ADAPTIVE ESTIMATION OF REGRESSION MODELS

VIA MOMENT RESTRICTIONS

Whitney K. Newey
Princeton University

Econometric Research Program
Research Memorandum No. 330

February 1987

This work was partially supported by National Science Foundation grants SES-8410249 and SES-8606545. Two referees, J. Hausman, D. McFadden, J. Powell, H. Pratt-Ferguson, R. Quandt, K. Singleton, K. West, and participants of the Harvard-MIT, Princeton, and Yale econometrics seminars have provided helpful comments.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, NJ 08544

ABSTRACT

B1-Newey, Whitney K.--

B2-Adaptive Estimation of Regression Models Via Moment Restrictions.

C2-This paper considers adaptive estimation of regression models by means of generalized method of moments estimators. Two models are considered, that with an i.i.d. disturbance that is independent of the regressors and that with a conditionally symmetric but (possibly) heteroskedastic disturbance. For both cases the paper develops linearized estimators that are asymptotically efficient if the number and variety of moment conditions is allowed to grow at an appropriate rate with the sample size. In the general symmetric case no other adaptive estimator has yet been proposed. Also, results of a small Monte Carlo study indicate that in the independence case the small sample performance of the generalized method of moments estimator can be quite good vis-a-vis other estimators previously proposed.

Keywords: Adaptive Estimation, Generalized Method of Moments, Efficiency, Independence, Symmetry.

JEL Classification: 210, 211, 212.

1. Introduction

Regression models are of fundamental importance in econometrics. Methods of estimating and interpreting the parameters of such models remain an important topic of research. Of particular relevance to this study is recent work concerning efficient estimation of regression parameters under weak distributional assumptions. Such estimators are attractive in that they have efficiency properties that do not depend on a particular parametric specification for the conditional distribution of the dependent variable.

It is possible to estimate regression parameters under a variety of semiparametric restrictions (i.e. restrictions involving both parametric and nonparametric assumptions) concerning the conditional distribution of the dependent variable given the regressors. One type of restriction involves an assumption that some location measure for the conditional distribution, such as the mean or median, has a known functional form. For the model where the only restriction imposed is that the conditional mean has a known functional form, Chamberlain (1987) has shown that maximum attainable efficiency is that of the (heteroskedasticity corrected) generalized least squares estimator. Caroll (1982), Newey (1986), and Robinson (1987) have constructed estimators that are asymptotically efficient in this model without further functional form restrictions (i.e. concerning the form of heteroskedasticity). Newey and Powell (1987a) carry out a similar that applie to the conditional median case.

In this paper efficient estimation of linear regression parameters will be considered in two kinds of models where the regression function has a known functional form. The first kind has an i.i.d. disturbance that is independent of the regressors (referred to as the independence case henceforth) and the second a dependent variable symmetrically distributed around the regression function conditionally on the regressors (referred to as the symmetric case henceforth). These kinds of models are important despite the fact that the restrictions imposed are stronger than a single restriction concerning a location measure. For instance, these models are less open to criticism that the interpretation of the parameters depends crucially on the location measure chosen.

Bickel (1982) and Manski (1984) have shown that in both the independence and symmetric cases the maximum attainable efficiency for the regression parameters (except for the constant in the independence case) is that of the maximum likelihood estimator which could be obtained if the actual (unknown) functional form of the disturbance distribution were used in forming the likelihood. Bickel (1982) and Manski (1984) have constructed efficient estimators for the independence case, and Manski (1984) for the symmetric case where the regressors have a finite number of possible outcomes. These estimators involve nonparametric estimation of the conditional score function (derivative of the log density) of the disturbance.

In this paper a different approach is taken to efficient estimation. This approach is based on the observation that in both the independence and symmetric cases there are an infinite number of moment restrictions that can be used in estimation. In the independence case any function of the disturbance will be uncorrelated with any function of the regressors and in the symmetric case any odd function of the disturbance will be uncorrelated with any function of the regressors. These moment restrictions can be used to form generalized method of moments (GMM, Hansen (1982)) estimators of the regression parameters. Efficiency gains from these types of estimators in the independence case have been discussed by MaCurdy (1982), Chamberlain (1984), and Newey (1984). Here it is shown that asymptotically efficient estimators can be obtained from linearized GMM estimators by allowing the number and variety of moments used in estimation to grow appropriately with the sample size.

Efficient GMM estimators appear to offer some advantages over counterparts which involve direct, nonparametric estimation of score functions. The most important advantage is that the GMM method applies in a straightforward way to the symmetric model, even when the regressors are continuously distributed, a situation where no other efficient estimator is currently known to exist. Also, it is straightforward to extend the estimators considered here to multivariate regression models, although the extension has not been worked out here. The difficulty of handling these cases with the direct nonparametric estimation approach has been noted by Manski (1984). In addition, the GMM method provides a relatively transparent and parsimonious way of making use of information implied by the model to form efficient estimators. In the Monte Carlo results given here the GMM estimator performs very well relative to the nonparametric efficient estimator, which may reflect the parsimony of the GMM estimator. Finally, GMM estimators have a familiar form and may be more convenient to compute than their nonparametric counterparts.

In Section 2 of the paper the independence case is discussed. After presenting the form of the moment restrictions, the form of the GMM estimator is discussed and a linearized version presented. The estimator is interpreted as an approximation to the linearized maximum likelihood estimator. This interpretation is then used in showing that the estimator is asymptotically efficient when the number of moment restrictions grows appropriately with the sample size. Section 3 carries out a similar analysis for the symmetric case. Section 4 presents the results of a small sampling experiment. Section 5 offers some conclusions and discusses extensions of efficient GMM estimators to other models that are the topic of current and future research. The proofs of the results are gathered in an Appendix.

2. Independence of the Regressors and Disturbance

One familiar and important regression model is that with a disturbance which is independent of the regressors. Consider the model

$$(2.1) \qquad y_t = x_t'\beta_0 + \varepsilon_t, \qquad (t=1,2,\ldots),$$

where x_t is a $k \times 1$ vector of regressors, β_0 is a $k \times 1$ vector of regression slopes, and the disturbance ε_t is i.i.d. and distributed independently of the regressors. The independence of the disturbance and regressors means that $x_t'\beta_0$ summarizes entirely the dependence of any conditional location measure for y_t (e.g. conditional mean or median) on the regressors. It is convenient in what follows to avoid normalizing the location of the disturbance distribution, so it will be assumed throughout this section that any constant in the regression is absorbed by the disturbance.

The assumption that the disturbance is distributed independently of the regressors yields many conditional moment restrictions that can be used to estimate the slope coefficients β_0. Note that independence of ε_t and x_t implies that any function of ε_t should be uncorrelated with any function of x_t. This type of moment restriction can easily be used to form a generalized method of moments (GMM) estimator. Consider a sequence $\{\bar{m}_1(\varepsilon), \bar{m}_2(\varepsilon), \ldots\}$ of differentiable functions that have finite second moment and let $\bar{\mu}_{j0} = E[\bar{m}_j(\varepsilon_t)]$. For some positive integer J, let $\theta = (\theta_1', \theta_2')'$ be a $(k+J) \times 1$ vector with $\theta_1 = \beta$ and $\theta_2 = (\mu_1, \ldots, \mu_J)'$. Also, let

$$\bar{\rho}_j(z,\theta) = \bar{m}_j(y-x'\beta) - \mu_j, \quad (j=1,\ldots,J).$$

Independence of ε_t and x_t implies the conditional moment restriction

$$(2.2) \qquad E[\bar{\rho}_j(z_t,\theta_0)|x_t] = E[\bar{m}_j(\varepsilon_t)|x_t] - \bar{\mu}_{j0} = 0,$$

which in turn implies that $\bar{\rho}_j(z_t,\theta_0)$ will be uncorrelated with any function of x_t. Let $\bar{\rho}(z,\theta) = (\bar{\rho}_1(z,\theta),\ldots,\bar{\rho}_J(z,\theta))'$ and $\bar{g}(z,\theta) =$

$\bar{\rho}(z,\theta) \otimes a(x)$ for some $K \times 1$ vector $a(x)$ of functions of x such that $a(x_t)$ has finite second moments. Equation (2.2) implies that $\bar{g}(z,\theta)$ satisfies the population orthogonality condition

$$(2.3) \qquad E[\bar{g}(z_t,\theta_0)] = E\{E[\bar{\rho}(z_t,\theta_0)|x_t] \otimes a(x_t)\} = 0.$$

A GMM estimator that makes use of this orthogonality condition can be obtained by choosing $\hat{\theta}$ to solve

$$(2.4) \qquad \min_{\theta} \bar{g}_n(\theta)'W_n\bar{g}_n(\theta),$$

where $\bar{g}_n(\theta) = \Sigma_{t=1}^{n}\bar{g}(z_t,\theta)/n$ and W_n is a positive semi-definite matrix.

To better understand the moment restrictions used by this type of estimator, it is useful to consider some examples. Note that when $\bar{m}_j(\varepsilon) = \varepsilon^j$, the conditional moment restriction in equation (2.2) states that the first J conditional raw moments of the disturbance do not depend on the regressors. Thus, for this particular choice of $\rho(z,\theta)$ the GMM estimator $\hat{\theta}$ is making use of constancy of the higher order moments of the disturbance. This choice of moment restrictions to be used in estimation of regression parameters has been considered by MaCurdy (1982).

One drawback of an estimator that uses information concerning high order raw moments of the disturbance is that it may be sensitive to the thickness of the tails of the disturbance distribution. For example, such an estimator will be sensitive to the existence of raw moments. An alternative estimate that is less sensitive to thick-tailed distributions can be obtained by choosing $\bar{m}_j(\varepsilon) = [\bar{m}_0(\varepsilon)]^j$ for some bounded function $\bar{m}_0(\varepsilon)$, such as $\bar{m}_0(\varepsilon) = \varepsilon/[1+|\varepsilon|]$. Another estimate that does not require existence of moments of the disturbance can be obtained by choosing $\bar{m}_j(\varepsilon) = w(\varepsilon)\varepsilon^j$, where $w(\varepsilon)$ is some weight function that has the property that $w(\varepsilon)\varepsilon^j$ is bounded for each j, such as $w(\varepsilon) = \exp(-\varepsilon^2/2)$. The relative merits of different choices of moment functions will be further considered in Section 4.

There are several issues concerning the GMM estimator of equation (2.4) that need to be addressed. One issue concerns the choice of the matrix W_n and the vector of functions $a(x)$ which are used in the GMM estimator. Theorem 3.2 of Hansen (1982) implies that the optimal choice of the matrix W_n, in terms of minimizing the asymptotic covariance matrix of $\hat{\theta}$, is given by a consistent estimator of the inverse of

$$\Sigma_{t=1}^{n}E[\bar{g}(z_t,\theta_0)\bar{g}(z_t,\theta_0)']/n = \bar{\Sigma} \otimes \Sigma_{t=1}^{n}E[a(x_t)a(x_t)']/n,$$

where $\bar{\Sigma} = E[\bar{\rho}(z_t,\theta_0)\bar{\rho}(z_t,\theta_0)']$. Furthermore, note that the estimator $\hat{\theta}$ is formally identical to a nonlinear three-stage least squares estimator of an equation system with residual vector $\bar{\rho}(z_t,\theta)$. Thus, Amemiya's (1977) characterization of the optimal instruments for each residual of such a system implies that the optimal instrument vector for the j^{th} element of $\bar{\rho}(z_t,\theta)$ is given by

$$(2.5) \qquad E[\partial\bar{\rho}_j(z_t,\theta_0)/\partial\theta|x_t] = -(E[\partial\bar{m}_j(\varepsilon_t)/\partial\varepsilon]\circ x_t', \ e_j')',$$

where e_j is the j^{th} unit vector of dimension J. Since for each j this expression consists of zeros and a nonsingular linear combination of $X_t = (1,x_t')'$, it follows that X_t is an optimal choice of instrumental variables for each element of $\bar{\rho}(z_t,\theta)$. Thus, $a(x) = (1,x')'$ is optimal.

A second issue concerns the fact that the estimate of β_0 obtained from solving equation (2.4) is not location and scale equivariant, in the sense that the effect of a location and scale shift of the disturbance will involve more than a corresponding scale shift of $\hat{\beta}-\beta_0$. This unfortunate property of the GMM estimator can be fixed by basing the estimates on moment functions that have been adjusted for location and scale using preliminary estimates of location and scale parameters. Let $\tilde{\beta}$ be an initial location and scale equivariant estimate of β_0, such as could be obtained by ordinary least squares (OLS) or least absolute deviations (LAD) with a constant included in the regression, and let $\tilde{\varepsilon}_t = y_t-x_t'\tilde{\beta}$, $(t=1,\ldots,n)$ be the corresponding residuals. Let $\tilde{\alpha}$ be an estimate of some population location measure α_0 for ε_t, such as the sample mean or median of $\tilde{\varepsilon}_t$. Let $\tilde{\sigma}$ be an estimate of some population scale measure σ_0 of ε_t, such as the sample standard deviation or interquartile range of $\tilde{\varepsilon}_t$. Consider using $\tilde{m}_j(\varepsilon) = \bar{m}_j((\varepsilon-\tilde{\alpha})/\tilde{\sigma})$ in place of $\bar{m}_j(\varepsilon)$ in the minimization problem (2.4). The resulting estimator of β_0 will be location and scale equivariant. Furthermore, its asymptotic distribution will be identical to that obtained if $m_j(\varepsilon) = \bar{m}_j((\varepsilon-\alpha_0)/\sigma_0)$ were used in place of $\tilde{m}_j(\varepsilon)$. The absence of effect of the estimation of the location and scale parameters on the asymptotic distribution of $\hat{\beta}$ results from the fact that the first order condition for $\hat{\beta}$ involves the sample covariance between x and $\bar{m}_j((\varepsilon-\alpha)/\sigma)$, and the population covariance between these two quantities is zero for all possible values of α and σ.

A third issue concerns the computational burden of solving equation (2.4). This minimization problem is quite nonlinear in the regression slopes β, so that a one step alternative based on a version of equation (2.4) where $\bar{g}(z,\theta)$ is linearized around an initial estimate of the parameters will be much easier to solve. As in maximum likelihood contexts such an estimator will have the same asymptotic properties as an estimator obtained from solving equaiton (2.4) (e.g. see Newey, 1985). Note that an initial estimate $\tilde{\beta}$ of the regression slopes is readily available, e.g. from OLS or LAD. For each j an initial estimator of $\mu_{j0} = E[m_j(\varepsilon_t)] = E[\bar{m}_j((\varepsilon_t-\alpha_0)/\sigma_0)]$ can be obtained as

$$(2.6) \qquad \tilde{\mu}_j = \Sigma_{t=1}^{n}\tilde{m}_j(\tilde{\varepsilon}_t)/n = \Sigma_{t=1}^{n}\bar{m}_j((y_t-x_t'\tilde{\beta}-\tilde{\alpha})/\tilde{\sigma})/n,$$

where $\tilde{\varepsilon}_t = y_t - x_t'\tilde{\beta}$. An initial estimate of $\theta_0 = (\beta_0', \mu_{10}, \ldots, \mu_{J0})'$ is then given by $\tilde{\theta} = (\tilde{\beta}', \tilde{\mu}_1, \ldots, \tilde{\mu}_J)'$. Define the location and scale adjusted function $\tilde{\rho}_j(z_t,\theta) = \tilde{m}_j(y_t-x_t'\beta) - \mu_j$. Also, let $\tilde{m}_{j\varepsilon}(\varepsilon) = d\tilde{m}_j(\varepsilon)/d\varepsilon$ and $\tilde{M}_j = \Sigma_{t=1}^{n}\tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)/n$. A first order expansion of $\tilde{\rho}_j(z_t,\theta)$ around $\tilde{\theta}$ gives

$$(2.7) \qquad \tilde{\rho}_j(z_t,\theta) \cong \tilde{\rho}_j(z_t,\tilde{\theta}) + [\partial\tilde{\rho}_j(z_t,\tilde{\theta})/\partial\theta]'(\theta - \tilde{\theta})$$

$$= \tilde{m}_j(\tilde{\varepsilon}_t) - \tilde{\mu}_j - [\tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)x_t'](\beta-\tilde{\beta}) - (\mu_j-\tilde{\mu}_j)$$

$$= \tilde{m}_j(\tilde{\varepsilon}_t) + \tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)x_t'\tilde{\beta} - \tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)x_t'\beta - \mu_j$$

$$\cong \tilde{m}_j(\tilde{\varepsilon}_t) + \tilde{M}_j x_t'\tilde{\beta} - \tilde{M}_j x_t'\beta - \mu_j,$$

where the replacement of $\tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)$ by $\tilde{M}_j$ will not affect the asymptotic properties of the linearized estimator because of the independence of the regressors and the disturbance. Let

$$\tilde{\Sigma} = \Sigma_{t=1}^{n}[\tilde{\rho}_1(z_t,\tilde{\theta}),\ldots,\tilde{\rho}_J(z_t,\tilde{\theta})]'[\tilde{\rho}_1(z_t,\tilde{\theta}),\ldots,\tilde{\rho}_J(z_t,\tilde{\theta})]/n.$$

