this respect relative to the least squares estimator which has unbounded asymptotic variance in the t_1 and t_2 models.



4. Conclusion

We began our investigation with two questions; how does estimation of the covariance matrix V_n affect the performance of least-squares based minimum distance (GMM) estimators? In particular, how does estimating V_n affect the robustness of minimum distance methods to long-tailed error situations? These are questions that familiar first-order asymptotic methods are unable to answer. In the first-order asymptotics, $\hat{V}_n$ is as good as V_n as long as $\hat{V}_n \to V_n$ in probability.

Answers must be sought in "the garden of higher expansions" where the paths are arduous, but the vistas are rewarding at the end of the day. Perhaps the most surprising result is the robustifying effect of the Eicker-White covariance matrix estimate for the least-squares estimator $\hat{\alpha}_n$ when the errors are iid with kurtosis greater than 5. One might think that since V_n is more difficult to estimate in such situations, that the use of $\hat{V}_n$ would inflate the variance of $\hat{\alpha}_n$ relative to its performance when $\hat{V}_n$ is replaced by the correct V_n. This is certainly true in the Gaussian case as the elegant sufficiency arguments of Rothenberg (1982) show. But when $k(u) > 5$, even a poor estimate of $\hat{V}_n$ is valuable, because it effectively serves to downweight observations with large error realizations. Nevertheless, this effect is transitory, and provides a fragile straw to grasp when least-squares methods are confronted by serious outliers.

This notorious lack of robustness of least-squares methods leads us to consider a class of minimum distance estimators based on preliminary M-estimation. Here the technical aspects of the expansions are more challenging and we require an inversion of a cubic expansion of the defining equation (3.2) for the M-estimator. This three term expansion to $O_p(n^{-1})$ for the preliminary M-estimation, $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$, may be of independent interest. We may note at this point that we have treated scale as known throughout these expansions for the M-estimator; in subsequent work we hope to report on joint expansions incorporating scale. As in the least-squares case we focus the interpretation of the M-estimator expansions for $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$

and $\sqrt{n}\,(\tilde{\alpha}_n - \alpha_0)$ on the two-term, $O(n^{-1})$, expansions for the asymptotic variance of these quantities when the errors are iid. In this case the design contributions to the $O(n^{-1})$ terms may be greatly simplified by treating the design observations as iid as in Box and Watson (1962). This facilitates a rather simple graphical presentation of the results and reveals also the effects of the dimensionality of the initial parameter vector β, and the kurtosis of the design observations: both tend to inflate the scale of the $O(1/n)$ term of the variance expansion.

Taken as a whole our results reenforce the familiar robustness critique of least-squares based methods. But by explicitly computing the $O(n^{-1})$ term in the asymptotic variance expansion of the MDE we are able to address a variety of questions unresolvable by traditional first-order asymptotic methods: How does estimation of the covariance matrix of the preliminary estimator $\hat{\beta}_n$ effect the performance of the final estimator $\hat{\alpha}_n$? How does the dimensionality of $\hat{\beta}_n$ affect the performance of $\hat{\alpha}_n$? What design conditions are necessary to insure the validity of the usual first-order asymptotic normal approximation for $\hat{\alpha}_n$? Finally, we may note that the methods employed here apply quite directly to a broad range of estimators used in econometrics and we hope to explore other examples in future work.

References

Altonji, J.G., Martins, A.P. and Siow, A. (1987) "Dynamic Factor Models of Consumption, Hours and Income," NBER Working Paper 2155.

Bickel, P.J. and Duksum, K. (1977) *Mathematical Statistics*, Holden Day: San Francisco.

Billingsley, P. (1979) *Probability and Measure*, Wiley: New york.

Box, G.E.P. and Walson, G.S. (1962) Robustness to non-normality of regression tests," *Biometrika*, 49, 93-106.

Carroll, R.J., Wu, C.F.J. and Rupport, D. (1988) "The effect of estimating weights in weighted least squares," *JASA*, 83, 1045-1054.

Chamberlain, G. (1982) "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5-46.

Chamberlain, G. (1987) "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

Gallant, A.R. (1987) *Nonlinear Statistical Models*, Wiley: New York.

