Title
Probit with Dependent Obervations

Authors
Poirier, Dale J.
Ruud, Paul A.

Publication Date
1987-03-06

Peer reviewed

UNIVERSITY OF CALIFORNIA, BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper 8734

PROBIT WITH DEPENDENT OBSERVATIONS

Dale J. Poirier and Paul A. Ruud

March 6, 1987

Key words: ARMA, limited dependent variables, probit, generalized method of moments, autocorrelation.

Abstract

Estimation of limited dependent variable models, and binary probit in particular, is examined. Asymptotic distribution theory is provided for the consistency and asymptotic normality of quasi-maximum likelihood estimators. A family of relatively efficient estimators, called generalized conditional moment estimators, is proposed. Probit with a first-order autoregressive error structure is given as an example.

JEL Classification: 210,211

PROBIT WITH DEPENDENT OBSERVATIONS

Dale J. Poirier and Paul A. Ruud[1]

1. Introduction

Estimation of limited dependent variable (LDV) models based on dependent observations has received relatively little attention due to the computational complexity of obtaining the maximum likelihood estimator (MLE). This paper considers an alternative method of estimation which is computationally attractive and relatively efficient. For the sake of brevity, this paper considers only the probit model, but the approach developed here generalizes to other LDV models (e.g., the tobit model).

In many situations ignoring dependency among observations and simply computing the pseudo MLE, conditional on the false assumption of independence, yields a consistent, asymptotically normal but inefficient, estimator. A familiar case in point is the ordinary least squares (OLS) estimator in the context of the general linear model. A general treatment of this problem is given by Levine (1983). Levine's treatment, however, assumes the observed dependent variable is continuous, thus ruling out LDV models. It is widely recognized that even for the case of independent sampling, standard theorems concerning MLE's cannot be simply invoked in the case of LDV models. Levine's intuitively sensible results have been generalized, for example, by White (1984b, Cor. 2.9). We present arguments for Probit models.

In a most important contribution, Robinson (1982) showed under suitable regularity conditions, that the ordinary tobit (OT) estimator predicated on

---

[1] The authors are Professor of Economics at the University of Toronto and Assistant Professor of Economics at the University of California, respectively. The authors gratefully acknowledge the helpful comments of Angelo Melino and two anonymous referees.

independent observations remains consistent and asymptotically normal when the disturbances are in fact serially dependent. While reminiscent of the familiar, uncensored regression case, the proofs of Robinson's results are nontrivial. In an earlier related paper, Robinson (1980) showed how to consistently estimate the autocorrelations of a censored Gaussian process and how to use such estimates to test for serial dependence.

Robinson's results provide the required asymptotic theory for use of the OT estimator in the presence of serially correlated disturbances. Although it is most likely inefficient compared to the true MLE, it has decided computational advantages. Dagenais (1982) considered computation of the MLE for the tobit model with first-order autoregressive disturbances and showed that it is intractable unless the numbers of consecutive limit-point observations are small, since these numbers equal the dimensions of the integrals involved in the likelihood function.[2] Furthermore, the asymptotic theory for the MLE in the presence of serially corelated disturbances has not been rigorously developed.

Gourieroux, Monfort, and Trognon (1982) consider the probit model with autoregressive-moving-average (ARMA) disturbances and show explicitly how Robinson's results of consistency and asymptotic normality extend to the ordinary probit (OP) estimator and to a related nonlinear least squares estimator. The authors also derive the autocorrelation score (Lagrange multiplier) test statistics assuming a first-order autoregressive disturbance term and show that it is identical to that for a first-order moving-average disturbance. The score test for a first-order autoregressive disturbance in a

---

[2] In the related literature on markets in disequilibrium with no sample partitioning information, Quandt (1981, p. 59) notes that in his empirical analysis computational costs were about 125 times costlier for models involving serial correlation than in those that did not.

probit model was also derived by Poirier and Ruud (1980). Jarque and Bera (1981) derived the score test for autoregressive or moving-average disturbances of arbitrary order in censored (tobit) and truncated regression models, with lagged dependent variables possibly present. Kiefer (1982) derived the score test for zero-covariance restrictions in probit models based on panel data. More recently, Robinson, Bera, and Jarque (1985) also derive score tests for the tobit model against autoregressive disturbances. The motivation in all of these cases for the score test is that it requires estimation only under the null hypothesis of independence, and hence it is computationally tractable.

