LOCALLY EFFICIENT, RESIDUAL-BASED ESTIMATION OF NONLINEAR SIMULTANEOUS EQUATIONS

Whitney K. Newey
Princeton University and Bellcore

Econometric Research Program
Research Memorandum No. 351

October 1989

Econometric Research Program
Princeton University
204 Fisher Hall
Princeton, NJ 08544-1021, USA

Locally Efficient, Residual-Based Estimation of Nonlinear Simultaneous Equations

Abstract

Nonlinear simultaneous equations models are important in both statistics and econometrics. They are useful in econometric applications, such as estimation of supply and demand systems, and have been the subject of an extensive literature. The special case where there is one dependent variable is a transformation model that has long been of interest in statistics, including as a special case the Box-Cox transformation.

It is well known that maximum likelihood methods for these model can be sensitive to distributional assumption. A number of distribution free methods have been proposed, including instrumental variables and transformations to symmetry or homoskedasticity. This paper considers estimation when disturbances are independent of regressors or conditionally symmetric around zero given the regressors.

Distribution-free estimators that are semiparametric efficient for particular parametric families are developed. Each of these estimators is residual based, being formed from a V-statistic in the independent case and an antithetic symmetrization in the conditionally symmetric case. An empirical and Monte Carlo transformation example is considered. In the Monte Carlo examples these estimators perform much better than recently suggested alternative transformations to symmetry and homoskedasticity.

Keywords: Simultaneous Equations, Transformation Models, Semiparametric Efficiency, V-statistics, Antithetic Variates.

1. Introduction

Nonlinear simultaneous equations models are important in both econometrics and statistics. They are useful in econometric applications, such as estimation of nonlinear demand and supply systems, and have been the subject of an extensive literature; e.g. see Amemiya (1985, Chapter 8). The special case where there is one dependent variable is a transformation model that has long been of interest in statistics; e.g. see Ruppert and Aldershof (1989) for a recent list of references.

The consistency of maximum likelihood estimators for such models can be sensitive to the distributional assumption for the disturbances; see Amemiya (1977) and Phillips (1982). One class of estimators that does not have this problem are nonlinear instrumental variables estimators, as considered by Sargan (1959), Kelejian (1971), Amemiya (1974, 1977), and Amemiya and Powell (1981). Consistency of these estimators depends only on the disturbances having conditional mean zero given exogenous variables, and not on the functional form of the disturbance distribution. For transformation models, an analogous class of estimators based on conditional median assumptions are developed by Carroll and Ruppert (1984) and Powell (1990). Also included among previously suggested distribution-free estimators are those of Hinkley (1975), MaCurdy (1982), Taylor (1985), and Ruppert and Aldershof (1989). Consistency of these estimators only depends on the disturbances being independent of the exogenous variables and/or conditional symmetry of the disturbances.

The efficiency of distribution-free estimators is of potential concern. For example, some estimators of parameters of transformation models can have large variances; see Section 6. This concern motivates the present paper. The purpose of this paper is development of methods that efficiently use information from independence of disturbances and exogenous variables and/or conditional symmetry of the disturbance given exogenous variables. The estimators considered here will be efficient among the class of estimators that use only such information, when the disturbance distribution is a member of particular parametric families.

The relevant efficiency standard for distribution-free estimators is the semiparametric efficiency bound, general theory for which is developed in Stein (1956), Koshevnik and Levit (1976), Pfanzagl (1982), Begun, Hall, Huang, and Wellner (1983), and Bickel, Klaassen, Ritov, and Wellner (1989). Under the conditional mean zero assumption, which is the assumption required for consistency of instrumental variables estimators, Chamberlain (1987) showed that an instrumental variables estimator with optimal instruments is efficient; nonparametric estimation of the optimal instruments has been considered in Newey (1987). Here, the semiparametric efficiency bounds for nonlinear simultaneous equations model with independent or conditionally symmetric disturbances are derived.

For each model, the form of the bound motivates the specification of a particular class of estimators. Each of the bounds is the inverse covariance matrix of the efficient score. In each case a distribution-free m-estimator can be obtained by replacing the piece of the efficient score that depends on the disturbance distribution by that corresponding to some parametric family of distributions, estimating the remaining nuisance functions and/or parameters, and choosing the parameters of interest to set the sample average to be close to zero.

In the independence case, the efficient score contains a conditional expectation. Because of the independence of disturbances and exogenous variables, this conditional expectation can be estimated by an average over the entire sample. An estimating equation can be obtained by averaging the result over the sample. The resulting estimating equation has a V-statistic form, so that it is natural to refer to the estimator as a V-estimator. Similar estimates of conditional expectations have been used for forecasting by Duan (1983) and Brown and Mariano (1984). Robinson (1989) has previously suggested their use in estimation in the context of optimal nonlinear instrumental variables.

In the conditionally symmetric case the efficient score is derived from the score for parameters of interest by subtracting the score evaluated at the negative of the residual, i.e. by antithetic symmetrization. When applied to the score for a known family of distributions antithetic symmetrization yields an m-estimator that is consistent under conditional symmetry and efficient if the true distribution is a member of the specified family. It is also possible to construct an estimator that uses both independence and symmetry, from a V-statistic involving antithetic variates. This estimator will be considered in less detail than the others.

The resulting estimators attain the semiparametric efficiency bound when the distribution of the disturbances is a member of the parametric family used in their construction. That is, when the disturbance has a particular form they are efficient in the class of estimators that are asymptotically normal and sufficiently well behaved under the corresponding assumption on the disturbances. For example, if the score from a T or Gaussian distribution is used to form the estimator, then it will be efficient under a T or Gaussian distribution, respectively.

It is also possible to construct estimators that are efficient for all possible distributions of the disturbances, i.e. that are globally efficient; see Newey (1989a) for the independence case. This construction is not carried out here. In comparison with globally efficient estimators, locally efficient estimators like those discussed here are relatively computationally simple, hopefully have better small sample properties by depending on fewer nuisance parameters, and allow the investigator more control over the way residuals affect the estimator, at the expense of giving up on asymptotic efficiency for many distributions.

Section 2 of the paper defines the nonlinear simultaneous equations model and discusses transformation models as special cases. Section 3 derives the the semiparametric efficiency bounds. Sections 4 and 5 consider V-statistic and antithetic variate estimators respectively, and a brief discussion of how the two might be combined is also given in Section 5. Section 6 gives an empirical and Monte Carlo example, and Section 7 offers some concluding remarks.

2. The Models

A central feature of each model to be considered is an equation of the form

$$\rho(z,\beta_0) = \varepsilon, \qquad (2.1)$$

where $\rho(z,\beta)$ is an $s \times 1$ vector of residuals that depend on a data vector z and a $q \times 1$ vector of parameters β, with true value β_0. It will be assumed throughout that $z = (y',x')'$, where y is an $s \times 1$ vector of endogenous variables and x a $k \times 1$ vector of exogenous variables. Here the term "exogenous" refers to the fact that each model will impose conditions on the conditional distribution of ε given x. Attention will be restricted to the case where the data observations z_i, $(i=1,\ldots,n)$, are i.i.d., although some of the results should extend easily to other cases.

It will be assumed throughout that equation (2.1) defines a one-to-one relationship between y and ε for any realizable x and β in some set of parameters B, so that y can be solved for via a reduced form

$$(2.2) \qquad y = \pi(\varepsilon, x, \beta).$$

In the econometrics literature such a model is known as a full information, nonlinear simultaneous equations model, where the "full information" term refers to the fact that there are as many scalar equations in (2.1) as there are elements of y; "limited information" models are those where there are fewer equations in (2.1) than elements of y. The estimators considered here only work for full information models, because of the need for a reduced form. Of course, it might be possible to complete a limited information equation system by adding equations, but the resulting estimators can be sensitive to the specification of the additional equations.

Two types of restrictions on the conditional distribution of ε are considered. The first is the independence restriction

(2.3) ε and x are statistically independent.

The second is the conditional symmetry restriction

(2.4) $\varepsilon|x$ is symmetrically distributed around zero.

These restrictions are complementary. Independence allows the shape of the distribution of the disturbances to be very general at the expense of ruling out heteroskedasticity in ε. Conditional symmetry allows for heteroskedasticity at the expense of imposing the symmetry restriction.

It should be noted that constant terms in $\rho(z,\beta)$ will be treated differently in these two cases. In the independence case it is convenient to impose no location restriction on the disturbances, so that the constant term is absorbed in ε. In the symmetry case, the location of ε is restricted to be zero, so that it is important to include constant terms in $\rho(z,\beta)$.

A special case of the nonlinear simultaneous equations model is a transformation model, where y consists of a single variable. A general form for such a moael is

$$h(y,\lambda_0) = f(x,\beta_0) + \varepsilon, \tag{2.5}$$

where $\beta = (\lambda,\beta_2')'$, λ is a transformation parameter (or parameters) and $h(y,\lambda)$, is a one-to-one function of y for each λ in some set of possible values. Included as special cases are the Box and Cox (1964) transformation of a linear model, where $h(y,\lambda) = y^{(\lambda)} = (y^{\lambda} - 1)/\lambda$ and $f(x,\beta) = x'\beta_2$, the transform both sides model of Carroll and Ruppert (1984), where $h(y,\lambda)$ is as before and $f(x,\beta) = f(x,\beta_2)^{(\lambda)}$ for some function $f(x,\beta_2)$, as well as other transformations considered by Burbidge, Magee, and Robb (1988). In this example, equations (2.3) and (2.4) correspond to transformations to independence or conditional symmetry, similar in spirit to transformations considered in Hinkley (1975), Taylor (1985), and Ruppert and Aldershof (1989).

3. The Semiparametric Efficiency Bounds

It is helpful to briefly review semiparametric efficiency bounds, as developed by Stein (1956), Koshevnik and Levit (1976), Pfanzagl (1982), Begun, Hall, Huang, and Wellner (1983), and Bickel, Klaassen, Ritov, and Wellner (1989) (BKRW henceforth). Define a *parametric submodel* to be one that satisfies the semiparametric assumptions and contains the truth. Any semiparametric estimator must have an asymptotic variance that is no smaller than the Cramer-Rao bound for every parametric submodel, giving Stein's

(1956) result:

The asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramer-Rao bounds for all parametric submodels, denoted V.

Regularity conditions are needed to make this statement precise. The parametric submodels be regular in that they are *smooth* in the mean-square sense (see the Appendix), have nonsingular information matrices, and satisfy other regularity conditions appropriate to the model. A precise definition of V is that it is the supremum of Cramer-Rao bounds for regular parametric submodels. The estimators must be regular in the following sense. For a parametric submodel with Euclidean parameter vector θ let $\beta(\theta)$ be the parameters of interest. A local data generating process (LDGP) is one where the for each sample size n the data is distributed according to θ_n, with $\sqrt{n}(\theta_n-\theta_0)$ bounded. An estimator $\hat{\beta}$ is said to be *regular* if for each regular parametric submodel and LDGP, $\sqrt{n}(\hat{\beta}-\beta(\theta_n))$ has a limiting distribution that does not depend on the sequence $\{\theta_n\}$ or the parametric submodel. That V is an asymptotic variance bound for regular estimators follows from semiparametric extensions of Hajek's (1970) representation theorem, e.g. Begun et. al. (1983). A vector version of Theorem 2 i) of Chamberlain (1986) is

If $\hat{\beta}$ *is regular then the limiting distribution of* $\sqrt{n}(\hat{\beta}-\beta_0)$ *is equal to the distribution of* $Y + U$, *where* $Y \sim N(0,V)$ *and* U *is independent of* Y.

An efficient semiparametric estimator is one that is asymptotically normal with covariance matrix V and is regular.