Feller, W. (1971) *Introduction to Probability Theory and Its Applications*, Wiley: New York.

Hansen, L.P. (1982) "Large sample properties of generalized method of moment estimators," *Econometrica*, 50, 1029-1054.

Hansen, L.P. and Singleton, K.J. (1983) "Stochastic consumption, risk aversion, and the temporal behavior of stock market returns," *Journal of Political Economy*, 91, 249-265.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics*, Wiley: New York.

Huber, P.J. (1964) "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, 36, 1753-1758.

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

Jurečková, J. (1985) "Robust estimators of location and their second-order asymptotic relations," in A.C. Atkinson and S.E. Fienberg (eds.), *A Celebration of Statistics*, Springer-Verlag: New York.

Koenker, R., Skeels, C.L. and Welsh, A.H. (1990) "A Note on Amemiya's form of the weighted least squares estimator," preprint.

Mosteller, F. and Tukey, J.W. (1977) *Data Analysis and Regression*, Addison-Wesley: Reading, Mass.

Newey, W. (1985) "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229-256.

Newey, W. (1988) "Asymptotic equivalence of closest moments and GMM estimators," *Econometric Theory*, 4, 336-340.

Rothenberg, T.J. (1984) "Approximate normality of generalized least squares estimates," *Econometrica*, 52, 811-825.

Rousseeuw, P.J. and Leroy, A.M. (19) *Robust Regression and Outlier Detection*, Wiley: New York.

Tauchen, G.E. (1986) "Statistical properties of generalized method of moments estimators of structural parameters obtained from financial market data," *Journla of Business and Economic Statistics*, 4, 397-424 (with discussion).

Tukey, J.W. (1960) "A survey of sampling from contaminated normal distributions," in J. Olkin (ed.) *Contributions to Probability and Statistics*, Stanford University Press: Stanford, Calif.

Welsh, A.H. (1989) "On M-process and M-estimation," *Annals of Statistics*, 17, 337-361.

Wolfram, S. (1988) *Mathematica*, Addison-Wesley: Reading, Mass.

Appendix A

Proof (of Theorem 2.1) A direct proof is straightforward but tedious. Instead we may simply substitute $\psi(u) = u$ into the expansion of Theorem 3.3 to obtain the result.

Proof (of Theorem 2.2) Clearly $V(A_{0n}) = Q_n^{-1}$ so it remains to compute the $O(n^{-1})$ term which is, like our scalar example, given by

$$n^{-1}[V(A_{1n}) + 2n^{1/2}Cov(A_{0n}, A_{1n}) + 2Cov(A_{0n}, A_{2n})].$$

Consider

$$V(A_{1n}) = Q_n^{-1}GH_nJ_n^{-1}n^{-2}\textstyle\sum_{ijkl}x_ix_i'K_nx_jx_k'K_nx_lx_l'$$

$$\cdot [E(u_i^2 - \sigma_i^2)(u_k^2 - \sigma_k^2)u_ju_l]J_n^{-1}H_nG'Q_n^{-1}$$

The expectation in square brackets is $\sigma_i^4\sigma_j^2(k_i - 1)$ if $i = k$ and $j = l$, otherwise the expectation is zero, thus noting $K_nJ_nK_n = K_n$,

$$V(A_{1n}) = Q_n^{-1}GH_nJ_n^{-1}n^{-1}\textstyle\sum\sigma_i^4(k_i - 1)x_ix_i'K_nx_ix_i'J_n^{-1}H_nG'Q_n^{-1} + o_p(n^{-1}).$$

To illustrate the remaining computation, write $A_{2n} = \sum_{i=1}^{3} A_{2in}$, and consider,

$$Cov(A_{0n}, A_{22n}) = -Q_n^{-1}GH_nJ_n^{-1}L_nJ_n^{-1}H_nGQ_n^{-1}$$

where

$$L_n = n^{-3}\textstyle\sum_{ijklm}x_lx_l'H_n^{-1}x_ix_j'H_n^{-1}x_lx_l'K_nx_kx_m' \cdot Eu_iu_ju_ku_m.$$