A linearized version of the optimal GMM estimator can now be obtained by choosing $a(x_t) = X_t$ and $W_n = \tilde{\Sigma}^{-1} \otimes (X'X/n)^{-1}$, where $x = (x_1,\ldots,x_n)'$, e is an $n \times 1$ vector of ones, and $X = [e,X]$, and then replacing $\bar{\rho}_j(z_t,\theta)$ in the definition of $\bar{g}_n(\theta)$ with the location and scale adjusted, linearized function $\tilde{m}_j(\tilde{\varepsilon}_t) + \tilde{M}_j x_t'\tilde{\beta} - \tilde{M}_j x_t'\beta - \mu_j$. To be specific, let $\tilde{M} = (\tilde{M}_1,\ldots,\tilde{M}_J)'$, $\tilde{Z} = [\tilde{M}\otimes x, I_J\otimes e]$, where I_J is a J-dimensional identity matrix. Stacking equation (2.7) then gives

$$(2.8) \qquad \tilde{\rho}(\theta) \cong \tilde{Y} - \tilde{Z}\theta,$$

where $\tilde{\rho}(\theta) = (\tilde{\rho}_1(z_1,\theta),\ldots,\tilde{\rho}_1(z_n,\theta),\ldots,\tilde{\rho}_J(z_1,\theta),\ldots,\tilde{\rho}_J(z_n,\theta))$ and $\tilde{Y} = \tilde{\rho}(\tilde{\theta})+\tilde{Z}\tilde{\theta}$. Replacing $\bar{g}_n(\theta)$ with $(I_J \otimes X)'(\tilde{Y}-\tilde{Z}\theta)/n$ in equation (2.4) and solving then yields the linearized GMM (LGMM) estimator

$$(2.9) \qquad \hat{\theta}_s = (\tilde{Z}'(\tilde{\Sigma}^{-1}\otimes X(X'X)^{-1}X')\tilde{Z})^{-1}\tilde{Z}'(\tilde{\Sigma}^{-1}\otimes X(X'X)^{-1}X')\tilde{Y}.$$

This estimator is formally identical to a three-stage least squares (3SLS) estimator of a system with J equations, residuals $\tilde{m}_j(\tilde{\varepsilon}_t) + \tilde{M}_j x_t'\tilde{\beta} - \tilde{M}_j x_t'\beta - \mu_j$, $(j=1,\ldots,J)$, instrumental variables X_t for each equation, linear cross-equation restrictions corresponding to the elements of β, and initial estimate of the system covariance matrix obtained from the residuals $\tilde{Y} - \tilde{Z}\tilde{\theta} = \tilde{\rho}(\tilde{\theta})$. Thus, this linearized version of the GMM estimator has a familiar form, and can even be computed using some standard software (e.g. TSP).

In order to better understand the nature of this LGMM estimator it is useful to compare the LGMM estimator of β_0 with a corresponding linearized maximum likelihood estimator (LMLE) that could be obtained if the distribution of the disturbance were known up to location. Suppose that, instead of having no knowledge concerning the form of the distribution of the disturbance, an investigator knew that the unknown density $f(\varepsilon)$ of ε_t took the form $\bar{f}(\varepsilon-\alpha_0^*)$, where $\bar{f}(u)$ is a known density function and α_0^* is an unknown location parameter. The likelihood of the t^{th} observation would then be known to take the form $\bar{f}(y_t-\alpha-x_t'\beta)$. Let $\bar{s}(u) = [d\bar{f}(u)/du]/\bar{f}(u)$ be the score corresponding to this density function and $\mathcal{I} = E[\bar{s}(\varepsilon_t-\alpha_0^*)^2]$ be the information, and let $Q_n = \Sigma_{t=1}^n E[X_tX_t']/n$. In this notation the score for the parameters $b = (\alpha,\beta')'$ and the t^{th} observation is $-X_t\bar{s}(y_t-X_t'b)$, and the information matrix for a sample of size n is $n\mathcal{I}Q_n$. Thus, for an initial estimate $\tilde{b}$ of b, the LMLE would be given by

$$(2.10)\qquad \hat{b}_s^* = \tilde{b} - (\hat{\mathcal{I}}\hat{Q})^{-1}\Sigma_{t=1}^n X_t\tilde{s}(\tilde{\varepsilon}_t)/n,$$

where $\tilde{s}(\varepsilon) = \bar{s}(\varepsilon-\tilde{\alpha})$, $\tilde{\varepsilon}_t = y_t-x_t'\tilde{\beta}$, $\hat{Q} = X'X/n$, and $\hat{\mathcal{I}}$ is an estimate of $\mathcal{I}$ (e.g. $\hat{\mathcal{I}} = \Sigma_{t=1}^n[\tilde{s}(\tilde{\varepsilon}_t)]^2/n$). Solving for the slope coefficient LMLE $\hat{\beta}_s^*$ then yields

$$(2.11)\qquad \hat{\beta}_s^* = \tilde{\beta} - (\hat{\mathcal{I}}\hat{Q}_x)^{-1}\Sigma_{t=1}^n(x_t-\bar{x})\tilde{s}(\tilde{\varepsilon}_t)/n,$$

where $\bar{x} = \Sigma_{t=1}^n x_t/n$ and $\hat{Q}_x = (\Sigma_{t=1}^n x_tx_t'/n) - \bar{x}(\bar{x})'$. Also, solving equation (2.9) for the LGMM estimator $\hat{\beta}_s$ of β_0 yields

$$(2.12)\qquad \hat{\beta}_s = \tilde{\beta} - (\hat{\mathcal{I}}_J\hat{Q}_x)^{-1}\Sigma_{t=1}^n(x_t-\bar{x})\hat{s}_J(\tilde{\varepsilon}_t)/n,$$

where $\hat{\mathcal{I}}_J = \tilde{M}'\tilde{\Sigma}^{-1}\tilde{M}$ and $\hat{s}_J(\varepsilon) = -\tilde{M}'\tilde{\Sigma}^{-1}[\tilde{m}_1(\varepsilon)-\tilde{\mu}_1,\ldots,\tilde{m}_J(\varepsilon)-\tilde{\mu}_J]'$. Equations (2.11) and (2.12) are remarkably similar. Assuming that the same initial estimator $\tilde{\beta}$ is used in both the only difference between these equations is that $\hat{\vartheta}_J$ appears in place of $\hat{\vartheta}$ and $\hat{s}_J(\varepsilon)$ in place of $\tilde{s}(\varepsilon)$ in the formula for the linearized GMM estimator.

The similarity of equations (2.11) and (2.12) leads to an interpretation of the LGMM estimator as an implicit approximation to the LMLE. This interpretation arises from interpreting $-E[m_{j\varepsilon}(\varepsilon_t)]$ as the covariance between $m_j(\varepsilon_t)$ and the unknown disturbance score $s(\varepsilon_t)$, where $m_j(\varepsilon) = \bar{m}_j((\varepsilon-\alpha_0)/\sigma_0)$, $s(\varepsilon) = f_\varepsilon(\varepsilon)/f(\varepsilon)$, $f(\varepsilon)$ is the density of ε_t, and the ε subscript denotes the partial derivative with respect to ε. To give this interpretation and state the other results of this section it is useful to assume that the density $f(\varepsilon_t)$ of ε_t is regular in the sense of Hajek and Sidak (1967).

Assumption 2.1: The density $f(\varepsilon)$ of ε_t is absolutely continuous and has Radon-Nikodym derivative $f_\varepsilon(\varepsilon)$ such that $\int(f_\varepsilon{}^2/f)(\varepsilon)d\varepsilon$ is finite.

The following result is a consequence of this assumption:

Lemma 2.1: If Assumption 1 is satisfied and $m(\varepsilon)$ is a function such that $E[m(\varepsilon+\alpha)^4]$ is bounded in a neighborhood of α^*, then $E[m(\varepsilon_t+\alpha)]$ is continuously differentiable on a neighborhood of α^* with $dE[m(\varepsilon_t+\alpha)]/d\alpha = -E[m(\varepsilon_t+\alpha)s(\varepsilon_t)]$.

When the hypotheses of this Lemma are satisfied, $m_j(\varepsilon)$ $(j=1,\ldots,J)$ are differentiable, and the order of integration and differentiation can be interchanged, then from the conclusion of this Lemma with $\alpha = 0$ and from $E[s(\varepsilon_t)] = 0$ (which is another consequence of Assumption 2.1; Hajek and Sidak (1967, p.20)) it follows that for $M = E[(m_{1\varepsilon}(\varepsilon_t),\ldots,m_{J\varepsilon}(\varepsilon_t))']$,

$$(2.13) \qquad M = -E\{[m_1(\varepsilon_t),\ldots,m_J(\varepsilon_t)]'s(\varepsilon_t)\}$$

$$= -E\{[m_1(\varepsilon_t),\ldots,m_J(\varepsilon_t)]'s(\varepsilon_t)\} + E\{[\mu_{10},\ldots,\mu_{J0}]'s(\varepsilon_t)\}$$

$$= -E[\rho(z_t,\theta_0)s(\varepsilon_t)],$$

where $\rho(z_t,\theta) = [m_1(y_t-x_t'\beta) - \mu_j,\ldots,m_J(y_t-x_t'\beta) - \mu_J]'$ and $\mu_{j0} = E[m_j(\varepsilon_t)]$. Therefore, $-\tilde{M}$ is an estimate of $E[\rho(z_t,\theta_0)s(\varepsilon_t)]$. Furthermore, the matrix $\tilde{\Sigma}$ is an estimate of $\Sigma = E[\rho(z_t,\theta_0)\rho(z_t,\theta_0)']$, so that $\hat{d}_J = -\tilde{\Sigma}^{-1}\tilde{M}$ is an estimate of the $d_J = \Sigma^{-1}E[\rho(z_t,\theta_0)s(\varepsilon_t)]$, which is the vector of least squares coefficients of the projection of $s(\varepsilon_t)$ on $\rho(z_t,\theta_0)$. It follows that $s_J(\varepsilon_t) = \rho(z_t,\theta_0)'d_J$ is the population projection of $s(\varepsilon_t)$ on $\rho(z_t,\theta_0)$ and that $\mathcal{J}_J = M'\Sigma^{-1}M = d_J'\Sigma d_J = E[s_J(\varepsilon_t)^2]$ is the variance of this projection. Consequently, $\hat{s}_J(\varepsilon_t) = \hat{d}_J'\rho(z_t,\theta_0)$ is an estimate of the minimum mean square error projection of the actual, unknown score function on linear combinations of the elements of $\rho(z_t,\theta_0)$, and $\hat{\mathcal{J}}_J$ is the estimated variance of this minimum mean square error approximation. We now see that the LGMM estimator is an approximation to the LMLE estimator involving implicit estimated mean square error approximations to the unkown components of the LMLE. A similar interpretation of the (unlinearized) GMM estimator can also be given.

This interpretation is very suggestive concerning the efficiency of the LGMM estimator. It suggests that the better that linear combinations of the moment functions can approximate the score in mean square the closer the LGMM estimator will be to being as efficient as the LMLE. Furthermore, if it is possible to approximate the score arbitrarily well, in terms of mean square, by choosing the number of moment functions, J, to be sufficiently large, it should be possible to obtain a LGMM estimator of the slope coefficients that is as efficient as the LMLE, i.e. is adaptive in the sense of Bickel (1982), by letting J grow with the sample size. As J grows the linearized optimal GMM estimator should closely approximate the LMLE.

The possibility of adaptation via the LGMM estimator depends crucially on a linear combination of a sufficient number of moment functions providing an arbitrarily good approximation to the unknown score, i.e. on $\lim_{J\to\infty}E\{[s(\varepsilon_t)-d_J'\rho(z_t,\theta_0)]^2\} = 0$. To understand conditions under which such an approximation is available, consider the special case that arises when $\bar{m}_j(\varepsilon) = [\bar{m}_0(\varepsilon)]^j$. In this case the LGMM estimator is using the constancy of the raw moments of $\bar{m}_0((\varepsilon-\alpha_0)/\sigma_0)$. Let $m_0(\varepsilon) = \bar{m}_0((\varepsilon-\alpha_0)/\sigma_0)$.

Lemma 2.2: Suppose that Assumption 2.1 is satisfied and $m_0(\varepsilon)$ is an increasing, continuously differentiable function such that $m_{0\varepsilon}(\varepsilon) > 0$ for all ε. Also suppose that for each positive integer j there is a neighborhood N of zero such that $E[\sup_{\alpha\in N}|m_0(\varepsilon_t-\alpha)|^{4j}]$ and $E[\sup_{\alpha\in N}|\partial\{m_0(\varepsilon_t-\alpha)\}^j/\partial\varepsilon|]$ are finite. If the moments of $m_0(\varepsilon_t)$ characterize the distribution of $m_0(\varepsilon_t)$ then $0 = \lim_{J\to\infty}E\{[s(\varepsilon_t)-d_J'\rho(z_t,\theta_0)]^2\}$.

One condition that is sufficient for moments of $m_0(\varepsilon_t)$ to characterize the distribution of $m_0(\varepsilon_t)$ is that the moment generating function of $m_0(\varepsilon_t)$ is well defined (e. g. see Billingsley (1979) p. 345). Thus, if $m_0(\varepsilon)$ is bounded and has a sufficiently well behaved derivative, the hypotheses of this Lemma will be satisfied for all distributions of ε_t satisfying Assumption 2.1.

The mean square error approximation of the score by the sequence of moment functions can be used as the basis of showing that the linearized GMM estimator is adaptive when J is allowed to grow at an appropriate rate with the sample size. To obtain an appropriate rate of growth for J it is useful to impose further regularity conditions. For a matrix $A = [a_{ij}]$ let $|A| = \max_{i,j}|a_{ij}|$. Also let $X_t = (1,x_t')'$.

Assumption 2.2: x_t is independently not (necessarily) identically distributed and $\Sigma_{t=1}^n E[|x_t|^{2+\delta}]/n$ exists and is bounded for some $\delta > 0$. Also, $Q = \lim_{n\to\infty}\Sigma_{t=1}^n E[X_tX_t']/n$ exists and is nonsingular.

This is a standard type of regularity condition for the regressors that allows for either fixed or random regressors.

Assumption 2.3: The function $\bar{m}_0(\varepsilon)$ is continuously differentiable. Also, for any $\sigma_0 > 0$ and α_0 there exists a neighborhood N of (α_0,σ_0), measurable functions $B_1(\varepsilon)$ and $B_2(\varepsilon)$, and $\tau > 0$ such that $E[\exp(\tau B_1(\varepsilon_t))]$ and $E[B_2(\varepsilon_t)^4(1+\varepsilon_t^4)]$ exist and for all ε and $(\alpha,\sigma) \in N$,

$$(2.14)\quad \sup_N|\bar{m}_0((\varepsilon-\alpha)/\sigma)| \le B_1(\varepsilon), \quad \sup_N|\bar{m}_{0\varepsilon}((\varepsilon-\alpha)/\sigma)| \le B_2(\varepsilon),$$

$$|\bar{m}_{0\varepsilon}((\varepsilon-\alpha)/\sigma) - \bar{m}_{0\varepsilon}((\varepsilon-\alpha_0)/\sigma_0)| \le B_2(\varepsilon)|(\alpha,\sigma) - (\alpha_0,\sigma_0)|.$$

The choice of moment functions and Assumption 2.3 can impose some restrictions on the distribution of ε_t. For example, if $\bar{m}_0(\varepsilon) = \varepsilon$ then Assumption 2.3 implies existence of the moment generating function of ε_t. Of course such a choice of moment function would not be appropriate in general. It is possible to choose $\bar{m}_0(\varepsilon)$ such that Assumption 2.3 is satisfied for all distributions of ε_t by choosing $\bar{m}_0(\varepsilon)$ to be a bounded function with sufficiently well-behaved first derivative. For example, the function $\bar{m}_0(\varepsilon) = \varepsilon/[1+|\varepsilon|]$ will satisfy Assumption 2.3 for any distribution of ε_t.