The task of developing a computationally simple estimator that improves upon the OP estimator in the presence of serial correltation has been undertaken by Avery, Hansen, and Hotz (1983), Poirier and Ruud (1980, 1981a, 1981b), and Ruud (1981). Avery *et al.* (1983) consider the case of panel data in which there is temporal but not cross-sectional dependency in the observations.[3] The authors propose "orthogonality condition estimators" that are members of the class of generalized method-of-moment (GMM) estimators developed by Hansen (1982). The efficiency gains of such estimators are, however, based on asymptotics in which the number of cross-sectional units grows. Here, as in our previous unpublished work, we consider only a single time series, unlike the case studied by Avery *et al.* (1983). Our primary goal is to develop a computationally simple estimator that improves upon the OP estimator in the presence of serial correlation. All asymptotic distribution theory considered here involves the length of a single time series growing.

---

[3] For other discussions of probit analysis based on panel data, see Heckman (1981a, 1981b) and Ochi and Prentice (1984).

As will become clear in the subsequent sections, the problem studied here requires different techniques than that employed by Avery *et al.* (1983).

The basic outline of this paper is as follows. Section 2 explores ML and quasi-ML estimation in probit models derived from latent variable regressions involving correlated disturbances. The development emphasizes the nature of the orthogonality conditions that generate such estimators. Section 3 builds upon the motivations outlined in Section 2 and develops alternative estimators in the case of a known disturbance covariance matrix. For reasons that will become clear later, such alternative estimators are referred to as generalized conditional moment (GCM) estimators. An ordering of GCM estimators according to their asymptotic relative efficiency is developed in Section 4. Section 5 extends the discussion to cover the case of an unknown covariance matrix based on first-order autoregressive disturbances. To expedite reading of this inherently complicated material, all proofs are confined to appendices, except where we consider the proofs to be insightful. Section 6 gives an illustrative example of a GCM estimator for a first-order autoregressive process. Finally, Section 7 provides a few concluding remarks.

2. MLE and Quasi-MLE

Consider the linear regression model

$$y_t^* = x_t'\beta_0 + u_t \qquad (t=1,\ldots,T) \tag{2.1}$$

where y_t^* is a latent dependent variable, x_t is a K-dimensional column vector of fixed explanatory variables, β_0 is a K-dimensional column vector of unknown parameters, and u_t is a stochastic disturbance. In the probit model only the sign of y_t^*, as indicated by

$$(2.2) \qquad y_t = \left\{ \begin{array}{ll} 0 , & \text{if } y_t^* \leq 0 \\ 1 , & \text{if } y_t^* > 0 \end{array} \right\} \qquad (t=1,\ldots,T)$$

is observed. Letting the notation $[\cdot]$ denote a stacking of the argument, (2.1) can be written in matrix notation as

$$(2.3) \qquad y^* = X\beta_0 + u$$

where $y^* \equiv [y_t^*]$ and $u \equiv [u_t]$ are $T \times 1$ vectors and $X \equiv [x_t']$ is a $T \times K$ matrix.

In connection with (2.1)-(2.3) the following assumptions are made.

Assumption 1: The disturbances u_t $(t=1,2,\ldots)$ are a stationary process with zero means and covariance matrix $\Gamma_0 \equiv \Gamma(\gamma_0) \equiv E(uu')$, where Γ_0 denotes a positive definite matrix of known functional form three times differentiable in the finite-dimensional vector γ_0. The variances of u_t $(t=1,2,\ldots)$ are all normalized to unity. Also suppose:

a) u_t $(t=1,2,\ldots)$ is α-mixing with mixing coefficient $\alpha(m)$. [See Definitions A.1 and A.2 of Appendix 1.]

b) $\alpha(m) = O(m^{-c})$ where $c > 1$; that is, $m^c\alpha(m)$ is bounded for all $m = 0,1,2,\ldots$.

c) u has a nonsingular multivariate normal distribution for any finite T.