As above, the expectations are zero except when the indices are pairwise equal. When $i = j \neq k = m$,

$$n^{-2}\textstyle\sum_l\sum_i\sigma_i^2x_lx_l'H_n^{-1}x_ix_i'H_n^{-1}x_lx_l'K_nn^{-1}\sum_{j\neq i}\sigma_j^2x_jx_j'$$

$$= n^{-2}\textstyle\sum_{li}\sigma_i^2x_lx_l'H_n^{-1}x_ix_i'H_n^{-1}x_lx_l'K_nJ_n + O(n^{-1}).$$

But $K_n J_n J^{-1} H_n G Q_n^{-1} = 0$ so the preceding contribution is negligible. When $i = k \neq j = m$ and $i = m \neq j = k$, the contributions are identical, and take the form,

$$n^{-3}\sum_l \sum_i \sum_{j \neq i} \sigma_i^2 \sigma_j^2 x_l x_l' H_n^{-1} x_i x_j' H_n^{-1} x_l x_l' K_n x_i x_j'$$

$$= n^{-1}\sum_l x_l x_l' K_n J_n H_n^{-1} x_l x_l' H_n^{-1} J_n + O(n^{-1}).$$

The remaining terms may be computed similarly.

Proof (of Theorem 3.1) Applying the inversion lemma and its notation to (3.7) we have for n sufficiently large, by D(iii),

$$\Gamma = [n^{-1}\sum x_i x_i' \psi_i']^{-1} = \tilde{H}_n^{-1}$$

$$\Delta_m = (\sum_{j=1}^{q} \Gamma_{mj} \Gamma B_j)\Gamma.$$

Since, by hypothesis $E\psi''(u_i) = 0, B_j = n^{-1}\sum x_i x_i' x_{ij} \psi_i'' = o_p(1)$, and hence

$$\Lambda_{jkl} = \sum A^{mj} A^{nk} A^{ol} c_{mno} + o_p(1)$$

$$= -\frac{1}{6}\Gamma \sum_{mno} \Gamma_{mj} \Gamma_{nk} \Gamma_{ol} n^{-1} \sum x_{im} x_{in} x_{io} x_i \psi_i'''.$$

Approximating $\tilde{\Gamma} = -\tilde{A}^{-1}$ as in (2.8) we have,

$$\tilde{A}^{-1} = A^{-1} =- A^{-1}(\tilde{A} - A)A^{-1} + A^{-1}(\tilde{A} - A)A^{-1}(\tilde{A} - A)A^{-1} + O_p(\|\tilde{A} - A\|^3).$$

Since $\tilde{A} - A = n^{-1}\sum x_i x_i'(\psi' - E\psi_i') = O_p(n^{-1/2})$ the preceeding remainder is of the desired order $O_p(n^{-3/2})$.

For the quadratic terms $h'\Delta_i h = O_p(n^{-1})\Delta_i$ so we may neglect the $o_p(n^{-1/2})$ terms in the expansion of Δ_i. Thus we may write,

$$\Delta_s = -\sum_j A^{sj} A^{-1} B_j A^{-1} + o_p(n^{-1/2})$$

$$= \frac{1}{2}[n^{-1}\sum x_i x_i E\psi_i']^{-1}\{\sum(n^{-1}\sum x_i x_i' E\psi_i')_{sj}^{-1}\}$$

$$\cdot \{\sum(n^{-1}\sum x_i x_i' E\psi_i')_{sj}(n^{-1}\sum x_i x_i' x_{ij}\psi_i'')\}$$

$$\cdot [n^{-1}\sum x_i x_i E\psi_i']^{-1} + o_p(n^{-1/2}).$$

Similarly, in the cubic term we need only,

$$\Lambda_{jkl} = -\frac{1}{6}H_n^{-1}\sum_{mno} H_n^{mj}H_n^{nk}H_n^{ol}[n^{-1}\sum_i x_{im}x_{ik}x_{il}x_i\psi_i'''] + o_p(1)$$

and assembling terms yields the expansion ■

Proof (of Theorem 3.2) The argument proceeds exactly like that for Theorem 2.2. The computations are similar except for the covariance of B_{on} with the last term of B_{2n} which we give here explicitly. Denote the latter as B_{23n},

$$Cov(B_{0n}, B_{23n}) = -\frac{1}{6}n^{-3}\sum_{ijklr} H_n^{-1}\xi_{ijkl}\psi_i\psi_j\psi_k E\psi_l'''\psi_r x_r' H_n^{-1}.$$