The following result gives a growth rate for J such that $\hat{\beta}_s$ is adaptive. Note that the block of the inverse of the limit of the average of the information matrix corresponding to β is $(\mathcal{J}Q_x)^{-1}$ where $(Q_x)^{-1}$ is the lower right $k \times k$ block of Q^{-1}.

Theorem 2.3: If Assumptions 2.1 - 2.3 are satisfied, $\sqrt{n}(\tilde{\beta}-\beta_0)$, $\sqrt{n}(\tilde{\alpha}-\alpha_0)$, and $\sqrt{n}(\tilde{\sigma}-\sigma_0)$ are bounded in probability, and $J = J(n)$ is chosen such that $J(n) \to \infty$ and $J(n)^2\ln(J(n))/\ln(n) \to 0$, then

$$(2.15)\qquad \sqrt{n}(\hat{\beta}_s-\beta_0) \xrightarrow{d} N(0,(\mathcal{J}Q_x)^{-1}), \quad (\hat{\mathcal{J}}_J\hat{Q}_x)^{-1} \xrightarrow{p} (\mathcal{J}Q_x)^{-1}.$$

The growth rate for the number of moment functions that is specified in this theorem is quite slow, being slower than the square root of the natural log of the sample size. One suspects that faster growth rates may also give asymptotic efficiency, although the task of obtaining such rates has not been attempted here.

The adaptive estimation result of Theorem 2.1 has attractive features relative to some of the results that have been previously presented in the literature. In particular, there is no sample splitting or discretization of the parameter space (e.g. see Bickel (1982) or Manski (1984)). Also, the result allows for a fixed design (i.e. nonrandom regressors), which appears not to have been allowed in previous results.

Although local regularity of the estimator $\hat{\beta}_s$ for families of regular likelihoods (e.g. Bickel (1982)) has not been shown here, it is expected that $\hat{\beta}_s$ will be locally regular, under additional regularity conditions (including local regularity of $\tilde{\beta}$, $\tilde{\alpha}$, and $\tilde{\sigma}$). In particular, it is shown in the proof of Theorem 2.3 that $\sqrt{n}(\hat{\beta}_s-\hat{\beta}_s^*) \xrightarrow{p} 0$, so that by the contiguity property of data generated from a sequence of regular likelihoods (see Hajek and Sidak (1967), Ch. 6) local regularity of $\hat{\beta}_s$ will be a corollary of local regularity of $\hat{\beta}_s^*$.

The adaptive estimation result of Theorem 2.3 is specific to moment functions of the form $\bar{m}_j(\varepsilon) = [\bar{m}_0(\varepsilon)]^j$. It is expected that this result will extend to other types of moment functions, although this extension has not been attempted here. Also, although only the linear regression case has been discussed here, the extension to nonlinear regression models is straightforward. This extension can be accomplished by appropriate modifications of Assumption 2.2 and the definition of the LGMM estimator.

3. Conditional Symmetry of the Disturbance

Another important regression model is that with a disturbance which is symmetrically distributed around zero conditionally on the regressors. Consider the model

$$(3.1) \qquad y_t = X_t'b_0 + \varepsilon_t, \qquad (t=1,2,\ldots),$$

where X_t is a $k \times 1$ vector of regressors that may include a constant, b_0 is a $k \times 1$ vector of parameters, (X_t', ε_t) is i.i.d., and the disturbance ε_t is symmetrically distributed around zero conditionally on the regressors. In this model the symmetry of the distribution of the disturbance implies that $X_t'b_0$ is the single natural measure of the center of location of the conditional distribution of the dependent variable y_t. Note that this model allows for (conditional) heteroskedasticity, in that the conditional distribution of ε_t is allowed to depend on X_t.

The assumption that the disturbance is symmetrically distributed yields many conditional moment restrictions that can be used in the estimation of the regression coefficients b_0. Conditional symmetry of the disturbance implies that any odd function $m(\varepsilon_t)$ (i.e. $m(-\varepsilon) = -m(\varepsilon)$) of the disturbance should be uncorrelated with any function of the regressors. This type of moment restriction can be used to form a GMM estimator. Consider a sequence $\{\bar{m}_1(\varepsilon), \bar{m}_2(\varepsilon), \ldots\}$ of odd, differentiable functions that have finite second moment. For some positive integer J, let $z = (y,X)$, $\bar{\rho}_j(z,b) = \bar{m}_j(y-X'b)$, and $\rho(z,b) = (\bar{\rho}_1(z,b),\ldots,\bar{\rho}_J(z,b))'$. Conditional symmetry of ε_t implies that for $\theta_0 = b_0$ the conditional moment restriction of equation (2.2) will be satisfied, which in turn implies that $\bar{\rho}_j(z,b_0)$ will be uncorrelated with any function of X_t. It follows that equation (2.3) will be satisfied for $\theta_0 = b_0$ and $\bar{g}(z,b) = \bar{\rho}(z,b) \otimes a(X)$, where $a(X)$ is a $K \times 1$ vector of functions of the regressors. A GMM estimator of b_0 can be based on the moment restrictions of equation (2.3) exactly as discussed in Section 2.

As in Section 2 it is useful to consider an optimal, scale adjusted, linearized version of the GMM estimator. In general the construction of an optimal GMM estimator of the type considered in Section 2 is problematical in the symmetric case, because heteroskedasticity must be allowed for. For given choice of $a(X)$ the optimal weighting matrix W_n is still as given in Section 2, but the optimal form of $a(X_t)$ now involves the unknown data generating process. In particular, the optimal form of $a(X_t)$ depends on $E[\partial\bar{\rho}(z_t,b_0)/\partial b|X_t] = -X_tE[\bar{m}_{j\varepsilon}(\varepsilon_t)|X_t]$, as well as on the unknown conditional covariance matrix of $\bar{\rho}(z_t,b_0)$, as discussed by Chamberlain (1987). An alternative strategy for efficient GMM estimation that will be adopted in this section is to allow the number of components of $a(X)$ to grow with the sample size with the idea of obtaining an estimator that behaves in large samples like the one that uses an optimal choice of $a(X)$. The feasibility of this approach is suggested by Chamberlain's (1987) observations concerning the near optimality of a GMM estimator with a large number of components in $a(X)$.

Concerning adjustment for the location and scale of the GMM estimator, note that when X_t includes a constant variable the estimator is already location equivariant. The GMM estimator can be made scale equivariant by replacing $\bar{m}_j(\varepsilon)$ with $\tilde{m}_j(\varepsilon) = \bar{m}_j(\varepsilon/\tilde{\sigma})$, where $\tilde{\sigma}$ is an estimator of a population scale parameter σ_0. As in Section 2 , the asymptotic distribution of the resulting GMM estimator of b will be identical to that obtained if the population value of the scale parameter was used in the formation of $\tilde{m}_j(\varepsilon)$, i.e. to that obtained if $m_j(\varepsilon) = \bar{m}_j(\varepsilon/\sigma_0)$ were used in place of $\tilde{m}_j(\varepsilon)$. This lack of effect from estimation of σ_0 occurs because $\partial m_j(\varepsilon_t/\sigma)/\partial\sigma = -m_{j\varepsilon}(\varepsilon_t/\sigma)\varepsilon_t/(\sigma^2)$ is an odd function of ε_t and so has conditional expectation zero.

A scale adjusted, LGMM estimator that uses the optimal weighting matrix can be obtained by linearizing as in Section 2. Define the scale adjusted function $\tilde{\rho}_j(z_t,b) = \tilde{m}_j(y_t-X_t'b)$. Let $\tilde{b}$ be an initial estimator of b_0 and $\tilde{\varepsilon}_t = y_t - X_t'\tilde{b}$. A first order expansion of $\tilde{\rho}_j(z_t,b)$ around $\tilde{b}$ gives

$$(3.2)\qquad \tilde{\rho}_j(z_t,b) \cong \tilde{\rho}_j(z_t,\tilde{b}) + [\partial\tilde{\rho}_j(z_t,\tilde{b})/\partial b]'(b - \tilde{b})$$

$$= \tilde{m}_j(\tilde{\varepsilon}_t) + \tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)X_t'\tilde{b} - \tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)X_t'b.$$

Let

$$(3.3)\qquad \tilde{V} = [\Sigma_{t=1}^{n}\tilde{\rho}(z_t,\tilde{b})\tilde{\rho}(z_t,\tilde{b})'\otimes a(X_t)a(X_t)']/n.$$

A LGMM estimator with optimal weighting matrix can now be obtained by choosing $W_n = \tilde{V}^{-1}$, and replacing $\bar{\rho}_j(z_t,\theta)$ in the definition of $\bar{g}_n(\theta)$ with the scale adjusted, linearized function $\tilde{m}_j(\varepsilon_t) + \tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)X_t'\tilde{b} - \tilde{m}_{j\varepsilon}(\tilde{\varepsilon}_t)X_t'b$. To be specific let $\tilde{Z} = [\tilde{m}_{1\varepsilon}(\tilde{\varepsilon}_1)X_1,\ldots,\tilde{m}_{1\varepsilon}(\tilde{\varepsilon}_n)X_n,\ldots,\tilde{m}_{J\varepsilon}(\tilde{\varepsilon}_1)X_1,\ldots,\tilde{m}_{J\varepsilon}(\tilde{\varepsilon}_n)X_n]'$, $\tilde{Y} = (\tilde{m}_1(\tilde{\varepsilon}_1),\ldots,\tilde{m}_1(\tilde{\varepsilon}_n),\ldots\tilde{m}_J(\tilde{\varepsilon}_1),\ldots,\tilde{m}_J(\tilde{\varepsilon}_n))' + \tilde{Z}\tilde{b}$, and $\tilde{X} = I_J\otimes[a(X_1),\ldots a(X_n)]'$. Then a LGMM estimator is given by

$$(3.4)\qquad \hat{b}_s = [\tilde{Z}'\tilde{X}(\tilde{V}^{-1})\tilde{X}'\tilde{Z}]^{-1}\tilde{Z}'\tilde{X}(\tilde{V}^{-1})\tilde{X}'\tilde{Y}.$$

This estimator is formally identical to a system version of White's (1982) two stage instrumental variables estimator which partially corrects for heteroskedasticity of unknown form.

The LGMM estimator can be interpreted as an approximation to the LMLE that could be obtained if the conditional distribution of the disturbance were completely known. This interpretation is similar to that obtained in the independence case. Let $f(\varepsilon|X)$ denote the conditional density function of ε_t given X_t. Let $s(\varepsilon,X) = f_\varepsilon(\varepsilon|X)/f(\varepsilon|X)$ be the conditional score corresponding to this density function. For $z = (X',y)$ the score for b is then given by $S(z,b) = -Xs(y-X'b,X)$, and the information matrix for a single observation by $\Omega = E[S(z_t,b_0)S(z_t,b_0)'] = E[s(\varepsilon_t,X_t)^2X_tX_t']$. Thus, for an initial estimator $\tilde{b}$ and the outer product estimate of the information matrix, a LMLE can be obtained as

$$(3.5) \qquad \hat{b}^*_s = \tilde{b} + [\Sigma_{t=1}^n S(z_t,\tilde{b})S(z_t,\tilde{b})'/n]^{-1}\Sigma_{t=1}^n S(z_t,\tilde{b})/n.$$

Also, let $\hat{D} = \tilde{V}^{-1}\tilde{X}'\tilde{Z}/n$ and $\hat{S}(z_t,b) = \hat{D}'[\tilde{\rho}(z_t,b)\otimes a(X_t)]$. Note that $\tilde{Z}'\tilde{X}(\tilde{V}^{-1})\tilde{X}'\tilde{Z} = n^2\hat{D}'\tilde{V}\hat{D} = n\Sigma_{t=1}^n\hat{S}(z_t,\tilde{b})\hat{S}(z_t,\tilde{b})'$ and $\tilde{Z}'\tilde{X}(\tilde{V}^{-1})\tilde{X}'(\tilde{Y}-\tilde{Z}\tilde{b}) = n\hat{D}'\tilde{X}'(\tilde{Y}-\tilde{Z}\tilde{b}) = \Sigma_{t=1}^n\hat{S}(z_t,\tilde{b})$. Applying these definitions to equation (3.4) yields

$$(3.6) \qquad \hat{b} = \tilde{b} + [\Sigma_{t=1}^n\hat{S}(z_t,\tilde{b})\hat{S}(z_t,\tilde{b})'/n]^{-1}\Sigma_{t=1}^n\hat{S}(z_t,\tilde{b})/n.$$

From equations (3.5) and (3.6) it is apparent that the LMLE and the LGMM estimator are identical when the same preliminary estimate $\tilde{b}$ is used to form each, except that the LGMM estimator uses $\hat{S}(z_t,b)$ in its formation rather than $S(z_t,b)$. An interpretation of the LGMM estimator as an approximation to the LMLE can therefore be obtained by interpreting $\hat{S}(z_t,b)$ as an approximation to $S(z_t,b)$. Let $\rho(z,b) = (m_1(y-X'b),\ldots,m_J(y-X'b))' = (\bar{m}_1((y-X'b)/\sigma_0),\ldots,\bar{m}_J((y-X'b)/\sigma_0))'$, $g(z,b) = \rho(z,b)\otimes a(X)$, and $M(\varepsilon) = (m_{1\varepsilon}(\varepsilon),\ldots,m_{J\varepsilon}(\varepsilon))'$, and note that $\partial g(z,b)/\partial b = -M(y-X'b)\otimes(a(X)X')$ and that $\tilde{X}'\tilde{Z}/n$ is an estimator of $-E[\partial g(z_t,b_0)/\partial b]$. If the conclusion of Lemma 2.1 applies to the conditional distribution of ε_t then it follows that $E[M(\varepsilon_t)|X_t] = -E[\rho(z_t,b_0)s(\varepsilon_t,X_t)|X_t]$, so that $\tilde{X}'\tilde{Z}/n$ also estimates $E[-E[\rho(z_t,b_0)s(\varepsilon_t,X_t)|X_t]\otimes a(X_t)X_t'] = E[g(z_t,b_0)S(z_t,b_0)']$, and $\hat{D}$ will estimate

$$\begin{aligned}(3.7) \qquad D &= -\{E[g(z_t,b_0)g(z_t,b_0)']\}^{-1}E[\partial g(z_t,b_0)/\partial b] \\ &= \{E[g(z_t,b_0)g(z_t,b_0)']\}^{-1}E[E[M(\varepsilon_t)|X_t]\otimes a(X_t)X_t'] \\ &= \{E[g(z_t,b_0)g(z_t,b_0)']\}^{-1}E[g(z_t,b_0)S(z_t,b_0)'],\end{aligned}$$

which is the matrix of coefficients from the least squares projection of each component of the score $S(z_t,b_0)$ on the components of $g(z_t,b_0)$. Thus, when equation (3.7) holds $\hat{S}(z_t,b)$ can be interpreted as an estimate of the minimum mean square error approximation to the score $S(z_t, b_0)$ obtained by projecting the score on the space spanned by the components of the moment function vector $g(z_t, b_0)$. Consequently, if linear combinations of products of functions of the regressors with odd functions of the disturbance can provide an arbitrarily good mean square error approximation to $X_t s(\varepsilon_t|X_t)$, it should be possible to obtain an adaptive LGMM estimator by letting the number of components of $a(X)$ and $\rho(z,b)$ grow with the sample size.

The following assumption is useful to guarantee that equation (3.7) and the other results given in this section hold:

Assumption 3.1: (X_t', ε_t) is i.i.d. and its distribution is absolutely continuous with respect to the Cartesian product of the probabilty measure for X_t and Lebesgue measure, and has (conditional) density $f(\varepsilon|X)$. For almost all X_t, $f(\varepsilon|X_t)$ is absolutely continuous and $E[\int[f_\varepsilon(\varepsilon|X_t)^2/f(\varepsilon|X_t)]d\varepsilon]$ exists. There is an interval N and a positive constant c such that $\inf_{\varepsilon \in N} f(\varepsilon|X_t) \geq c$ with probability one.

The assumption that the conditional density is bounded away from zero on some interval uniformly in X_t restricts somewhat the heterogeneity of the conditional distribution of ε_t. For example, if $f(\varepsilon|X_t) = f(\varepsilon/\sigma(X_t))$ for some positive function $\sigma(X_t)$ of X_t, then this assumption implies that $\sigma(X_t)$ is bounded and bounded away from zero.

