NEARLY EFFICIENT MOMENT RESTRICTION ESTIMATION OF REGRESSION MODELS WITH NONNORMAL DISTURBANCES

Whitney K. Newey

Econometric Research Program
Research Memorandum No. 315

August 1984

This paper is based upon work partially supported by the National Science Foundation under Grant No. SES-8410249. Jerry Hausman, James Powell, Helaman Pratt-Ferguson, Kenneth Singleton, Ken West, and participants at the Harvard-MIT and Princeton econometrics seminars provided helpful comments.

Econometric Research Program
PRINCETON UNIVERSITY
207 Dickinson Hall
Princeton, NJ 08544

ABSTRACT

This paper considers generalized method of moments estimators which use disturbance moment restrictions in nonlinear regression models when the disturbance distribution has an unknown form. A convenient linearized estimator is proposed. It is shown that if the unknown density satisfies certain tail behavior restrictions, then the asymptotic variance of the estimator approaches the Cramer-Rao bound in two cases. The first case involves independently and identically distributed disturbances and the second case involves a disturbance which is symmetrically distributed conditionally on the regressors. Results of a sampling experiment which investigates the finite sample performance of the proposed estimators are also presented.

1. INTRODUCTION

If the form of the distribution of the disturbance is known maximum likelihood can be used to estimate the parameters of a regression model. When the disturbance has an unknown distribution it is often the case that other kinds of information can be used to estimate the parameters. This paper considers estimation of nonlinear regression models using disturbance moment restrictions. For example, when the disturbance is independently and identically distributed (i.i.d.), functions of the disturbance will be uncorrelated with the regressors, and this fact can be used to form generalized method of moments (GMM; Hansen, 1982) estimators of the regression parameters. This approach to estimation of regression models is similar to that taken by MaCurdy (1982), who considers GMM estimators which use first, second, and higher-order raw moments of the dependent variable.

Failure of the disturbance to be normally distributed can have serious consequences for the efficiency of least squares. The efficiency cost of using least squares in the presence of nonnormal disturbances has been well documented in the robustness literature, for example in Huber (1981) and the references cited therein. When there is no reason to believe that the disturbance is normally distributed it is prudent to use estimators other than least squares. Recently it has been shown by Bickel (1982) and Manski (1984) that when the disturbance is i.i.d. it is possible to obtain an adaptive estimator of the slope coefficients of a regression model, that is, an estimator which does not use any knowledge of the density of the disturbance but which is as efficient, asymptotically, as the maximum likelihood estimator would be if the distribution of the disturbance were known. Such an estimator provides an efficient alternative to least

squares. In this paper it is shown that GMM estimators which use disturbance moment restrictions also provide efficient alternatives to least squares, and do so in a larger variety of circumstances than those which allow the use of the adaptive estimators of Bickel (1982) and Manski (1984). Under two different sets of conditions it is possible to show that GMM estimators are nearly efficient, that is, as the number of moment conditions used in estimation grows the asymptotic distribution of the GMM estimator approaches the best attainable distribution. One set of conditions involves an i.i.d. disturbance. The other set of conditions involves a disturbance which is symmetrically distributed around zero conditionally on the regressors, and is possibly heteroskedastic, where the adaptive estimators of Bickel (1982) and Manski (1984) do not apply.

In Section 2 the asymptotic properties of estimators which use disturbance moment restrictions are discussed and a computationally convenient linearized GMM estimator is presented. In Section 3 the i.i.d. disturbance case is considered and Section 4 deals with the symmetric case. Section 5 presents the results of a sampling experiment which shows that efficiency gains can be achieved in finite samples using a linearized GMM estimator, and that these estimators compare very favorably with adaptive estimators. Section 5 offers some conclusions and directions for future research.

2. LINEARIZED MOMENT CONDITION ESTIMATION

In this section notation will be defined, the asymptotic properties of generalized method of moments estimators will be discussed, and a linearized, one-step estimator will be proposed. Consider the following nonlinear regression model.

Assumption 1: The sequence of observations (y_t, x_t) is independently distributed and satisfies

$$y_t = f(x_t, b_0) + e_t , \qquad (t = 1,2,\ldots) \qquad (1)$$

where y_t is a scalar, x_t is a $1\times p$ vector, b_0 is a $k\times 1$ vector, and $E(e_t | x_t) = 0$.

Note that since the conditional mean of e_t is zero, equation (1) gives a correctly specified regression model. Correct specification of the regression function will be a maintained assumption throughout this paper.

It is useful to impose regularity conditions on $f(x,b)$.

Assumption 2: The parameter vector b_0 is an element of the interior of a compact set B. For each x in X, $f(x,b)$ is twice continuously differentiable in b for b in B, and $f(x,b)$ and its derivatives are continuous in (x,b) on $X\times B$, where X is a compact set and $\text{Prob}(x_t \varepsilon X) = 1$, $(t = 1,2,\ldots)$.

The continuity of $f(x,b)$ in x and uniform boundedness of the support of x_t are stronger than needed for the asymptotic distribution theory of this section, but will prove to be useful in the discussion of asymptotic efficiency in the heteroskedastic case of Section 4.

In the absence of exact a priori knowledge of the form of the conditional distribution of the disturbance, moment restrictions can be used to estimate b_0. For example, the restriction that the conditional mean of the disturbance is zero can be used to estimate b_0 by nonlinear least squares (NLS). More generally, consider moment restrictions of the following form.

Assumption 3: The sequence (x_t, e_t) satisfies

$$E[p_j(e_t, a_0) | x_t] = 0 , \quad (j = 1,\ldots,J;\ t = 1,2,\ldots) \qquad (2)$$

where a_0 is a $\ell \times 1$ vector of parameters.

Equation (2) says that certain functions of a known form have conditional expectation zero. For example, Assumption 1 implies that equation (2) is satisfied for $p_0(e) = e$. More generally, these functions will be allowed to depend on some parameters a_0 which may be estimated along with the regression parameters. Additional specific examples will be given in Sections 3 and 4.

It is useful to assume that each moment function $p_j(e,a)$ is regular, in the following sense. Let

$$p_{je}(e,a) = \partial p_j(e,a)/\partial e , \quad p_{ja}(e,a) = \partial p_j(e,a)/\partial a ,$$

$$M = \sup_{X\times B\times B} |f(x,b) - f(x,b')| , \quad I = [-M,M] .$$

Assumption 4: For each j, $p_j(e,a)$ is continuously differentiable on $R\times A$. The vector a_0 is an element of the interior of a compact set A and for each j there exist finite constants δ, $N>0$, such that

$$E(\sup_{I\times A}|p_j(e_t + m,a)|^{2+\delta}) \leq N , \quad E(\sup_{I\times A}|p_{je}(e_t + m,a)|^{1+\delta}) \leq N$$

$$E(\sup_{I\times A}|p_{ja}(e_t + m,a)|^{1+\delta}) \leq N , \qquad (t = 1,2,\ldots) .$$

Throughout most of this paper it will be the case that the conditional mean restriction $E(e_t|x_t) = 0$ is used in estimation, so that it is useful to have available a hypothesis which will guarantee that Assumption 4 is satisfied for $p_0(e) = e$.

Assumption 5: There exists a finite constant, $N > 0$, such that

$$E(|e_t|^{2+\delta}) < N , \qquad (t = 1,2,\ldots) .$$

The conditional moment restrictions of Assumptions 1 and 3 imply that for each j, $p_j(e_t,a_0)$ will be uncorrelated with functions of x_t. A generalized method of moments (GMM) estimator (e.g. Hansen, 1982) can therefore be formed by minimizing a quadratic form in sample averages of products of $p_j(y_t - f(x_t,b),a)$ with functions of x_t. Specifically, let

$$\theta = (b',a')' , \quad \textcircled{H} = B \times A ,$$

and let $z(x,\theta)$ be a $m\times 1$ vector of functions which satisfy the following assumption.

Assumption 6: For each x in X, $z(x,\theta)$ is a continuously differentiable function of θ on $\textcircled{H}$, $z(x,\theta)$ is measurable in x for each θ, and

$$\sup_{X\times\textcircled{H}}|z(x,\theta)| < +\infty , \quad \sup_{X\times\textcircled{H}}|\partial z(x,\theta)/\partial\theta| < +\infty .$$

Let the sample size be denoted by n, and let

$$p(e,a) = (e,p_1(e,a), \ldots , p_J(e,a))',$$

$$h(e,x,\theta) = p(e,a) \otimes z(x,\theta),$$

$$h_n(\theta) = \sum_{t=1}^{n} h(y_t - f(x_t,b),x_t,\theta)/n,$$

so that $h_n(\theta)$ is a $(J + 1)\cdot m$ dimensional vector of sample moments. Note that J is a positive integer which equals the number of functions in addition to $p_0(e) = e$ which are used to form $h_n(\theta)$. A GMM estimator $\hat{\theta}$ of $\theta_0 = (b_0',a_0')$ can now be obtained by solving

$$\min_H h_n(\theta)'D_n h_n(\theta) \tag{3}$$

where D_n is a conformable positive semi-definite matrix.

It is possible to proceed to obtain the asymptotic distribution of the GMM estimator $\hat{\theta}$ in a manner almost identical to that used by Hansen (1982) or Burgete, Gallant, and Souza (1982). However, a different and more convenient approach is available. This approach is based on obtaining an initial consistent estimator of θ_0 and linearization of the GMM minimization problem around the initial consistent estimator.

An initial estimator of the regression parameters b_0 is readily available, namely NLS. Under an additional identification assumption, White's (1980) analysis of the asymptotic properties of NLS with independent observation yields the consistency and asymptotic normality of NLS in the context of this paper. Let

$$f_b(x,b) = \partial f(x,b)/\partial b, \quad Q = \lim \sum_{t=1}^{n} E[f_b(x_t,b_0)f_b(x_t,b_0)']/n,$$

where it is assumed here and below that, for notational convenience, moment matrices converge. It will also be assumed that b_0 is identified.