Assumption 2:

a) The parameter space $\Theta \equiv B \times G$ for $\theta \equiv [\beta',\gamma']'$ is compact.

b) β_0 lies in the interior of B.

c) γ_0 lies in the interior of G.

Assumption 3:

a) x_t $(t=1,2,\ldots)$ is non stochastic and lies in the compact space $\mathfrak{X}$.

b) The empirical distribution of x_t $(t=1,2,\ldots,T)$ converges to a limit distribution denoted by H .

c) The empirical distribution of x_t $(t=a+1,a+2,\ldots,a+T)$ converges to a limit distribution denoted by H , uniformly in a .

d) For any m , the empirical distribution of (x_t,x_{t+m}) $(t=a+1,a+2,\ldots,a+T)$ converges to a limit distribution denoted by H_m , uniformly in a .

Assumption 4: If $x'\beta = x'\beta_0$ almost surely (according to distribution function H) then $\beta = \beta_0$.

Assumptions 2-4 are essentially the same as those used by Gourieroux *et al*. (1982). Our Assumption 1 is slightly weaker than their corresponding assumption that u_t follow a stationary, invertible Gaussian ARMA process since the latter implies Assumption 1, but not conversely. In fact, for purposes of Theorems 1 and 2 below, Assumption 1(c) can be weakened to only assuming that the marginal distribution $F(\cdot)$ is known and twice differentiable. Then replacing the normal distribution function $\Phi(\cdot)$ in Theorems 1 and 2 by $F(\cdot)$, these theorems can be extended to cover quasi-ML estimation in such cases as binary logit with dependent data. We do not explicitly consider such extensions in this paper because the class of estimators we propose in Sections 3-5 rely explicitly on the normality assumption and important properties of the multivariate normal distribution. These suggested extensions of Theorems 1 and 2, however, are interesting since they imply that mixing properties alone are sufficient to investigate the asymptotic properties of quasi-ML estimation applied to binary time series for joint distributions that do not have the convenient correlation structure of

the multivariate normal distribution. In this sense, such extensions are very much in the spirit of the analysis of Levine (1983).

Our Theorems 1 and 2 below for the OP estimator (quasi-MLE) are more or less similar in scope to Theorems 1 and 2 of Gourieroux *et al.* (1982), except for our slightly weaker Assumption 1, but our approach to proof is different than the approach taken by Gourieroux *et al.* (1982). This difference again reflects the convenience of working directly with the mixing properties of the disturbances. We use the mixing properties of the disturbance u_t to bound directly the time dependency in the observed data, whereas Gourieroux *et al.* (1982) employ a more cumbersome approach (in our opinion) requiring bounding correlations of the disturbances. The former approach turns out to be quite natural in binary problems since the mixing coefficient $\alpha(m)$ is defined in terms of the covariances of observable binary data [see Definition A.1 in Appendix 1]. The approach of Gourieroux *et al.* (1982) follows closely the approach of Robinson (1982) which appears to be more natural for tobit models than for probit models.

Dispensing with issues of approach, we now consider the quasi-MLE corresponding to maximizing the log-likelihood function predicated on independence of observations. This estimator corresponds to the ordinary probit (OP) estimator $\hat{\beta}^{OP}$ [under Assumption 1(c)] obtained by maximizing

$$(2.4) \qquad \mathcal{L}^{OP}(\beta;y) \equiv \sum_{t=1}^{T} (1-y_t)\cdot ln[\Phi(-x_t'\beta)] + y_t \cdot ln[\Phi(x_t'\beta)]$$

where $y \equiv [y_t]$. The asymptotic properties of $\hat{\beta}^{OP}$ are given in the following two theorems, the proofs of which are contained in Appendices 1 and 2.

Theorem 1: Under Assumptions 1(a,b,c), 2(a), 3(a,b), and 4, $\hat{\beta}^{OP}$ is a strongly consistent estimator of β_0 as $T \longrightarrow \infty$.