Assumption 3.1 implies that $\int[f_\varepsilon(\varepsilon|X_t)^2/f(\varepsilon|X_t)]d\varepsilon$ is finite with probability one. Then, if with probability one there is a neighborhood of zero on which $E[m_j(\varepsilon_t+\alpha)^4|X_t]$ is bounded in α and the order of differentiation and integration can be interchanged, the conclusion of Lemma 2.1 implies that $E[m_{j\varepsilon}(\varepsilon_t)|X_t] = -E[m_j(\varepsilon_t)s(\varepsilon_t, X_t)|X_t]$ with probability one and equation (3.7) will hold.

The nature of the implicit approximation of the LGMM estimator to the LMLE is somewhat different in the symmetric case than the independence case. Note that the approximation involves approximating each element of the score for the regression parameters rather than just the score for the disturbance. More importantly, the approximation is a multivariate one, involving not only the disturbance but also the regressors. Thus, constructing an adaptive estimator will involve an appropriate growth rate for the number of components of $a(X)$ as well as for the number of components of $\bar{\rho}(z,b)$. To obtain such growth rates it is useful to impose a restriction on the distribution of the regressors and to be specific about the form of $a(X)$ and $\bar{\rho}(z,b)$.

The form for $\bar{\rho}(z,b)$ considered in this section will be analogous to that considered in Section 2, with $\bar{m}_j(\varepsilon) = [\bar{m}_0(\varepsilon)]^{2j-1}$, $(j=1,\ldots,J)$, where $\bar{m}_0(\varepsilon)$ is an odd, monotonic increasing function of ε. It will also be assumed that $\bar{m}_0(\varepsilon)$ satisfies the regularity conditions of Assumption 2.3, an assumption that will be made explicit in the statement of the theorem below.

The restriction that will be imposed on the distribution of the regressors is that they consist of functions of discrete and continuous components. Let w_{1t} be a random variable with finite support. The set of possible realizations of w_{1t} is meant to represent the set of possible outcomes for the discrete components of the regressors. For example, if the discrete components of X_t consist of two dummy variables then w_{1t} will have four possible outcomes. Let $w_{2t} = (w_{2t}^1,\ldots,w_{2t}^{\mathcal{L}})'$ be a $\mathcal{L} \times 1$ vector of continuously distributed random variables which help determine the value of the regressors, and let $w_t = (w_{1t},w_{2t}')'$.

Assumption 3.2: $X_t = X(w_t)$ for some measurable, one to one function $X(w)$: $\mathbb{R}^{\mathcal{L}+1} \to \mathbb{R}^k$ and for some $\delta > 0$ $E[|X_t|^{2+\delta}]$ is finite. Also, the support of w_{1t} is a finite set and for each element of this support there is an open subset $\mathcal{O}(w_{1t})$ of $\mathbb{R}^{\mathcal{L}}$ such that on $\mathcal{O}(w_{1t})$ the conditional distribution of w_{2t} given w_{1t} is absolutely continuous with density function bounded away from zero.

The form of $a(X)$ that will be considered is analogous to the form considered for $\bar{\rho}(z,b)$. It will be assumed that $a(X) = a_1(w_1)\otimes a_2(w_2)$, where $a_2(w_2)$ is made up of products of powers of monotonic functions $a_0^{\ell}(w_2^{\ell})$, $(\ell=1,\ldots,\mathcal{L})$, of the individual components of w_2. It will be assumed that there are $\mathcal{J}$ such terms, i.e. that

$$a_2(w_2) = (\Pi_{\ell=1}^{\mathcal{L}}[a_0^{\ell}(w_2^{\ell})]^{\nu_{\ell 1}},\ldots,\Pi_{\ell=1}^{\mathcal{L}}[a_0^{\ell}(w_2^{\ell})]^{\nu_{\ell\mathcal{J}}}),$$

where $\nu_{\ell j}$, $(\ell=1,\ldots,\mathcal{L}, j=1,\ldots,\mathcal{J})$, are all nonnegative integers. Conditions under which such a choice of $a(X)$ can assist in providing an arbitrarily good approximation to the likelihood score are analogous to those for $\rho(z_t,b_0)$. First of all, for K large enough, $a_1(w_{1t})$ must span the possible realizations for w_{1t}. For example, if the discrete components of X_t consist of two dummy variables it suffices to let $a_1(w_{1t})$ consist of a constant, each dummy variable, and the product of the two dummy variables. In addition, for each possible value of w_{1t} the vector $a_2(w_{2t})$ must be able to form an arbitrarily good mean square approximation to functions of w_{2t} when K is large enough. For this spanning condition to hold it is enough to assume that the moment generating function of each $a_0^{\ell}(w_{2t}^{\ell})$ exists and that all cross-products of all powers of each $a_0^{\ell}(w_2^{\ell})$ are eventually used (i.e. for large enough K) as components of $a_2(w_2)$. Let ν_* be the smallest nonnegative integer such that $a_2(w_2)$ includes all terms of the form $\Pi_{\ell=1}^{\mathcal{L}}[a_0^{\ell}(w_2^{\ell})]^{\nu_\ell}$ for $\Sigma_{\ell=1}^{\mathcal{L}}\nu_\ell \le \nu^*$.

Assumption 3.3: For each ℓ, $a_2^{\ell}:\mathbb{R}\rightarrow\mathbb{R}$ is a monotonic increasing, continuously differentiable function with everywhere positive derivative, such that for some $\tau > 0$, $E[\exp(\tau|a_2^{\ell}(w_{2t}^{\ell})|)]$ is finite. Also, the vector $a(X)$ is chosen in such a way that for all K large enough, $a(X_t) = a_1(w_{1t})\otimes a_2(w_{2t})$ such that $E[a_1(w_{1t})a_1(w_{1t})']$ is nonsingular, the number of components of $a_1(w_{1t})$ is equal to the number of elements in the support of w_{1t}, and ν_* goes to inifinity with K.

Note that if each a_0^{ℓ} is chosen to be a bounded function, then the above condition that the moment generating function of $a_0^{\ell}(w_{2t})$ exists is not restrictive.

Under these regularity conditions it is possible to give growth rates for J and K as a function of n such that the LGMM estimator is adaptive in the symmetric case considered here. Let ν^* denote the smallest integer such that $\Sigma_{\ell=1}^{\mathcal{L}} \nu_{\ell j} \le \nu^*$, $(j=1,\ldots,\mathcal{J})$ (i.e. largest order of the components of $a_2(w_{2t})$).

Theorem 3.1: Suppose that Assumptions 2.3 and 3.1 - 3.3 are satisfied, $\sqrt{n}(\tilde{b}-b_0)$ and $\sqrt{n}(\tilde{\sigma}-\sigma_0)$ are bounded in probability, and the information matrix $\Omega = E[X_tX_t's(\varepsilon_t|X_t)^2]$ is nonsingular. If $J = J(n)$ and $K = K(n)$ are chosen such that $J(n) \to \infty$, $K(n) \to \infty$, and $[J(n)\nu^*(n)]^3[\nu^*(n)]^{\mathcal{L}-1}\ln[J(n)\nu^*(n)]/\ln(n) \to 0$, then

$$(3.8) \qquad \sqrt{n}(\hat{b}_s-b_0) \xrightarrow{d} N(0,\Omega^{-1}), \quad [\Sigma_{t=1}^{n}\hat{S}(z_t,\tilde{b})\hat{S}(z_t,\tilde{b})'/n]^{-1} \xrightarrow{p} \Omega^{-1}.$$

To interpret the specified growth rate for $J(n)$ and $K(n)$, note that if terms are added to $a_2(w_{2t})$ by adding all terms of a given order before increasing the order, then K and $(\nu^*)^{\mathcal{L}}$ are of the same order, i.e. $(\nu^*)^{\mathcal{L}} = O(K)$. In this case $J(\nu^*)^{\mathcal{L}}$ and the dimension of $g(z,b)$ (which is $J\circ K$) are of the same order, so that (since $3\mathcal{L} \ge 3 + \mathcal{L} - 1$) a growth rate for the total number of moment functions slightly slower than the cube root of the natural log of sample size will suffice. The drop in the growth rate from Theorem 2.3 to Theorem 3.1 results from the use of incomplete order polynomial terms (e.g. using only odd powers of $\bar{m}_0(\varepsilon)$), but is probably not really necessary.

The remarks following Theorem 2.3 concerning the growth rate for the number of moment functions, local regularity of the estimator, and extensions of the result also apply to Theorem 3.1.

4. Sampling Experiments

To obtain information concerning the small sample performance of the LGMM estimator, two sampling experiments were carried out. Attention was restricted to the independence case to allow comparison with the results for the nonparametric adaptive maximum likelihood (AML) estimator reported by Hsieh and Manski (1987) (HM henceforth).

The first experiment involved the same model, sample size, and a subset of the distributions considered by HM. The model was $y_t = \alpha_0 + \beta_0 x_t + \varepsilon_t$, where x_t is a binomial random variable with $\text{Prob}(x_t = 0) = \text{Prob}(x_t = 1) = 1/2$, $\alpha_0 = -1$, and $\beta_0 = 1$. The distributions considered were: A. Standard normal; B. Variance contaminated mixture of normals with relative scale of nine, being $.1N(0,9) + .9N(0,1/9)$; C. Bimodal symmetric mixture of normals, being $.5N(-3,1) + .5N(3,1)$; D. Lognormal, being $\exp(u)$ where u is distributed as standard normal. Where necessary ε_t was normalized to have mean zero and variance one. The sample size was 50.

Computations were performed using GAUSS on a microcomputer. Table One reports the root mean square error (RMSE) of several different estimators of β_0 for each of the four distributions, estimated from 2000 replications. The estimators considered were ordinary least squares (OLS), least absolute deviations (LAD), AML, and two different LGMM estimators for choices of J (the number of moments used in estimation) between 2 and 7. One LGMM estimator, referred to in the tables as TRANSFORMED, used $\bar{m}_j(\varepsilon) = [\varepsilon/(1+|\varepsilon|)]^j$ and the other, referred to as WEIGHTED, used $\bar{m}_j(\varepsilon) = \exp(-\varepsilon^2/2)\varepsilon^j$. The initial estimator of β_0 used to form the AML and LGMM estimators was the OLS estimator. The location and scale parameters used in both the LGMM and the AML estimators were the sample mean and standard deviation of the OLS residuals. The trimming and smoothing parameters used in the AML were fixed at the values that minimize the RMSE of the AML for each of the four distributions. These values were obtained by HM. In their notation these values were $t1 = t2 = 8$ and $t3 = \exp(-(8)^2/2)$ throughout, and $s = 2$, .25, .25, and .10 respectively for the four distributions.

Because of the location and scale adjustment the relative magnitudes of the RMSE results in Table One are invariant to the choice of α_0, β_0, and the scale of the disturbance. Also, because of the use of a symmetric (normal) kernel for the AML estimator and powers of an odd function, weighted by an even function for the LGMM estimator, the difference of each estimator and β_0 is an odd function of the disturbance realizations. Therefore, when the disturbance is symmetrically distributed each estimator will be symmetrically distributed around β_0, which occurs for each of the first three distributions. The estimators also appear to be unbiased for all the distributions. In calculations not reported here it was found that the average deviation of the estimators from β_0 was typically two orders of magnitude smaller than the RMSE.

Turning to the results reported in Table One, note that in the normal case the RMSE of the LGMM estimator was estimated to be 8 - 12 percent larger than that of the OLS and AML estimators for the majority of J values. However, in the nonnormal cases the RMSE of the AML estimator was found to be 50-100 percent larger than that of the LGMM estimators in many cases. Furthermore, for most values of J the RMSE of the LGMM estmators is smaller than that of the LAD estimator for all the distributions. Thus, the LGMM estimator seems to perform quite well relative to the other estimators in terms of its efficiency.

The performance of the LGMM estimators relative to the AML estimator in the nonnormal cases is quite suprising, but may be due in part to the more parsimonius score function estimate that is implicit in the LGMM estimator. The LGMM estimator involves an implicit approximation of the score by a linear combination of J functions while the AML estimator uses a kernel estimator of the unknown density function. It should be noted that the performance of the AML estimator reported in

Table One is slightly worse than that reported in HM, which may be due to the use here of the sample standard deviation of the OLS residuals for scale adjusting the AML estimator.

The value of J that gives the smallest RMSE for the LGMM estimator is $J = 3$ throughout Table One. The RMSE seems to rise sharply as J is decreased below 3 and more gradually as J rises above 3. The rise in the RMSE for J below 3 is particularly pronounced for the bimodal, symmetric mixture of normals. This result probably occurs because the score function for this density is shaped like a cubic polynomial and so is not very well approximated by $m_1(\varepsilon)$ alone (in the symmetric case $m_2(\varepsilon)$, which is an even power of an odd function, will be of no help in forming a mean square error approximation to the score). This result also suggests that the outstanding performance of the LGMM estimator in the nonnormal cases may be due in part to the simple nature of the score functions for the distributions considered here, which appear to be approximated reasonably well by a linear combination of the first three moment functions. The LGMM estimator may perform less well for distributions with more complicated score functions (e.g. a trimodal mixture of normals). Of course, the AML will also probably work less well in small samples when the score function has a more complicated shape.

As a function of J the RMSE of the WEIGHTED LGMM estimator has a lower but sharper minimum than the RMSE of the TRANSFORMED LGMM estimator. The best value of J gives a slightly lower RMSE for the WEIGHTED estimator but the RMSE rises more sharply as J departs from its best value. The relative insensitivity to J of the TRANSFORMED estimator is probably preferable to the slightly better RMSE performance of the best WEIGHTED estimator. For this reason attention will be restricted to the TRANSFORMED estimator in the second experiment.

Table Two reports the ratio of root mean square of the estimated standard errors to the RMSE for the LGMM estimators for the same cases considered in Table One. The estimated standard errors are taken from the formula given in the conclusion of Theorem 2.3. For $J = 3$ the ratio is within five percent of one in all the nonnormal cases and within 12 percent in the normal case. However, this ratio rapidly departs from one as J increases. Also, the ratio seems to be closer to one for the TRANSFORMED estimator in most cases.

To obtain some idea of how to choose J as a function of the sample size in small to medium size samples an additional experiment was performed. The model considered was $y_t = \alpha_0 + \beta_{10}x_{t1} + \beta_{20}x_{t2} + \varepsilon_t$, where x_{t1} is a binomial random variable with $\text{Prob}(x_{t1} = 0) = \text{Prob}(x_{t1} = 1) = 1/2$, x_{t2} is uniformly distributed on $(0,1)$, x_{t1} and x_{t2} are independent, $\alpha_0 = -1$, and $\beta_{10} = \beta_{20} = 1$. A case with two regressors was considered in order that the experimental model might correspond more closely to situations that arise in practice. The distribution of ε_t considered was the variance contaminated mixture normals considered in the first experiment (distribution B). The sample sizes considered were 50, 100, and 200.

Table Three reports ratios of the RMSE for the TRANSFORMED LGMM estimator ($\bar{m}_j(\varepsilon) = [\varepsilon/(1+|\varepsilon|)]^j$) to the RMSE of the OLS estimator in the second experiment. Suprisingly, the RMSE of the LGMM estimator is minimized at $J = 3$ for all the sample sizes considered here. Note that the RMSE as a function of J does flatten out rapidly as the sample size increases. The loss in efficiency which results from moving from $J = 3$ to $J = 7$ is only a few percentage points in the 100 observations case and still less in the 200 observations case.

Table Four reports ratios of the root mean square of the estimated standard errors to actual RMSE in the second experiment. This ratio is within five percent of one for J up to 4 for a sample size of 50, for J up to 6 for a sample size of 100, and for J up to 7 for a sample size of 200.

In summary, in the examples considered here the LGMM estimators perform well relative to the AML in terms of efficiency. All of the efficiency gain available from the LGMM estimator is obtained for $J = 3$

for the distributions considered here, which have simply behaved score functions. Standard error estimates approximate the actual variability of the estimates reasonably well in small samples for $J = 3$, and the performance for larger J improves substantially as the sample size increases.

The extreme sensitivity of the efficiency of the LGMM estimator to choosing J less than three in the bimodal mixture of normals case shows that the efficiency of the LGMM estimator can be very sensitive to the choice of J. Thus, it is important to have available a method of choosing J in particular applications. Choosing J so that the estimated standard error from the asymptotic formula is minimized will not work, because this estimated standard error declines as J increases. One method would be to choose J to minimize standard errors estimated by the bootstrap. A similar method of choosing the trimming and smoothing parameters for the AML has been considered by HM. Note that this method would be particularly simple for LGMM estimators since the choice involves the single, integer valued variable J. Assesing the small sample performance of LGMM estimators which use the bootstrap to choose J is beyond the scope of this paper but is an interesting topic for further study.