Assumption 7: For any neighborhood C of b_0 there exists finite constants δ, $n_0 > 0$, such that

$$\sup_{B\backslash C} \sum_{t=1}^{n} E[(f(x_t,b) - f(x_t,b_0))^2]/n \geq \delta$$

if $n \geq n_0$. Also Q is nonsingular.

Let $\tilde{b}$ be the NLS estimator obtained from solving

$$\min_{B} \sum_{t=1}^{n} (y_t - f(x_t,b))^2. \tag{4}$$

Proposition 2.1: If Assumptions 1-7 are satisfied then $\sqrt{n}(\tilde{b}-b_0)$ converges in distribution to a multivariate normal random vector.

This and the other propositions of this section will not be proved here, since the asymptotic theory is straightforward. An appendix which contains these proofs is available from the author upon request.

In the general formulation presented here it is not as obvious how an initial consistent estimator $\tilde{a}$ of a can be obtained. One method which should work in the general case to obtain $\tilde{a}$ is to solve

$$\min_{A} h_n(\tilde{b},a)'\tilde{D}_n h_n(\tilde{b},a) \tag{5}$$

for some choice of $\tilde{D}_n$. That is, by solving the GMM minimization problem (3) for a only, while replacing b by $\tilde{b}$. In the context of the specific situations considered later, simple methods of obtaining initial consistent estimators will be available. For now, the following assumption will be imposed.

Assumption 8: An estimator $\tilde{a}$ exists such that $\sqrt{n}(\tilde{a}-a_0)$ converges in distribution .

In order to see how the minimization problem (3) can be linearized at the initial consistent estimator $\tilde{\theta} = (\tilde{b}',\tilde{a}')'$, so that a one-step GMM estimator can be obtained, consider the first order Taylor's expansion

$$z(x_t,\theta)p_j(y_t - f(x_t,b),a) = z(x_t,\tilde{\theta})[p_j(\tilde{e}_t,\tilde{a}) - p_{je}(\tilde{e}_t,\tilde{a})f_b(x_t,\tilde{b})'(b-\tilde{b})$$

$$+ p_{ja}(\tilde{e}_t,\tilde{a})'(a-\tilde{a})] + r_{jt} = z(x_t,\tilde{\theta})(\tilde{p}_{jt} - W_{jt}\gamma) + r_{jt} \qquad (6)$$

where

$$\tilde{e}_t = y_t - f(x_t,\tilde{b}), \quad \tilde{p}_{jt} = p_j(\tilde{e}_t,\tilde{a}),$$

$$W_{jt} = [p_{je}(\tilde{e}_t,\tilde{a})f_b(x_t,\tilde{b})',-p_{ja}(\tilde{e}_t,\tilde{a})'], \gamma = (\theta-\tilde{\theta}),$$

and r_{jt} is a remainder term which includes $\partial z(x_t,\tilde{\theta})/\partial\theta\cdot\tilde{p}_{jt}(\theta-\tilde{\theta})$ as well as higher order terms. The term $\partial z(x_t,\tilde{\theta})/\partial\theta\cdot\tilde{p}_{jt}(\theta-\tilde{\theta})$ will turn out to be negligible, asymptotically, because $\tilde{\theta}$ is a consistent estimator of θ_0 and

$$E[\partial z(x_t,\theta_0)/\partial\theta\cdot p_j(e_t,a_0)] = E[\partial z(x_t,\theta_0)/\partial\theta\cdot E[p_j(e_t,a_0)\ x_t]] = 0.$$

Averaging equation (6) over the observations yields

$$\sum_{t=1}^{n} z(x_t,\theta)p_j(y_t - f(x_t,b),a)/n = \tilde{Z}'(\tilde{p}_j - W_j\gamma)/n + u'r_j/n,$$

where u is an n×1 vector of ones and

$$\tilde{Z} = [z(x_1,\tilde{\theta})',\ldots,z(x_n,\tilde{\theta})']', \quad \tilde{p}_j = (\tilde{p}_{j1},\ldots,\tilde{p}_{jn})',$$

$$W_j = [W_{j1}',\ldots,W_{jn}']', \quad r_j = (r_{j1},\ldots,r_{jn})'.$$

Stacking by j, (j = 0,1,...,J), gives

$$h_n(\theta) = Z'(\tilde{p} - W\gamma)/n + (I_{J+1} \otimes u)'r/n$$

where

$$Z = I_{J+1} \otimes \tilde{Z}, \quad \tilde{p} = (\tilde{p}_0', \ldots, \tilde{p}_J')$$

$$W = [W_0', \ldots, W_J']', \quad r = (r_0', \ldots, r_J')' .$$

By ignoring the remainder term $(I_{J+1} \otimes u)'r/n$, a version of the GMM minimization problem (3) which is linearized around $\tilde{\theta}$ can be obtained. Let $\hat{\theta}$ solve

$$\min_{\theta} (p - W\gamma)'ZD_nZ'(p - W\delta).$$

Then $\hat{\theta}$ is given by

$$\hat{\theta} = \tilde{\theta} + \hat{\gamma} , \quad \hat{\gamma} = (W'ZD_nZ'W)^{-1}W'ZD_nZ'\tilde{p} . \tag{8}$$

That is, the estimator $\hat{\theta}$ is equal to the initial consistent estimator $\tilde{\theta}$ plus a step $\hat{\gamma}$. The step $\hat{\gamma}$ is formed by what is essentially one Newton-Raphson iteration from $\tilde{\theta}$ toward the solution of (3), with asymptotically negligible terms deleted. Note that the step has a form which is familiar to most econometricians. The step $\hat{\gamma}$ equals an instrumental variables estimator of a system of equations,

$$\tilde{p}_j = W_j\gamma + r_j , \quad (j = 0,\ldots,J) \tag{9}$$

with instrumental variables $\tilde{Z}$ for each equation, γ constrained to be equal across equations, and D_n used as a distance matrix. This is an arcane sort of instrumental variables estimator, since the variables x_t which appear in the original regression equation (1) are exogenous, but this instrumental variables interpretation facilitates computation, as illustrated below.

It remains to show that the remainder $Z'r/n$ is asymptotically negligible when θ equals $\hat{\theta}$, and to derive the asymptotic covariance matrix of $\hat{\theta}$. In order to do this it is helpful to define some further notation. Let

$$
\begin{aligned}
v_t &= (e_t, p_1(e_t, a_0), \ldots, p_J(e_t, a_0))' \\
\Sigma(x_t) &= E[v_t v_t' | x_t] \\
P(x_t) &= -E[(1, p_{1e}(e_t, a_0), \ldots, p_{Je}(e_t, a_0))' | x_t] \\
R(x_t) &= E[[p_{0a}(e_t, a_0), \ldots, p_{Ja}(e_t, a_0)]' | x_t] \\
\Omega &= \lim \sum_{t=1}^{n} E[\Sigma(x_t) \otimes z(x_t, \theta_0) z(x_t, \theta_0)']/n \\
H_b &= \lim \sum_{t=1}^{n} E[P(x_t) \otimes z(x_t, \theta_0) f_b(x_t, b_0)']/n \\
H_a &= \lim \sum_{t=1}^{n} E[R(x_t) \otimes z(x_t, \theta_0)]/n, \quad H = [H_b, H_a].
\end{aligned}
\tag{10}
$$

An extra t subscript on Σ, P, and R is dropped for notational convenience. In the specific situations considered in Sections 3 and 4, the conditional distribution of e_t given x_t will be the same for each t.

One additional assumption, which is essentially a local identification condition, is required.

Assumption 9: plim $D_n = D$ and $H'DH$ is nonsingular.

Theorem 2.3: If Assumptions 1-9 are satisfied, then $\sqrt{n}(\hat{\theta}-\theta)$ converges in distribution to a multivariate normal random vector with covariance matrix

$$(H'DH)^{-1}H'D\Omega DH(H'DH)^{-1}.$$

Inspection of the asymptotic covariance matrix of $\hat{\theta}$ and the work of Hansen (1982) and Burgete, Gallant, and Souza (1982) leads to the conclusion that $\hat{\theta}$ has the same asymptotic distribution as the GMM estimator $\bar{\theta}$ obtained from solving (3). In fact, it is the case that $\sqrt{n}(\bar{\theta}-\hat{\theta})$ converges to zero in probability, so that the linearized estimator $\hat{\theta}$ is asymptotically equivalent to the GMM estimator $\bar{\theta}$ obtained from solving the complicated nonlinear minimization problem (3).

The asymptotic ovariance matrix of $\hat{\theta}$ depends on the choice of distance matrix D_n. As shown by Hansen (1982), this covariance matrix is minimized (in the positive semi-definite sense) when $D = \Omega^{-1}$. When $D = \Omega^{-1}$, the asymptotic covariance matrix of $\hat{\theta}$ reduces to

$$V = (H'\Omega^{-1}H)^{-1}.$$

To compute an optimal $\hat{\theta}$ with the smallest covariance matrix V, $D_n = \hat{\Omega}_n^{-1}$ can be chosen, where $\hat{\Omega}_n$ is a consistent estimator of Ω. Such an estimator is readily available. Let

$$\hat{\Omega}_n = Z'(\text{diag}[\tilde{p}_{01}^2,\ldots,\tilde{p}_{0n}^2,\ldots,\tilde{p}_{J1}^2,\ldots,\tilde{p}_{Jn}^2])Z/n.$$

The matrix $\hat{\Omega}_n$ is the system analogue of White's (1980) heteroskedasticity consistent covariance matrix.

Proposition 2.3: If Assumptions 1-8 are satisfied, then

$$\text{plim } \hat{\Omega}_n = \Omega.$$

When $D_n = \hat{\Omega}_n^{-1}$ is used to form $\hat{\theta}$, the step $\hat{\gamma}$ is given by

$$\hat{\gamma} = (W'Z\hat{\Omega}_n^{-1}Z'W)^{-1}W'Z\hat{\Omega}_n^{-1}Z'\tilde{p}.$$

The step $\hat{\gamma}$ then has the form of the generalization of three stage least squares (3SLS) to the heteroskedastic case which was obtained by Chamberlain (1982). The initial estimate of γ in equation (9) which is used here to form $\hat{\Omega}_n$ is the zero vector.