The projection form of the bound developed by Begun et. al. (1983) and BKRW will prove useful here. Let the data consist of i.i.d. observations $z_1,\ldots,z_n$. Consider a regular parametric submodel with parameters $\theta =$

$(\beta', \eta')'$ and likelihood function $\ell(z|\theta)$ for a single observation z_i. The $q \times 1$ vector of parameters of interest is β and the η parameters correspond to the nonparametric part of the model. Let $S_\theta = (S_\beta', S_\eta')'$ be the score for θ for a single observation, evaluated at the true parameter values, where typically $S_\theta = \partial \ln \ell(z|\theta_0)/\partial\theta$ (see the Appendix for a precise definition). The z argument may be suppressed for notational convenience, as here. Define the *tangent set* $\mathcal{T}$ to be the mean-square closure of $q \times 1$ linear combinations of scores S_η for the nonparametric component:

$$\mathcal{T} = \{t \in \mathbb{R}^q : E[\|t\|^2] < \infty,\ \exists\ B_j,\ S_{\eta j} \text{ with } \lim_{j\to\infty} E[\|t - B_j S_{\eta j}\|^2] = 0\},$$

where each B_j is a matrix of constants. Consider S_β as an element of, and $\mathcal{T}$ as a subset of, the Hilbert space of $q \times 1$ random vectors r with inner product $E[r_1' r_2]$. If $\mathcal{T}$ is linear then the residual from the projection of S_β on $\mathcal{T}$ exists, and is the unique vector S satisfying

(3.1) $\quad S_\beta - S \in \mathcal{T}, \quad E[S't] = 0$ for all $t \in \mathcal{T}$.

A version of Corollary 3.4.1 of BKRW (see Newey, 1990a, Theorem 3.2) is

If $\ell(z|\beta)$ is regular with score S_β, $\mathcal{T}$ is linear, and $E[SS']$ is nonsingular, then $V = (E[SS'])^{-1}$.

The vector S is referred to as the *efficient score.*

It will be useful for the estimation results discussed below to carry through this calculation for both the independence and symmetry cases. In both cases the parameters of interest are those of the residual $\rho(z,\beta)$. A parametric submodel corresponds to a parametric family of density functions $f(\varepsilon, x|\eta)$ (with respect to a carrier measure) for (ε_i, x_i) such that $f(\varepsilon, x|\eta_0)$ is the true density for some η_0 and $f(\varepsilon, x|\eta)$ satisfies the restrictions implied by independence and or symmetry. In the independence case $f(\varepsilon,x|\eta) = f_1(\varepsilon|\eta)f_2(x|\eta)$, under conditional symmetry $f(-\varepsilon,x|\eta) = f(\varepsilon,x|\eta)$, and under both $f_1(-\varepsilon|\eta) = f_1(\varepsilon|\eta)$. The likelihood and score vectors for a parametric submodel are

$$\ell(z|\theta) = |\det(\partial\rho(z,\beta)/\partial y)|\cdot f(\rho(z,\beta),x|\eta), \tag{3.2}$$

$$S_\beta = J_\beta(z,\beta_0) + \rho_\beta(z,\beta_0)'s(\varepsilon,x), \quad S_\eta = \partial \ln f(\varepsilon,x|\eta_0)/\partial\eta,$$

where $J(z,\beta) = \ln|\det(\partial\rho(z,\beta)/\partial y)|$, $s(\varepsilon,x) = f_0(\varepsilon,x)^{-1}\partial f_0(\varepsilon,x)/\partial\varepsilon$, and the β subscripts denote the partial derivatives.

Independence of ε and x implies a restriction on the form of the scores, and hence on the tangent set. By independence, $S_\eta = S_{\eta 1}(\varepsilon)+S_{\eta 2}(x)$, where $S_{\eta 1}$ is the score for the marginal density of ε and $S_{\eta 2}$ is that for the marginal density of x. This should be the only restriction on the scores, except for the usual mean zero property, so that we expect the tangent set to be

$$\mathcal{T} = \{t_1(\varepsilon)+t_2(x) : E[t_1(\varepsilon)] = E[t_2(x)] = 0\}. \tag{3.3}$$

To avoid notational clutter it is here and henceforth assumed that second moments exist whenever needed. For a $q \times 1$ random vector $R(z)$ let

$$\bar{R} = E[R|\varepsilon]-E[R] + E[R|x]-E[R]. \tag{3.4}$$

Note that $\bar{R}$ is an element of $\mathcal{T}$, and independence of ε and x implies $E[(R-\bar{R})'t] = 0$ for any element t of $\mathcal{T}$. It follows that $\bar{R} = \mathrm{Proj}(R|\mathcal{T})$, the projection of R on $\mathcal{T}$. Thus, since x is ancillary for β, implying $E[S_\beta|x] = 0$, the efficient score should be

$$S = S_\beta - \mathrm{Proj}(S_\beta|\mathcal{T}) = S_\beta - E[S_\beta|\varepsilon] - E[S_\beta|x] + 2E[S_\beta] = S_\beta - E[S_\beta|\varepsilon] \tag{3.5}$$

$$= J_\beta - E[J_\beta|\varepsilon] + \{\rho_\beta - E[\rho_\beta|\varepsilon]\}'s(\varepsilon).$$

The following regularity condition is useful in making this result rigorous. Let a subscript on an expectation denote the expectation as a function of the true parameter value.

Assumption 3.1: $f(z|\beta)$ *is smooth with* S_β *given in equation (3.2),* $\rho(z,\beta)$ *is continuously differentiable in* β *in a neighborhood* N *of* β_0, $\sup_N\|\rho_\beta(z,\beta)\| = M(z)$ *satisfies* $E_\beta[M(z)^2]$ *continuous, with probability one* $\rho(\circ,x,\beta_0)$ *is a one-to-one function and* $\partial\rho(z,\beta_0)/\partial y$ *is nonsingular.*

Theorem 3.1: Suppose that ε *and* x *are independent, Assumption 3.1 is satisfied, each parametric submodel correspond to a density* $f(\varepsilon,x|\eta) = f(\varepsilon|\eta)f(x|\eta)$ *such that* $f(\varepsilon|\eta)$ *and* $f(x|\eta)$ *are smooth, and* $E[SS']$ *is nonsingular for* S *from equation (3.5). Then* S *is the efficient score.*

This result verifies a conjecture of Newey (1990a).

Consider next the conditional symmetry case. Here, since the density is an even function of ε, the score is also an even function of ε, i.e. $S_\eta(-\varepsilon,x) = S_\eta(\varepsilon,x)$. Since this is the only restriction implied by conditional symmetry, we expect the tangent set to be

(3.6) $$\mathcal{T} = \{t(\varepsilon,x) : t(-\varepsilon,x) = t(\varepsilon,x), \quad E[t(\varepsilon,x)] = 0\}.$$

For a $q \times 1$ random vector $R(z) = R(y,x) = R(\pi(\varepsilon,x,\beta_0),x)$ let

(3.7) $$\bar{R} = [R(z) + R(\pi(-\varepsilon,x,\beta_0),x)]/2 - E[R].$$

By construction and symmetry (which imply $E[R(\pi(-\varepsilon,x,\beta_0),x)] = E[R]$), $\bar{R}$ is an element of the tangent set. Symmetry also implies $E[(R-\bar{R})'t] = E[\{R(z)-R(\pi(-\varepsilon,x,\beta_0),x)\}'t]/2 = 0$, since $\{R(z)-R(\pi(-\varepsilon,x,\beta_0),x)\}'t(\varepsilon,x)$ is an odd function of ε. Thus, the efficient score should be

$$(3.8) \qquad S = [S_\beta(z) - S_\beta(\pi(-\varepsilon,x,\beta_0),x)]/2$$

$$= \{J_\beta(z) - J_\beta(\pi(-\varepsilon,x,\beta_0),x)$$

$$+ \{\rho_\beta(z) + \rho_\beta(\pi(-\varepsilon,x,\beta_0),x)\}'s(\varepsilon,x)\}/2.$$

The following result makes this calculation rigorous.

Theorem 3.2: Suppose that ε *is symmetrically distributed around zero conditional on* x, *Assumption 3.1 is satisfied, and* $E[SS']$ *is nonsingular for* S *from equation (3.8). Then* S *is the efficient score.*

By combining the preceding calculations is is possible to obtain a result for the case where both independence and symmetry hold. The tangent set, projection, and efficient score for this case are

$$(3.9) \qquad \mathcal{T} = \{t_1(\varepsilon)+t_2(x) : E[t_1(\varepsilon)] = E[t_2(x)] = 0, \quad t_1(-\varepsilon) = t_1(\varepsilon)\},$$

$$\bar{R} = E[R|x]-E[R] + (E[R|\varepsilon] + E[R|\varepsilon]|_{\varepsilon=-\varepsilon})/2 - E[R],$$

$$S = S_\beta(z) - (E[S_\beta|\varepsilon] + E[S_\beta|\varepsilon]|_{\varepsilon=-\varepsilon})/2,$$

a result verified in:

Theorem 3.3: Suppose ε *and* x *are independent,* ε *is symmetrically distributed around zero, Assumption 3.1 are satisfied, each parametric submodel corresponds to a density* $f(\varepsilon,x|\eta) = f(\varepsilon|\eta)f(x|\eta)$ *such that* $f(\varepsilon|\eta)$ *and* $f(x|\eta)$ *are smooth, and* $E[SS']$ *is nonsingular for* S *from equation (3.9). Then* S *is the efficient score.*

To show that an estimator is efficient in the sense discussed above, which includes regularity of the estimator, it is useful to impose on each parametric submodel the additional regularity condition of continuity of $E_{\theta}[\|S\|^2]$ at θ_0. The following result verifies that this additional regularity condition does not change the bound in any case discussed above.

Theorem 3.4 For each of Theorems 3.1 - 3.3, if $E_{\beta}[\|S\|^2]$ *is continuous at* β_0, *then the conclusion of the Theorems are unchanged if the parametric submodels are restricted to those with* $E_{\theta}[\|S\|^2]$ *is continuous at* θ_0.

4. V-Statistic Estimation for the Independence Case

The type of estimators to be considered in this paper satisfy a general estimating equation of the form

(4.1) $$\hat{m}_n(\hat{\theta}) = o_p(1/\sqrt{n}),$$

where for each θ, $\hat{m}_n(\theta)$ is a statistic. The idea is that it is known that $\hat{m}_n(\theta_0) = o_p(1)$, and $\hat{\theta}$ is being chosen so that $\hat{m}_n(\theta)$ is close to the true limiting value of zero. Here θ includes the parameters of interest β as well as additional nuisance parameters.

A choice of $\hat{m}_n(\theta)$ with high asymptotic efficiency is motivated by the form of the efficient score. The idea is to base an estimating equation on a sample average of the efficient score score for some particular parametric families of distributions of ε. For reasons to be discussed below, the consistency of such an estimator should not depend on the assumption about the distribution, and the estimator should be efficient when the true distribution is a member of the class, because it is based on the efficient score.

The justification of this procedure arises from the fact that the efficient score has mean zero, suggesting that an estimator could be based on the sample average of the efficient score. Of course, the form of some components of the efficient score may not be known, so that estimation of these components would be required for such an m-estimator. Rather than following this procedure, which could be complicated and involve a proliferation of nuisance parameters, one might fix some components of the efficient score at known values. The convex model results of Bickel (1982), BKRW, and Newey (1990a) provide conditions under which such a procedure is valid. As defined by Bickel (1982), a convex model is one where the nonparametric components are taken from a convex set. BKRW showed that if the model is convex, then when the efficient score is evaluated at some false distribution it retains the zero mean property. Newey (1990a) extended this result to show that the same conclusion holds componentwise in nonparametric components. Thus, when a model is convex in some nonparametric component, that component in the efficient score can be replaced by a fixed value without affecting the mean zero property.

Consider the efficient score, in equation (3.5), for the independence case. Replacing the true location score $s(\varepsilon)$ by a known, parametric function $\tilde{s}(\varepsilon,\eta)$ gives

$$(4.2) \qquad m \equiv \tilde{S}(z,\theta_0) - E[\tilde{S}|\varepsilon], \quad \tilde{S}(z,\theta) \equiv J_\beta(z,\beta) + \rho_\beta(z,\beta)'\tilde{s}(\rho(z,\beta),\eta),$$

where $\theta = (\beta',\eta')'$. Note that $E[m] = 0$ holds by the definition of m, a result that follows from convexity of this model in the density function of ε.