Note that

$$E\psi_i\psi_j\psi_k\psi_r = \begin{cases} E\psi^4 & \text{if } i = j = k = m \\ (E\psi^2)^2 & \text{if } i = j \neq k = r,\ i = k \neq j = r,\ i = r \neq j = k \\ 0 & \text{otherwise.} \end{cases}$$

As before we may neglect the first case since it contributes a term of $O(n^{-1})$. Symmetry in the indices implies that the three cases of pairwise equality are identical, so it suffices to consider the case when $i = j \neq k = m$. Note first that with $\{u_i\}$ iid,

$$n^{-1}\sum_l \xi_{ijkl} E\psi_r''' \equiv \frac{E\psi'''}{(E\psi')^3} n^{-1}\sum_l \sum_{mnostv} x_{is}x_{jt}x_{kv}H_n^{ms}H_n^{nt}H_n^{ov}x_{lm}x_{ln}x_{lo}x_l$$

so

$$n^{-3}\sum_{jiklr}\xi_{ijkl}x_r' = E\psi'''n^{-3}\sum_i\sum_{j\neq i}\sum_l \xi_{iijl}x_j'$$

$$= \frac{E\psi'''}{(E\psi')^3}\sum_{mno}\sum_{stv} H_n^{ms} H_n^{nt} H_n^{ov} w_{mnostv}$$

where

$$w_{mnostv} = n^{-3}\sum_i \sum_{j \neq i} \sum_l x_{is} x_{it} x_{jv} x_{lm} x_{ln} x_{lo} x_l x_j'$$

$$= (n^{-1}\sum_i x_{is} x_{it})(n^{-1}\sum_j x_{jv} x_j - n^{-1} x_{iv} x_i) \cdot (n^{-1}\sum_l x_{lm} x_{ln} x_{lo} x_l')$$

$$= (H_n)_{st}(H_n)_{\cdot v}(n^{-1}\sum_l x_{lm} x_{ln} x_{lo} x_l') + O(n^{-1})$$

Since $\sum_t H_n^{nt}(H_n)_{st} = \delta_{ns}$ and $\sum_v H_n^{ov}(H_n)_{\cdot v} = H_n H_n^{\cdot o} = e_o$ where δ_{ns} is Kronecker's delta, and e_o is the o^{th} unit vector of $\mathbf{R}^q$, respectively. Thus,

$$n^{-3} \sum_{i=j \neq j=ijklr} \xi_{ijkl} x_r' = \frac{E\psi'''}{(E\psi')^3} n^{-1}\sum_l \sum_{mno} D_n^{mn} x_{lm} x_{ln} x_{lo} e_o x_l'$$

$$= \frac{E\psi'''}{(E\psi')^3} n^{-1}\sum_i x_i x_i' D_n^{-1} x_i x_i'$$

since $\sum_{o=1}^{q} x_{lo} e_o x_l' = x_l x_l'$ and $\sum_{mn=1}^{q} H_n^{mn} x_{lm} x_{ln} = x_l' H_n^{-1} x_l$. Thus, since there are three such terms,

$$Cov(B_{0n}, B_{23n}) = -\frac{1}{2}\frac{(E\psi^2)^2 E\psi'''}{(E\psi')^5} D_n^{-1}\sum_i x_i x_i' D_n^{-1} x_i x_i' D_n^{-1}.$$

Appendix D

On the Design Effect

Consider $\delta_{zx} = \sum z_i'(Z'Z)^{-1}z_i x_i'(X'X)^{-1}x_i$ and suppose provisionally that $Z'Z$ and $X'X$ are diagonal. Then

$$z_i'(Z'Z)^{-1}z_i = \sum_{j=1}^{p}(z_{ij}^2/\sum_{i=1}^{n} z_{ij}^2) = \sum_{j=1}^{p}\left[\frac{\sum_r\sum_s g_{jr}g_{js}x_{ir}x_{is}}{\sum_s g_{js}^2 \sum_i x_{is}^2}\right]$$

and thus

$$\delta_{zx} = \sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{q}\sum_{mm=1}^{q}\left[\frac{g_{jl}g_{jm}\sum_{i=1}^{n} x_{ik}^2 x_{il}x_{im}}{\sum_i x_{ik}^2 \sum_{s=1}^{p}(g_{js}^2\sum_i x_{is}^2)}\right].$$