When x_t is independent of e_t, and e_t is i.i.d., the computation of $\hat{\gamma}$ can be further simplified to simple 3SLS. Independence of x_t and e_t implies that $\Sigma(x_t)$, $P(x_t)$, and $R(x_t)$ are each constant. Let

$$\tilde{p}_t = (p_0(\tilde{e}_t,\tilde{a}),\ldots,p_J(\tilde{e}_t,\tilde{a}))', \qquad \tilde{\Sigma} = \sum_{t=1}^{n} \tilde{p}_t\tilde{p}_t'/n,$$

$$\tilde{P}_j = -\sum_{t=1}^{n} p_{je}(\tilde{e}_t,\tilde{a})/n, \qquad \tilde{R}_j = \sum_{t=1}^{n} p_{ja}(\tilde{e}_t,\tilde{a})/n,$$

$$\tilde{W}_{jb} = \tilde{P}_j[f_b(x_1,\tilde{b}),\ldots,f_b(x_n,b)]', \qquad \tilde{W}_j = [W_{jb},\tilde{R}_j' \otimes u],$$

where u is the $n\times 1$ vector which has 1 for each element. In this case, with independence of e_t and x_t, an alternative method of forming $\hat{\gamma}$ is

$$\hat{\gamma} = (\tilde{W}'(\tilde{\Sigma}^{-1}\otimes\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}')\tilde{W})^{-1}\tilde{W}'(\tilde{\Sigma}^{-1}\otimes\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}')\tilde{p}, \tag{11}$$

where $\tilde{W} = [\tilde{W}_0',\ldots,\tilde{W}_J']$. This step size is equal to the 3SLS estimator of γ in the equation system

$$\tilde{p}_j = \tilde{W}_j\gamma + \tilde{r}_j, \qquad (j = 0,\ldots,J), \tag{12}$$

where the initial estimator of γ used to form the disturbance covariance matrix estimator $\tilde{\Sigma}$ is the zero vector, and where γ is constrained to be equal across equations. Such estimators are readily available in several standard regression packages, such as TSP.

It is straightforward to obtain a consistent estimator of the asymptotic covariance matrix of $\hat{\theta}$ when the optimal distance matrix, $D_n = \hat{\Omega}_n^{-1}$ is used. One possibility for a consistent estimator of V is

$$\hat{V} = T^2(W'Z\hat{\Omega}^{-1}Z'W)^{-1}.$$

Except for the factor T^2, this matrix is used to form Chamberlain's (1982) generalization of 3SLS, and $\hat{V}$ is thus easily computed along with $\hat{\gamma}$. It is also consistent under the above assumptions.

Proposition 2.4: If Assumptions 1-9 are satisfied and Ω is nonsingular then

$$\text{plim } \hat{V} = V.$$

In the case when e_t and x_t are mutually independent, a consistent estimator of V can be obtained via the usual 3SLS formula. Let

$$\hat{V} = T(\tilde{W}'(\tilde{\Sigma}^{-1} \otimes \tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z})\tilde{W})^{-1}. \tag{13}$$

Except for the factor T this matrix is the usual formula for the 3SLS covariance matrix for the equation system (12) when $\gamma = 0$ is used to form the disturbance covariance matrix estimator. Further, since plim $\hat{\gamma}$ = plim$(\hat{\theta}-\tilde{\theta}) = 0$, the estimate of the disturbance covariance matrix which uses $\hat{\gamma}$ can also be used in place of $\tilde{\Sigma}$ to form $\hat{V}$. When this alternative estimate of Σ is used, the resulting $\hat{V}$ will be the usual 3SLS covariance matrix output by a regression package, except for the factor T, and will be consistent for $\hat{V}$.

The overidentifying restrictions implied by the moment conditions can be tested using a statistic of the form discussed by Hansen (1982). Let

$$\hat{p} = (p_0(\hat{e}_1,\hat{a}),\ldots,p_0(\hat{e}_n,\hat{a}),\ldots,p_J(\hat{e}_1,\hat{a}),\ldots,p_J(\hat{e}_n,\hat{a}))$$

where $\hat{e}_t = y_t - f(x_t,\hat{b})$ is the residual formed from the optimal one step estimator. Then our assumptions will be sufficient for the statistic

$$\hat{p}'Z\hat{\Omega}_n^{-1}Z'\hat{p}/n$$

to have an asymptotic chi-squared distribution with $(J+1)\cdot m - (\ell+k)$ degrees of freedom, while if the moment conditions of equation (2) are violated this statistic will often be far from zero.

3. INDEPENDENCE OF THE DISTURBANCE AND THE REGRESSORS

When there is no heteroskedasticity or any other form of dependence of the distribution of e_t on x_t and when e_t is identically distributed across observations, there are many conditional moment restrictions which can be used in estimation. Any function $c(e)$ with finite expectation yields, for $a_c = E[c(e_1)]$,

$$E[c(e_t) - a_c | x_t] = E[c(e_t)] - a_c = 0, \quad (t = 1,2,\ldots) \tag{14}$$

Since independence generates such a large class of conditional moment restrictions, one might conjecture that it will be possible to obtain a GMM estimator which is nearly efficient, even though the distribution of u_t is unknown, by choosing an appropriate set of moment conditions. In this section it is verified that a nearly efficient GMM estimator of the slope coefficients of a nonlinear regression model does exist when the unknown disturbance density satisfies certain tail behavior restrictions. Also, particular examples of moment functions will be presented and discussed.

Consider the following specialization of the nonlinear regression model to the independence case.

Assumption I1: Assumption 1 is satisfied, x_t and e_t are mutually stochastically independent, e_t has a density function $g(e)$ which is differentiable on the real line, and $b = (\alpha,\beta')'$, where α is a scalar and

$$f(x,b) = \alpha + f_1(x,\beta). \tag{15}$$

Besides specifying that the disturbance is i.i.d., this assumption restricts the unknown density function to be differentiable and the regression function

to include a constant. Both of these restrictions are important in the work on adaptive estimation by Bickel (1982) and Manski (1984). The differentiability of $g(e)$ is an important regularity condition and the inclusion of a constant is important for the existence of an adaptive estimator of the slope coefficients β.

Conditional moment restrictions which are generated by the i.i.d. disturbance can easily be put in the GMM estimation framework. Let $(c_1(e),\ldots,c_J(e))'$ be a vector of functions of the disturbance and let $a = (a_1,\ldots,a_J)'$ be a $J\times 1$ vector of constants. If a_0 is defined by

$$a_0 = E(c_1(e_1),\ldots,c_J(e_1))'$$

and $p_j(e,a)$ by

$$p_j(e,a) = c_j(e) - a_j, \qquad (j = 1,\ldots,J),$$

then the conditional moment restriction $E[p_j(e_t,a_0)|x_t] = 0$ is satisfied. In this context, the vector a_0 gives the expectations of the moment functions $c_j(e)$ evaluated at the true disturbance.

Since the i.i.d. disturbance case fits into the GMM estimation framework, one can think of implementing the one step GMM estimator by using the step $\hat{\gamma}$ in equation (11), which can be computed by using 3SLS. Several questions about such a procedure immediately come to mind. One question concerns conditions which will be sufficient for the one step GMM estimator to be asymptotically normal. The following assumption will imply that $p_j(e,a) = c_j(e)-a_j$ will satisfy the regularity conditions of Assumption 4.

Assumption I2: For each j, $c_j(e)$ is continuously differentiable on the real line and there exists $\delta > 0$ such that

$$E(\sup_I |c_j(e_1+m)|^{2+\delta}) < +\infty, \qquad E(\sup_I |c_j'(e_1+m)|^{1+\delta}) < +\infty.$$

A second question concerns a method of obtaining an initial estimator of $\theta_0 = (b_0', a_0')'$. As previously discussed, b_0 can be estimated by NLS. Also, since a_0 is the vector of expectations of the moment functions $c_j(e)$, a reasonable estimator of a is to choose

$$\tilde{a}_j = \sum_{t=1}^{n} c_j(\tilde{e}_t)/n, \qquad (j = 1,\ldots,J)$$

which is the sample mean of $c_j(e)$ evaluated at the NLS residuals. Such an estimator will be consistent and asymptotically normal under previous conditions.

A third question concerns the choice of "instruments" $z(x,\theta)$ to use in forming the one-step GMM estimator. One would like to choose $z(x,\theta)$ so that the asymptotic covariance matrix $V = (H'\Omega^{-1}H)^{-1}$ of the optimal one-step GMM estimator is as small as possible. The condition that e_t is i.i.d. can be used to derive an explicit, known form for the optimal $z(x,\theta)$. To see how this is done, note that the independence of x_t and e_t implies that V has the form of the asymptotic covariance matrix of a nonlinear 3SLS estimator of $\theta_0 = (b_0', a_0')'$ in the system of equations

$$p_j(y_t - f(x_t,b_0), a_0) = v_{jt}, \qquad (j = 0,\ldots,J). \tag{16}$$

By the exogeneity of x_t, and e_t i.i.d., Amemiya's (1977) derivation of the best nonlinear 3SLS implies that the optimal choice of "instruments" for the j^{th} equation should be

$$E[\partial p_j(y_t - f(x_t,b_0),a_0)/\partial\theta|x_t] \tag{17}$$

$$= E[-p_{je}(e_t,a_0)\cdot f_b(x_t,b_0)', p_{ja}(e_t,a_0)'|x_t]'$$

$$= (P_j\cdot f_b(x_t,b_0)', R_j'),$$

where $P_j = -E[p_{je}(e_1,a_0)]$ is a scalar and $R_j = E[p_{ja}(e_1,a_0)]$ is a $J\times 1$ vector. Since a constant is included in the regression, so that the first element of $f_b(x_t,b_0)$ is 1, the entire vector $(P_j\cdot f_b(x_t,b_0)', R_j')'$ is a linear combination of $f_b(x_t,b_0)$. Therefore an optimal choice of $z(x,\theta)$ should be, and is, equal to $f_b(x,b)$.