To construct a feasible estimator, it is necessary to estimate the conditional expectation in (4.2), and to choose a value of the nuisance parameters η. The literature on residual based prediction for nonlinear simultaneous equations models, e.g. Duan (1983), Brown and Mariano (1984), Robinson (1989), suggests a useful estimator of this conditional expectation. Because of independence of exogenous variables and disturbances, the conditional expectation can be estimated by averaging $\tilde{S}(z,\beta_0)$ over all the sample values of x holding $\rho(z,\beta_0)$ fixed. This fact suggests that this same average with β_0 replaced by β might be used in the construction of an estimator. Such an average can be constructed via the reduced form $\pi(\varepsilon,x,\beta)$. Recall that $z = (y',x')' = (\pi(\rho(z,\beta),x,\beta)',x')$. Substituting for z in $\tilde{S}(z,\beta)$ and averaging over x gives

$$(4.3) \qquad \hat{E}[\tilde{S}|\rho(z,\beta)] \equiv \textstyle\sum_{j=1}^{n}\tilde{S}(\pi(\rho(z,\beta),x_j,\beta),x_j,\beta)/n.$$

Let η be nuisance parameters corresponding to, say, location and scale of ε, and let $\theta = (\beta',\eta')'$. Substituting $\hat{E}$ for $E[\tilde{S}|\varepsilon]$ in equation (4.2), averaging, and adding equations for estimation of the nuisance parameters gives

$$(4.4) \qquad \hat{m}_n(\theta) = \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}b(z_i,z_j,\theta)/n^2, \quad b(z,\tilde{z},\theta) = (b_1(z,\tilde{z},\theta)',\chi(z,\theta))',$$

$$b_1(z,\tilde{z},\theta) = \tilde{S}(z,\theta) - \tilde{S}(\pi(\rho(z,\beta),\tilde{x},\beta),\tilde{x},\theta)$$

$$= J_\beta(z,\beta) - J_\beta(\pi(\rho(z,\beta),\tilde{x},\beta),\tilde{x},\beta)$$

$$+ [\rho_\beta(z,\beta) - \rho_\beta(\pi(\rho(z,\beta),\tilde{x},\beta),\tilde{x},\beta)]'\tilde{s}(\rho(z,\beta),\eta),$$

where the elements of $\hat{m}_n(\theta)$ corresponding to $\chi(z,\theta)$ are sample averages used in the estimation of the nuisance parameters η. Here $\hat{m}_n(\theta)$ is a V-statistic (e.g. see Serfling, 1981), so that a natural name for this estimator is a V-estimator.

Detailing a specific example may help in understanding the nature of this estimator. Consider the transformation model of equation (2.5), with $f(x,\beta,\lambda) = x'\beta_1$. Let $h^{-1}(\circ,\lambda)$ denote the inverse of the transformation function, and let $\eta = (\mu,\sigma^2)'$, $\tilde{s}(\varepsilon,\eta) = -(\varepsilon-\mu)/\sigma^2$, and $\chi(z,\theta) =$

$(h(y,\lambda)-\mu-x'\beta_2, \sigma^2-[h(y,\lambda)-\mu-x'\beta_2]^2)'$. Then for $J_\lambda(y,\lambda) \equiv h_{y\lambda}(y,\lambda)/h_y(y,\lambda)$, with subscripts on h denoting partial derivatives, and $u_{ij}(\beta) \equiv x_j'\beta_2 + \rho(z_i,\beta) = (x_j-x_i)'\beta_2 + h(y_i,\lambda)$, the estimating equation (4.1) with $\hat{m}_n(\theta)$ as in equation (4.4) is

$$(4.5) \quad \textstyle\sum_{i=1}^n\{ J_\lambda(y_i,\hat{\lambda})/n - \sum_{j=1}^n J_\lambda(h^{-1}(u_{ij}(\hat{\beta}),\hat{\lambda}),\hat{\lambda})/n \}/n$$

$$- \textstyle\sum_{i=1}^n\{ h_\lambda(y_i,\hat{\lambda})/n - \sum_{j=1}^n h_\lambda(h^{-1}(u_{ij}(\hat{\beta}),\hat{\lambda}),\hat{\lambda})/n \}[h(y_i,\hat{\lambda})-\hat{\mu}-x_i'\hat{\beta}_2]/n\hat{\sigma}^2$$

$$= o_p(1/\sqrt{n}),$$

$$\textstyle\sum_{i=1}^n(1,x_i')'[h(y_i,\hat{\lambda})-\hat{\mu}-x_i'\hat{\beta}_2]/n = o_p(1/\sqrt{n}),$$

$$\hat{\sigma}^2 - \textstyle\sum_{i=1}^n[h(y_i,\hat{\lambda})-\hat{\mu}-x_i'\hat{\beta}_2]^2/n = o_p(1/\sqrt{n}),$$

where the subscripts on h denote partial derivatives and the second equality follows by combining terms corresponding to estimating equations for β_2 and μ. With the exception of the first equation, these equations are the usual likelihood equations for (quasi) maximum likelihood estimation of transformation models with normal disturbances. The first corresponds to the likelihood score, e.g. including a Jacobian term, but its validity does not depend on the distribution having any particular form.

Throughout this paper the Gaussian score is used mainly for illustrative purposes. In practice, for robustness reasons it may be desireable to use a bounded function of the disturbance in the estimator, as will be described below.

To relate this estimator to previous results, it may be helpful to point out the way in which other estimators can be formulated as V-estimators. Of course, since means are V-statistics, any m-estimator is also a V-estimator. For example, instrumental variables estimators correspond to choosing

(4.6) $\qquad b(z,\tilde{z},\theta) = A(x)[\rho(z,\beta) - \eta],$

where $A(x)$ is a $(q+s) \times q$ matrix of instrumental variables and η parameterizes $E[\varepsilon]$. M-estimators such as those of MaCurdy (1982), Taylor (1985), and Ruppert and Aldershof (1989) correspond to replacing $\rho(z,\beta)$ by some vector of functions $r(\rho(z,\beta))$ and η by corresponding mean parameters. Robinson (1989) has previously suggested a V-estimator that is not a simple m-estimator. This estimator has the form of the instrumental variables estimator corresponding to (4.6), only $A(x)$ is replaced by a residual-based estimator of the optimal (asymptotic variance minimizing) instruments. An iterative version of Robinson's (1989) two-step estimator could be obtained from the solution of equation (4.1) with $\eta = (\mu',\text{hvec}(\Sigma)')'$, where μ and Σ parameterize the mean and covariance matrix of ε, respectively, $\text{hvec}(\cdot)$ denotes the usual vectorization of a symmetric matrix, and

(4.7) $\qquad b(z,\tilde{z},\theta) = (b_1(z,\tilde{z},\theta)',\chi(z,\theta))'$

$$b_1(z,\tilde{z},\theta) = [\rho_\beta(\pi(\rho(\tilde{z},\beta),x,\beta),x,\beta),-I]'\Sigma^{-1}[\rho(z,\beta)-\mu]$$

$$\chi(z,\theta) = \text{hvec}(\Sigma - [\rho(z,\beta)-\mu][\rho(z,\beta)-\mu]').$$

Although this estimator has a nontrivial V-statistic structure, it does not correspond to subtracting an residual-based estimate of a conditional expectation, and so is quite different than the one proposed above.

In comparison with these other estimators, one expects high efficiency efficiency for $\hat{\beta}$ obtained with $b(z,\tilde{z},\beta)$ as specified in equation (4.4), at least for some distributions, because it more closely matches the efficient score. For example, consider $\eta = (\mu',\text{vech}(\Sigma)')'$ and

$$(4.8)\qquad \tilde{s}(\varepsilon,\eta) = -\Sigma^{-1}(\varepsilon-\mu),$$

$$\chi(z,\theta) = (\ (\rho(z,\beta) - \mu)',\ \text{hvec}(\Sigma - [\rho(z,\beta)-\mu][\rho(z,\beta)-\mu]')'\ )'.$$

As discussed below, the V-estimator obtained via equation (4.4) will be an efficient semiparametric estimator under normality, i.e. when the disturbances are normal it will be efficient in the class of estimators that do not depend on the distribution of the disturbances. More robust and/or partially adaptive alternatives could be obtained by applying existing ideas to the choice of $\tilde{s}(\varepsilon,\eta)$ and estimating equations for η. It should be possible to make the efficiency of the estimator less sensitive to the true distribution by replacing $-\Sigma^{-1}(\varepsilon-\mu)$ by $\Sigma^{-1/2\prime}\psi(\Sigma^{-1/2}(\varepsilon-\mu))$, where $\psi(\circ)$ is a bounded function, and replacing $\chi(z,\theta)$ by corresponding location and scale estimating equations; see Huber (1981). Also, one might enlarge the class of distributions for which the estimator is efficient by including in η and $\chi(z,\eta)$ shape parameters and corresponding estimating functions.

To make asymptotic inferences concerning the parameters it is necessary to construct an estimator of the asymptotic variance of $\hat{\theta}$. As usual, the asymptotic variance will be of the form $M^{-1}\Omega M^{-1\prime}$, where $M = \text{plim}[\partial\hat{m}_n(\theta_0)/\partial\theta]$ and Ω is the asymptotic variance of $\sqrt{n}\hat{m}_n(\theta_0)$; an estimator $\hat{M}^{-1}\hat{\Omega}\hat{M}^{-1\prime}$ can be constructed from corresponding estimators $\hat{M}$ and $\hat{\Omega}$. As estimator of M can easily be constructed as

$$(4.9)\qquad \hat{M} = \partial\hat{m}_n(\hat{\theta})/\partial\theta.$$

Estimation of Ω is more difficult, but can be carried out via V-statistic theory. The V-statistic projection theorem (e.g. Serfling, 1980) and $E[b(z_1,z_2,\theta_0)] = 0$ give

(4.10) $$\sqrt{n}\hat{m}_n(\theta_0) = \sum_{i=1}^n u_i/\sqrt{n} + o_p(1),$$

$$u_i \equiv E[b(z_j,z_i,\theta_0)|z_i] + E[b(z_i,z_j,\theta_0)|z_i], \quad i \neq j.$$

Thus, by the central limit theorem $\Omega = E[u_i u_i']$. Since the observations are independent, the conditional expectations in (4.10) can be estimated by replacing θ_0 by $\hat{\theta}$ and summing over all the other observations. Then Ω can be estimated as the sample average of the estimates of u_i, yielding

(4.11) $$\hat{\Omega} = \sum_{i=1}^n \hat{u}_i \hat{u}_i'/n, \qquad \hat{u}_i = \sum_{j=1}^n [b(z_i,z_j,\hat{\theta}) + b(z_j,z_i,\hat{\theta})]/n.$$

It is difficult to be fully primitive concerning regularity conditions at the level of generality considered here. The following assumption gives a set of identification and dominance conditions, verification of which may require substantial work in particular models. For a matrix A let $\|A\| = [\text{trace}(A'A)]^{1/2}$, and let the dimension of θ be p.

Assumption 4.1: i) θ_0 is an element of the interior of some compact set Θ; ii) θ_0 is the unique solution of $E[b(z_1,z_2,\theta)] = 0$ for $\theta \in \Theta$. iii) $b(z_1,z_2,\theta)$ is twice continuously differentiable on Θ with probability one and there exists $B(z,\tilde{z})$ such that for all $\theta \in \Theta$, $\|b(z_1,z_j,\theta)\|^2 \le B(z_1,z_j)$, $\|\partial b(z_1,z_j,\theta)/\partial\theta\|^2 \le B(z_1,z_j)$, $\|\partial^2 b(z_1,z_j,\theta)_k/\partial\theta\partial\theta'\| \le B(z_1,z_j)$, $E[B(z_1,z_j)] < \infty$, $(j = 1,2;\ k = 1,\ldots,p)$; iv) M and Ω are nonsingular.