Now assuming that the $\{x_{ij}\}$ are iid, symmetric about zero we have,

$$En^{-1}\sum x_{ik}^2 x_{il}x_{im} = \begin{cases} Ex_{ij}^2 \equiv \mu_4(x) & j = l = m \\ (Ex_{ij}^2)^2 \equiv \sigma^4(x) & j \neq l = m \\ 0 & \text{otherwise} \end{cases}$$

so

$$\delta_{zx} = n^{-1}\sum_j^p\sum_k^q\left(\frac{\mu_4(x)g_{jk}^2}{\sigma^4(x)\sum_{s=1}^{q} g_{js}^2}\right) + \sum_{j-1}^{p}\sum_{k=1}^{1}\sum_{l \neq k}^{q}\left(\frac{g_{jl}^2}{\sum_{s=1}^{q} g_{js}}\right) = n^{-1}\left[pk(x) + \sum_j\sum_k\left(\frac{\sum_l g_{il}^2 - g_{jk}^2}{\sum_{s=1}^{q} g_{js}^2}\right)\right]$$

$$= n^{-1}[pk(x) + p(q-1)]$$

Similarly,

$$\delta_{zz} = \sum(z_i'(Z'Z)^{-1}z_i)^2 = \sum_{j=1}^{p}\left[\frac{n^{-2}\sum_i z_{ij}^4}{(n^{-1}\sum_i z_{ij}^2)^2}\right] + \sum_{j-1}^{p}\sum_{k \neq j}^{p}\left[\frac{n^{-2}\sum z_{ij}^2 z_{ik}^2}{n^{-2}\sum z_{ij}^2\sum z_{ik}^2}\right].$$

Now $z_{ij}^4 = \sum_{rstu}^q g_{ir} g_{js} g_{jt} g_{ju} x_{ir} x_{is} x_{it} x_{iu}$ and

$$Ex_{ir} x_{is} x_{it} x_{iu} = \begin{cases} \mu_4(x) & r = s = t = u \\ \sigma^4(x) & (r = s \neq t = u), (r = t \neq s = u), (r = u \neq s = t) \\ 0 & \text{otherwise} \end{cases}$$

so

$$n^{-1} \sum_{i=1}^{n} z_{ij}^4 \to \mu_4(x) \sum_{k=1}^{q} g_{jk}^4 + 3\sigma^4(x) \sum_{k=1}^{q} \sum_{l \neq k} g_{jk}^2 g_{jl}^2,$$

$$n^{-1} \sum_{i} z_{ij}^2 \to \sigma^2(x) \sum_{k=1}^{q} g_{jk}^2,$$

and for $j \neq k$,

$$n^{-1} \sum_{i} z_{ij}^2 z_{ik}^2 \to \mu_4(x) \sum_{l=1}^{q} g_{jl}^2 g_{kl}^2 + \sigma^4(x) \sum\sum_{l \neq m} g_{jl}^2 g_{km}^2 + \sigma^4(x) 2 \sum\sum_{l \neq m} g_{jl} g_{kl} g_{kl} g_{km}.$$

Writing, as in Box and Watson,

$$S_4^k = \sum_{l=1}^{1} g_{kl}^4 \, ; \quad S_2^k = \sum_{l=1}^{q} g_{kl}^2$$

and

$$S_2^{kl} = \sum_{m=1}^{q} g_{km}^2 g_{lm}^2 \, ; \quad S_1^{kl} = \sum_{m=1}^{q} g_{km} g_{lm}$$

we have finally, assembling these expressions,

$$\delta_{zz} = n^{-1} \{ p(P + 2) + (k(x) - 3) \left[\sum_{k=1}^{p} \frac{S_4^k}{(S_2^k)^2} + \sum\sum_{k \neq l} \frac{S_2^{kl}}{S_2^k S_2^l} \right]$$

$$+ 2 \sum\sum_{k \neq l} (S_1^{kl})^2 / (S_2^k S_2^l) \}.$$