Finally, it would be interesting to know if it is possible to obtain a GMM estimator which is nearly efficient, and if so what conditions are sufficient for the existence of such an estimator. Manski (1984) has shown that Bickel's (1982) necessary conditions for adaptive estimation are not satisfied for the constant α but are satisfied for the slope coefficients β. Since there is thus no hope of obtaining an adaptive (or a nearly efficient) estimator of the constant, attention will be confined to analysis of the efficiency of the GMM estimator $\hat{\beta}$ of the slope coefficients.

To obtain an expression for the asymptotic covariance matrix of the slope coefficient estimator, note that e_t i.i.d. implies that $\Sigma(x_t)$, $P(x_t)$, and $R(x_t)$ are constant. Then substituting $f_b(x,b)$ for $z(x,\theta)$ in the definitions (10) gives

$$\Omega = \Sigma \otimes Q, \quad H_b = P \otimes Q, \quad H_a = R \otimes F, \quad H = [H_b, H_a]$$

where $F = \lim \Sigma_{t=1}^{n} E[f_b(x_t,b_0)]/n$. For the moment it will be assumed that Σ is nonsingular, although the specific forms of $c_j(e)$ to be considered

later will imply that Σ is nonsingular. Also, note that since $p_j(e,a) = c_j(e) - a_j$, $(j = 1,\ldots,J)$, and $p_0(e,a) = e$, it follows that the first row of R is zero, and for $i = 2,\ldots,J+1$ the i^{th} row of R has -1 in the i-1 position and zeros elsewhere. Therefore $R'\Sigma^{-1}R$ is also nonsingular. By partitioned inversion the asymptotic covariance matrix of $\hat{b}$, which is the upper-left k×k block of $V = (H'\Omega^{-1}H)^{-1}$, is given by

$$V(\hat{b}) = (P'\Sigma^{-1}P\cdot Q - P'\Sigma^{-1}R(R'\Sigma^{-1}R)^{-1}R'\Sigma^{-1}P\cdot FF')^{-1}.$$

By partitioned inversion once again the asymptotic covariance matrix of $\hat{\beta}$ is

$$V(\hat{\beta}) = (1/I_J)\cdot(Q_1 - F_1F_1')^{-1}$$

where $I_J = P'\Sigma^{-1}P$ and for $f_{1\beta}(x_t,\beta_0) = \partial f_1(x_t,\beta_0)/\partial\beta$

$$Q_1 = \lim \sum_{t=1}^{n} E[f_{1\beta}(x_t,\beta_0)f_{1\beta}(x_t,\beta_0)']/n,$$

$$F_1 = \lim \sum_{t=1}^{n} E[f_{1\beta}(x_t,\beta_0)]/n.$$

One useful property of the asymptotic covariance matrix of $\hat{\beta}$ is immediately apparent. Note that if a_0 were known rather than estimated the asymptotic covariance matrix of $\hat{b}$ would be

$$(H_b'\Omega^{-1}H_b)^{-1} = (I_J\cdot Q)^{-1}.$$

The lower right block of this matrix is $V(\hat{\beta})$. That is, estimation of the nuisance parameters a_0 does not affect efficiency of the slope coefficient estimator $\hat{\beta}$. This result in no way depends on the way in which a enters the

moment function $p(e,a)$. Consequently, there is a simple method of making a single adjustment. If is a preliminary estimate of the scale of e_t (e.g. the sample standard deviation) such that $\sqrt{n}(\tilde{\sigma}-\sigma_0)$ converges in distribution for some $\sigma_0 > 0$, then the one-step GMM estimator of the slope coefficients which is formed by using $c_j(e/\hat{\sigma})$ will have the same asymptotic distribution as the GMM estimator with $c_j(e/\sigma_0)$. For brevity, a formal proof is omitted.

Another useful property of $V(\hat{\beta})$ is that it has a simple relationship to the Cramer-Rao (CR) lower bound. The CR bound referred to here means the asymptotic covariance matrix of the maximum likelihood estimator which would be obtained if the distribution of the disturbance was known. To compute the CR bound, it will be assumed for the moment that $g(e)$ has an associated finite information constant. Let $s(e) = g'(e)/g(e)$ for $g(e)$ positive and $s(e) = 0$ otherwise, and let $I^* = E[s(e_1)^2]$. When $g(e)$ is known the maximum likelihood estimator (MLE) of b_0 is obtained by solving

$$\max_b \sum_{t=1}^{n} \ln g(y_t - f(x_t,b))$$

The asymptotic information matrix is therefore equal to $I^* \cdot Q$ and the asymptotic CR bound for the slope coefficients is the lower right-hand k-1 dimensional block of the inverse information matrix, which is

$$V^* = (1/I^*) \cdot (Q_1 - F_1F_1')^{-1}. \qquad (20)$$

Note that the difference between $V(\hat{\beta})$ and V^* depends only on the scalars I_J and I^*.

The asymptotic efficiency of $\hat{\beta}$ is determined by how close I_J is to I^*. There is an enlightening interpretation of I_J which clarifies its relationship to I^*. Since both $E[c_j(e_1)^2]$ and $E[s(e_1)^2]$ are finite, $E[c_j(e_1)] = \int c_j(e)g(e)de$ is finite and integration by parts gives

$$E[c_j'(e_1)] = \int c_j'(e)g(e)de = -\int c_j(e)g'(e)de = -E[c_j(e_1)s(e_1)]. \qquad (21)$$

Also, by the differentiability of $g(e)$ and $\int g(e) = 1 < +\infty$ (which implies $g(\infty) = g(-\infty) = 0$)

$$E[s(e_1)] = \int g'(e)de = \lim_{r\to\infty} \int_{-r}^{r} g'(e)de = \lim_{r\to\infty}[g(r)-g(-r)] = 0. \qquad (22)$$

Combining these two equations, it follows that

$$P = -E[v_1 s(e_1)]. \qquad (23)$$

From equation (23)

$$I_J = P'\Sigma^{-1}P = (-P'\Sigma^{-1})\Sigma(-\Sigma^{-1}P) = d^{*\prime}\Sigma d^* = E[(v_1'd^*)^2]$$

where $d^* = E[(v_1 v_1')]^{-1} E[v_1 s(e_1)]$ is the vector of coefficients in the linear projection of $s(e_1)$ on v_1. That is, I_J is the variance of the best linear prediction $v_1'd^*$ of the score $s(e_1)$, so that

$$I^* - I_J = I^* - d^{*\prime}\Sigma d^* = I^* - 2E[s(e_1)v_1']d^* + E[(v_1'd^*)^2] \qquad (24)$$

$$= \min_d E[(s(e_1) - v_1'd)^2].$$

The asymptotic efficiency of $\hat{\beta}$, relative to the CR bound, is determined by the goodness of fit of the best linear predictor $v_1'd^*$ to the disturbance score $s(e_1)$.

The one-step GMM estimator $\hat{\beta}$ will be asymptotically efficient if and only if $s(e_1)$ is a linear combination of v_1. For example, if the disturbance is normally distributed, then $s(e)$ is proportional to the first element of v_1, which is $p_0(e,a_0) = e$, and $\hat{\beta}$ is asymptotically efficient. Generally, though, it will not be possible to guarantee that $s(e_1)$ is a linear combination of v_1 when the form of the disturbance distribution is unknown. But one might suspect that by letting J grow or, in other words, increasing the number of moment functions, it might be possible to make $\hat{\beta}$ come arbitrarily close to being asymptotically efficient. Equation (24) provides the key to an analysis of this conjecture. If the moment functions are elements of a sequence $\{c_j(e);\ j = 1,2,\ldots\}$ which forms a basis for the Hilbert space of random variables (measurable functions of e_1) with finite second moment, or in other words the sequence $\{c_j(e_1)\}$ is complete, then a nearly efficient GMM estimator can be obtained by letting J grow.

The completeness of the sequence $\{c_j(e)\}$ requires some discussion in the current context because the disturbance density can be positive on the entire real line. The viewpoint of this paper is that the form of the distribution of the disturbance is unknown, while the moment functions are known and are used in estimation. As far as asymptotic efficiency is concerned, it would be useful to know what sort of moment functions would turn out to yield a complete sequence for a wide variety of disturbance distributions. For example, if the moment functions consist of powers of e, $c_j(e) = e^{j+1}$, $(j \geq 1)$, then the sequence $\{c_j(e_1)\}$ will fail to be complete when some raw moments of the disturbance do not exist (i.e. $E(|e_1|^J)$ is infinite for some positive integer J). In general, choice of a particular sequence of moment functions will imply that the GMM estimator is nearly efficient when the unknown disturbance distribution belongs to a class of unknown distributions. A complete characterization of this class of disturbance distributions will not be given here for any sequence of moment functions. It is possible to obtain a characterization of a subset of the class of distributions for which a nearly efficient GMM estimator exists when the sequence of moment functions has a particular kind of form.

Consider functions which have the form

$$c_j(e) = [c(e)]^j \qquad (j = 1,2,\ldots) \tag{25}$$

where c(e) satisfies the following assumption.

Assumption I3: The function c(e) is continuously differentiable and $c'(e) > 0$ on the entire real line.

The functions of equation (25) are powers of a monotonic function. It is useful to note that when the moment functions take this form, nonsingularity of the covariance matrix Σ of the moment functions is, essentially, a consequence of the disturbance having a continuous distribution. Let $\tilde{d}$ have a zero for a first element. Then

$$\tilde{d}'\Sigma\tilde{d} = E[(\sum_{j=1}^{J} \tilde{d}_{j+1}\{[c(e_1)]^j - a_{0j}\})^2] > 0$$

because $c(e)$ is strictly monotonic and any nonzero polynomial has at most a finite number of roots. Nonsingularity of Σ then follows as long as e_1 is not a linear combination of a finite number of powers of $c(e_1)$.

It is well known that the sequence of functions $\{u^j \exp(-u^2/2)\}$, linear combinations of which form Hermite polynomials, provide a basis for $L_2(-\infty,+\infty)$, the space of square integrable functions on the real line. A slight generalization of this fact can be used to obtain a characterization of a subset of unknown densities for which a nearly efficient GMM estimator exists when powers of $c(e)$ are used as moment functions.