Theorem 2: Under Assumptions 1(a,b,c), 2(a,b), 3(a,c,d), and 4, $T^{\frac{1}{2}}(\hat{\beta}^{OP} - \beta_0) \xrightarrow{D} N(0,\Omega^{OP})$ as $T \longrightarrow \infty$, where

$$\Omega^{OP} \equiv \Omega^{OP}(\beta_0,\gamma_0) = \Delta_0^{-1}\left[\Delta_0 + 2\sum_{m=1}^{\infty}\Delta_m\right]\Delta_0^{-1}, \tag{4.5}$$

$$\Delta_m \equiv \Delta_m(\beta_0,\gamma_0)$$

$$= E_x\left[(x\ x_{+m}')\left[\frac{\phi(x'\beta_0)\phi(x_{+m}'\beta_0)\{\Phi_2(x'\beta_0,x_{+m}'\beta_0;\rho_m) - \Phi(x'\beta_0)\Phi(x_{+m}'\beta_0)\}}{\Phi(x'\beta_0)\Phi(-x'\beta_0)\Phi(x_{+m}'\beta_0)\Phi(-x_{+m}'\beta_0)}\right]\right], \tag{2.6}$$

$(m=0,1,\ldots)$ where E_x denotes the expectation assocatiated with the limiting distribution H_m of x and $x_{+m} \equiv [x_{t+m}]$, $H_0 \equiv H$, $\rho_m = \rho_m(\gamma_0) \equiv \mathrm{Cov}(u_t,u_{t+m})$, $\phi(\cdot)$ and $\phi_2(\cdot,\cdot;\rho)$ denote the univariate and bivariate standard normal probability density functions, respectively, and

$$\Phi_2(h,k;\rho) \equiv \int_{-\infty}^{h}\int_{-\infty}^{k}\phi_2(s,t;\rho)\,dt\,ds \tag{2.7}$$

denotes the bivariate standard normal distribution function.

Theorems 1 and 2 imply that the ordinary probit estimator $\hat{\beta}^{OP}$ is still consistent and asymptotically normal when the observations are dependent, but the standard asymptotic covariance matrix Δ_0^{-1} in (2.5) must be replaced by the computationally more demanding Ω^{OP} in (2.5). Not surprisingly, Ω^{OP}

depends on the covariance parameters γ_0 through Δ_j in (2.6). If γ_0 is unknown, then a consistent estimator can be obtained and used in (2.5) to obtain a consistent estimator of Ω^{OP}. Detailed discussion of estimation of γ_0 is postponed, however, until Section 5.

The importance of Theorems 1 and 2 can be appreciated by considering the complexity involved in obtaining the MLE. Let $\phi_T(\cdot;\mu,\Sigma)$ denote the probability density function of a T-dimensional multivariate normal random variable with mean μ and covariance matrix Σ. Then the likelihood for the dependent probit model is given by the T-dimensional integral

$$L(\beta,\gamma_0;y) = \int_{a_1}^{b_1} \cdots \int_{a_T}^{b_T} \phi_T(y^*;X\beta,\Gamma_0)\, dy_T^* \cdots dy_1^* , \tag{2.8}$$

where

$$(a_t,b_t) = \left\{ \begin{array}{ll} (-\infty,0) , & \text{if } y_t = 0 \\ (0,\infty) , & \text{if } y_t = 1 \end{array} \right\} \qquad (t=1,\ldots,T). \tag{2.9}$$

For even moderate values of T, (2.8) is unmanageable for commonly encountered structures of Γ_0. Furthermore, the asymptotic properties of the MLE based on (2.8) have not been determined.