5. Conclusion

An adaptive estimator based on moment conditions has been presented for the regression model with a disturbance distributed independently of the regressors. Also, it has been demonstrated that this estimator can perform very well in small samples. In addition, an adaptive estimator has been presented for the regression model with a symmetric, heteroskedastic disturbance with (possibly) continuously distributed regressors, a model where no other adaptive estimator is currently known to exist.

Estimators arising from moment conditions offer a promising avenue of approach to efficient estimation in other environments. In Newey (1987b) it is shown that the interpretation of GMM estimators as approximations to efficient estimators holds quite generally. This interpretation and the asymptotic theory developed here can be used to obtain efficient estimators for models with conditional moment restrictions (Newey (1987a)), nonlinear simultaneous equations models with independent or conditionally symmetric disturbances (Newey (1987b)), and censored or truncated regression models with conditionally symmetric latent disturbances (Newey and Powell (1987b)). This method should also prove useful for constructing efficient instrumental variables estimators in time series models (e.g. Hansen (1985)). Of course, the availability of such estimators is limited to models where moment conditions based on known functions of the data and parameters can be specified, which is not the general case by any means (e.g. no such functions are known to exist for the binary choice model). Nevertheless, this class of models includes many interesting cases.

APPENDIX

Some Lemmas will first be stated and their proofs sketched. More detailed, handwritten proofs of some of the results are available upon request from the author. Throughout the appendix C will denote a generic positive constant that does not depend on the variable indices and need not be the same in different uses.

Lemma A1: Suppose that $w = (w_1,\dots,w_{\mathcal{L}})'$ is a $\mathcal{L} \times 1$ vector of random variables such that on some open set $\mathcal{O} \subset \mathbb{R}^{\mathcal{L}}$ w is absolutely continuous with density bounded away from zero. Also suppose that $a_0^{\ell}:\mathbb{R}\to\mathbb{R}$, $(\ell=1,\dots,\mathcal{L})$, are increasing, continuously differentiable functions with everywhere nonzero derivatives. Let $u = (u_1,\dots,u_{\mathcal{L}})'$, $p_{\mathcal{L}}(u_{\mathcal{L}}) = (1,u_{\ell},\dots,(u_{\ell})^{\nu})'$, $p(u) = p_1(u_1)\otimes\circ\circ\circ\otimes p_{\mathcal{L}}(u_{\mathcal{L}})$, and $a_0(w) = (a_0^1(w_1),\dots,a_0^{\mathcal{L}}(w_{\mathcal{L}}))'$. Then there is a constant C independent of ν such that for ν large enough,

$$(A.1) \qquad \det\{E[p(a_0(w))p(a_0(w))']\} \geq \nu^{-C\nu^{\mathcal{L}+1}},$$

and for any subvector $p^*(u)$ of $p(u)$

$$(A.2) \qquad \det\{E[p^*(a_0(w))p^*(a_0(w))']\} \geq \nu^{-C\nu^{\mathcal{L}+2}},$$

Proof: Assume without loss of generality (w.l.g.) that $\mathcal{O} = \mathcal{O}_1 \times \circ\circ\circ \times \mathcal{O}_{\mathcal{L}}$ for open intervals $\mathcal{O}_{\ell}$. Let ψ be a $\mathcal{L} \times 1$ vector and $\phi = \mathrm{diag}[\phi_1,\dots,\phi_{\mathcal{L}}]$ a $\mathcal{L} \times \mathcal{L}$ diagonal matrix with $\phi_{\ell} > 0$ such that $\{\phi a_0(w) + \psi : w \in \mathcal{O}\}$ contains $\mathcal{U} = (0,1)\times\circ\circ\circ\times(0,1)$. By the components of $a_0(w)$ monotonic and continuously differentiable, the change of variables $u = \phi a_0(w)+\psi$ yields a random vector u that has an absolutely continuous distribution with density bounded away from zero on $\mathcal{U}$. By this change of variables it follows that $E[p(a_0(w))p(a_0(w))'] \geq c\Sigma$, where c is a lower bound for the density of u on $\mathcal{U}$, $\Sigma = \int_{\mathcal{U}} p(\phi^{-1}(u-\psi))p(\phi^{-1}(u-\psi))'du$, and the inequality denotes the positive semi-definite partial order. Note that for $\Sigma_\ell = \int_0^1 p_\ell((u_\ell-\psi_\ell)/\phi_\ell)p_\ell((u_\ell-\psi_\ell)/\phi_\ell)'du_\ell$ we have $\Sigma = \Sigma_1\otimes\circ\circ\circ\otimes\Sigma_\ell$. Note also that $[(u_\ell-\psi_\ell)/\phi_\ell]^j$ is a j^{th} order polynomial with a coefficient of ϕ_ℓ^{-j} for u_ℓ^j, so that there is a lower triangular matrix L_ℓ with j^{th} diagonal element $L_{jj} = \phi_\ell^{-j+1}$ which does not depend on u_ℓ such that $p_\ell((u_\ell-\psi_\ell)/\phi_\ell) = L_\ell p_\ell(u_\ell)$. Thus, $\Sigma_\ell = L_\ell[\int_0^1 p_\ell(u)p_\ell(u)'du_\ell]L_\ell' = L_\ell H_{\nu+1}L_\ell'$, where $H_{\nu+1}$ is the Hilbert matrix of order $\nu+1$ (the ij^{th} element of the Hilbert matrix is $1/(i+j-1)$). Let $H^{-1} = [H^{ij}]$ denote its inverse. By repeated application of the formula $\det(B) = \det(B_{11})/\det(B^{22})$ for the determinant of a partitioned matrix $B = [B_{ij}]$, $(i,j=1,2)$ and its inverse $[B^{ij}]$, and by $(H_{\nu+1})_{11} = 1$, it follows that $\det(H_{\nu+1}) = 1/[\Pi_{j=2}^{\nu+1}H^{jj}]$. The closed form expression for the elements of the inverse Hilbert matrix given by Gregory and Karney (1969, p. 34) then yields

$$\det(H_{\nu+1}) = \Pi_{j=2}^{\nu+1} \frac{(2j-1)[(j-1)!]^4[(\nu+1-j)!]^2}{[(\nu+j)!]^2}$$

$$\geq \Pi_{j=2}^{\nu+1} 1/[(\nu+j)!]^2 \geq 1/[(2\nu)^{4\nu^2}] \geq \nu^{-8\nu^2}.$$

Also, note that $\det(L_\ell) = \Pi_{j=1}^{\nu}\phi_\ell^{-j+1} \geq (1+\phi_\ell)^{-\nu^2} \geq \nu^{-C\nu^2}$. Thus, $\det(\Sigma_\ell) = \det(L_\ell)^2\det(H_{\nu+1}) \geq \nu^{-C\nu^2}\det(H_{\nu+1}) \geq \nu^{-C\nu^2}$ for large enough ν. By the formula for the determinant of a Kronecker product it follows that $\det(\Sigma) = \Pi_{\ell=1}^{\ell}[\det(\Sigma_\ell)]^{\nu^{\ell-1}} \geq ((\nu^{-C\nu^2})^{\nu^{\ell-1}})^{\ell} \geq \nu^{-C\nu^{\ell+1}}$. Equation (A.1) then follows from the fact that for two matrices A and B, $A \geq B$ implies $\det(A) \geq \det(B)$.

Next, note that $|\Sigma_\ell| \leq \sup_{u\in(0,1)}|p_\ell((u-\psi_\ell)/\phi_\ell)|^2 \leq (1+\phi_\ell^{-1})^{\nu}(1+|\psi_\ell|)^{2\nu} \leq \nu^{C\nu}$, so that by the extremum characterization of the largest eigenvalue, $\max_{\zeta'\zeta=1}\zeta'\Sigma_\ell\zeta \leq (\nu+1)|\Sigma_\ell|$, and the fact that the determinant is equal to the product of the eigenvalues, it follows that the smallest eigenvalue of Σ_ℓ is no smaller than $\det(\Sigma_\ell)/[(\nu+1)|\Sigma_\ell|]^{\nu} \geq \nu^{-C\nu^2}[(\nu+1)\nu^{C\nu}]^{-\nu} \geq \nu^{-C\nu^2}$. Also, note that the eigenvalues of a Kronecker product consist of products of the eigenvalues of the component matrices, so that the smallest eigenvalue of Σ is no smaller than $(\nu^{-C\nu^2})^{\mathcal{L}} \ge \nu^{-C\nu^2}$. Therefore, by the extremum characterization of the smallest eigenvalue, the smallest eigenvalue of $E[p^*(a_0(w))p^*(a_0(w))']$ is bounded below by the smallest eigenvalue of $E[p(a_0(w))p(a_0(w))']$, which is bounded below by the smallest eigenvalue of $c\Sigma$, which is bounded below by $\nu^{-C\nu^2}$. Then eq. (A.2) follows from equality of the determinant and the product of the eigenvalues and the dimension of $E[p^*(a_0(w))p^*(a_0(w))']$ bounded above by $\nu^{\mathcal{L}}$, which gives $\det\{E[p^*(a_0(w))p^*(a_0(w))']\} \ge (\nu^{-C\nu^2})^{\nu^{\mathcal{L}}}$ for ν big enough.

Let $O_p(a_n)$ and $o_p(a_n)$ denote the usual order in probability notation.

Lemma A2: Consider a sequence of $\{h_i(z)\}_{i=1}^{\infty}$ of measurable functions and a sequence $\{z_t\}_{t=1}^{\infty}$ of independent random variables. For each integer I let $h_{tI} = (h_1(z_t),\ldots,h_I(z_t))'$. For some $1 < \omega \le 2$ and $\{B(I)\}_{I=1}^{\infty}$ let $\Sigma_{i=1}^{I}\sup_n\{\Sigma_{t=1}^{n}E[|h_i(z_t)|^{1+\omega}]/n\} = O(B(I))$. Then

$$(A.3) \qquad |\Sigma_{t=1}^{n}E(h_{tI})/n| = O_p(B(I)^{1/(1+\omega)}), \quad |\Sigma_{t=1}^{n}h_{tI}/n| = O_p(B(I)^{1/(1+\omega)}),$$

and for $\omega < 1$ $(\omega = 1)$ and any sequence $a_n \rightarrow \infty$ $(a_n = 1)$

$$(A.4) \qquad |\Sigma_{t=1}^{n}[h_{tI} - E(h_{tI})]/n| = O_p(a_n n^{-\omega/(1+\omega)}B(I)^{1/(1+\omega)}).$$

Proof: Eq. (A.3) follows by the definition of $B(I)$ and the Holder and Markov inequalities. For $\omega = 1$ eq. (A.4) follows by Chebyshev's inequality, and for $\omega < 1$ eq. (A.5) follows by the Chebyshev and Holder inequalities and the truncation argument used to prove Markov's law of large numbers.

Lemma A3: Let $\{h_i(z,\gamma)\}_{i=1}^{I}$ be a sequence of functions and $\{z_t\}_{t=1}^{\infty}$ a sequence of random variables where γ is a Euclidean vector. Suppose that there is a neighborhood Γ of γ_0 and a sequence of measurable functions $\{\zeta_i(z)\}_{i=1}^{\infty}$ such that for all γ in Γ the Lipschitz condition $|h_i(z_t,\gamma) - h_i(z_t,\gamma_0)| \le \zeta_i(z_t)|\gamma-\gamma_0|$ holds for all i with probability one. Let $\{\tilde{\gamma}_{tn}\}_{t=1,n=1}^{n\quad\infty}$ be a triangular array of random variables such that $\sup_{1\le t\le n}|\tilde{\gamma}_{tn}-\gamma_0| = O_p(n^{-\epsilon})$ for some $\epsilon >$ 0. Then for any $\{B(I)\}$ such that $\Sigma_{i=1}^{I}\sup_n\{\Sigma_{t=1}^{n}E[\zeta_i(z_t)]/n\} = O(B(I))$, it follows that for $h_{tI}(\gamma) = (h_1(z_t,\gamma),\ldots,h_I(z_t,\gamma))'$,

$$(A.5)\qquad |\Sigma_{t=1}^{n}[h_{tI}(\tilde{\gamma}_{tn})-h_{tI}(\gamma_0)]/n| = O_p(n^{-\epsilon}B(I)).$$

Proof: By $\epsilon > 0$ we have $\text{plim}_{n\to\infty}[\sup_{1\le t\le n}|\tilde{\gamma}_{tn}-\gamma_0|] = 0$, so that with probability approaching one $\tilde{\gamma}_{tn}$ lies in Γ for all $t \le n$. Note that $\Sigma_{i=1}^{I}\Sigma_{t=1}^{n}\zeta_i(z_t)/n = O_p(B(I))$ by Markov's inequality. The Lipshitz condition then gives

$$(A.6)\qquad |\Sigma_{t=1}^{n}[h_{tI}(\tilde{\gamma}_{tn})-h_{tI}(\gamma_0)]/n| \le \Sigma_{i=1}^{I}|\Sigma_{t=1}^{n}[h_{ti}(\tilde{\gamma}_{tn})-h_{ti}(\gamma_0)]/n|$$
$$\le \Sigma_{i=1}^{I}\Sigma_{t=1}^{n}\zeta_i(z_t)|\tilde{\gamma}_{tn}-\gamma_0|/n \le \sup_{t\le n}|\tilde{\gamma}_{tn}-\gamma_0|\Sigma_{i=1}^{I}\Sigma_{t=1}^{n}\zeta_i(z_t)/n$$
$$= O_p(n^{-\epsilon})O_p(B(I)) = O_p(n^{-\epsilon}B(I)).$$

Lemma A4: Let $h = (\text{vec}(H)',h_2')'$ be a $(J^2+J) \times 1$ vector, where H is a $J \times J$ matrix and h_2 is a $J \times 1$ vector. Consider a random vector $\hat{h}_n$ and a nonrandom vector h_n such that $\hat{h}_n = (\text{vec}(\hat{H}_n)',\hat{h}_{2n}')'$ and $h_n = (\text{vec}(H_n)',h_{2n}')'$. For $J = J(n)$ a function of the sample size n suppose that for sequences of positive constants $\{a_n\}$ and $\{b_n\}$ that are bounded away from zero,

$$(A.7)\qquad |h_n| = O(a_n),\quad |\hat{h}_n - h_n| = O(b_n),\quad b_n = O(a_n).$$

Also suppose that H_n is nonsingular for all n and that for some sequence $\{d_n\}$ of positive constants $1/\det(H_n) = O(d_n)$. Finally, suppose that $\{J(n)\}$ is chosen so that for some sequence of constants $\{c_n\}$ such that $c_n \to \infty$, $(J!)J(c_na_n)^{J-1}d_nb_n = o(1)$. Then

$$(A.8a)\quad |\hat{H}_n^{-1}\hat{h}_{2n} - H_n^{-1}h_{2n}| = o_p((d_nJ!)^2J^3(c_na_n)^{J-1}a_nb_n)$$

(A.8b) $|H_n^{-1}h_{2n}| = o(d_nJ!(c_na_n)^{J-1}a_n)$, $|\hat{H}_n^{-1}\hat{h}_{2n}| = o_p(d_nJ!(c_na_n)^{J-1}a_n)$.

Proof: Let p index the possible permutations of the integers from 1 to J. Then for $(H)_{k\ell}$ denoting the k,ℓ^{th} element of H it follows from the definition of the determinant that

(A.9) $$|\det(\hat{H}_n) - \det(H_n)|$$

$$\le |\Sigma_p(\Pi_{j=1}^{J}(\hat{H}_n)_{k(p,j)\ell(p,j)} - \Pi_{j=1}^{J}(H_n)_{k(p,j)\ell(p,j)}|$$

$$\le \Sigma_p[J\max(|\hat{H}_n|,|H_n|)^{J-1}|\hat{H}_n-H_n|] \le (J!)J[|\hat{h}_n|^{J-1}+|h_n|^{J-1}]|\hat{h}_n-h_n|$$

$$= (J!)J[o_p((c_na_n)^{J-1}) + o((c_na_n)^{J-1})]o_p(b_n) = o_p(J!J(c_na_n)^{J-1}b_n),$$

where the first equality follows from (A.7) via the fact that $|h_n| = o_p(a_nc_n)$ and $|\hat{h}_n| \le |\hat{h}_n-h_n| + |h_n| = o_p(a_nc_n)$, which implies $|\hat{h}_n/(a_nc_n)| \le |\hat{h}_n/(a_nc_n)|^{J-1}$ with probability approaching one, which in turn implies $|\hat{h}_n|^{J-1} = o_p((a_nc_n)^{J-1})$ and $|h_n|^{J-1} = o((a_nc_n)^{J-1})$. From (A.9) and the hypotheses it follows that $O(d_n)|\det(\hat{H}_n)-\det(H_n)| = o_p(d_nJ!J(c_na_n)^{J-1}b_n) = o_p(1)$, so that

(A.10) $$1/\det(\hat{H}_n) = 1/[\det(H_n) + \det(\hat{H}_n) - \det(H_n)] =$$

$$O(d_n)/[1 + O(d_n)\{\det(\hat{H}_n) - \det(H_n)\}] = O(d_n)/[1 + o_p(1)] = O_p(d_n).$$

It follows from eq. (A.10) and calculations similar to those for eq. (A.9) applied to the cofactor formula for the inverse of a matrix that

(A.11) $|H_n^{-1}| = o((J-1)!(c_na_n)^{J-1}d_n)$, $|\hat{H}_n^{-1}| = o_p((J-1)!(c_na_n)^{J-1}d_n)$.