Theorem 3.1: If Assumptions I1-I3 are satisfied and there exist finite constants $\gamma > 1$, N, $\delta > 0$ such that

$$g(e) \leq Nc'(e)\exp(-\delta|c(e)|^{\gamma}), \quad \int s(e)^2 c'(e)\exp(-\delta|c(e)|^{\gamma})de < +\infty , \qquad (26)$$

then $\lim_{J\to\infty} I_J = I^*$.

Proof: By hypothesis $c(e)$ is invertible; its range is an open interval. By a change of variables (Halmos, 1950, p. 164) $u = c(e)$ and, by (26),

$$\int s(c^{-1}(u))^2 \exp(-\delta|u|^{\gamma})du = \int s(e)^2 \exp(-\delta|c(e)|^{\gamma})c'(e)de < +\infty , \qquad (27)$$

where the first integral is over the range of c(e). The inequality $\gamma > 1$ allows a simple modification of the proof of Theorem 5, p. 57, of Helmberg (1965) to show that the functions

$$u^j \exp(-\frac{\delta}{2}|u|^{\gamma}), \qquad j \geq 0$$

form a basis for the Hilbert space of square (Lebesgue) integrable functions on the range of c(e). By equation (27) $s(c^{-1}(u))\exp(-\frac{\delta}{2}|u|^{\gamma})$ is such a square integrable function. Therefore, for each J there exists $\tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_J$ (an additional J subscript is suppressed for convenience) such that

$$0 = \lim_{J\to\infty}\int [s(c^{-1}(u))\cdot\exp(-\frac{\delta}{2}|u|^{\gamma}) - \sum_{j=0}^{J}\tilde{d}_j u^j \exp(-\frac{\delta}{2}|u|^{\gamma})]^2 du \qquad (28)$$

$$= \lim_{J\to\infty}\int [s(c^{-1}(u)) - \sum_{j=0}^{J}\tilde{d}_j u^j]^2 \exp(-\delta|u|^{\gamma})du$$

$$= \lim_{J\to\infty}\int [s(e) - \sum_{j=0}^{J}\tilde{d}_j c(e)^j]^2 \exp(-\delta|c(e)|^{\gamma})c'(e)de ,$$

where the last equality is obtained by another change of variables. To account for the presence of a constant term $\tilde{d}_0$ in equation (28), note that by $E[s(e_1)] = 0$, the least squares projection of $s(e_1)$ on v_1 equals the least squares projection of $s(e_1)$ on $(1, e_1, c_1(e_1), \ldots, c_J(e_1))'$. Then, by the dominance of g(e) in equation (26)

$$0 \leq I^* - I_J = \min_{d} E[(s(e_1) - v_1'd)^2] \qquad (29)$$

$$= \min_{d_{-1}, d_0, \ldots, d_J} \int [s(e) - d_{-1} - d_0 e - \sum_{j=1}^{J} d_j c(e)^j]^2 g(e)de$$

$$\leq \int [s(e) - \sum_{j=0}^{J}\tilde{d}_j c(e)^j]^2 g(e)de$$

$$\leq N \int [s(e) - \sum_{j=0}^{J} \tilde{d}_j c(e)^j]^2 \exp(-\delta|c(e)|^\gamma) c'(e) de$$

The conclusion then follows from equation (28).

The hypotheses of this theorem restrict the tail behavior of the unknown distribution of the disturbance. The first inequality in (26) says that the disturbance density must decline to zero at least as fast as the function $c'(e)\exp[-\delta|e|^\gamma]$ as $|e|$ increases. The second inequality restricts the rate of growth of the disturbance score $s(e)$. In order to further interpret the meaning of condition (26) and to see what choices of moment functions might work well in practice, it is useful to examine some particular forms of moment functions.

Consider choosing moment functions to be positive integer powers of e,

$$c_j(e) = e^{j+1}, \quad j \geq 1, \quad p_0(e) = e.$$

Using moment functions which are powers of e amounts to using information about raw moments of the disturbance to help in estimating the regression parameters. MaCurdy's (1982) method will be asymptotically equivalent to the one-step GMM estimator proposed in Section 2, when the heteroskedasticity which is induced by using raw moments of the dependent variable is corrected for.

When powers of the disturbance are used so that $c'(e) = 1$, the conditions (26) for the existence of a nearly efficient GMM estimator become

$$g(e) \leq N\exp(-\delta|e|^\gamma), \quad \int s(e)^2 \exp(-\delta|e|^\gamma) de < +\infty \qquad (30)$$

The first inequality is quite strict, specifying that the unknown density

has tails which decline exponentially fast. Such a restriction is a natural sufficient condition for the existence of all raw moments. The second inequality is not very strict, specifying that $s(e)^2$ grows at less than an exponential rate as $|e|$ grows. Most familiar families of distributions would satisfy this restriction. It would be interesting to know what other choices of moment functions would lead to looser restrictions on the "thickness" of the tail of the unknown distribution without making the second condition too onerous.

There are some heuristic reasons to think that choosing something other than powers of the disturbance as moment functions might be a good thing to do in many situations. The kinds of departures from normality which are often of interest and those which appear to have the most serious consequences for the efficiency of least squares involve thick-tailed distributions. In such cases higher-order moments are notoriously difficult to estimate, and in finite samples the efficiency of the estimator of the slope coefficients may be adversely affected. Also, in terms of the asymptotic efficiency theory, using polynomials to approximate the disturbance score, as is done indirectly through the GMM estimation when powers of e are used, does not seem like a good idea where thick-tailed distributions are a concern. In most familiar families with thicker tails than the normal distribution, such as the t distribution, $|s(e)|$ grow less rapidly with $|e|$ than $|e|$. Polynomials are particularly poor approximants to such functions in the tails, where particularly heavy weight is given in the mean square error calculation in thick-tailed cases. Finally, the GMM estimator is likely to be quite sensitive to outliers in the residuals since the moment conditions involve residuals raised to positive integer powers.

One way to try to deal with the potential of thick-tailed departures from normality is to choose c(e) to be a bounded function. When c(e) is bounded, $\exp(-\delta|c(e)|^{\gamma})$ is bounded away from zero and the condition (26) for the existence of a nearly efficient GMM estimator reduces to

$$g(e) \leq Nc'(e), \quad \int s(e)^2 c'(e)de < +\infty.$$

To interpret these inequalities, suppose for the moment that c(e) is a cumulative distribution function which has positive density on the entire real line. Then the first inequality says that the unknown disturbance distributions have no thicker tails than the distribution c(e). The second inequality requires that $s(e)^2$ have finite expectation if the disturbance were actually distributed as c(e). To be specific, suppose that c(e) is the cymulative distribution function of a t-distribution with r degrees of freedom. Then the first of these inequalities will be satisfied as long as $g(e)|e|^{r+1}$ is bounded and $s(e)^2/|e|^{r+1}$ is integrable. For $r \geq 2$ these restrictions will be satisfied for most of the common families of distributions, including the normal distribution, the t-distribution (with at least two degrees of freedom), and the Box-Tiao family. A function which has similar tail behavior to the t-distribution with r degrees of freedom, but which is easier to compute and therefore might prove useful in practice, is

$$c(e) = \text{sign}(e)[(1+|e|)^r - 1]/[(1+|e|)^r + 1], \qquad (31)$$

where sign(e) equals 1 for $e \geq 0$ and equals -1 otherwise.

There are certainly other choices of moment functions which would be interesting to consider. This subject remains an important topic for future research. In Section 5, the relative merits of using raw moments of the

disturbance, or a bounded function like c(e) in equation (31), will be examined in some Monte Carlo experiments.

4. THE HETEROSKEDASTIC CASE

When the distribution of the disturbance is symmetrically distributed around zero, conditionally on the regressors, the assumption that the disturbance is i.i.d. can be relaxed. Remarkably, it is possible to obtain a GMM estimator of the regression coefficients which is arbitrarily close to being as efficient, asymptotically, as the maximum-likelihood estimator would be if the entire conditional distribution of the disturbance were known. In this section, conditions sufficient for the existence of such a nearly efficient GMM estimator are obtained.

The symmetry hypothesis is important. Manski (1984) has shown that in the absence of symmetry adaptive estimation is not possible when heteroskedasticity is allowed for. Also, Chamberlain (1983) has provided lower bounds for the asymptotic covariance matrix of estimators which utilize conditional moment restrictions and has shown that if the only a priori information available is that the disturbance has a conditional mean of zero, then the best that can be done is the generalized least squares estimator.

The symmetry hypothesis generates many conditional moment restrictions. Any odd function $c(e)$ (i.e. $c(-e) = -c(e)$) with finite expectation will satisfy

$$E[c(e_t)|x_t] = 0 \tag{32}$$

when the disturbance e_t is symmetrically distributed conditional on the regressor x_t.

Consider the following specialization of the nonlinear regression model to the symmetric case.

Assumption H1: Assumption 1 is satisfied, and for each t the distribution fo e_t conditional on x_t has a conditional density $g(e|x)$ which is symmetric around zero, $(g(-e|x) = g(e|x))$, for each x in X. Further, for each x in X, $g(e|x)$ is differentiable in e on the entire real and, for each e, $g(e|x)$ and $g'(e|x)$, $(= \partial g(e|x)/\partial e)$, are continuous in x.

Note that a hypothesis of Assumption H1 is that the conditional distribution of the disturbance is stationary across observations. This hypothesis does not rule out nonstationary observations since the distribution of x_t can vary across observations. For example, a sample which is stratified on the exogenous variables is allowed for. Also note that the regression function is not restricted to contain a constant term. The symmetry hypothesis means that a nearly efficient GMM estimator of all the regression coefficients will exist, as might be surmised from Manski's (1984) necessary conditions for adaptive estimation.

Conditional restrictions which are generated by conditional symmetry can easily be put in the GMM estimation framework. Let $(c_1(e),\ldots,c_J(e))$ be a vector of odd functions of the disturbance and let

$$p_j(e) = c_j(e), \quad (j = 1,\ldots,J), \quad p_0(e) = e.$$

To make sure that the regularity conditions for asymptotic normality are satisfied, the following assumption can be made.