The next section develops a family of estimators that includes both $\hat{\beta}^{OP}$ and $\hat{\beta}^{ML}$ as special cases. In order to provide the reader with intuition about this family, here we consider the first-order conditions which give rise to $\hat{\beta}^{OP}$ and $\hat{\beta}^{ML}$. The first-order conditions for (2.4) are

$$(2.10) \qquad \left.\frac{\partial \mathcal{L}^{OP}(\beta;y)}{\partial \beta}\right|_{\beta=\hat{\beta}^{OP}} = X'\hat{u}^{OP} = 0 ,$$

where $\hat{u}^{OP} \equiv [\hat{u}_t^{OP}]$ and

$$(2.11) \qquad \hat{u}_t^{OP} \equiv E(u_t|y_t;\hat{\beta}^{OP}) = \frac{[y_t - \Phi(x_t'\hat{\beta}^{OP})]\cdot\phi(x_t'\hat{\beta}^{OP})}{[1 - \Phi(x_t'\hat{\beta}^{OP})]\cdot\Phi(x_t'\hat{\beta}^{OP})} \quad (t=1,\ldots,T).$$

The first-order conditions corresponding to the logarithm of (2.8) are

$$(2.12) \qquad \left.\frac{\partial \ln L(\beta,\tau_0;y)}{\partial \beta}\right|_{\beta=\hat{\beta}^{ML}} = X'\Gamma_0^{-1}\hat{u}^{ML} = 0 ,$$

where $\hat{u}^{ML} \equiv [\hat{u}_t^{ML}]$ and

$$(2.13) \qquad \hat{u}_t^{ML} \equiv \hat{u}_t^{ML}(y;\hat{\beta}^{ML},\tau_0) = E(u_t|y;\hat{\beta}^{ML},\tau_0) .$$

As written, first-order conditions (2.10) and (2.12) comprise "orthogonality conditions" involving the regressors X and residuals (2.11) and (2.13), respectively. Note that one difference between (2.11) and (2.13) is that the latter expecatation is conditional on the <u>entire sample</u> of y_t's : $y \equiv [y_t]$. The orthogonality conditions are reminescent of the first-order conditions defing the OLS and generalized least squares (GLS) estimators:

$$(2.14) \qquad X'\hat{u}^{OLS} = 0 ,$$

$$(2.15) \qquad X'\Gamma_0^{-1}\hat{u}^{GLS} = 0 ,$$

where $\hat{u}^{OLS} \equiv y^* - X\hat{\beta}^{OLS}$, $\hat{\beta}^{OLS} = (X'X)^{-1}X'y^*$, $\hat{u}^{GLS} \equiv y^* - X\hat{\beta}^{GLS}$, and $\hat{\beta}^{GLS} \equiv (X'\Gamma_0^{-1}X)^{-1}X'\Gamma_0^{-1}y^*$.

The lesson to be learned from these analogies is the following. Since y^* is not observed, (2.14) and (2.15) are non-operational. The operational counterparts (2.10) and (2.12) replace the residuals $\hat{u}^{OLS}$ and $\hat{u}^{GLS}$ with their expectations based on (2.11) and (2.13), respectively. This, of course, is the interpretation that the EM algorithm gives to ML estimation in many LDV models with normal disturbances.[4]

Since the residual (2.11) is conditioned only on y_t, $\hat{\beta}^{OP}$ is expected to be less efficient than $\hat{\beta}^{ML}$. While (2.13) takes into account more information than (2.11), the former also involves a T-dimensional integral, whereas the latter involves only a univariate integral. This simple observation suggests a possible trade-off between efficiency and computational complexity, and specifically, use of residuals depending on more than one observation, but not all observations. Investigating such a conjecture is the goal of the remainder of this paper.

3. Estimation with Known Covariance Matrix

Building on the motivation of the last section, we now propose estimators that exploit orthogonality conditions with residuals similar to (2.10) and (2.12). Building on the work of Hansen [1982], we propose an optimal choice of orthogonality conditions formed from linear combinations of

---

4 Dempster, Laird, and Rubin (1977) describe the properties of the EM algorithm. Hartley (1974) proposed the algorithm for ordinary probit and Fair (1977) proposed its use for the ordinary tobit model. Also, Kiefer (1980) applied the EM algorithm to estimation of the switching regression model.

the residuals. Throughout this section, the covariance parameter vector γ_0 is assumed known; the case of unknown γ_0 will be treated in Section 5.

First, we wish to enlarge the family of residuals that will be considered. Ordinary Probit and ML Probit residuals are polar examples of a larger family. While each OP residual is the conditional expectation of a latent disturbance u_t given only the corresponding realization of y_t, each ML residual is the conditional expectation of a latent disturbance given the entire sample y. Intermediate kinds of Probit residuals might condition on subsets of the sample of observed y_t that contain more than one observation, but fewer than the whole sample.