Under this regularity condition a V-estimator has the usual properties:

Theorem 4.1: If Assumption 4.1 is satisfied then with probability approaching one there exists a unique solution $\hat{\theta}$ *of* $\sum_{i=1}^{n}\sum_{j=1}^{n} b(z_i, z_j, \hat{\theta})/n^2 = 0$ *over* Θ *such that*

$$(4.12) \qquad \sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0, M^{-1}\Omega M^{-1\prime}), \qquad \hat{M}^{-1}\hat{\Omega}\hat{M}^{-1\prime} \xrightarrow{p} M^{-1}\Omega M^{-1\prime}.$$

It is useful to note that the asymptotic variance of $\hat{\beta}$ is unaffected by estimation of the nuisance parameters, leading to a simplification in the variance formula. Let M, $\hat{M}$, Ω, and $\hat{\Omega}$ be partitioned conformably with $b(z, \tilde{z}, \theta)$ in equation (4.4) and $\theta = (\beta', \eta')'$, e.g. $M_{11} = E[\partial b_1(z_1, z_2, \theta_0)/\partial\beta]$.

Corollary 4.2: If $b(z_1, z_2, \theta)$ *is given in equation (4.4), and satisfies Assumption 4.1, and* $\tilde{s}(\varepsilon, \eta)$ *is differentiable in* η, *then*

$$(4.13) \qquad \sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, M_{11}^{-1}\Omega_{11}M_{11}^{-1\prime}), \qquad \hat{M}_{11}^{-1}\hat{\Omega}_{11}\hat{M}_{11}^{-1\prime} \xrightarrow{p} M_{11}^{-1}\Omega_{11}M_{11}^{-1\prime}.$$

As promised, the estimator $\hat{\beta}$ is semiparametric efficient if the true disturbance score is a member of the parametric family $\tilde{s}(\varepsilon, \eta)$ and $\hat{\eta}$ is a consistent estimator of the true nuisance parameters.

Theorem 4.3: Suppose that Assumptions 3.1 and 4.1 are satisfied and that $E_\beta[B(z_1, z_2)]$ *is continuous in a neighborhood of* β_0. *If* $s(\varepsilon) = \tilde{s}(\varepsilon, \eta_0)$ *then the asymptotic variance of* $\hat{\beta}$ *equals the semiparametric bound. Furthermore,* $\hat{\beta}$ *is regular if* $E_\beta[\|S\|^2]$ *is continuous at* β_0 *and parametric submodels are restricted to those with* $E_\theta[\|S\|^2]$ *continuous at* θ_0.

As noted in Theorem 3.4, continuity of $E_\theta[\|S\|^2]$ is a regularity condition that does not affect the form of the bound.

5. Antithetic Variate Estimation for the Symmetric Case

Motivated by efficiency considerations similar to those for the independence case, an estimating equation can be obtained from the efficient score by replacing the unknown disturbance score by a known function. Consider the efficient score in equation (3.8). Replacing $s(\varepsilon,x)$ by a known function $\tilde{s}(\varepsilon,x,\eta)$ that is an odd function of ε given x and η, and replacing θ_0 by θ gives

$$(5.1) \qquad m = \tilde{S}(z,\theta_0), \quad \tilde{S}(z,\theta) = [J_\beta(z,\beta) - J_\beta(\pi(-\rho(z,\beta),x,\beta),x,\beta)]/2 + [\rho_\beta(z,\beta) + \rho_\beta(\pi(-\rho(z,\beta),x,\beta),x,\beta)]'\tilde{s}(\rho(z,\beta),x,\eta)/2.$$

Note that this procedure has the effect of making $\tilde{S}(z,\theta)$ an odd function of $\rho(z,\beta)$ given x, so that $E[m] = 0$. Averaging over the sample and adding equations for estimation of η gives

$$(5.2) \qquad \hat{m}_n(\theta) = \textstyle\sum_{i=1}^n b(z_i,\theta)/n, \quad b(z,\theta) = (\tilde{S}(z,\theta)',\chi(z,\theta)')'.$$

Here the solution to equation (4.1) is an m-estimator based on an antithetic variate symmetrization of a function of the data (and parameters).

Detailing a specific example may help in understanding the nature of this estimator. Consider the transformation model as in Section 4, and let $\eta = \sigma^2$, $\tilde{s}(\varepsilon,x,\eta) = -\varepsilon/\sigma^2$, and $\chi(z,\theta) = \sigma^2 - [h(y,\lambda)-x'\beta]^2$. Then for $J_\lambda(y,\lambda)$ as before and $u_i(\beta) \equiv x_i'\beta_2 - \rho(z_i,\beta) = 2x_i'\beta_2 - h(y_i,\lambda)$, the estimating equation (4.1) with $\hat{m}_n(\theta)$ as in equation (5.2) is

$$(5.3)\quad \sum_{i=1}^{n}\{ J_\lambda(y_i,\hat\lambda)/n - J_\lambda(h^{-1}(u_i(\hat\beta),\hat\lambda),\hat\lambda) \}/2n$$

$$- \sum_{i=1}^{n}\{ h_\lambda(y_i,\hat\lambda)/n + h_\lambda(h^{-1}(u_i(\hat\beta),\hat\lambda),\hat\lambda) \}[h(y_i,\hat\lambda)-x_i'\hat\beta_2]/2n\hat\sigma^2$$

$$= o_p(1/\sqrt{n}),$$

$$\sum_{i=1}^{n}x_i[h(y_i,\hat\lambda)-x_i'\hat\beta_2]/n = o_p(1/\sqrt{n}),$$

$$\hat\sigma^2 - \sum_{i=1}^{n}[h(y_i,\hat\lambda)-x_i'\hat\beta_2]^2/n = o_p(1/\sqrt{n}).$$

With the exception of the first equation, these equations are the usual likelihood equations for quasi maximum likelihood estimation of a transformation model with normal disturbances. The first equation includes a Jacobian term, as do likelihood equations, but unlike the corresponding likelihood equation its validity depends only on conditional symmetry, and not on the distribution having any particular form.

These estimators are related to previously suggested estimators in the sense that they are based on odd (antisymmetric) functions of the disturbances given x. For example, Taylor (1985) and Ruppert and Aldershof (1989) have considered estimators based on the cubed residual.

As shown below, this estimator will attain the semiparametric efficiency bound for the conditionally symmetric case if the true disturbance score is $\tilde s(\varepsilon,x,\eta_0)$ where η_0 is the limit of $\hat\eta$. For example, consider $\eta = \mathrm{vech}(\Sigma)$,

$$(5.4)\quad \tilde s(\varepsilon,x,\eta) = -\Sigma^{-1}\varepsilon, \quad \chi(z,\theta) = \mathrm{hvec}(\Sigma - \rho(z,\beta)\rho(z,\beta)').$$

The corresponding estimator obtained via equations (5.1) and (5.2) will be efficient under normality and homoskedasticity of ε. More robust and/or partially adaptive estimators could be obtained by applying well known ideas to the choice of $\tilde s(\varepsilon,x,\eta)$ and $\chi(z,\theta)$. For example, some heteroskedasticity might be allowed for by specifying Σ to be a function of x and unknown parameters, and replacing $\chi(z,\theta)$ by squared residual estimating equations for these parameters. More robust versions could also be obtained by applying ideas of Carroll and Ruppert (1982).

These antithetic estimators are m-estimators of the Huber (1967) form so that asymptotic inference procedures can be carried out in the corresponding way. The asymptotic variance of $\hat{\theta}$ will be of the form

$$(5.5) \qquad M^{-1}\Omega M^{-1\prime}, \quad M = E[\partial b(z,\theta_0)/\partial\theta], \quad \Omega = E[b(z,\theta_0)b(z,\theta_0)'],$$

which can be estimated by replacing expectations by sample averages and true parameter values by estimates, as

$$\hat{M}^{-1}\hat{\Omega}\hat{M}^{-1\prime}, \quad \hat{M} = \textstyle\sum_{i=1}^{n}\partial b(z_i,\hat{\theta})/\partial\theta/n, \quad \hat{\Omega} = \textstyle\sum_{i=1}^{n} b(z_i,\hat{\theta})b(z_i,\hat{\theta})'/n.$$

As in Section 4, a high level set of regularity conditions can be specified. Here these are of a standard type, because the estimator $\hat{\theta}$ is just an m-estimator.

Assumption 5.1: i) θ_0 is an element of the interior of some compact set Θ; ii) θ_0 is the unique solution of $E[b(z,\theta)] = 0$ for $\theta \in \Theta$. iii) $b(z,\theta)$ is continuously differentiable on a neighborhood N of θ_0 with probability one and there exists $B(z)$ such that for all $\theta \in \Theta$, $\|b(z,\theta)\|^2 \le B(z)$ and for all $\theta \in N$, $\|\partial b(z,\theta)/\partial\theta\| \le B(z)$, $E[B(z)] < \infty$.
iv) $M = E[\partial b(z,\theta_0)/\partial\theta]$ and $\Omega = E[b(z,\theta_0)b(z,\theta_0)']$ are nonsingular.

Theorem 5.1: If Assumption 5.1 is satisfied then with probability approaching one there exists a unique solution $\hat{\theta}$ *of* $\sum_{i=1}^{n} b(z_i,\theta)/n = 0$ *over* Θ *such that*

$$(5.6) \qquad \sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0, M^{-1}\Omega M^{-1\prime}), \qquad \hat{M}^{-1}\hat{\Omega}\hat{M}^{-1\prime} \xrightarrow{p} M^{-1}\Omega M^{-1\prime}.$$

As in Section 4, the asymptotic variance of the estimator of the parameters of interest simplifies. Let M, $\hat{M}$, Ω, and $\hat{\Omega}$ be partitioned conformably with $b(z,\theta)$ in equation (5.2) and $\theta = (\beta',\eta')'$, e.g. $M_{11} = E[\partial b_1(z_1,z_2,\theta_0)/\partial\beta]$.

Corollary 5.2: If $b(z,\theta)$ *is given in equation (5.2), and satisfies Assumption 5.1, and* $\tilde{s}(\varepsilon,x,\eta)$ *is differentiable in* η, *then*

$$(5.7) \qquad \sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,\ M_{11}^{-1}\Omega_{11}M_{11}^{-1\prime}), \qquad \hat{M}_{11}^{-1}\hat{\Omega}_{11}\hat{M}_{11}^{-1\prime} \xrightarrow{p} M_{11}^{-1}\Omega_{11}M_{11}^{-1\prime}.$$

As promised, the estimator $\hat{\beta}$ is semiparametric efficient if the true disturbance score is a member of the parametric family $\tilde{s}(\varepsilon,x,\eta)$ and $\hat{\eta}$ is a consistent estimator of the true nuisance parameters.

Theorem 5.3: Suppose that Assumptions 3.1 and 5.1 are satisfied and that $E_\beta[B(z)]$ *is continuous in a neighborhood of* β_0. *If* $s(\varepsilon,x) = \tilde{s}(\varepsilon,x,\eta_0)$ *then the asymptotic variance of* $\hat{\beta}$ *equals the semiparametric bound. Furthermore,* $\hat{\beta}$ *is regular if* $E_\beta[\|S\|^2]$ *is continuous at* β_0 *and parametric submodels are restricted to those with* $E_\theta[\|S\|^2]$ *is continuous at* θ_0.

It also possible to construct an estimator that uses both independence and symmetry, by combining the previous estimators in a way analogous to that in which the efficient scores are combined. Let $\tilde{s}(\varepsilon,\eta)$ be an odd function of ε given η. Consider a V-estimator with

$$(5.8) \qquad b(z,\tilde{z},\theta) = (b_1(z,\tilde{z},\theta)',\chi(z,\theta))',$$

$$b_1(z,\tilde{z},\theta) = J_\beta(z,\beta) - \rho_\beta(z,\beta)'\tilde{s}(\rho(z,\beta),\eta)$$

$$- [J_\beta(\pi(\rho(z,\beta),\tilde{x},\beta),\tilde{x},\beta) - \rho_\beta(\pi(\rho(z,\beta),\tilde{x},\beta),\tilde{x},\beta)]'\tilde{s}(\rho(z,\beta),\eta)]/2$$

$$- [J_\beta(\pi(-\rho(z,\beta),\tilde{x},\beta),\tilde{x},\beta) + \rho_\beta(\pi(-\rho(z,\beta),\tilde{x},\beta),\tilde{x},\beta)]'\tilde{s}(\rho(z,\beta),\eta)]/2.$$

This estimator mimics the form of the efficient score. It will be consistent when the distribution of ε is independent of x and symmetric around zero, and attain the bound from Theorem 3.3 when the true score is a member of the assumed family $\tilde{s}(\varepsilon,\eta)$ and the nuisance parameters are estimated consistently. For brevity, formal results are omitted.