Assumption H2: For each j, $c_j(e)$ is an odd function which is continuously differentiable on the real line and there exist finite constants N, $\delta > 0$, such that

$$E(\sup_I |c_j(e_t+m)|^{2+\delta}) < N, \quad E(\sup_I |c_j'(e_t+m)|^{2+\delta}) < N.$$

Note that no nuisance parameter vector appears in the moment functions specified above. It may still be desirable to have a scale parameter which adjusts for the scale of the disturbance. Suppose that σ_0 is a positive constant and $\tilde{\sigma}$ an estimator for which $\sqrt{n}(\tilde{\sigma}-\sigma_0)$ converges in distribution. Differentiating a scale adjusted moment function $c(e/\sigma)$ with respect to the scale parameter σ gives

$$dc(e/\sigma)/d\sigma = -c'(e/\sigma)(e/\sigma^2) = (-1/\sigma^2)c'(e/\sigma)e.$$

Since $c(e/\sigma)$ is odd, it follows that $c'(e/\sigma)e$ is also odd and that $E[c'(e_t/\sigma)e_t] = 0$. Therefore the usual kind of asymptotic Taylor's expansion of the first order conditions for the GMM estimator implies that an estimator with moment functions $c_j(e/\tilde{\sigma})$ would have the same asymptotic distribution as an estimator with moment functions $c_j(e/\sigma_0)$.

To obtain the asymptotic covariance matrix of the GMM estimator, consider the notation of equation (10). Note that the first element of $P(x_t)$ is 1, so that if $f_b(x,b)$ is included among the elements of $z(x,b)$, H_b will have full rank by the nonsingularity of Q. It will also be assumed that Ω is nonsingular. Since there is no nuisance parameter vector, the asymptotic covariance matrix of the optimal one-step GMM estimator is given by

$$V(\hat{b}) = (H_b'\Omega^{-1}H_b)^{-1}. \tag{33}$$

Unfortunately, in the heteroskedastic case the form of the optimal functions $z(x,b)$ of the exogenous variables which minimize $V(\hat{b})$ depends on the form of the unknown conditional distribution of the disturbance.

Chamberlain (1983) has derived a lower bound for the covariance matrix of any GMM estimator which uses a fixed set of conditional moment restrictions. It will be assumed for the moment that the conditional covariance matrix of the moment functions, $\Sigma(x_t)$, is nonsingular. Specific forms of moment conditions considered later will imply $\Sigma(x_t)$ is nonsingular. To state Chamberlain's (1983) lower bound for the case considered here, let

$$I_J(x) = P(x)'\Sigma(x)^{-1}P(x).$$

From equation (4.3) of Chamberlain (1983) the form of a lower bound should be

$$V_J = \{\lim \sum_{t=1}^{n} E[I_J(x_t)f_b(x_t,b_0)f_b(x_t,b_0)']/n\}^{-1} \qquad (34)$$

where it is assumed that the limit is nonsingular. Chamberlain's (1984) Theorem 3 does not apply directly in the current context because the distribution of the exogenous variables is allowed to be nonstationary. The essential hypothesis is stationarity of the conditional distribution, which does hold here, and it can be shown $V(\hat{b}) - V_J$ is positive semi-definite in the current context, although, for brevity, no proof will be given here.

Chamberlain (1983) has also shown that the lower bound can be approximately obtained by choosing $z(x,b)$ appropriately. This will also be true in the current context, so that for the moment attention will be restricted to comparing the lower bound V_J with the CR bound.

When the conditional density $g(e|x)$ of the disturbance is known, the maximum likelihood estimator of b_0 is given by the solution to

$$\max \sum_{t=1}^{n} \ln g(y_t - f(x_t,b)|x_t). \qquad (35)$$

Let $s(e,x)$ be the conditional disturbance score, which is $g'(e|x)/g(e|x)$ when $g(e|x)$ is positive and is equal to zero otherwise. Let $I^*(x)$ be the conditional information

$$I^*(x) = \int s(e,x)^2 g(e|x)de,$$

where it is assumed for the moment that this integral exists for all x in X. It will also be assumed that the asymptotic information matrix for b is nonsingular.

Assumption H3: The matrix

$$\lim_{t=1}^{n} \sum E[I^*(x_t) f_b(x_t,b_0) f_b(x_t,b_0)']/n$$

is nonsingular.

The asymptotic CR bound for estimators of b_0 is then given by

$$V^* = \{\lim_{t=1}^{n} \sum E[I^*(x_t) f_b(x_t,b_0) f_b(x_t,b_0)']/n\}^{-1}.$$

The comparison of the lower bound V_J for conditional moment restriction estimators and the CR bound V* parallels closely the asymptotic efficiency discussion of Section 3. The relationship of V_J to V* depends entirely on the scalar functions $I_J(x)$ and $I^*(x)$. Note that

$$v = (e,c_1(e),\ldots,c_J(e))',$$

so that integration by parts gives

$$P(x) = -\int vs(e,x)g(e|x)de = -E[vs(e,x)|x].$$

It follows that

$$d^*(x) = -\Sigma(x)P(x) = [E(vv'|x)]^{-1}E[vs(e,x)|x]$$

is the vector of coefficients of the linear projection of $s(e,x)$ on v when e has density $g(e|x)$, so that

$$I^*(x_t) - I_J(x_t) = \min_d E[(s(e_t,x_t) - v_t d)^2|x_t] \tag{4.4}$$

(36)

That is, the difference between $I^*(x_t)$ and $I_J(x_t)$ is the minimum mean square from using $v'd^*(x)$ as the best linear predictor (at x_t) of the conditional disturbance score $s(e,x)$.

When the form of the conditional distribution of the disturbance is unknown, it will generally not be possible to guarantee that $I_J(x)$ is equal to $I^*(x)$ for any finite J. As in Theorem 3.1, it is possible to characterize a subset of conditional densities for which $I_J(x)$ converges to $I^*(x)$ as J grows, so that V_J converges to V^*. By combining choice of large J with choice of $z(x,b)$ so that the lower bound V_J is nearly attained for a one-step GMM estimator, it will be possible to obtain an estimator which is nearly as efficient as the maximum-likelihood estimator would be if the entire conditional distribution of the disturbance was known.

Consider choosing a sequence of moment functions $\{c_j(e)\}$ which satisfy

$$c_j(e) = [c(e)]^{2j-1}, \quad (j = 1,2,\ldots),$$

where $c(e)$ satisfies the following assumption.

Assumption H4: The function $c(e)$ is an odd function which is continuously differentiable on the entire real line, with $c'(e) > 0$.

That is, the sequence of moment functions consists of odd powers of the monotonic, odd function $c(e)$. Such functions will also be odd. The following result characterizes a subset of unknown conditional distributions of the disturbance for which a nearly efficient GMM estimator exists.

Theorem 4.1: If Assumptions H1-H4 are satisfied, $f_b(x,b)$ is a subvector of $z(x,b)$, $z(x,b)$ is chosen so that Ω is nonsingular, and there exist finite constants $\gamma > 1$, n, $\delta > 0$ such that

$$\sup_X g(e|x) \leq Nc'(e)\exp(-\delta|c(e)|^\gamma), \tag{37}$$

$$\int \sup_X s(e,x)^2 c'(e)\exp(-\delta|c(e)|^\gamma)de < +\infty,$$

then for each J there exists $z(x,b)$ such that

$$\lim_{J\to\infty} V(\hat{b}) = V^*.$$

Proof: First, to show that V_J converges to V^*, note that as in the proof of Theorem 3.1, the inequalities (37) imply that for each x in X, for each J there exists $\tilde{d}_0,\ldots,\tilde{d}_{2J-1}$ (depending on x and J) such that

$$\lim_{J\to\infty} \int [s(e,x) - \sum_{j=0}^{2J-1} \tilde{d}_j c(e)^j]\exp(-\delta|c(e)|^\gamma)c'(e)de = 0. \tag{38}$$

Let $w = (1,c(e)^2,\ldots,c(e)^{2J-2})'$ be a $J\times 1$ vector of even powers of $c(e)$. Note that $s(e,x)$ is an odd function of e, because $g(e|x)$ is an even function of e, ($g(-e|x) = g(e|x)$, and $g'(e|x)$ is odd). Since a product of an even and an odd function is also odd, $s(e,x)\cdot w$ is a vector of odd functions, and wv'

is a matrix of odd functions. Therefore, by symmetry and the inequalities (which imply these integrals are finite)

$$\int s(e,x)wg(e|x)de = 0, \quad \int wv'g(e|x)de = 0.$$

It follows that the coefficients of w in the linear projection of $s(e,x)$ on w and v are zero when e has density $g(e|x)$, so that by the density dominance condition of (37)

$$0 \leq I^*(x) - I_J(x) = \min_d \int [s(e,x) - v'd]^2 g(e|x)de \tag{39}$$

$$= \min_{d,f} \int [s(e,x) - v'd - w'f]^2 g(e|x)de$$

$$\leq \int [s(e,x) - \sum_{j=0}^{2J-1} \tilde{d}_j c(e)^j]^2 g(e|x)de$$

$$\leq N \int [s(e,x) - \sum_{j=0}^{2J-1} \tilde{d}_j c(e)^j]^2 \exp(-\delta|c(e)|^\gamma)c'(e)de.$$

Equations (38) and (39) imply that for each x in X,

$$\lim_{J\to\infty} I_J(x) = I^*(x).$$

Furthermore, it can also be shown that this convergence is uniform in x. From (37)

$$s(e,x)^2 g(e|x) \leq N \sup_X s(e,x)^2 \exp(-\delta|c(e)|^\gamma)c'(e),$$

so that (37) and the dominated convergence theorem imply that $I^*(x)$ is continuous on X. Similarly, $\Sigma(x)$ and $P(x)$ are continuous on X, implying $I_J(x)$ is continuous on X. Then uniform convergence follows by X compact,

$$I_J(x) \geq I_{J+1}(x),$$

and Rudin's (1976) Theorem 7.13. Convergence of V_J to V* then follows from boundedness of $f_b(x,b)$, the uniform convergence of $I_J(x)$, and Assumption H3.