In order to keep track of such subsets of the sample, we introduce some new notation. Let n be an integer, $1 \leq n \leq T$, such that $T \leq \binom{T}{n}$ and let $\mathfrak{s}_{ni}$ $(i=1,\ldots,T)$ denote T subsets of size n from the set $\mathfrak{s} \equiv \{1,\ldots,T\}$ of observation subscripts. For the moment, we postpone the choice of the specific $\mathfrak{s}_{ni}$ from among the $\binom{T}{n}$ possibilities. Each subset $\mathfrak{s}_{ni}$ is an indexing set to observations and will be used to index vectors and matrices in the following way:

$$a_{ni} \equiv [a_t; t \in \mathfrak{s}_{ni}] \quad \text{for any} \quad \{a_1, \ldots, a_T\} \; ; \tag{3.1}$$

Thus a_{ni} is an extract of observations from a sample put into a vector (if the a_t are scalars) or a matrix (if the a_t are vectors). All double subscripts have the meaning endowed by (3.1). An important example of (3.1) is y_{ni} which is a vector of n observations on y_t that will serve as a conditioning set intermediate to a single observation (OP) and the sample (ML).

We will refer to the size of $\mathfrak{s}_{ni}$, n, as the conditioning level.

The residuals that arise from admitting such conditioning sets are then defined as follows:

$$\lambda_{ni} \equiv \lambda_{ni}(\beta) \equiv E[u_{ni}|y_{ni};\beta,\gamma_0] \quad . \tag{3.2}$$

The analytical form of (3.2) is given in Appendix 3.[5] The case of OP corresponds to $n=1$: $\binom{T}{n} = T$, and $\mathfrak{s}_{ni} = \{i\}$ so that $\lambda_{ni} = \lambda_i \equiv E[u_i|y_i]$. The case of ML would correspond to $n = T$ except that it has been ruled out from present consideration by the condition $T \leq \binom{T}{n}$ since $\binom{T}{T} = 1$.

The primary goal is to obtain estimators of β , treating γ_0 as known, by focusing on orthogonality conditions analogous to (2.10) and (2.12) involving (3.2) as residuals. Specifically, we consider orthogonality conditions of the form

$$X'A_n\hat{u}_n = 0 \tag{3.3}$$

where $\hat{u}_n \equiv \hat{u}_n(\beta) = [\lambda_{ni}(\beta)]$ and A_n is a $T \times nT$ matrix. Again, comparisons with the previous estimators are straightforward. OP ($n=1$) sets A_n to the identity matrix and ML sets $A_n = \Gamma_0^{-1}$.

We are guided in our search for an optimal A_n by Hansen's (1982) generalized method of moments estimators. Roughly speaking, he has shown that for <u>finite dimensional</u> $\hat{u}_n$ (and certain regularity conditions), such

---

5 This expectation is derived by Tallis (1961). Note that in (3.2) the elements u and y have identical subscripts. Thus, all the u_t for which $t \notin \mathfrak{s}_{ni}$ are excluded. This is because $E[u|y_{ni}] = E\{E[u|u_{ni}]|y_{ni}\}$ $= E\{E(u\,u_{ni}')\ Var(u_{ni})^{-1}u_{ni}|y_{ni}\} = E(u\,u_{ni}')\ Var(u_{ni})^{-1}\lambda_{ni}$, so that the excluded residuals are linearly dependent on the included ones.

estimators as defined by (3.3) are consistent and asymptotically normal with mean β_0 and covariance matrix

$$(3.4) \qquad plim\ T^{-1}\left[X'A_n \frac{\partial \hat{u}_n}{\partial \beta'}\right]^{-1}\left[X'A_n Var(\hat{u}_n)\ A_n'X\right]\left[\frac{\partial \hat{u}_n'}{\partial \beta} A_n'X\right]^{-1}$$

and that as a result, an optimal A_n is given by

$$(3.5) \qquad X'A_n = E\left[\frac{\partial \hat{u}_n'}{\partial \beta}\right] Var(\hat{u}_n)^{-1} \quad or \quad A_n = E\left[\frac{\partial \hat{u}_n'}{\partial X\beta}\right] Var(\hat{u}_n)^{-1}\ .$$