6. A Box-Cox Example

An empirical and Monte Carlo example may provide some information concerning the performance of the estimators in practice. The model considered was the Box-Cox version of the equation (2.5) transformation model. The data was the Engel curve data described in Koenker and Bassett (1982), which consists of 224 observations on household income and expenditure. The object of the empirical and Monte Carlo exercises was the determination of a relationship between income and expenditure. To this end, the dependent variable of the Box-Cox model was specified to be the level of food expenditure and the regressors to be equal to a constant and the natural log of income. Also, in keeping with this objective, and to avoid the usual difficulty of interpreting parameters of transformation models, results for the regression coefficients are not reported. Instead, we consider the income elasticity of expenditure, which is defined as the derivative of a location measure for the conditional distribution of the natural log of income with respect to log-income, evaluated at the mean of log income. The location measure chosen here is the conditional median, which is convenient for transformation models, because it is equivariant with respect to monotonic transformations. For the Box-Cox model the conditional median elasticity is

(6.1) $$e \equiv \beta_{22}/[1 + \lambda\cdot(\bar{x}'\beta_2)],$$

assuming that the disturbance has conditional median zero. The empirical and Monte Carlo results were found to be insensitive to the assumption of conditional median zero (adjusting for the true median of the disturbances only affected the results in the fourth decimal place), so that for convenience the zero conditional median assumption was imposed in the elasticity estimation.

Table One reports several different estimates, corresponding to the V-statistic and antithetic variate estimators as well as others. The "log regression" rows are from a regression of the log of expenditure on the log of income. The other rows are obtained from various Box-Cox estimators. Each of these estimates was computed by concentrating out the regression and variance parameters by least squares regression of $h(y_i,\lambda)$ on x_i, which is taken throughout to include a constant, and using the resulting estimates $\hat{\beta}_2(\lambda)$, $\sigma^2(\lambda)$ to solve for $\hat{\lambda}$ from a scalar equation of the form

(6.2) $$\textstyle\sum_{i=1}^{n}\hat{m}(z_i,\hat{\lambda},\hat{\beta}_2(\hat{\lambda}),\hat{\sigma}^2(\hat{\lambda}))/n = 0.$$

The rows labeled "QMLE" are for the Box and Cox (1964) quasi maximum likelihood estimator where

$$\hat{m}(z,\lambda,\beta_2,\sigma^2) = \ln(y) - h_\lambda(y,\lambda)\cdot[h(y,\lambda)-x'\beta_2]/\sigma^2.$$

The rows labeled "V-stat" and "Antithetic" took $\hat{m}$ to be the estimate of the Gaussian efficient score for λ under independent and conditionally symmetric disturbance respectively, as described in Sections 4 and 5. For the rows labeled "IV," $\hat{m}$ was specified as

$$\hat{m}(z,\lambda,\beta_2,\sigma^2) = (x'\beta_2)^2[h(y,\lambda)-x'\beta_2],$$

which gives an nonlinear instrumental variables estimator in the class considered by Amemiya and Powell (1981). The algorithm for solving equation (6.2) was a grid search over $[-.31,1.2]$ for changes of sign of the left-hand side of equation (6.2) followed by Newton-Raphson (with numerical derivative). This algorithm never failed to produced a unique root. For the antithetic estimator, it was necessary to reduce the grid to $[-.31,.8]$, because the symmetrized reduced form $\pi(-\rho(z,\beta),x,\beta)$ was not well defined for some observations for larger values of λ. Throughout, computations were performed via GAUSS on a microcomputer.

As in Newey (1987), the choice of covariate $(x'\beta_2)^2$ is motivated by efficiency considerations. From Amemiya (1974) it is known that the optimal (asymptotic variance minimizing) function $a(x)$ in a choice of m as $a(x)[h(y,\lambda)-x'\beta_2]$ is $E[h_\lambda(y,\lambda_0)|x] + A + x'B$ for any constant A and vector B. If the disturbance has conditional mean zero and constant conditional variance, then for $\lambda_0 = 0$,

(6.3) $$E[h_\lambda(y,\lambda_0)|x] = E[\{\ln(y)\}^2|x]/2 = [(x'\beta_2)^2 + \sigma_\varepsilon^2]/2,$$

so that $(x'\beta_2)^2$ is optimal. Furthermore, it appears that $(x'\beta_2)^2$ is an approximately optimal instrument for many cases where $\lambda_0 \neq 0$. For example, for $(\lambda,\beta_2') = (.23,-11,3.7)$, ε distributed as $N(0,.36)$, and x distributed as $N(6.7,.19)$, which (except for the normality assumption) correspond to the IV values in Table One, the ratio of asymptotic standard error for the optimal IV estimator of λ to that of the estimator with instrument $(x'\beta_2)^2$ is .98.

It should be noted here, and in the context of the Gaussian cases discussed below, that, strictly speaking, the Box-Cox model is not well defined for Gaussian disturbances, because of the positivity constraint on y. However, for the Engel data and the parameter configurations considered here, a violation of the positivity condition requires that either log income or the disturbance is 8 standard deviations below their respective means. Thus, violation o. the positivity constraint is a very unlikely event, and ignoring this constraint, as was done here, will have little affect on the results.

Table One reports specification robust asymptotic standard error estimates, calculated from the Huber (1967) type formulae discussed earlier, except for the QMLE and V-stat standard error estimates. A numerical problem was encountered in calculating the QMLE and V-stat standard error estimates. The derivative matrix of the sample moments is ill-conditioned, with small changes in its elements leading to large changes in its inverse, and thus in the associated standard error estimates. To circumvent this problem the standard errors for QMLE and V-stat were estimated by the bootstrap. The QMLE standard error was the Monte Carlo standard error of the estimates from 224 replications of 224 observations drawn randomly from the empirical distribution of (y_i, x_i). The V-stat standard error was calculated in the same way, except that to lower computation time only 75 replications were used. Resampling from the empirical distribution of the entire vector would seem to be an appropriate method for obtaining specification robust standard error estimates. The other estimators did not suffer from this numerical problem, suggesting that it may be related to the fact that $\hat{\lambda}$ is near zero.

The estimates of λ in Table One vary substantially across the different methods, with the IV and antithetic estimates being greater than the others. There is much less variation in the elasticity estimates, although the IV and antithetic variate estimates appear to be slightly higher than the others.

Motivated by the nonlinearity of the model, the Monte Carlo design was based on the empirical example. Because there are few invariant features of the model, a comprehensive design would be quite large, and is beyond the scope of this paper. A design based on an empirical example at least gives some information concerning how the estimator might perform in an application of interest.

Both independent and conditionally symmetric disturbance distributions were considered. In each case the number of replications was 200. Table Two reports results for the independence case. Here the sample size was taken to be 75, in order to facilitate repeated computation of the V-statistic estimator. For each replication the regressor observations were drawn as an i.i.d. sample from the empirical distribution of the 224 data observations. In one case the disturbances were Gaussian, and in the other, labeled "Empirical ε" in the table, they were drawn from the empirical distribution of residuals calculated using the true values given at the beginning of the table. These true values were a value for λ corresponding to the IV estimates of Table One and a value for β_2 and σ_ε^2 obtained from a least squares regression of $h(y,\lambda)$ on x in the data. The IV estimate was chosen as the true value because its consistency only requires a conditional mean zero disturbance, so that it would be consistent under either independence or conditional symmetry. An additional estimator, labeled "Heterosked" in Table Two, was also computed. It has

$$\hat{m}(z,\lambda,\beta_2,\sigma^2) = (x'\beta_2)\{[h(y,\lambda)-x'\beta_2]^2 - \sigma^2\},$$

and corresponds to a transformation to homoskedasticity as in Ruppert and Aldershof (1989). The use of $x'\beta_2$ as a co-variate is motivated by similar considerations to the choice of $(x'\beta_2)^2$ as the IV co-variate; note that at $\lambda_0 = 0$, $E[\partial\{[h(y,\lambda)-x'\beta_2]^2 - \sigma^2\}/\partial(\lambda,\beta_2',\sigma^2)|x] = E[(2\varepsilon[\ln(y)]^2,-2x'\varepsilon,-1)'|x]$ $= (2E[\varepsilon^3](x'\beta_2)^2+4\sigma_\varepsilon^2(x'\beta_2),0',-1) = (4\sigma_\varepsilon^2x'\beta_2,0',-1)$ if ε is independent of x and symmetric around zero. The numbers reported in both Tables Two and

Three are the Monte-Carlo estimates of bias (BIAS), standard error (STE), root-mean square error (RMS), median (MED), and median absolute error (MAE) for each of the estimators discussed above.

The Table Two results for λ in the Gaussian case show the expected pattern, with QMLE having smaller variance than V-stat, which in turn has smaller variance than IV. It is interesting to note that the V-stat variance is quite close to that of the QMLE, and substantially less than that of the IV estimator. An analogous ranking holds for Antithetic in relation to QMLE and IV, but Antithetic has higher variance than V-stat. It should be noted that the efficiency bound calculation does not imply a ranking of the asymptotic variances of Antithetic and V-stat, because the models are not nested.

Two other features of the Gaussian results are the tiny biases of the estimators of λ and the tiny difference in performance of the elasticity estimators. The elasticity efficiency rankings are similar to those for λ by the differences across estimators are much smaller. Also, Homosked performs poorly in comparison with the other estimators. Its standard error is substantially larger than that of V-stat and it has a very large bias. This bias is not an artifact of computational error. Modifying the computer program by simply changing the number of observations to 450 wipes out most of the bias.

The Table Two results for the empirical disturbances case show a similar pattern to those for the Gaussian case, except that Antithetic performs worse, including the presence of some bias in the estimate of λ. The QMLE estimator also exhibits evidence of bias. For the IV and V-stat estimators, efficiency rankings like those of Table 3 hold, but the efficiency gains from V-stat are smaller.

Table Three reports results for the conditionally symmetric case. The number of observations given here corresponds to the 224 of the empirical example. The Gaussian data was generated exactly as for the independence case. The other data generation process, labeled "Empirical Antithetic ε," was conditionally symmetric and possibly heteroskedastic, and corresponded to drawing each observation pair (x_i, ε_i) at random from a distribution of $2x224 = 448$ points consisting of the original data points and corresponding points with the residual replaced by its negative. An additional estimator, labeled "Sym" in Table Three, was also computed. It has

$$\hat{m}(z, \lambda, \beta_2, \sigma^2) = [h(y, \lambda) - x'\beta_3]^3,$$

and corresponds to a transformation to symmetry as in Taylor (1985) and Ruppert and Aldershof (1989).

The Table Three variance results for λ in the Gaussian case show the expected pattern, with QMLE having smaller variance than V-stat, which in turn has smaller variance than IV. However, there is evidence of bias in the QMLE, suggesting that the QMLE bias is a nonmonotonic function of the sample size (compare with the Table Two Gaussian results). The properties of the elasticity estimators are like those of Table Two. Also, the Sym estimator has a much larger variance than the others. Even the elasticity estimates for Sym are substantially less efficient than for the other estimators.

The Table Three results for the empirical disturbances case show a similar pattern to those for the Gaussian case, except that QMLE is much more biased. There is even a moderate amount of bias in the elasticity estimator. One might conjecture that heteroskedasticity in the disturbances of the empirical distribution is the source of this bias. This conjecture was checked by running a regression of the squared residuals u^2, with parameters set at the population values of Table Three, on 1, log of income $\ln(I)$, and $\ln(I)^2$. The estimate was,

$$u^2 = 13.25 - 4.18\cdot\ln(I) + .33\cdot[\ln(I)]^2, \quad F(2,222) = 18.77,$$

providing evidence of heteroskedasticity.

In Table Three, the efficiency improvement over IV for the Antithetic estimator is smaller than the corresponding improvement for V-stat in Table Two. The Table Three Antithetic root-mean square error was about 10 percent smaller than that for IV, while the Table Two V-stat root-mean square error was about 25 percent smaller.