To show that for any J there exists a choice of a known vector $z(x,b)$ of functions for which $V(\hat{b})$ is arbitrarily close to V_J, note that x is a p-dimensional vector. Let $\{h_\ell(x)\}$ be a sequence of functions such that for any continuous function f(x) on X there exist $d_1,\ldots,d_L$ (d_ℓ depending on L) such that

$$\lim_{L\to\infty} \Sigma d_\ell h_\ell(x) = f(x),$$

where convergence is uniform in x. For example, by the Weirstrass Theorem $\{h_\ell(x)\}$ will have this property if its elements include all crossproducts of nonnegative integer powers of the elements of x. Since $f(x,b_0)P(x)'\Sigma(x)^{-1}$ is a $k\times(J+1)$ matrix of continuous functions, for each L there exists a $k\times[(J+1)\cdot L]$ matrix D_L such that for $z(x,b) = (h_1(x),\ldots,h_L(x))$

$$\lim_{L\to\infty} D_L[I \otimes z(x,b)] = f(x,b_0)P(x)'\Sigma(x)^{-1}$$

where I is a J+1 dimensional identity matrix and convergence is uniform in x. Consequently, by boundedness of $f(x,b_0)$, $P(x)$, and $\Sigma(x)$,

$$\lim_{L\to\infty} D_L[P(x) \otimes z(x,b_0)f_b(x,b_0)'] = \lim_{L\to\infty} D_L[\Sigma(x) \otimes z(x,b_0)z(x,b_0)']D'_L$$

$$= I_J(x)f_b(x,b_0)f_b(x,b_0)' \qquad (40)$$

where convergence is uniform in x. Consequently, for H_b and Ω as defined in (10) with $z(x,b_0) = (h_1(x),\ldots,h_L(x))$,

$$\lim_{L\to\infty} D_L H_b = \lim_{L\to\infty} D_L'\Omega D_L = \lim \sum_{t=1}^{n} E[I_J(x_t)f_b(x_t,b_0)f_b(x_t,b_0)']/n. \quad (41)$$

By uniform convergence of $I_J(x)$ to $I^*(x)$ and Assumption H3, the matrix of the right-hand side of the second equality is nonsingular for large J, so that $D_L H_b$ is nonsingular for large L and J. Then, by equation (41),

$$\lim_{L\to\infty} (D_L H_b)^{-1} D_L \Omega D_L' (H_b' D_L')^{-1} = V_J. \quad (42)$$

From Hansen's (1982) Theorem (3.2)

$$(D_L H_b)^{-1} D_L \Omega D_L (H_b' D_L')^{-1} - (H_b'\Omega^{-1} H_b)^{-1}$$

is positive semidefinite, so that, by equation (42)

$$\lim_{L\to\infty} (H_b'\Omega^{-1} H_b)^{-1} = V_J. \quad (43)$$

To make sure that $z(x,b)$ satisfies the hypotheses, include $f_b(x,b)$ among the elements of $z(x,b)$ and delete linearly dependent elements from the sequence $\{h_\ell(x)\}$ (see Chamberlain, 1983, p. 27).

Finally, to prove the results use equation (43) and convergence of V_J to V^*. For $J = 1$, let $z(x,b) = f_b(x,b)$ and let $L_1 = 0$. For $J > 1$ define L_J recursively as the maximum of L_{J-1} and an L from equation (43), such that $(H_b'\Omega^{-1} H_b)^{-1}$ is no more than $1/J$ in distance from V_J, and let $z(x,b) = (f_b(x,b), h_1(x), \ldots, h_{L_J}(x))'$. Then, since $(H_b'\Omega^{-1} H_b)^{-1}$ is monotonically decreasing in L, the conclusion follows.

To interpret the hypotheses of this theorem, suppose that $c(e)$ is a bounded function. Then, since $\exp(-\delta|c(e)|^\gamma)$ is bounded away from zero, the conditions (37) reduce to

$$\sup_X g(e|x) \leq Nc'(e), \quad \int \sup_X s(e,x)^2 c'(e)de < +\infty. \qquad (44)$$

These conditions are tail behavior restrictions which are specified to hold uniformly in x. To see the role of the uniformity in x, suppose that the conditional density has the heteroskedastic form

$$g(e|x) = [\bar{g}(e/\sigma(x))]/\sigma(x),$$

where $\sigma(x)$ is positive and continuous on X and $\bar{g}(e)$ is a symmetric, continuously differentiable density on the real line. In this case the dependence of e_t on x_t is limited to scale changes in the distribution of e_t. Let

$$S = [\min_X \sigma(x), \max_X \sigma(x)].$$

In this case the inequalities (44) become

$$\sup_S [\bar{g}(e/\sigma)] \leq Nc'(e), \quad \int \sup_S [\bar{g}'(e/\sigma)/\bar{g}(e/\sigma)]c'(e)de < +\infty. \qquad (45)$$

Because S is bounded away from zero and is compact, these inequalities will be satisfied, for example, if $\bar{g}(e)$ declines monotonically to zero at a faster rate than $c'(e)$ for large values of $|e|$ and $\bar{g}'(e)/\bar{g}(e)$ is bounded. The uniformity in x restriction does not seem to be very onerous.

This result also applies to more general forms of dependence of the distribution of the disturbance of the regressors. The shape of the distribution can also change with x, as long as it changes in a continuous fashion and the inequalities (37) are satisfied. Also, the distribution of x_t is allowed to be continuous, in which case changing x_t would trace out an uncountably infinite number of different distributions of the disturbance.

Use of two dimensions moment conditions, i.e. the moment functions and the elements of $z(x,b)$, yields a nearly efficient GMM estimator. Manski's (1984) heteroskedastic case, which involves a finite number of unknown conditional densities, can also be subsumed in Theorem 4.1 by allowing x_t to include a set of dummy variables, each of which is 1 when the conditional distribution takes on a particular (unknown) value, and zero otherwise.

Examples of moment functions which can be used in the symmetric case include odd, positive powers of e, which corresponds to $c(e) = e$, and odd, positive integer powers of some bounded function of e, such as that given in equation (31). Since a known "instrument" vector $z(x,\theta)$ which is optimal does not generally exist in the heteroskedastic case, one must also choose $z(x,\theta)$. An example of a choice of $z(x,\theta)$ is $z(x,\theta) = (f_b(x,b)', h_1(x), \ldots, h_L(x))$, where $h_\ell(x)$ is some product of positive integer powers of the elements of x, or $h_\ell(x)$ is a sine or cosine of linear combinations of the elements of x, as discussed in another context by Gallant (1981). The choice of moment functions and "instruments" which will work well in practice in the heteroskedastic case remains a topic for future research.

5. A SAMPLING EXPERIMENT

A small sampling study was conducted to determine to what extent the efficiency gains can be realized in finite samples using the one-step GMM estimator. Attention was restricted to the i.i.d. disturbance case to allow comparisons with the adaptive maximum-likelihood (AML) estimators of Bickel (1982) and Manski (1984). The model which was used is identical to that used in the Monte Carlo experiments reported by Manski (1984).

Consider the model $y = \alpha + \beta x + e$ with $\alpha = 1$, $\beta = -1$, x distributed uniformly on [-1,1], and e i.i.d. with mean zero and variance one. In the experiments three alternative densities were used to draw the realizations of e. These include (1) normal; (2) contaminated normal, being the convolution $.9N(0,1/9) + .1N(0,9)$; and (3) lognormal.

Given each density, a random sample of observations was drawn. Results were obtained for two sample sizes, $n = 25$ and $n = 100$. The ordinary least squares (OLS) estimator was computed and used as the initial estimator in the one-step GMM estimator. For $n = 25$ an AML estimator was computed. The AML estimator was identical to the one in Manki's (1984) experiments which used the entire sample at each stage of computation (with $\sigma = .08$, $b = 4.0$, $c = .004$, $d = 30.0$ in Manski's (1984) notation). Six one-step GMM estimators were also computed, three using raw moments (RAW) and three using the bounded function given in equation (31) (OTH) with $r = 4$. For both types of moment functions the 3SLS stepsize of equation (11) was used to form estimators for $J = 2,3$, and 4 extra moment functions. This seemed to be a reasonable range for J. For $n = 100$, the AML estimator was also identical to the estimator considered by Manski (1984) (with $\sigma = .06$, $b = 5.0$, $c = .002$, $d = 36.0$). For $n = 100$, $J = 4$, 5, and 6 extra moment functions were used. Each experiment

consisted of 500 replications in which a sample was drawn and the estimators computed. For brevity, and because only the slope coefficient estimator is guaranteed to be nearly efficient, results are reported only for the estimator of β. Table 1 presents results on the precision of the estimators, as measured by root mean square errors.

In the normal case, the OLS estimator slightly outperforms the AML and GMM estimators. The GMM estimators perform almost the same as the AML estimator. In half the 12 cases (where a case is a choice of moment function and J) the GMM estimator outperforms the AML estimator and in no case is the difference very large.

In the nonnormal cases the one-step GMM estimators almost always outperform the OLS estimator by a wide margin. The GMM estimator also outperforms the AML estimator in 16 of the 24 nonnormal cases. The best performance occurs for the GMM estimator which uses raw moments. For low values of J . this estimator outperfoms the AML estimator by more than the AML estimator outperforms least squares, except in the 100 observations lognormal case. This exception may be significant since the lognormal distribution has thicker tails than $\exp(-\delta|e|^{\gamma})$ for any $\delta > 0$, $\gamma > 1$, so that the sufficient conditions of Theorem 3.1 for the existence of a nearly efficient GMM estimator are not satisfied for the lognormal case when raw moments are used.