In this case, (3.4) reduces to

$$(3.6) \qquad plim\ T^{-1}\left[X'E\left[\frac{\partial \hat{u}_n'}{\partial X\beta}\right] Var(\hat{u}_n)^{-1}\ E\left[\frac{\partial \hat{u}_n'}{\partial X\beta}\right]' X\right]\ .$$

Although $\hat{u}_n$ actually contains nT elements, so that the number of orthogonality conditions is growing with sample size and Hansen's results do not apply, we make (3.5) our choice for A_n. Given the central role of conditional moments and the importance of the weighting matrix A_n in our family of estimators, we will call them <u>generalized conditional moments</u> (GCM) estimators. In describing the asymptotic properties of our GCM estimators, we will establish results analogous to those of Hansen on consistency, asymptotic normality, and optimality. The analytical form of the expresions in (3.5) and (3.6) are also in Appendix A.3.

3.1 One-Step Estimation

In practice, one will rarely solve the problem in (3.3) by direct, and naive, numerical methods. Given both its consistency and its easy computability, the OP estimator will always serve as a starting value for any algorithm to compute the actual GCM estimator for any conditioning level. In addition, the computation for conditioning levels greater than 2 or 3 is great because of the repeated evaluation of multivariate normal integrals. Because of this, iterative minimization algorithms will be prohibitively expensive in many applications. The popular alternative in such circumstances is a linearized one-step estimator based on an initial consistent estimator [cf. Rothenberg and Leenders (1964)]. This estimation method will be applied to our problem (3.3) in Section 3.3. In this section, we review the consistency and asymptotic normality of the linearized one-step estimator for general cases.

From the perspective of proving estimator consistency, this approach has theoretical appeal as well. Without the aid of an initial consistent estimator to identify the neighborhood of the true parameter value, the possibility of several local solutions to the GCM estimation problem makes it difficult to choose a solution in finite samples that is guaranteed to converge asymptotically to the true parameter value.

The linearized one-step estimator is defined as follows. In place of the solution to the system of equations

$$\beta^*: \quad f_T(\beta^*, \gamma_0) = 0 \,, \tag{3.7}$$

(where $f_T(\cdot)$ is given in (3.3) in our case), compute the linear approximant $\hat{\beta}$:

$$\hat{\beta} = \hat{\beta}^{OP} - \left[\partial f_T(\hat{\beta}^{OP},\gamma_0)/\partial\beta\right]^{-1} f_T(\hat{\beta}^{OP},\gamma_0) \quad . \tag{3.8}$$

(Note that the level of conditioning is held fixed in this section and therefore it is left out as a subscript.) This estimator successfully mimics β^* because it possible to approximate the function $f_T(\beta,\gamma_0)$ closely near $\beta = \beta_0$ with a first order Taylor series expansion around $\beta = \hat{\beta}^{OP}$. $\hat{\beta}$ is the linear solution using this approximation of $f_T(\beta,\gamma_0)$.

The linearized one-step estimator is frequently used to calculate estimators that are asymptotically MLE's, when the function f_T is the score of the log-likelihood function. Such estimators are called linearized maximum likelihood estimators (LMLE's). Rothenberg and Leenders (1964) first applied these estimators to linear simultaneous equations. Our statement of the properties of the linearized one-step estimator is a modest generalization of theirs and does not require proof:

Lemma 1: *Consider a continuously differentiable function* $f_T(\beta)$ *with the property that* $E[f_T(\beta_0)] = 0$ *and the linearized one-step estimator*

$$\hat{\beta} = \tilde{\beta} + [\partial f_T(\tilde{\beta})/\partial\beta]^{-1} f_T(\tilde{\beta}) \quad . \tag{3.9}$$