Eq. (A.8b) now follows from eqs. (A.7) and (A.11). Also, since

$$|\hat{H}_n^{-1} - H_n^{-1}| = |(\hat{H}_n^{-1})(\hat{H}_n - H_n)(H_n^{-1})| \le$$

$$\le J^3|\hat{H}_n^{-1}||\hat{H}_n - H_n||H_n^{-1}| = o_p(J^3[(J-1)!(c_na_n)^{J-1}d_n]^2b_n),$$

eq. (A.8b) also follows from eq. (A.7).

Lemma A5: If $\Sigma_{t=1}^{n}E[|x_t|^{2+\delta}]/n$ is bounded and $\sqrt{n}(\tilde{\beta}-\beta_0)$ is bounded in probability then $\max_{1\leq t\leq n}|x_t'(\tilde{\beta}-\beta_0)| = o_p(n^{-\epsilon})$ for $\epsilon = \delta/[2(2+\delta)]$.

Proof: By the Boole and Markov inequalities

$$\text{Prob}(\max_{1\leq t\leq n}|x_t|/(n^{1/(2+\delta)}) \geq \eta) \leq \Sigma_{t=1}^{n}\text{Prob}(|x_t| \geq \eta n^{1/(2+\delta)})$$

$$\leq \Sigma_{t=1}^{n}E[|x_t|^{2+\delta}]/[\eta^{2+\delta}n] \leq \eta^{-(2+\delta)}\Sigma_{t=1}^{n}E[|x_t|^{2+\delta}]/n,$$

so that $\max_{1\leq t\leq n}|x_t| = O_p(n^{1/(2+\delta)})$. The conclusion then follows from $|\tilde{\beta}-\beta_0| = O_p(n^{-1/2})$ and $\max_{1\leq t\leq n}|x_t'(\tilde{\beta}-\beta_0)| \leq C|\tilde{\beta}-\beta_0|\max_{1\leq t\leq n}|x_t|$.

Proof of Lemma 2.1: By the translation invariance of Lebesgue measure, $E[m(\varepsilon_t+\alpha)] = \int m(u)f(u-\alpha)du$. A consequence of Assumption 2.1 is that $f(u-\alpha)^{1/2}$ is differentiable in α, in the mean square sense (Hajek and Sidak (1967)). It follows from this fact, the moment condition in the hypotheses of this result, and some straightforward but tedious calculation that $\int m(u)f(u-\alpha)du$ is contin- uously differentiable with derivative $-\int m(u)s(u-\alpha)f(u-\alpha)du$ so that the conclusion follows from the translation invariance of Lebesgue measure.

Proof of Lemma 2.2: Note that the hypotheses of Lemma 2.1 are satisfied by the dominance condition for $m_j(\varepsilon_t+\alpha)$. Also, $m_j(\varepsilon_t+\alpha)$ is continuously differentiable at $\alpha = 0$ with probability one, so that by the dominance condition for its derivative the order of differentiation and integration can be interchanged (e.g. Corollary 5.9 of Bartle (1966)). The conclusion of Lemma 2.1 then yields $M = -E[\rho(z_t,\theta_0)s(\varepsilon_t)]$, so that $d_J'\rho(z_t,\theta_0)$ is the least squares projection of $s(\varepsilon_t)$ on $\rho(z_t,\theta_0)$. Theorem 4.3 of Freud (1971) states that the raw moments of $u_t = m_0(\varepsilon_t)$ characterize its distribution (if and) only if the nonnegative integer powers of u_t to form a basis for the Hilbert space of measurable functions of u_t that have finite squared expectation.

Therefore, since $E[(m_0^{-1}(u_t))^2] = E[s(\varepsilon_t)^2]$ is finite there exists a triangular array $\{\tilde{c}_{jJ}\}_{j\leq J}$ such that $E[\{s(\tilde{m}_0^{-1}(u_t)) - \Sigma_{j=0}^{J}\tilde{c}_{jJ}u_t^{\,j}\}^2] = E[\{s(\varepsilon_t) - \Sigma_{j=0}^{J}\tilde{c}_{jJ}m_0(\varepsilon)^j\}^2] \to 0$ as $J \to \infty$. Also, since $E[s(\varepsilon_t)] = 0$ it follows that the least squares projection of $s(\varepsilon_t)$ on $(1,\ m_0(\varepsilon_t),\ \ldots,\ [m_0(\varepsilon_t)]^J)$ equals the least squares projection of $s(\varepsilon_t)$ on $\rho(z_t,\theta_0)$. Therefore, the conclusion follows from

$$E[\{s(\varepsilon_t)-d_J'\rho(z_t,\theta_0)\}^2] = \min_{c_0,\ldots,c_J}E[\{s(\varepsilon_t) - \Sigma_{j=0}^{J}c_j[m_0(\varepsilon_t)]^j\}^2]$$

$$\leq E[\{s(\varepsilon_t) - \Sigma_{j=0}^{J}\tilde{c}_{jJ}[m_0(\varepsilon_t)]^j\}^2].$$

Proof of Theorem 2.1: A mean value expansion of $\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{s}_J(\tilde{\varepsilon}_t)$ around β_0 gives

(A.12) $$\sqrt{n}(\hat{\beta}_s-\beta_0)$$

$$= [I_k - \{\ (1/\hat{\mathcal{J}}_J)\hat{Q}_x^{\ -1}[\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'\hat{d}_J'\mathring{M}_t/n\ \}]\sqrt{n}(\mathring{\beta}-\beta_0)$$

$$+ (1/\hat{\mathcal{J}}_J)\hat{Q}_x^{\ -1}\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{s}_J(\varepsilon_t)/\sqrt{n}.$$

where $X_t = (1,x_t')'$, $\hat{Q} = \Sigma_{t=1}^{n}X_tX_t'/n$, $\hat{Q}_x = (\Sigma_{t=1}^{n}x_tx_t'/n) - \bar{x}(\bar{x}')$, $\mathring{\beta}$ denotes the mean value, and $\mathring{M}_t = (\tilde{m}_{1\varepsilon}(y_t-x_t'\mathring{\beta}),\ldots,\tilde{m}_{J\varepsilon}(y_t-x_t'\mathring{\beta}))'$. By the weak law of large numbers, $\hat{Q}_x$ converges in probability to $Q_x = \lim_{n\to\infty}\{\Sigma_{t=1}^{n}E(x_tx_t')/n - E(\bar{x})E(\bar{x})'\}$. By nonsingularity of Q, Q_x is nonsingular and $\text{plim}(\hat{Q}_x^{\ -1}) = Q_x^{\ -1}$. Also, for $\hat{A}_x = [-\bar{x},I_k]$ note that $\text{plim}(\hat{A}_x) = A_x = [-\lim_{n\to\infty}E(\bar{x}),I_k]$, so that by the Liapunov central limit theorem, $\Sigma_{t=1}^{n}(x_t-\bar{x})s(\varepsilon_t)/\sqrt{n} = [-\bar{x},I_k]\Sigma_{t=1}^{n}X_ts(\varepsilon_t)/\sqrt{n} \xrightarrow{d} A_xN(0,\mathcal{J}Q) \overset{d}{=} N(0,\mathcal{J}Q_x)$. Therefore by eq. (A.12) and $\sqrt{n}(\mathring{\beta}-\beta_0)$ bounded in probability is suffices to show that

(A.13a) $$\hat{\mathcal{J}}_J \xrightarrow{p} \mathcal{J},$$

(A.13b) $$\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'\hat{d}_J'\mathring{M}_t/n \xrightarrow{p} \mathcal{J}Q_x,$$

(A.13c) $$\Sigma_{t=1}^{n}(x_t-\bar{x})[\hat{s}_J(\varepsilon_t)-s(\varepsilon_t)]/\sqrt{n} \xrightarrow{p} 0.$$

Eq. (A.13) will be shown to hold by proving several more primitive results. Note that $\tau^j|m|^j/(j!) \le \exp[\tau|m|]$, so that

(A.14) $$|m|^j \le (j!)\exp[\tau|m|]/(\tau^j) \le C(j!)^2\exp[\tau|m|].$$

Eq. (A.14) and Assumption 2.3 then imply that the hypotheses of Lemma 2.2 are satisfied. Then for $s_J(\varepsilon_t) = d_J'\rho(z_t,\theta_0)$ and $Q_n = \Sigma_{t=1}^n E[X_tX_t']/n$, it follows from Assumption 2.2, independence of the observations and X_t and ε_t, and $E[X_t\{s_J(\varepsilon_t)-s(\varepsilon_t)\}] = 0$, that

(A.15) $$E[(\Sigma_{t=1}^n X_t\{s_J(\varepsilon_t)-s(\varepsilon_t)\}/\sqrt{n})(\Sigma_{t=1}^n X_t\{s_J(\varepsilon_t)-s(\varepsilon_t)\}/\sqrt{n})']$$
$$= E[\{s_J(\varepsilon_t)-s(\varepsilon_t)\}^2]Q_n \to 0.$$

The Markov inequality then implies that

(A.16) $$\Sigma_{t=1}^n(x_t-\bar{x})\{s_J(\varepsilon_t)-s(\varepsilon_t)\}/\sqrt{n} = \hat{A}_x\Sigma_{t=1}^n X_t\{s_J(\varepsilon_t)-s(\varepsilon_t)\}/\sqrt{n} = o_p(1).$$

Next, note that for any positive constants C and ϵ it follows from the specification of the growth rate for $J = J(n)$ that $\ln[J^{CJ^2}n^{-\epsilon}] = CJ^2\ln(J) - \epsilon\ln(n) = \ln(n)\{C[J^2\ln(J)/\ln(n)] - \epsilon\} \to -\infty$, so

(A.17) $$J^{CJ^2}n^{-\epsilon} \to 0.$$

Next, let γ be a three dimensional vector with $\gamma_0 = (0,\alpha_0,\sigma_0)'$ and $\tilde{\gamma}_{tn} = (x_t'(\tilde{\beta}-\beta_0),\tilde{\alpha},\tilde{\sigma})$. By the hypotheses of Theorem 2.3, Assumption 2.2, and Lemma A5 we have $\max_{t\le n}|\tilde{\gamma}_{tn}-\gamma_0| = o_p(n^{-\epsilon})$ for $\epsilon = \delta/[2(2+\delta)]$. For $\varepsilon = y - x'\beta_0$ and $I = 4J+1$ let $h_I(\varepsilon,\gamma)$ be the vector with components $\bar{m}_0((\varepsilon-\gamma_1-\gamma_2)/\gamma_3)^j$, $(j = 0,\ldots,2J)$, $\partial[\bar{m}_0((\varepsilon-\gamma_1-\gamma_2)/\gamma_3)^j]/\partial\gamma_1$, $(j=1,\ldots,J)$, and $\partial[\bar{m}_0((\varepsilon-\gamma_1-\gamma_2)/\gamma_3)]^j/\partial\gamma_3$, $(j=1,\ldots,J)$. By Assumption 2.3 and eq. (A.14), it is straightforward to show that the hypotheses of Lemma A3 are satisfied with $z_t = \varepsilon_t$, $B(I) = J^{CJ}$, and $\epsilon = \delta/[2(2+\delta)]$. For example by Assumption 2.3 and eq. (A.14) it follows by a mean value expansion that, for an open convex set Γ such that $(\gamma_1+\gamma_2,\gamma_3) \in N$ implies $\gamma \in \Gamma$ and γ_3 bounded away from zero, and for any $j \le 2J$,

$$\text{(A.18)}\quad |\bar{m}_0((\varepsilon-\gamma_1-\gamma_2)/\gamma_3)^j - m_0(\varepsilon)^j|$$

$$= j|\bar{m}_0((\varepsilon-\mathring{\gamma}_1-\mathring{\gamma}_2)/\mathring{\gamma}_3)^{j-1}\bar{m}_{0\varepsilon}((\varepsilon-\mathring{\gamma}_1-\mathring{\gamma}_2)/\mathring{\gamma}_3)/\mathring{\gamma}_3|\circ$$

$$|(\gamma_1+\gamma_2-\alpha_0) + (\varepsilon-\mathring{\gamma}_1-\mathring{\gamma}_2)(\gamma_3-\sigma_0)/\mathring{\gamma}_3|$$

$$\le CB_1(\varepsilon)^{2J}B_2(\varepsilon)(|\gamma_1| + |\gamma_2-\alpha_0| + C(|\varepsilon|+C)|\gamma_3-\sigma_0|)$$

$$\le CB_1(\varepsilon)^{2J}B_2(\varepsilon)[|\varepsilon|+C]|\gamma-\gamma_0|$$

$$\le C[B_1(\varepsilon)^{4J} + B_2(\varepsilon)^2(1+|\varepsilon|^2)]|\gamma-\gamma_0|,$$

$$\text{(A.19)}\quad E[B_1(\varepsilon_t)^{4J}] \le ((4J)!)^2E[\exp\{\tau B_1(\varepsilon_t)\}],$$

where without loss of generality $B_1(\varepsilon)$ and $B_2(\varepsilon)$ are taken to be bounded below by 1. Similarly, it is also the case that

$$\text{(A.20)}\quad \Sigma_{i=1}^{I}E[(h_I(\varepsilon_t,\gamma_0))_i^2] = O(J^{CJ}),$$

so that the hypotheses of Lemma A2 are satisfied with $\varepsilon_t = z_t$, $h_i(z_t) = (h_I(\varepsilon_t,\gamma_0))_i$, and $\omega = 2$. Then by $\in < 1/2$, the conclusions of Lemmas A2 and A3 and the triangle inequality yield

$$\text{(A.21a)}\quad |[\Sigma_{t=1}^{n}h_I(\varepsilon_t,\tilde{\gamma}_{tn})/n] - E[h_I(\varepsilon_t,\gamma_0)]| = O_p(n^{-\in}J^{CJ}),$$

$$\text{(A.21b)}\quad |E[h_I(\varepsilon_t,\gamma_0)]| = O(J^{CJ}), \quad |\Sigma_{t=1}^{n}h_I(\varepsilon_t,\tilde{\gamma}_{tn})/n| = O_p(J^{CJ}).$$

For $\hat{h}_I = \Sigma_{t=1}^{n}h_I(\varepsilon_t,\tilde{\gamma}_{tn})/n$ and $h_I = E[h_I(\varepsilon_t,\gamma_0)]$ let $\hat{h}_n = \hat{h}_I \otimes \hat{h}_I$ and $h_n = h_I \otimes h_I$. Note that $|\hat{h}_n-h_n| \le |(\hat{h}_I-h_I)\otimes\hat{h}_I| + |h_I\otimes(\hat{h}_I-h_I)| \le [|\hat{h}_I|+|h_I|]|\hat{h}_I-h_I| = [O_p(J^{CJ})+O(J^{CJ})]O_p(n^{-\in}J^{CJ}) = O_p(n^{-\in}J^{CJ})$, and similarly that $|h_n| = O(J^{CJ})$ and $|\hat{h}_n| = O_p(J^{CJ})$. Furthermore, note that the elements of $\tilde{\Sigma}$ and $\tilde{M}$ consist of components of $\hat{h}_n$ and differences of components of $\hat{h}_n$ so that

$$\text{(A.22a)}\quad |\tilde{\Sigma} - \Sigma| = O_p(n^{-\in}J^{CJ}), \quad |\tilde{M} - M| = O_p(n^{-\in}J^{CJ})$$

$$\text{(A.22b)}\quad |\Sigma| = O(J^{CJ}), \quad |M| = O(J^{CJ}), \quad |\tilde{\Sigma}| = O_p(J^{CJ}), \quad |\tilde{M}| = O_p(J^{CJ}).$$

Note that by Assumption 2.1 the density $f(\varepsilon)$ is continuous so that there is an interval on which it is bounded away from zero. Then by Lemma A1, eq. (A.1), and $\det(B) = \det(B_{11})/\det(B^{22})$, it follows that for $\bar{p}(\varepsilon) = (1,\ldots,\bar{m}_0(\varepsilon)^J)'$,

$$\text{(A.23)} \qquad \det(\Sigma) = \det(E[\bar{p}(\varepsilon_t)\bar{p}(\varepsilon_t)']) \geq J^{-CJ^2}.$$

By eqs. (A.17), (A.22), and (A.23) the hypotheses of Lemma A4 are satisfied with $\hat{H}_n = \tilde{\Sigma}$, $\hat{h}_{n2} = \tilde{M}$, $H_n = \Sigma$, $h_{2n} = M$, $a_n = J^{CJ}$, $b_n = n^{-\epsilon}J^{CJ}$, $c_n = J$, and $d_n = J^{CJ^2}$, so that

$$\text{(A.24a)} \qquad |\hat{d}_J - d_J| = o_p((J^{CJ^2}J!)^2 J^3 (JJ^{CJ})^{J-1} J^{CJ} n^{-\epsilon} J^{CJ}) = o_p(n^{-\epsilon}J^{CJ^2}),$$

$$\text{(A.24b)} \qquad |d_J| = O(J^{CJ^2}), \quad |\hat{d}_J| = O_p(J^{CJ^2}).$$

It follows from eq. (A.15) (and $(Q_n)_{11} = 1$) that $-d_J'M = E[s_J(\varepsilon_t)^2] \to \mathcal{I}$, so that eq. (A.13a) follows from eqs. (A.17), (A.22), and (A.24).