7. Conclusion

This paper has presented estimators for nonlinear simultaneous equations models with consistency properties that do not depend on the form of the disturbance distribution and that are efficient when the true disturbance is a member of a specified parametric family. Models with independent and/or conditionally symmetric disturbances were considered. In each case the estimator is residual based, being obtained from a V-statistic in the independence case and antithetic residuals in the symmetric case.

For the special case of transformation models, the estimators here correspond to efficient transformations to homoskedasticity or symmetry, as recently considered by Ruppert and Aldershof (1989). In the Monte Carlo study, Gaussian efficient transformations perform much better than some previously suggested. It also would be interesting to investigate the finite sample properties of robust versions of these estimators.

Appendix: Proofs of Theorems

The following notation and terminology will be useful. Mean-square (m.s.) continuity of functions of θ will taken to be continuity in θ for the mean-square error norm. Similarly, m.s.-differentiability will mean Frechet differentiability in this norm. Also, throughout c and C will denote generic positive constants that need not be the same in different expressions. The conditions for smoothness and regularity of parametric submodels are like those of Ibragimov and Hasminskii (1981, Ch. 7), which reference is referred to as IH henceforth. Suppose that $\mathcal{P}_\theta = \{\ell(z|\theta) : \theta \in \Theta\}$ is a family of densities $\ell(z|\theta)$ with respect to some measure, and let dz denote integration with respect to that measure.

Definition A.1: $\mathcal{P}_\theta$ is *smooth* if Θ is open and i) $\ell(z|\theta)$ is continuous on Θ a.s.; ii) $\ell(z|\theta)^{1/2}$ is m.s. differentiable with respect to θ on Θ with derivative $\psi(z,\theta)$, i.e. $\int\|\psi(z,\theta)\|^2dz$ is finite on Θ and for each θ and $\theta_i \longrightarrow \theta$, $\int[\ell(z|\theta_i)^{1/2}-\ell(z|\theta)^{1/2}-\psi(z,\theta)'(\theta_i-\theta)]^2dz/\|\theta_i-\theta\|^2 \longrightarrow 0$; iii) $\psi(z,\theta)$ is m.s. continuous. Also, for smooth $\mathcal{P}_\theta$ the score is defined by $S_\theta \equiv 2\cdot 1(\ell(z|\theta)>0)\psi(z,\theta)/\ell(z|\theta)^{1/2}$ and the information matrix by $\int S_\theta S_\theta'\ell(z|\theta)dz$. $\mathcal{P}_\theta$ is *regular* if it is smooth and the information matrix is nonsingular on Θ.

See IH for further details.

The following pair of Lemmas are useful for proving Theorems 3.1 - 3.4. They are proved in Newey (1990b) as Lemmas C.4 and C.6, respectively.

Lemma A.1: *Suppose* $\ell(z|\beta)$ *is smooth, with score* S_β *at* β_0. *For* $\theta \equiv (\beta',\eta')'$, *let* $\Delta(z,\theta)$ *be bounded, bounded away from zero, continuously differentiable in an open ball* Θ *containing* $\theta_0 \equiv (\beta_0',0')'$, *with* $\|\partial\Delta(z,\theta)/\partial\theta\| \le b(z)$ *for* $\theta \in \Theta$, *such that* $\int b(z)^2\ell(z|\beta)d\mu$ *exists and is continuous on* Θ, $\Delta(z,\beta,0) = 1$, *and* $\int \ell(z|\beta)\Delta(z,\theta)d\mu = 1$. *Then* $\ell(z|\theta) \equiv \ell(z|\beta)\Delta(z,\theta)$ *is smooth with score* $S_\theta = (S_\beta', \Delta_\eta')'$ *at* θ_0, *where* $\Delta_\eta = 1(\ell(z|\beta_0)>0)\partial\Delta(z,\theta_0)/\partial\eta$.

Lemma A.2: *Consider vectors of random variables* $\mathcal{Y}$, $\mathcal{X}$, *and* $\mathcal{Z} = (\mathcal{Y}',\mathcal{X}')'$. *Suppose that* $s(\mathcal{Z})$ *and* $(q_1(\mathcal{Z}),\ldots,q_m(\mathcal{Z}))$ *have finite variance,* $E[s(\mathcal{Z})q(\mathcal{Z})'|\mathcal{X}] = 0$, *and* $E[q(\mathcal{Z})q(\mathcal{Z})'|\mathcal{X}]$ *is nonsingular* (a.s.). *Then there exists* $\{s_K(\mathcal{y},\mathcal{X})\}_{K=1}^{\infty}$ *such that* $E[s_K(\mathcal{Y},\mathcal{X})q_j(\mathcal{Z})|\mathcal{X}] = 0$, $E[\{s(\mathcal{Z})-s_K(\mathcal{Y},\mathcal{X})\}^2]$ 0, $s_K(\mathcal{y},\mathcal{X})$ *is bounded, and the partial derivatives of* $s_K(\mathcal{y},\mathcal{X})$ *with respect to* $\mathcal{y}$ *of all orders are bounded. Furthermore, if* $s(\mathcal{Z})$ *and* $q(\mathcal{Z})$ *are even functions of a subvector* $\tilde{\mathcal{Y}}$ *of* $\mathcal{Y}$ *and the conditional distribution of* $\tilde{\mathcal{Y}}$ *given the other components of* $\mathcal{Z}$ *is symmetric around zero, then* $s_K(\mathcal{y},\mathcal{X})$ *can be chosen as an even function of* $\mathcal{y}$.

The following Lemma and the hypotheses of Theorem 3.1 are useful for verifying the form of the tangent set conjectured in the text. For now, let $z = (\varepsilon, x)$.

Lemma A.3: *If* $f(\varepsilon|\eta)$ *and* $f(x|\eta)$ *are smooth with scores* $S_\eta(\varepsilon)$ *and* $S_\eta(x)$ *respectively, then* $f(\varepsilon|\eta)f(x|\eta)$ *is smooth with score* $S_\eta(z) = 1(\ f(\varepsilon|\eta)f(x|\eta) > 0)\{S_\eta(\varepsilon) + S_\eta(x)\}$.

Proof: Consider $\psi(\varepsilon,\eta) = f(\varepsilon|\eta)^{1/2}S_\eta(\varepsilon,\eta)/2$ and $\psi(x,\eta) = \cdot 1/2f(x|\eta)^{1/2}S_\eta(x,\eta)/2$. A standard result is that the m.s. derivative of the root-density is zero at η if the density is zero at η (e.g. see the appendix of BKRW), so that these are the m.s. derivatives of the respective densities. By the Cauchy-Schwartz and triangle inequalities, the Fubini theorem, and $\int f(x|\tilde{\eta})dx = 1$, it follows that $\int\|f(x|\tilde{\eta})^{1/2}\psi(\varepsilon,\tilde{\eta}) - f(x|\eta)^{1/2}\psi(\varepsilon,\eta)\|^2dz \le 2\int\|f(x|\tilde{\eta})^{1/2}-f(x|\eta)^{1/2}\|^2dx\cdot\int\|\psi(\varepsilon,\eta)\|^2dz + 2\int\|\psi(\varepsilon,\tilde{\eta})- \psi(\varepsilon,\eta)\|^2d\varepsilon$, so that m.s.-continuity of $f(x|\eta)^{1/2}\psi(\varepsilon,\eta)$ follows by m.s.-continuity of $f(x|\eta)^{1/2}$ and $\psi(\varepsilon,\eta)$. Similarly, $f(\varepsilon|\eta)^{1/2}\psi(x,\eta)$ is m.s.-continuous, so that by the triangle inequality $\psi(z,\eta) = f(\varepsilon|\eta)^{1/2}f(x|\eta)^{1/2}S_\eta(z,\eta)/2 = f(x|\eta)^{1/2}\psi(\varepsilon,\eta)/2 + f(\varepsilon|\eta)^{1/2}\psi(x,\eta)/2$ is also. The fact that $\psi(z,\eta)$ is the m.s.-derivative of $f(\varepsilon|\eta)$ follows similarly:

$$\int|[f(\varepsilon|\tilde{\eta})f(x|\tilde{\eta})]^{1/2}-[f(\varepsilon|\eta)f(x|\eta)]^{1/2}-\psi(z,\eta)'(\tilde{\eta}-\eta)|^2dz/\|\tilde{\eta}-\eta\|^2$$

$$\le 4\int|f(x|\tilde{\eta})^{1/2}-f(x|\eta)^{1/2}-\psi(x,\eta)'(\tilde{\eta}-\eta)|^2dx/\|\tilde{\eta}-\eta\|^2$$

$$+ 4\int|f(\varepsilon|\tilde{\eta})^{1/2}-f(\varepsilon|\eta)^{1/2}|^2d\varepsilon\cdot\int|\psi(x,\eta)'(\tilde{\eta}-\eta)|^2dx/\|\tilde{\eta}-\eta\|^2$$

$$+ 2\int|f(\varepsilon|\tilde{\eta})^{1/2}-f(\varepsilon|\eta)^{1/2}-\psi(\varepsilon,\eta)'(\tilde{\eta}-\eta)|^2d\varepsilon/\|\tilde{\eta}-\eta\|^2. \quad \blacksquare$$

Proof of Theorem 3.1: Since S was shown in Section 3 to be the residuals of the projection of S_β on $\mathcal{T}$ of eq. (3.3), by the projection interpretation of the bound it suffices to show this set is the tangent set, as claimed, and to verify regularity of $\ell(z|\beta)$. First note that $\ell(z|\beta)$ is smooth by hypothesis. Nonsingularity of its information matrix will follow from nonsingularity of $E[SS']$, since as noted in Newey (1990a, Lemma A1), $E[St']$ = 0 for S satisfying eq. (3.1) and all $t \in \mathcal{T}$, implying positive semi-definiteness of $E[S_\beta S_\beta']-E[SS']$. Then by smoothness (implying continuity of the information matrix as a function of the parameters), there will be a neighborhood of β_0 for which $\ell(z|\beta)$ is regular.

Next, to verify the form of the tangent set, consider a smooth parametric submodel. To calculate the score for η, it suffices to fix β at β_0 when calculating the score for η. Then by a change of variables $(y,x) \longrightarrow (\varepsilon,x)$ (recall $\varepsilon = \rho(z,\beta_0)$), the likelihood can be written as $f(\varepsilon|\eta)f(x|\eta)$. By Lemma A.3, $AS_\eta = AS_\eta(\varepsilon) + AS_\eta(x) \in \mathcal{T}$, by the mean zero properties of scores, so the tangent set is a subset of $\mathcal{T}$ by $\mathcal{T}$ closed.