The outstanding performance of the raw moments estimator for some nonnormal cases is offset by the sensitivity of this estimator with respect to the number of higher-order moments used. Its performance deteriorates rapidly as more moments are used, probably reflecting the fact that it is very difficult to estimate high-order raw moments of thick-tailed distribution. The other GMM estimator, which uses a bounded moment function, is much

less sensitive to the choice of J. It outperforms the AML estimator in all cases but the 25 observations and the contaminated normal case, and its performance relative to both the AML and raw moment GMM estimators improved with the increase in sample size. Also, OTH does best in the lognormal case, which is the case with the thickest tail. Given the remarkable performance of the raw moment GMM estimator for some cases and the insensitivity of the other GMM estimator to the choice of J, one suspects that there are probably other choices of moment functions which would do better than OTH without being very sensitive to the choice of J.

Since experiments were performed for two different sample sizes, one can obtain a rough idea of what an appropriate rate of growth of J might be. For each distribution of the disturbances and each GMM estimator, let J_1 be the J which gave the best performance of the GMM estimator (in the range of J values examined) when n = 25 and let J_2 be the best J when n = 100. Then, excluding the raw moment estimator for the lognormal distribution, the average ratio of total moment functions used $(J_2+1)/(J_1+1)$ was 1.8, suggesting that a rate of growth of J between $n^{1/3}$ and $n^{1/2}$ might be appropriate.

The OTH GMM estimator computed for Table 1 was not scale invariant and neither is the AML estimator. The RAW GMM estimator is scale invariant because positive integer powers of e are homogenous functions. To see what effect a scale adjustment would have on the GMM estimators, another version of OTH was also computed with the moment function $c_j(e)$ replaced by $c_j(e/\tilde{\sigma})$, $(j = 1,\ldots,J)$, where $\tilde{\sigma}$ was the sample standard deviation of the OLS residuals. Table 2 contains the resulting root mean square errors. Comparing Table 2 with the last three columns of Table 1 shows that using a scale parameter has some harmful effects for n = 25, particularly for the contaminated normal distribution, but that these effects disappear in the n = 100 observations case.

After examination of Table 1 it is clear that asymptotic standard error formulas must be misleading in some cases. For example, the dispersion of the raw moments estimator can increase dramatically as the number of moment functions increases, while asymptotic standard error estimates must fall as the number of moment functions increases as long as the estimated asymptotic standard errors are evaluated at the same parameter estimates. In the experiments, estimates of the standard error of $\hat{\beta}$ were computed using the asymptotic formula (13). Also, in each experiment an estimate of the AML standard error was computed by estimating the information constant for the unknown density by the average of the square of the estimated disturbance score for each observation. Table 3 reports the ratios of the root mean square (RMS) of the estimated standard errors to the actual standard errors of the slope coefficient estimators. The asymptotic standard error estimates, including those for the AML estimator, do indeed perform quite poorly in most cases, including the estimated AML standard errors. Performance always deteriorates as J increases and improves substantially with the increase in sample size from 25 to 100 observations. Performance also improves with increase in observations as J increases with sample size at a rate which is between $n^{1/3}$ and $n^{1/2}$.

One potential solution to the problem of estimating standard errors is to use the bootstrap method, see Efron (1982). To see if bootstrap standard error estimates for the one-step GMM estimator might work well, a separate experiment was performed. The model used was the same as for the previous experiment. The density used to draw realizations of e was the same log-normal density which was used above. Results were obtained for the raw moments GMM estimator with J = 4 and a sample size of 25.

The experiment consisted of 100 replications. For each replication the one-step GMM estimator $\hat{\beta}$ was computed and residuals were obtained by

$$\hat{e}_t = y_t - \bar{\alpha} - x_t\hat{\beta}, \quad \bar{\alpha} = \bar{y} - \bar{x}\hat{\beta}, \quad t = 1,\ldots,25.$$

For each such vector of residuals, 100 bootstrap replications of 25 observations were drawn from the resulting empirical distribution function of the residuals. For each bootstrap replication y values were generated as if $\hat{\beta}$ were the true value of β and the sample disturbances were the actual disturbances, and the one-step GMM estimator was recomputed. The bootstrap standard error estimate was taken to be the sample standard deviation of the GMM estimators over the 100 bootstrap replications. Table 4 reports the results of this experiment. The asymptotic standard error estimates are biased downward by a factor of two while the bootstrap standard error estimate gives a very accurate estimate, on average.

The performance of the bootstrap standard error estimate in this experiment is very promising. Since computation of the one-step GMM estimator requires no iteration, once an initial estimator of the regression parameters has been obtained bootstrap estimates of the standard errors should not be difficult to obtain in practice. Since the asymptotic standard error estimates may be biased downward by a substantial amount, the bootstrap estimator of the standard deviation should probably be used in practice.

6. CONCLUSION

In Sections 3 and 4 it has been shown that nearly efficient moment condition estimators of regression coefficients exist in the i.i.d. and symmetric, heteroskedastic disturbance cases. One can make these estimators adaptive by allowing J (and $z(x,\theta)$ in the heteroskedastic case) to grow slowly enough with the sample size. The argument of Amemiya (1973, Theorem 5) shows that this is the case. It has not been specified here what rate of growth of J is slow enough and this question remains a topic for future research.

Even knowing the appropriate rate for J would not help answer the question of what J to pick in a particular application. The Monte Carlo results of Section 5 do indicate that using several extra moment functions may give good results in the i.i.d. case, even with quite small samples. Also, the bootstrap standard error estimates could be used to select J. One could choose J by computing bootstrap standard errors for several dif- different values of J, choosing J equal to that value with the smallest bootstrap standard error. Of course, the bootstrap standard error for the particular J would then likely be biased downward somewhat if used as an estimate of the standard error of the resulting regression parameter estimates, because it does not account for varying J.

Moment condition estimators provide a particularly rich framework for future research into the attainment of asymptotic efficiency bounds. Systems of regression equations can be handled in the moment condition framework with some added notational complexity and Theorems 3.1 and 4.1 should extend without difficulty to the systems case to show that nearly efficient GMM estimators exist. By estimating the reduced form and using a minimum distance

estimator of the structural coefficients, nearly efficient estimators of a linear simultaneous equations system can also be constructed. It would also be interesting to know how efficient a moment condition estimator could be obtained in a nonlinear simultaneous equations system, where Manski (1984) has shown that adaptive estimators do not exist.

TABLE 1: ROOT MEAN SQUARE ERRORS

J	OLS	AML	RAW			OTH		
n=25			2	3	4	2	3	4
Normal	.3733	.3919	.4110	.4082	.3900	.4189	.3812	.3801
Cont. Normal	.3738	.3260	.2136	.2368	.3369	.3752	.3577	.3565
Lognormal	.3759	.3277	.1507	.2364	.3442	.2438	.3056	.3002

J	OLS	AML	RAW			OTH		
n=100			4	5	6	4	5	6
Normal	.1792	.1884	.1913	.1937	.1939	.1880	.1845	.1841
Cont. Normal	.1804	.1366	.0835	.0961	.1440	.1318	.1249	.1302
Lognormal	.1691	.1101	.0845	.1173	.1749	.0932	.1025	.0968

TABLE 2: RMSE OF SCALE ADJUSTED GMM ESTIMATOR

J	n = 25			n = 100		
	2	3	4	4	5	6
Normal	.4191	.3829	.3821	.1878	.1843	.1840
Cont. Normal	.3698	.3728	.3779	.1306	.1297	.1335
Lognormal	.2362	.3065	.3101	.0875	.0964	.0896

TABLE 3: RATIO OF RMS OF ESTIMATED STANDARD ERRORS

TO ACTUAL STANDARD ERRORS

J	AML	RAW			OTH		
n=25		2	3	4	2	3	4
Normal	.263	.666	.622	.386	.460	.279	.264
Cont. Normal	.321	.765	.526	.169	.581	.426	.362
Lognormal	.311	.808	.440	.176	.728	.496	.358

J	AML	RAW			OTH		
n=100		4	5	6	4	5	6
Normal	.228	.769	.726	.537	.340	.230	.227
Cont. Normal	.353	.791	.614	.309	.772	.599	.555
Lognormal	.380	.484	.308	.145	.743	.649	.593

TABLE 4: COMPARISON OF ACTUAL STANDARD ERROR WITH ROOT MEAN SQUARE OF BOOTSTRAP AND ASYMPTOTIC ESTIMATES

Actual SE	RMS Asymptotic Estimate	RMS Bootstrap Estimate
.2056	.1014	.2062

REFERENCES

Amemiya, T., 1973, Generalized Least Squares with an Estimated Autocovariance Matrix, Econometrica, 41, 723-732.

Amemiya, T., 1977, The Maximum Likelihood and Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model, Econometrica, 45, 955-968.

Burguete, J. F., A. R. Gallant, and G. Souza, 1982, On the Unification of the Asymptotic Theory of Nonlinear Econometric Models, Econometric Reviews, 1, 151-190.

Bickle, P., 1982, On Adaptive Estimation, Annals of Statistics, 10, 647-671.

Chamberlain, G., 1982, Multivariate Regression Models for Panel Data, Journal of Econometrics, 18, 5-46.

Chamberlain, G., 1983, Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, University of Wisconsin, Social Systems Research Institute, No. 8307.

Efron, B., 1982, The Jacknife, the Bootstrap, and Other Resampling Plans, Philadelphia: Society for Industrial and Applied Mathematics.

Gallant, A. R., 1981, On the Bias in Flexible Forms and an Essentially Unbiased Form, Journal of Econometrics, 15, 211-245.

Halmos, P. R., 1974, Measure Theory, New York: Springer-Verlag.

Hansen, L. P., 1982, Large Sample Properties of Generalized Method of Moments Estimators, Econometrica, 50, 1029-1054.

Helmberg, G., 1969, Introduction to Spectral Theory in Hilbert Space, Amsterdam: North-Holland.

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

MaCurdy, T. E., 1982, Using Information on the Moments of Disturbances to Increase the Efficiency of Estimation, Stanford University, unpublished manuscript.

Manski, C. F., 1984, Adaptive Estimation of Non-Linear Regression Models, Econometric Reviews, forthcoming.

Rudin, W., 1976, Principles of Mathematical Analysis, New York: McGraw-Hill.

White, H., 1980, Nonlinear Regression on Cross-Section Data, Econometrica, 48, 721-746.