Next, let γ be as defined above and let $\tilde{\gamma}_{tn} = (x_t'(\mathring{\beta}-\beta_0),\tilde{\alpha},\tilde{\sigma})$, where $\mathring{\beta}$ is the mean value from (A.13b). For $\varepsilon = y - x'\beta_0$ let $M(\varepsilon,\gamma) = d(\bar{m}_0((\varepsilon-\gamma_1-\gamma_2)/\gamma_3)^1,\ldots,\bar{m}_0((\varepsilon-\gamma_1-\gamma_2)/\gamma_3)^J)'/d\varepsilon$, so that $\mathring{M}_t = M(\varepsilon_t,\tilde{\gamma}_{tn})$ and $M = E[M(\varepsilon_t,\gamma_0)]$. For $z = (\varepsilon,x')'$ and $I = J[(k+1)^2]$ let $h_I(z,\gamma) = [M(\varepsilon,\gamma)-M]\otimes[\text{vec}(X_tX_t')]$. As in the discussion of eqs. (A.18) and (A.19) it is straightforward to check that the hypotheses of Lemma A3 are satisfied for $\epsilon = \delta/(2(2+\delta))$ and $B(I) = J^{CJ}$. By Also, it is the case that by independence of x_t and ε_t,

$$\text{(A.25)} \qquad \Sigma_{i=1}^{I}\sup_n\{\Sigma_{t=1}^{n}E[|(h_I(z_t,\gamma_0))_i|^{1+\delta/2}]/n\}$$
$$= O(J^{CJ})(1 + \sup_n\{\Sigma_{t=1}^{n}E[|x_t|^{2+\delta}]/n\}) = O(J^{CJ}),$$

so that the hypotheses of Lemma A2 are satisfied for $h_{tI} = h(z_t,\gamma_0)$, $\omega = \delta/2$, and $B(I) = O(J^{CJ})$. Let $\hat{h}_n = \Sigma_{t=1}^{n}h_I(z_t,\tilde{\gamma}_{tn})/n$, and note that that $E[h_I(z_t,\gamma_0)] = 0$. Then by the conclusion of Lemma A2, with $a_n = n^{\delta/(2(2+\delta))}$, and the conclusion of Lemma A3 it follows from the triangle inequality that for $\epsilon = \delta/(2(2+\delta))$

(A.26) $\quad |\hat{h}_n| = o_p(n^{-\epsilon}J^{CJ}).$

Note that $\Sigma_{t=1}^{n}[(\mathring{M}_t-M)\otimes(x_t-\bar{x})x_t']/n$ is a linear combination of the elements of $\hat{h}_n$, with linear combination coefficients consisting of elements of $\hat{A}_x = O_p(1)$. Also note that $\text{plim}[\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'/n] = Q_x$ and that for any random variable $\hat{C}$ it is the case that $\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{C} = 0.$. Eq. (A.13b) then follows from eqs. (A.26), (A.24), (A.22), (A.17), and (A.13a) by

(A.27) $$|\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'\hat{d}_J'\mathring{M}_t/n - \mathcal{J}Q_x|$$

$$\leq |(\hat{d}_J\otimes I_k)'\Sigma_{t=1}^{n}[(\mathring{M}_t-M)\otimes(x_t-\bar{x})x_t']/n| + |(\hat{d}_J'M)\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'/n - \mathcal{J}Q_x|$$

$$\leq JkO_p(J^{CJ^2})o_p(n^{-\epsilon}J^{CJ})$$

$$+ |\hat{d}_J'M-\hat{d}_J'\tilde{M}||\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'/n| + |\hat{\mathcal{J}}_J\Sigma_{t=1}^{n}(x_t-\bar{x})x_t'/n - \mathcal{J}Q_x|$$

$$\leq o_p(n^{-\epsilon}J^{CJ^2}) + O_p(J^{CJ^2})o_p(n^{-\epsilon}J^{CJ}) + o_p(1) = o_p(1).$$

For the remainder of the proof let $\gamma = (\alpha,\sigma)'$, $\gamma_0 = (\alpha_0,\sigma_0)'$, and $\tilde{\gamma} = (\tilde{\alpha},\tilde{\sigma})'$. Let $m(\varepsilon,\gamma) = (\bar{m}_0((\varepsilon-\alpha)/\sigma)^1,\ldots,\bar{m}_0((\varepsilon-\alpha)/\sigma)^J)'$ and $M(\varepsilon,\gamma) = \partial m(\varepsilon,\gamma)/\partial\gamma$. A mean value expansion around $(\alpha_0,\sigma_0)'$ then gives

(A.28) $$\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{s}_J(\varepsilon_t)/\sqrt{n} = \Sigma_{t=1}^{n}(x_t-\bar{x})[\hat{d}_J'm(\varepsilon_t,\tilde{\gamma})]/\sqrt{n}$$

$$= \Sigma_{t=1}^{n}(x_t-\bar{x})\hat{d}_J'm(\varepsilon_t,\gamma_0)/\sqrt{n} + \{\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{d}_J'M(\varepsilon_t,\mathring{\gamma})/n\}\sqrt{n}(\hat{\gamma}-\gamma_0)$$

$$= \Sigma_{t=1}^{n}(x_t-\bar{x})\hat{d}_J'\rho(z_t,\theta_0)/\sqrt{n}$$

$$+ \{\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{d}_J'\{M(\varepsilon_t,\mathring{\gamma})-E[M(\varepsilon_t,\gamma_0)]\}/n\}O_p(1).$$

where the mean value $\mathring{\gamma}$ is such that $\sqrt{n}(\mathring{\gamma}-\gamma_0)$ is bounded in probability. For $I = 2J(k+1)$ let $h_I(z_t,\gamma) = \text{vec}[M(\varepsilon,\gamma)-E[M(\varepsilon_t,\gamma_0)] \otimes X$. Then it follows in a straightforward way that the hypotheses of Lemmas A2 and A3 are satisfied with $\omega = 1$, $\tilde{\gamma}_{tn} = \mathring{\gamma}$, $\epsilon = 1/2$, and $B(I) = CJ^{CJ}$. Noting that $\Sigma_{t=1}^{n}[M(\varepsilon_t,\mathring{\gamma})-E[M(\varepsilon_t,\gamma_0)]\otimes(x_t-\bar{x})/n$ consists entirely of components that are linear combinations of $\hat{A}_x$ with $\Sigma_{t=1}^{n}h_I(z_t,\mathring{\gamma})/n$, it follows from the conclusions of Lemmas A2 and A3 and eq. (A.24b) that

$$(A.29)\qquad |\Sigma_{t=1}^{n}(x_t-\bar{x})\hat{d}_J'[M(\varepsilon_t,\mathring{\gamma})-E[M(\varepsilon_t,\gamma_0)]/n|$$

$$\le J|\hat{d}_J|O_p(1)|\Sigma_{t=1}^{n}h_I(z_t,\mathring{\gamma})/n| = o_p(J^{CJ^2})O_p(n^{-1/2}J^{CJ}) = o_p(1).$$

Also, by the fact that $|E[(\Sigma_{t=1}^{n}\rho(z_t,\theta_0)\otimes X_t/\sqrt{n})(\Sigma_{t=1}^{n}\rho(z_t,\theta_0)\otimes X_t/\sqrt{n})']| = |\Sigma\otimes Q_n| \le |\Sigma||Q_n| = O(J^{CJ})O(1)$, it follows from the Markov inequality that $|\Sigma_{t=1}^{n}\rho(z_t,\theta_0)\otimes X_t/\sqrt{n}| = O_p(J^{CJ})$. Similarly to eq. (A.29) we have

$$(A.30)\qquad |\Sigma_{t=1}^{n}(x_t-\bar{x})\{\hat{d}_J'\rho(z_t,\theta_0) - s_J(\varepsilon_t)\}/\sqrt{n}|$$

$$\le Jk|(\hat{d}_J-d_J)\otimes I_k||\Sigma_{t=1}^{n}\rho(z_t,\theta_0)\otimes(x_t-\bar{x})/\sqrt{n}|$$

$$= o_p(n^{-\epsilon}J^{CJ^2})O_p(J^{CJ}) = o_p(n^{-\epsilon}J^{CJ^2}) = o_p(1).$$

where eq. (A.24a) has been used. Eq. (A.13c) now follows from the triangle inequality and eqs. (A.30), (A.29), (A.28), and (A.16).

Proof of Theorem 3.1: Let $\tilde{G}(b) = \Sigma_{t=1}^{n}\partial\tilde{\rho}(z_t,b)/\partial b\otimes a(X_t)/n$, and note that $\Sigma_{t=1}^{n}\hat{S}(z_t,\tilde{b})\hat{S}(z_t,\tilde{b})'/n = -\hat{D}'\tilde{G}(\tilde{b})$ and $\partial[\Sigma_{t=1}^{n}\hat{S}(z_t,b)/n]/\partial b = \hat{D}'\tilde{G}(b)$. A mean value expansion of $\Sigma_{t=1}^{n}\hat{S}(z_t,\tilde{b})/n$ around b_0 then gives, with probability approaching one,

$$(A.31)\qquad \sqrt{n}(\hat{b}_S-b_0) = \{I_k - [\hat{D}'\tilde{G}(\tilde{b})]^{-1}\hat{D}'\tilde{G}(\mathring{b})\}\sqrt{n}(\tilde{b}-b_0)$$

$$- [\hat{D}'\tilde{G}(\tilde{b})]^{-1}\Sigma_{t=1}^{n}\hat{S}(z_t,b_0)/\sqrt{n},$$

where $\mathring{b}$ denotes the mean value. By the Lindberg-Levy central limit theorem $\Sigma_{t=1}^{n}S(z_t,b_0)/\sqrt{n} \xrightarrow{d} N(0,\Omega)$. Also, $\sqrt{n}(\mathring{b}-b_0)$ is bounded in probability. Then by inspection of eq. (A.31) it is apparent that it suffices to show that for any $\bar{b}$ such that $\sqrt{n}(\bar{b}-b_0)$ is bounded in probability, (A.32a) $\hat{D}'\tilde{G}(\bar{b}) \xrightarrow{p} \Omega$,

(A.32b) $\Sigma_{t=1}^{n}[\hat{S}(z_t,b_0)-S(z_t,b_0)]/\sqrt{n} \xrightarrow{p} 0$.

Let $G = E[\partial g(z_t,b_0)/\partial b]$. Eq. (A.14) and Assumption 2.3 imply that $E[m_j(\varepsilon_t+\alpha)^4|X_t] \le E[B_1(\varepsilon_t)^{4(2j+1)}|X_t]$ with probability one (w.p.1), so that the hypotheses of Lemma 2.1 are satisfied for the conditional distribution of ε_t w.p.1. (Henceforth the w.p.1 qualifier will be dropped.) Also, $m_j(\varepsilon_t+\alpha)$ is continuously differentiable in a neighborhood N_α of zero and by Assumption 2.3 and eq. (A.14) $E[\sup_{N_\alpha}|m_j(\varepsilon_t+\alpha)||X_t]$ is finite. Then by the conclusion of Lemma 2.1 and interchange of order of differentiation and integration it follows that $E[m_{j\varepsilon}(\varepsilon_t)|X_t] = -E[m_j(\varepsilon_t)s(\varepsilon_t|X_t)|X_t]$, so that $G = -E[g(z_t,b_0)S(z_t,b_0)']$. Let $V = E[g(z_t,b_0)g(z_t,b_0)']$ and $D = -V^{-1}G$, and note that the columns of D are the coefficients of the linear projection of each component of $S(z_t,b_0)$ on $g(z_t,b_0)$. For $\gamma = (\alpha,\sigma)'$ and $\gamma_0 = (0,\sigma_0)'$, let $\rho(\varepsilon,\gamma) = (\bar{m}_0((\varepsilon-\alpha)/\sigma),\ldots,[\bar{m}_0((\varepsilon-\alpha)/\sigma)]^{2J+1})$. Also, let $\rho_2(\varepsilon) = (1,[\bar{m}_0(\varepsilon/\sigma_0)]^2,\ldots,[\bar{m}_0(\varepsilon/\sigma_0)]^{2J})$, $\rho^*(\varepsilon) = (\rho(\varepsilon,\gamma_0)',\rho_2(\varepsilon)')'$, and $g^*(z_t) = \rho^*(\varepsilon_t)\otimes a(X_t)$. Note that by conditional symmetry $\rho_2(\varepsilon_t)\otimes a(X_t)$ is orthogonal to both $S(z_t,b_0)$ and $g(z_t,b)$, so that the linear projection of $S(z_t,b_0)$ on $g^*(z_t)$ equals the linear projection of $S(z_t,b_0)$ on $g(z_t,b_0)$. Let $\bar{S}(z_t,b_0) = D'g(z_t,b_0)$. Then to show that

(A.33) $\lim_{J,K\to\infty}E[\{S(z_t,b_0)-\bar{S}(z_t,b_0)\}'\{S(z_t,b_0)-\bar{S}(z_t,b_0)\}] = 0$,

it suffices to show that there exists a triangular array of 2JK x k matrices of constants $\tilde{C}_{JK}$ such that $\tilde{C}_{JK}'g^*(z_t)$ converges in mean square to $S(z_t,b_0)$. After the change of variables $u_t = (m_0(\varepsilon_t), a_0^1(w_{2t}^1), \ldots, a_0^{\mathcal{L}}(w_{2t}^{\mathcal{L}}))'$ the terms in $(\rho(\varepsilon_t,\gamma_0)',\rho_2(\varepsilon_t)')'\otimes a_2(w_{2t})$ become multivariate polynomial terms, with lowest order equal to $\min\{2J+1,\nu_*\}$. It follows from Assumptions 2.3 and 3.2 and Theorem 3 of

Gallant (1980) that such functions form a basis for the Hilbert space of square integrable functions of $(\varepsilon_t, w_{2t}')'$. Thus, since $E[|S(z_t,b_0)|^2|w_{1t}=w_1]$ is finite for each of the points of support w_1 of w_{1t}, for each such w_1 there exists $\tilde{C}_{JK}(w_1)$ such that $\tilde{C}_{JK}(w_1)'[\rho^*(\varepsilon_t)\otimes a_2(w_{2t})]$ converges in conditional (on $w_{1t} = w_1$) mean square to $S(z_t,b_0)$. Also, w.l.g. $a_1(w_{1t})$ can be take to be a vector of dummy variables, each of which is equal to one when w_{1t} is equal to one of its support points and zero otherwise. The existence of the specified $\tilde{C}_{JK}$ then follows from the conditional mean square error convergence for each w_{1t} and stacking $\tilde{C}_{JK}(w_1)'[\rho^*(\varepsilon_t)\otimes a_2(w_{2t})]$ appropriately.