Next, to show that any element of $\mathcal{T}$ can be approximated arbitrarily closely in mean square by the score for a regular parametric submodel, consider a submodel of the form $f(z|\beta)\Delta(z,\theta)$ for $\Delta(z,\theta) \equiv [1+\eta'h_1(\rho(z,\beta))]\circ[1+\eta'h_2(x)]$ where h_1 and h_2 are bounded, $E[h_1(\varepsilon)]=0$, $E[h_2(x)]=0$, and $h_1(\varepsilon)$ is continuously differentiable with $h_{1\varepsilon}(\varepsilon)$ bounded. That this is a probability density for η close enough to zero follows by the mean zero assumption and independence. By hypothesis, $\Delta(z,\theta)$ is continuously differentiable, with $\partial\Delta(z,\theta)/\partial\theta = \Big(\eta'h_{1\varepsilon}(\rho(z,\beta))\rho_\beta(z,\beta),\ h_1(\rho(z,\beta))'[1+\eta_2'h_2(x)] + h_2(x)'[1+\eta'h(\rho(z,\beta))]\Big)'$ bounded by $C\|\rho_\beta(z,\beta)\|$ on a neighborhood of θ_0. Then by Lemma A.1, $f(z|\theta)$ is smooth with $S_\eta = h_1(\varepsilon)+h_2(x)$. Further, by Lemma A.2, for any $t = t_1(\varepsilon) + t_2(x)$ there exists $h_1(\varepsilon)$, $h_2(x)$ such that $E[\|t_1(\varepsilon)-h_1(\varepsilon)\|^2]$ and $E[\|t_2(x)-h_2(x)\|^2]$ is arbitrarily small, implying that $E[\|t-S_\eta\|^2]$ can be made arbitrarily small. ∎

Proof of Theorem 3.2: As in the proof of Theorem 3.1, it suffices to show that $\mathcal{T}$ of equation (3.7) is the tangent set, as claimed, and to verify regularity of $\ell(z|\beta)$. Regularity of $\ell(z|\beta)$ follows as in the proof of Theorem 3.1. Next, consider a parametric submodel with score S_η for the nuisance parameters. Let $f(\varepsilon,x|\eta)$ denote the density of (ε,x). Because $f(-\varepsilon,x|\eta) = f(\varepsilon,x|\eta)$, it follows by taking an almost sure convergent subsequence of a mean-square convergent difference quotient that $S_\eta(-\varepsilon,x) = S_\eta(\varepsilon,x)$, so that the tangent set is a subset of $\mathcal{T}$. To show that any element of $\mathcal{T}$ can be approximated arbitrarily closely in mean square by the score for a regular parametric submodel, consider a submodel of the form $f(z|\beta)\Delta(z,\theta)$ for $\Delta(z,\theta) \equiv 1 + \eta'h(\rho(z,\beta),x)$ where h is bounded, $E[h(\varepsilon,x)]=0$, $h(-\varepsilon,x) = h(\varepsilon,x)$, and $h(\varepsilon,x)$ is continuously differentiable in ε with bounded derivative. It follows from Lemma A.1, as in the proof of Theorem 3.1, that this is a smooth parametric submodel, with score $h(\varepsilon,x)$. That such $h(\varepsilon,x)$

can approximate any element of $\mathcal{T}$ follows from Lemma A.2. ∎

Proof of Theorem 3.3: The logic of the proof is analogous to those of Theorems 3.1 and 3.2. Regularity of $\ell(z|\beta)$ follows as there. The fact that any linear combination of scores are elements of $\mathcal{T}$ follows by combining arguments the arguments of each proof. Consider parametric submodels of the form in the proof of Theorem 3.1, where $h_1(\varepsilon)$ is restricted to a vector of even functions. It follows as in the proof of Theorem 3.1 that the resulting submodel is regular, with score in $\mathcal{T}$. The fact that such scores can provide an arbitrarily good approximation then follows from Lemma A.2 by the same argument as in the proof of Theorem 3.1.

For this case, it remains to verify that the projection has the form given in eq. (3.9). Note that $\bar{R}$ is an element of $\mathcal{T}$ by construction. Thus, it suffices to show that $R-\bar{R}$ is orthogonal to $\mathcal{T}$. By symmetry, $E[a(\varepsilon)] = E[a(-\varepsilon)]$ for any $a(\varepsilon)$. Then for any $t = t(\varepsilon) + t(x) \in \mathcal{T}$, it follows by $t(-\varepsilon) = t(\varepsilon)$ that

$$E[(R-\bar{R})'t] = E[\{R-E[R|x]-(E[R|\varepsilon] + E[R|\varepsilon]|_{\varepsilon=-\varepsilon})/2\}'t] =$$

$$= E[\{R-E[R|x]\}'t(x)] + E[R-(E[R|\varepsilon] + E[R|\varepsilon]|_{\varepsilon=-\varepsilon})/2\}'t(\varepsilon)]$$

$$= E[\{R-E[R|\varepsilon]\}'t(\varepsilon)]/2 + E[\{R-E[R|\varepsilon]|_{\varepsilon=-\varepsilon}\}'t(\varepsilon)]/2$$

$$= \{E[R't(\varepsilon)] - E[E[R|\varepsilon]'t(-\varepsilon)]\}/2$$

$$= \{E[R't(\varepsilon)] - E[E[R|\varepsilon]'t(-\varepsilon)]\}/2 = 0. \quad ∎$$

Proof of Theorem 3.4: It suffices to show that the tangent set remains unchanged when continuity of $E_\theta[\|S\|^2]$ is imposed. Furthermore, since imposing this restriction can only shrink the tangent set, it suffices to show that $E_\theta[\|S\|^2]$ is continuous at θ_0 for the class of parametric submodels that were used above to approximate the elements of the tangent set, each of which satisfy the hypotheses of Lemma A.1. For a parametric submodel as in Lemma A.1,

$$(A.1) \qquad |E_\theta[\|S\|^2] - E_\beta[\|S\|^2]| \le \int \|S\|^2 |\Delta(z,\theta) - 1| \ell(z|\beta)dz.$$

Consider $\theta \to \theta_0$. By continuity of $\Delta(z,\theta)$ at θ_0 and regularity of $\ell(z|\beta)$, $\|S\|^2|\Delta(z,\theta) - 1|\ell(z|\beta) \to 0$ for $\theta \to \theta_0$, while by boundedness of $\Delta(z,\theta)$ and continuity of $E_\beta[\|S\|^2]$ at β_0, $\|S\|^2|\Delta(z,\theta) - 1|\ell(z|\beta) \le C\|S\|^2\ell(z|\beta)$ and $\int C\|S\|^2\ell(z|\beta)dz = CE_\beta[\|S\|^2] \to CE[\|S\|^2]$. Then by the dominated convergence theorem of Pitman (1979), $E_\theta[\|S\|^2] \to E[\|S\|^2]$. $\le C\|\eta\|E_\beta[\|S\|^2]$, by h_1, h_2 bounded. Thus, continuity of $E_\theta[\|S\|^2]$ at $\theta_0 = (\beta_0', 0)'$ follows by continuity of $E_\theta[\|S_\beta\|^2]$. ∎

The following Lemma is a standard one on the behavior of an estimator obtained from equation (4.1), and so its proof is omitted.

Lemma A.4: Suppose that i) θ_0 *is an element of the interior of a compact, convex set* Θ; ii) $\sup_{\theta\in\Theta}\|\hat{m}_n(\theta)-m_0(\theta)\| = o_p(1)$; iii) $m_0(\theta) = 0$ *has a unique solution on* Θ *at* θ_0; iv) $m_0(\theta)$ *is continuous on* Θ; v) $\hat{m}_n(\theta)$ *is continuously differentiable on a neighborhood* N *of* θ_0; vi) *for any* $\bar{\theta} = \theta_0 + o_p(1)$, $\partial\hat{m}_n(\bar{\theta})/\partial\theta = M + o_p(1)$; vii) M *is nonsingular;* viii) $\sqrt{n}\hat{m}_n(\theta_0) = \sum_{i=1}^n u_i/\sqrt{n} + o_p(1)$, $E[u_i] = 0$ *and* $E[u_iu_i']$ *is nonsingular. Then with probability approaching one there exists a unique solution* $\hat{\theta}$ *of* $\hat{m}_n(\theta) = 0$ *satisfying* $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0, M^{-1}E[u_iu_i']M^{-1\prime})$.

Proof of Theorem 4.1: The proof of the asymptotic distribution result consists of a verification of the hypotheses of Lemma A.4. Note i) holds by Assumption 4.1. For ii) and iv), note that $\hat{m}_n(\theta) = \sum_{i=1}^n\sum_{j=1}^n b(z_i,z_j,\theta)/n^2 = U_n(\theta) + T_n(\theta)$ for $U_n(\theta) = \sum_{i=1}^n\sum_{j>i}[b(z_i,z_j,\theta)+b(z_j,z_i,\theta)]/n(n-1)$ and $T_n(\theta)$

$= \hat{m}_n(\theta) - U_n(\theta)$. Note that $U_n(\theta)$ is a U-statistic with kernel $[b(z_i,z_j,\theta)+b(z_j,z_i,\theta)]/2$. Then it follows from Assumption 4.1 by the U-statistic uniform convergence result of Newey (1989b), Corollary 4.2, that $\sup_{\theta\in\Theta}|U_n(\theta)-m_0(\theta)| = o_p(1)$ for $m_0(\theta) = E[b(z_1,z_2,\theta)+b(z_2,z_1,\theta)]/2 = E[[b(z_1,z_2,\theta)]$, and that $m_0(\theta)$ is continuous, giving iv). Also, by Assumption 4.1 and the Markov inequality, $\sup_{\theta\in\Theta}\|\hat{m}_n(\theta)\| \le \sum_{i=1}^n\sum_{j=1}^n B(z_i,z_j)/n^2 = O_p(1)$, so that $\sup_{\theta\in\Theta}\|T_n(\theta)\| \le |1-(n^2/n(n-1))|\sup_{\theta\in\Theta}\|\hat{m}_n(\theta)\| + \sum_{i=1}^n B(z_i,z_i)/n(n-1) = o_p(1)$, so that ii) holds by the triangle inequality. Note that iii) and v) hold by hypothesis. Also, by an argument exactly like that just used to show ii), it follows that $\partial\hat{m}_n(\theta)/\partial\theta$ converges uniformly in probability to $E[\partial b(z_1,z_2,\theta)/\partial\theta]$ on a neighborhood of θ, which is continuous. Then vi) follows by a standard argument, while vii) holds by hypothesis. As noted in the text, viii) holds, with u_i given in eq. (4.9), by the V-statistic projection theorem, giving the final hypothesis for the asymptotic distribution result.

To show consistency of the variance estimator, note that $\hat{M} = M + o_p(1)$ follows by the definition of $\hat{M}$ in eq. (4.8), consistency of $\hat{\theta}$, and vi) of Lemma A.4, so that $\hat{M}^{-1} = M^{-1} + o_p(1)$ follows by nonsingularity of M and the Slutzky theorem. Therefore, by the Slutzky theorem, it suffices to show consistency of $\hat{\Omega}$. By the law of large numbers, consistency will follow by $\hat{\Omega} - \sum_{i=1}^n u_iu_i'/n = o_p(1)$, which in turn can be shown to follow from $\sum_{i=1}^n\|\hat{u}_i-u_i\|^2/n = o_p(1)$. Note that for $a_{ij} = b(z_i,z_j,\theta_0) + b(z_j,z_i,\theta_0)$ and $\tilde{u}_i = \sum_{j=1}^n a_{ij}/n$,

$$\text{(A.2)} \qquad \sum_{i=1}^n\|\hat{u}_i-u_i\|^2/n \le 2(T_1+T_2),\ T_1 = \sum_{i=1}^n\|\hat{u}_i-\tilde{u}_i\|^2/n,\ T_2 = \sum_{i=1}^n\|\tilde{u}_i-u_i\|^2/n.$$

With probability approaching one $\hat{\theta}$ is an element of any convex neighborhood of θ_0, so that by n mean value expansions, the Markov inequality, and consistency of $\hat{\theta}$,

$$T_1 = \sum_{i=1}^n \|\hat{u}_i - \tilde{u}_i\|^2/n = \sum_{i=1}^n \|\sum_{j=1}^n \{\partial[b(z_i,z_j,\bar{\theta}_i)+b(z_j,z_i,\bar{\theta}_i)]/\partial\theta/n\}(\hat{\theta}-\theta_0)\|^2/n$$

$$\le 4[\sum_{i=1}^n\sum_{j=1}^n B(z_i,z_j)^2/n^2]\|\hat{\theta}-\theta_0\|^2 = O_p(1)o_p(1) = o_p(1),$$

where $\bar{\theta}_i$ denote the mea' values, which each lie between $\hat{\theta}$ and θ_0. Also note that by the i.i.d. assumption and $E[\|a_{ij}\|^2] \le 2E[B(z_i,z_j)] < \infty$,

$$E[T_2] = E[E[\|\tilde{u}_1 - u_1\|^2|z_1]]$$

$$\le E[\ 4E[\|\sum_{j=2}^n(a_{1j}-E[a_{1j}|z_1])/(n-1)\|^2|z_1] + 4E[\|a_{11}\|^2|z_1]/(n-1)$$

$$+ 2|1-[n/(n-1)]|\sum_{j=1}^n E[\|a_{1j}\|^2|z_1]/n\ ]$$

$$\le 6E[\|a_{12}\|^2 + \|a_{11}\|^2]/(n-1) = o(1),$$

so that $T_2 = o_p(1)$ follows by the Markov inequality. Thus, it follows by eq. (A.2) and the triangle inequality that $\sum_{i=1}^n\|\hat{u}_i - u_i\|^2/n = o_p(1)$. ∎