Next, it follows from eq. (A.33), $J(n) \to \infty$, $K(n) \to \infty$, independence of the observations, and $E[S(z_t,b_0)] = E[\bar{S}(z_t,b_0)] = 0$ that $E[\{\Sigma_{t=1}^n[S(z_t,b_0)-\bar{S}(z_t,b_0)]/\sqrt{n}\}'\{\Sigma_{t=1}^n[S(z_t,b_0)-\bar{S}(z_t,b_0)]/\sqrt{n}\}] = E[\{S(z_t,b_0)-\bar{S}(z_t,b_0)\}'\{S(z_t,b_0)-\bar{S}(z_t,b_0)\}] \to 0$, so that

$$\text{(A.34)} \qquad \Sigma_{t=1}^n[S(z_t,b_0)-\bar{S}(z_t,b_0)]/\sqrt{n} = o_p(1).$$

Next, it follows by eq. (A.33) that

$$\text{(A.35)} \qquad -D'G = D'VD = E[\bar{S}(z_t,b_0)\bar{S}(z_t,b_0)'] \to E[S(z_t,b_0)S(z_t,b_0)'] = \Omega.$$

Let $\varepsilon = y-X'b_0$, $z = (\varepsilon,X')'$, $\tilde{\gamma}_{tn} = (X_t'(\tilde{b}-b_0),\tilde{\sigma})'$, and $\gamma_0 = (0,\sigma_0)$. Also, for $I = JKk + J^2K^2$, let $h_I(z,\gamma) = (\text{vec}[\rho(\varepsilon,\gamma)\rho(\varepsilon,\gamma)'\otimes a(X)a(X)']', \text{vec}[\partial\rho(\varepsilon,\gamma)/\partial\varepsilon\otimes a(X)X']')'$ and $h_I(z_t) = h_I(z_t,\gamma_0)$. It is straightforward to use Assumptions 2.3, 3.2, and 3.3 to verify that the hypotheses of Lemmas A2 and A3 are satisfied with $\omega = \delta/2$, $\epsilon = \delta/(2(2+\delta))$ and, for $r = J\nu^*$, $B(I) = r^{Cr}$. Then by the conclusion of Lemmas A2 and A3, with $a_n = n^{\delta/[2(2+\delta)]}$, it follows that for $\epsilon = \delta/[2(2+\delta)]$,

$$\text{(A.36)} \qquad |[\tilde{V},\tilde{G}(\tilde{b})] - [V,G]| = |(\Sigma_{t=1}^n h(z_t,\tilde{\gamma}_{tn})/n) - E[h(z_t,\gamma_0)]|$$

$$= O_p(n^{-\epsilon}r^{Cr}), \quad |[V,G]| = O(r^{Cr}).$$

Similarly, for any $\bar{b}$ such that $\sqrt{n}(\bar{b}-b_0) = O_p(1)$, by choosing $\tilde{\gamma}_{tn} = (X_t'(\bar{b}-b_0),\tilde{\sigma})'$ one obtains

$$\text{(A.37)} \qquad |\tilde{G}(\bar{b})-G| \le |(\Sigma_{t=1}^{n}h(z_t,\tilde{\gamma}_{tn})/n) - E[h(z_t,\gamma_0)]| = O_p(n^{-\epsilon}r^{Cr}).$$

By Assumption 3.1 there is a bounded interval N and a positive constant c such that

$$\text{(A.38)} \quad \Sigma(X_t) = E[\rho(\varepsilon_t,\gamma_0)\rho(\varepsilon_t,\gamma_0)'|X_t] \ge c\textstyle\int_N\rho(\varepsilon,\gamma_0)\rho(\varepsilon,\gamma_0)'d\varepsilon \equiv \underline{\Sigma},$$

and by eq. (A.2) of Lemma A1, applied to the uniform distribution on N with $\mathcal{L} = 1$, $\det(\underline{\Sigma}) \ge (2J+1)^{-C(2J+1)^3} \ge J^{-CJ^3}$ for J large enough. Also, note that the components of $a_2(w_{2t})$ make up a subset of the components of $(1,\ldots,[a_0^1(w_{2t}^1)]^{\nu^*})'\otimes\circ\circ\circ\otimes(1,\ldots,[a_0^{\mathcal{L}}(w_{2t}^{\mathcal{L}})]^{\nu^*})'$, so that by eq. (A.2) of Lemma A1 it follows that for each support point w_1 of w_{1t}, $\det\{E[a_2(w_{2t})a_2(w_{2t})'|w_{1t}=w_1]\} \ge (\nu^*)^{-C(\nu^*)^{\mathcal{L}+2}}$. Also, let W_1 denote the number of points of support for w_{1t} and let $\{w_1(1),\ldots,w_1(W_1)\}$ be the support. Assuming $a_1(w_{1t})$ is a vector of dummy variables as specified above, $E[a(X_t)a(X_t)']$ is a block diagonal matrix with i^{th} diagonal block given by $E[a_2(w_{2t})a_2(w_{2t})'|w_{1t}=w_1(i)] \circ \text{Prob}(w_{1t}=w_1(i))$. Since the determinant of a block diagonal matrix is the product of the determinants of the diagonal blocks, $\det(E[a(X_t)a(X_t)']) \ge (\nu^*)^{-C(\nu^*)^{\mathcal{L}+2}}$. Furthermore, by eq. (A.38)

$$\text{(A.39)} \qquad V = E[\Sigma(X_t) \otimes a(X_t)a(X_t)'] \ge \underline{\Sigma} \otimes E[a(X_t)a(X_t)'].$$

It follows from $a(X_t)$ having no more than $(C\nu^*)^{\mathcal{L}}$ components and $\det(\underline{\Sigma}) \le 1$ for large enough J that

$$\text{(A.40)} \qquad \det(V) \ge \det(\underline{\Sigma})^{(C\nu^*)^{\mathcal{L}}}\det(E[a(X_t)a(X_t)'])^J$$
$$\ge J^{-CJ^3(\nu^*)^{\mathcal{L}}}(\nu^*)^{-C(\nu^*)^{\mathcal{L}+2}J} \ge r^{-Cr^3(\nu^*)^{\mathcal{L}-1}}$$

Next, note that for any positive constants C and ϵ it follows from the specification of the growth rate for $J(n)$ and $K(n)$ that

$\ln[n^{-\epsilon}r^{-Cr^3(\nu^*)^{\mathcal{L}-1}}] = \ln(n)[Cr^3(\nu^*)^{\mathcal{L}-1}\ln(r)/\ln(n) - \epsilon] \rightarrow -\infty$, so

$$\text{(A.41)} \qquad n^{-\epsilon}r^{-Cr^3(\nu^*)^{\mathcal{L}-1}} \rightarrow 0.$$

Then it follows from eqs. (A.36), (A.40), (A.41) and Lemma A4 with $a_n = r^{Cr}$, $b_n = n^{-\epsilon}r^{Cr}$, $c_n = r$, and $d_n = r^{Cr^3(\nu^*)^{\mathcal{L}-1}}$, that

$$\text{(A.42)} \qquad |\hat{D}-D| = o_p(n^{-\epsilon}r^{Cr^3(\nu^*)^{\mathcal{L}-1}}),$$

$$|D| = o_p(r^{Cr^3(\nu^*)^{\mathcal{L}-1}}), \qquad |\hat{D}| = o_p(r^{Cr^3(\nu^*)^{\mathcal{L}-1}}).$$

Eq. (A.32a) now follows from eqs. (A.35), (A.38), (A.41), and (A.42).

Next, note that $\hat{S}(z_t,b_0) = \hat{D}'[\rho(\varepsilon_t,0,\tilde{\sigma})\otimes a(X_t)]$, so that a mean value expansion of $\Sigma_{t=1}^n\rho(\varepsilon_t,0,\tilde{\sigma})\otimes a(X_t)$ around γ_0 yields

$$\text{(A.43)} \qquad \Sigma_{t=1}^n[\hat{S}(z_t,b_0) - \bar{S}(z_t,b_0)]/\sqrt{n}$$

$$= (\hat{D}-D)'\Sigma_{t=1}^n g(z_t,b_0)/\sqrt{n} + \hat{D}'\{\Sigma_{t=1}^n[\partial\rho(\varepsilon_t,0,\mathring{\sigma})/\partial\sigma]\otimes a(X_t)/n\}\sqrt{n}(\tilde{\sigma}-\sigma_0).$$

It is also the case that $E[|g(z_t,b_0)|^2] = O(r^{Cr})$, so that by Lemma A2, $E[g(z_t,b_0)] = 0$, and eq. (A.41),

$$\text{(A.44)} \qquad |(\hat{D}-D)'\Sigma_{t=1}^n g(z_t,b_0)/\sqrt{n}| \leq JK|\hat{D}-D|\sqrt{n}|\Sigma_{t=1}^n g(z_t,b_0)/n|$$

$$= JKo_p(n^{-\epsilon}r^{Cr^3(\nu^*)^{\mathcal{L}-1}})O_p(\sqrt{n})O_p(r^{Cr}/\sqrt{n}) = o_p(1).$$

Let $\gamma = \sigma$, $\gamma_0 = \sigma_0$, $\tilde{\gamma}_{tn} = \mathring{\sigma}$, $z = (\varepsilon,X')'$, $I = JK$, and $h_I(z,\gamma) = [\partial\rho(\varepsilon,0,\sigma)/\partial\sigma]\otimes a(X)$. It is straightforward to use Assumptions 2.3, 3.2, and 3.3 to verify that the hypotheses of Lemmas A2 and A3 are satisfied for $h_{tI} = h_I(z_t,\gamma_0)$, $\omega = 1$, and $\epsilon = 1/2$. Furthermore,since $\bar{m}_j(\varepsilon/\sigma)$ is an odd function of ε for each σ, $\partial\bar{m}_j(\varepsilon/\sigma)/\partial\sigma$ is an odd function of ε for each σ, implying $E[h_I(z_t,\gamma_0)] = 0$. It then follows from the conclusions of Lemmas A2 and Lemma A3, and eqs. (A.41) and (A.42) that there is an $\epsilon > 0$ such that

$$(A.45)\qquad |\hat{D}'\{\Sigma_{t=1}^{n}[\partial\rho(\varepsilon_t,0,\mathring{\sigma})/\partial\sigma]\otimes a(X_t)/n\}\sqrt{n}(\tilde{\sigma}-\sigma_0)|$$

$$\leq JK|\hat{D}||\Sigma_{t=1}^{n}h_I(z_t,\tilde{\gamma})/n|\sqrt{n}|\tilde{\sigma}-\sigma_0|$$

$$\leq o_p(r^{Cr^3(\nu^*)^{\mathcal{L}-1}})O_p(n^{-\epsilon}r^{Cr})O_p(1) = o_p(1).$$

Eq. (A.32b) now follows from eqs. (A.43), (A.44), (A.45), and (A.34).

References

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

Bartle, R. G., 1966, The Elements of Integration, John Wiley and Sons, New York.

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

Billingsley, P., 1979, Probability and Measure, John Wiley and Sons, New York.

Carroll, R. J., 1982, Adapting for Heteroskedasticity in Linear Models, Annals of Statistics 10, 1224-1233.

Chamberlain, G., 1984, Comment on Adaptive Estimation of Nonlinear Regression Models, Econometric Reviews.

Chamberlain, G., 1987, Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, Journal of Econometrics, forthcoming.

Freud, G., 1971, Orthogonal Polynomials, Pergammon Press, Oxford.

Gregory, R. T., and D. L. Karney, 1969, A Collection of Matrices for Testing Computational Algorithms, John Wiley and Sons, New York.

Hajek, J., and Z. Sidak, 1967, Theory of Rank Tests, Academic Press, New York.

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

Hansen, L. P., 1985, Using Martingale Difference Approximations to Obtain Covariance Matrix Bounds for Generalized Method of Moments Estimators, University of Chicago/NORC, Working Paper 85-16.

Hsieh, D., and C. Manski, 1987, Monte-Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression, Annals of Statistics, forthcoming.

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

Manski, C., 1984, Adaptive Estimation of Nonlinear Regression Models, Econometric Reviews 3, 145-194.

Newey, W. K., 1984, Nearly Efficient Moment Restriction Estimation of Regression Models with Nonnormal Disturbances, Princeton University, Econometric Research Program Memo. No. 315.

Newey, W. K., 1985, Generalized Method of Moments Specification Testing, Journal of Econometrics 29, 229-256.

Newey, W. K., 1986, Efficient Estimation of Models with Conditional Moment Restrictions, Princeton University, manuscript.

Newey, W. K., and J. L. Powell, 1987a, Efficient Estimation of Type I Censored Regression Models Under Conditional Quantile Restrictions, manuscript, University of Wisconsin.

Newey, W. K., and J. L. Powell, 1987b, Efficient Estimation of Tobit Models Under Symmetry, Princeton University, manuscript.

Robinson, P., 1987, Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form, Econometrica, forthcoming.

White, H., 1982, Instrumental Variable Regression with Independent Observations, Econometrica 50, 483-499.

TABLE 1: ROOT MEAN SQUARE ERROR, WITH ONE REGRESSOR AND SAMPLE SIZE 50

A. NORMAL DISTRIBUTION

J	OLS	AML	LAD	2	3	4	5	6	7
TRANSFORMED	.28	.28	.35	.30	.29	.30	.31	.31	.31
WEIGHTED	——	——	——	.36	.31	.32	.31	.32	.31

B. VARIANCE-CONTAMINATED NORMAL DISTRIBUTION

J	OLS	AML	LAD	2	3	4	5	6	7
TRANSFORMED	.29	.19	.13	.13	.12	.12	.13	.14	.16
WEIGHTED	——	——	——	.11	.12	.12	.13	.15	.19

C. BIMODAL SYMMETRIC MIXTURE OF NORMAL DISTRIBUTIONS

J	OLS	AML	LAD	2	3	4	5	6	7
TRANSFORMED	.29	.18	.84	.38	.10	.11	.11	.13	.16
WEIGHTED	——	——	——	.65	.10	.11	.11	.13	.17

D. LOGNORMAL DISTRIBUTION

J	OLS	AML	LAD	2	3	4	5	6	7
TRANSFORMED	.28	.19	.16	.12	.09	.09	.11	.15	.19
WEIGHTED	——	——	——	.11	.09	.09	.15	.18	.22

TABLE 2: RATIO OF ROOT MEAN SQAURE OF ESTIMATED STANDARD ERRORS TO ACTUAL ROOT MEAN SQUARE ERROR, WITH ONE REGRESSOR AND SAMPLE SIZE 50.

A. NORMAL DISTRIBUTION

J	2	3	4	5	6	7
TRANSFORMED	.97	.88	.78	.72	.65	.57
WEIGHTED	.95	.88	.82	.73	.66	.57

B. VARIANCE-CONTAMINATED NORMAL DISTRIBUTION

J	2	3	4	5	6	7
TRANSFORMED	1.02	1.02	.95	.84	.72	.60
WEIGHTED	1.02	.94	.87	.77	.66	.45

C. BIMODAL SYMMETRIC MIXTURE OF NORMAL DISTRIBUTIONS

J	2	3	4	5	6	7
TRANSFORMED	.93	1.03	.95	.86	.73	.53
WEIGHTED	.96	1.04	.96	.83	.74	.45

D. LOGNORMAL DISTRIBUTION

J	2	3	4	5	6	7
TRANSFORMED	1.03	1.04	.97	.75	.49	.34
WEIGHTED	1.05	1.04	.93	.52	.40	.26

TABLE 3: RATIO OF ROOT MEAN SQUARE ERRORS OF TRANSFORMED LGMM AND OLS ESTIMATORS, FOR VARIANCE CONTAMINATED NORMAL AND TWO REGRESSORS.

n		J = 2	3	4	5	6	7
50	β_1	.49	.42	.43	.45	.49	.53
	β_2	.50	.42	.44	.47	.50	.55
100	β_1	.48	.40	.41	.41	.42	.43
	β_2	.48	.39	.40	.41	.41	.42
200	β_1	.47	.38	.38	.38	.39	.39
	β_2	.48	.39	.39	.39	.39	.39

TABLE 4: RATIO OF ROOT MEAN SQUARE OF ESTIMATED STANDARD ERRORS TO ACTUAL ROOT MEAN SQUARE ERROR, FOR TRANSFORMED LGMM ESTIMATOR WITH TWO REGRESSORS.

n		J = 2	3	4	5	6	7
50	β_1	1.03	1.03	.97	.88	.78	.65
	β_2	1.00	1.05	.96	.86	.77	.64
100	β_1	1.03	1.04	1.02	.99	.94	.88
	β_2	1.02	1.05	1.02	.99	.95	.88
200	β_1	1.05	1.07	1.06	1.04	1.03	.99
	β_2	.99	1.02	1.01	.99	.98	.95