Proof of Corollary 4.2: Note that by independence and differentiability of $\tilde{s}(\varepsilon,\eta)$,

$$M_{12} = E[\partial b_1(z_1,z_2,\theta_0)/\partial\eta] = E[\{\rho_\beta(z_1)-\rho_\beta(\pi(\varepsilon_1,x_2),x_2)\}'\partial\tilde{s}(\varepsilon_1,\eta_0)/\partial\eta] \quad \text{(A.3)}$$

$$= E[\{\rho_\beta - E[\rho_\beta|\varepsilon]\}'\partial\tilde{s}(\varepsilon,\eta_0)/\partial\eta] = 0.$$

The conclusion then follows by a partitioned inverse argument like that of Ruppert and Aldershof (1989). ∎

Proof of Theorem 4.3: Let $b(z_1,z_2,\beta) = b(z_1,z_2,\beta,\eta_0)$ and note that by Assumption 4.1, $b(z_1,z_2,\beta)$ is continuously differentiable in β and $\|\partial b(z_1,z_2,\beta)/\partial\beta\|^2$, $\|\partial b(z_1,z_2,\beta)/\partial\beta\| \le B(z_1,z_2) + 1$. By Lemma A.3 and Assumption 3.1, $f(z_1|\beta)f(z_2|\beta)$ is smooth. Differentiating the identity $E_\beta[b(z_1,z_2,\beta)] = 0$, it follows by Lemma C.2 of Newey (1990b) that

$$M_{11} = -E[b(z_1,z_2,\beta_0)\{S_\beta(z_1)+S_\beta(z_2)\}'] = -E[E[b(z_1,z_2,\beta_0)|z_1]S_\beta(z_1)']$$

$$- E[E[b(z_1,z_2,\beta_0)|z_2]S_\beta(z_2)] = -E[uS_\beta'].$$

Also, note that for $\tilde{s}(\varepsilon,\eta_0) = s(\varepsilon)$, $\tilde{S}(z,\beta_0) = S_\beta$. Furthermore, by the definition of $b(z_1,z_2,\beta)$, $E[b(z_1,z_2,\beta_0)|z_1] = S_\beta - E[S_\beta|\varepsilon]$ and $E[b(z_2,z_1,\beta_0)|z_1] = E[S_\beta]-E[S_\beta|x]$, so that $u = S_\beta - E[S_\beta|\varepsilon]-E[S_\beta|x]+E[S_\beta] = S$. The conclusion then follows by Corollary 4.2, which gives equality of the asymptotic variance of $\hat{\beta}$ and $M_{11}^{-1}\Omega_{11}M_{11}^{-1} = (E[SS'])^{-1}$. ∎

Proof of Theorem 5.1: Standard, and so omitted.

Proof of Corollary 5.2: By $\tilde{s}(\varepsilon,x,\eta)$ an odd function of ε given x and η, $\partial\tilde{s}(\varepsilon,x,\eta)/\partial\eta$ is also an odd function. Then by $\rho_\beta + \rho_\beta(\pi(-\varepsilon,x),x)$ an even function of ε, $\partial b_1(z,\theta_0)/\partial\eta = [\rho_\beta + \rho_\beta(\pi(-\varepsilon,x),x)]'\partial\tilde{s}(\varepsilon,x,\eta)/\partial\eta$ is an odd function. Therefore, by symmetry, $M_{12} = E[\partial b_1(z,\theta_0)/\partial\eta] = 0$, so the conclusion follows as in the proof of Corollary (4.2). ∎

Proof of Theorem 5.3: Let $b(z,\beta) = b(z,\beta,\eta_0)$ and note that by Assumption 4.1, $b(z,\beta)$ is continuously differentiable in β and $\|b(z,\beta)\|^2$, $\|\partial b(z,\beta)/\partial\beta\| \le B(z)+1$. By regularity of $f(z|\beta)$, Lemma C.2 of Newey (1989b) can be applied to the identity $E_\beta[b(z,\beta)] = 0$ to obtain $M_{11} = -E[b(z,\beta_0)S_\beta']$. Furthermore, by $b(z,\beta_0)$ and $t \in \mathcal{T}$ odd and even functions of ε given x, respectively, $E[b(z,\beta_0)t'] = 0$. Therefore, $M_{11} = -E[b(z,\beta_0)S']$. Furthermore, note that $b(z,\beta_0) = S$ if $\tilde{s}(\varepsilon,x,\eta_0) = s(\varepsilon,x)$, so that the conclusion follows as in the proof of Theorem 4.3. ∎

References

Amemiya, T. (1974): "The Non-linear Two-stage Least-Squares Estimator," *Journal of Econometrics*, 2, 105-110.

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

Amemiya, T. (1985): *Advanced Econometrics*, Cambridge: Harvard University Press.

Amemiya, T. and J.L. Powell (1981): "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Non-Linear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 17, 351-381.

Begun, J., W. Hall, W. Huang, and J. Wellner (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432-452.

Bickel, P. (1982): "On Adaptive Estimation," *Annals of Statistics*, 10, 647-671.

Bickel P., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1988): "Efficient and Adaptive Inference in Semiparametric Models" Forthcoming monograph, Johns Hopkins University Press.

Box, G.,E.P., and Cox, D.R. (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society, Ser. B,* 26, 211-246.

Brown, B.W. and R.S. Mariano (1984): "Residual-Based Procedures for Prediction and Estimation in a Nonlinear Simultaneous Equations Model," *Econometrica*, 52, 321-343.

Burbidge, J.B., L. Magee and A.L. Robb (1988): "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *Journal of the American Statistical Association*, 76, 296-311.

Carroll, R.J. and D. Ruppert (1984): "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Association,*" 79, 321-328.

Carroll, R.J. and D. Ruppert (1982): "Robust Estimation of Heteroskedastic Linear Models," *Annals of Statistics*, 10, 429-441.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 32, 189-218.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305-334.

Duan, N. (1983): "Smearing Estimate: a Nonparametric Retransformation Method," *Journal of the American Statistical Association*, 78, 605-610.

Hajek, J. (1970): "A Characterization of Limiting Distributions of Regular Estimates," Z. *Wahrscheinlichkeitstheorie verw. Geb.*, 14, 323-330.

Hinkley, D.V. (1975): "On Power Transformations to Symmetry," *Biometrika*, 62, 101-111.

Huber, P., (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.

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

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

Kelejian, H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in Parameters and Nonlinear in Endogenous Variables," *Journal of the American Statistical Association*, 66, 373-374.

Koenker, R. and G. Bassett Jr. (1982): "Robust Tests for Heteroskedasticity Based on Regression Quantiles," *Econometrica*, 50: 46-31.

Koshevnik and Levit (1976): "On a Non-parametric Analouge of the Information Matrix," *Theory of Probability and Applications*, 21, 738-753.

MaCurdy, T.E. (1982): "Using Information on the Moments of the Disturbance to Increase the Efficiency of Estimation," mimeo, Stanford University.

Newey, W.K. (1987): "Efficient Estimation of Models with Conditional Moment Restrictions," preprint, Department of Economics, Princeton University.

Newey, W.K. (1989a): "Efficient Estimation of Semiparametric Models Via Moment Restrictions," preprint, Department of Economics, Princeton University.

Newey, W.K. (1989b): "Uniform Convergence in Probability and Stochastic Equicontinuity," Econometric Research Program Memo No. 342, Princeton University.

Newey, W.K. (1990a): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, forthcoming.

Newey, W.K. (1990b): "Efficient Estimation of Tobit Models Under Conditional Symmetry," in W. Barnett, J. Powell, G. Tauchen eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, New York: Cambridge University Press, forthcoming.

Pfanzagl, J. (1982): *Contributions to a General Asymptotic Statistical Theory*, New York: Springer-Verlag.

Pitman, E.J.G. (1979): *Some Basic Theory for Statistical Inference*, London: Chapman and Hall.

Phillips, P.C.B. (1982): "On the Consistency of Nonlinear FIML," *Econometrica*, 50: 1307-1324.

Powell, J.L. (1990): "Estimation of Monotonic Regression Models Under Conditional Quantile Restrictions," in W. Barnett, J. Powell, G. Tauchen eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, New York: Cambridge University Press, forthcoming.

Robinson, P.M. (1989): "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," preprint, London School of Economics.

Ruppert, D. and Aldershof, B. (1989), "Transformations to Symmetry and Homoscedasticity," *Journal of the American Statistical Association*, 84, 437-446.

Sargan, J.D. (1959): "The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables," *Journal of the Royal Statistical Society, Series B*, 21 91-105,

Serfling, R.J. (1980): "Approximation Theorems of Mathematical Statistics," New York: Wiley.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press.

Taylor, J.M.G. (1985): "Power Transformations to Symmetry," *Biometrika*, 72, 145-152.

Table One
Engel Data Results

	λ	Elasticity	σ_{ε}^2
Log Regression	0	.847821 (.026499)	.018803
QMLE	.091999 (.090887)	.853324 (.022995)	.060726
IV	.231016 (.122779)	.859298 (.022390)	.363120
V-Stat	-.005792 (.096701)	.847477 (.022865)	.017474
Antithetic	.249620 (.116062)	.862951 (.023395)	.468415

Numbers in parentheses are estimated standard errors.

Table Two

Estimator Performance Under Independence, with 75 observations.

$\lambda = .23, \quad \beta_2 = (-10.855, 3.73109)', \quad \sigma^2_\varepsilon = .599269.$

	BIAS	STE	RMS	MED-BIAS	MAE
Gaussian ε.					
λ					
QMLE	.0116	.1407	.1412	.0105	.0883
IV	.0081	.1860	.1862	-.0022	.1168
V-Stat	-.0023	.1439	.1439	-.0021	.1025
Antithetic	-.0076	.1619	.1621	-.0083	.1013
Homosked	.2356	.1602	.2849	.2244	.2244
Elasticity					
QMLE	-.0018	.0441	.0442	-.0002	.0279
IV	-.0036	.0453	.0455	-.0021	.0303
V-Stat	.0011	.0441	.0441	-.0027	.0306
Antithetic	.0004	.0437	.0437	-.0037	.0330
Homosked	.0302	.0492	.0577	.0230	.0363
Empirical ε.					
λ					
QMLE	.0217	.1410	.1426	.0196	.0871
IV	.0023	.1785	.1786	-.0178	.1105
V-Stat	-.0001	.1451	.1451	-.0002	.0864
Antithetic	.0420	.1811	.1859	.0340	.1053
Homosked	.2414	.1831	.3030	.2209	.2209
Elasticity					
QMLE	.0015	.0463	.0463	-.0003	.0305
IV	-.0004	.0471	.0471	-.0021	.0292
V-Stat	-.0003	.0468	.0468	-.0026	.0315
Antithetic	.0024	.0463	.0463	-.0012	.0273
Homosked	.0258	.0513	.0574	.0198	.0362

Table Three

Estimator Performance Under Symmetry, with 224 observations.

$\lambda = .23, \quad \beta_2 = (-10.855, 3.73109)', \quad \sigma_\varepsilon^2 = .599269.$

	BIAS	STE	RMS	MED-BIAS	MAE
Gaussian ε					
λ					
QMLE	.0683	.0733	.1002	.0707	.0740
IV	-.0079	.0966	.0969	-.0066	.0629
Antithetic	-.0055	.0905	.0906	-.0098	.0637
Sym	.0286	.3184	.3196	.0470	.2257
Elasticity					
QMLE	.0086	.0233	.0249	.0098	.0163
IV	.0001	.0238	.0238	-.0010	.0165
Antithetic	.0003	.0237	.0237	.0002	.0155
Sym	.0057	.0439	.0443	.0040	.0302
Empirical Antithetic ε					
λ					
QMLE	-.2057	.0875	.2235	-.2119	.2119
IV	.0158	.1399	.1407	.0067	.0835
Antithetic	.0126	.1249	.1256	.0109	.0788
Sym	.0169	.2191	.2197	.0113	.1508
Elasticity					
QMLE	-.0272	.0244	.0365	-.0257	.0267
IV	-.0018	.0216	.0216	-.0012	.0141
Antithetic	-.0019	.0220	.0220	-.0011	.0149
Sym	.0007	.0385	.0385	.0022	.0244