If

i) $\tilde{\beta} \xrightarrow{a.s.} \beta_0$ *as* $T \to \infty$,

ii) $f_T(\beta) \xrightarrow{a.s.} E[f_T(\beta)]$ *uniformly in* β *as* $T \to \infty$, *and*

iii) $\partial f_T(\beta)/\partial\beta$ *is bounded and non-singular for all* $\beta \in B$,

then $\hat{\beta} \xrightarrow{a.s.} \beta_0$. *If, in addition,*

iv) $T^{1/2} f_T(\beta_0) \xrightarrow{D} z$ *as* $T \to \infty$ *and*

v) $\partial f_T/\partial\beta \xrightarrow{a.s.} E[\partial f_T/\partial\beta] \equiv M(\beta)$ *uniformly in* β *as* $T \to \infty$, $M(\beta_0)$ *is nonsingular,*

then $T^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} M(\beta_0)^{-1}z$. *If there exists a sequence* $\{\beta^*\}$ *such that* $f_T(\beta^*) = 0$ *and* $\beta^* \xrightarrow{a.s.} \beta_0$, *then* $T^{1/2}(\hat{\beta} - \beta^*) \xrightarrow{a.s.} 0$.

Note that (1) Lemma 1 essentially shows how to pick the correct root in situations where $f_T(\beta)$ has several roots, (2) it applies to such quasi-likelihood situations as those considered by White (1982) and Hansen (1982), (3) the distribution of the estimators follows from that of $f(\beta_0)$ and not the initial consistent estimator $\hat{\beta}_T$, and (4) one can replace the stochastic matrix function $\partial f_T(\beta)/\partial\beta$ with its expectation in (3.9) without affecting the asymptotic behavior of the resulting estimator. Note also that the estimator might be modified to take a "step" other than unity from the initial consistent estimator, say the step which minimizes the length of the vector $f(\hat{\beta}_T)$ [see Newey (1984)].

3.2 Feasible Estimation in the Presence of Nuisance Parameters

Often, an estimation procedure must also handle nuisance parameters in order to obtain an optimal estimator. For example in the general linear model, the parameters of the covariance matrix must be estimated consistently in order to calculate a *feasible* Aitken estimator for the slope coefficients that is asymptotically efficient among linear estimators. This is also a feature of our estimation problem because the weighting matrix A in (3.3) and (3.5) is a function of the unknown parameter vector β_0 . In these cases, the distribution of the estimator that replaces unknown nuisance parameters in the weighting matrix with consistent estimators is of interest. The next lemma describes the asymptotic behavior of such estimators for general cases.

Lemma 2: *Reconsider the situation in Lemma 1 with the additional structure*

(3.10) $$f_T(\beta) \equiv f_T(\beta,\delta) = T^{-1} X(\delta)'\hat{u}(\beta)$$

and $E[\hat{u}(\beta_0)|X(\delta)] = 0$ *uniquely for* $\beta = \beta_0$, *where* δ *is a finite-dimensional vector of constants and* f_T *is also continuously differentiable in* δ . *If*

i) $\hat{\delta} \xrightarrow{a.s.} \delta_0$ *as* $T \to \infty$,

ii) $f_T(\beta,\delta)$ *converges a.s. to its expectation, and*

iii) $\partial f_T/\partial\beta$ *and* $\partial f_T/\partial\delta$ *are bounded,* $\partial f_T/\partial\beta$ *is non-singular,*

then $\{\hat{\beta} : f_T(\hat{\beta},\hat{\delta}) = 0\}$ *converges a.s. to* β_0 . *If, in addition,*

iv) $T^{1/2}(\hat{\delta} - \delta_0)$ *converges in distribution,*

v) $T^{1/2} f_T(\beta_0,\delta_0) \xrightarrow{D} z$, *and*

vi) $\partial f_T/\partial\beta$ *and* $\partial f_T/\partial\tau$ *converge a.s. uniformly in* (β,δ) *to their expectations* $M_\beta(\beta,\delta)$ *and* $M_\delta(\beta,\delta)$ *respectively,*

then

vii) $M(\beta_0,\delta) = 0$ for all δ ,

viii) $T^{1/2}[f_T(\beta_0,\delta_0) - f_T(\beta_0,\hat{\delta})] \xrightarrow{a.s.} 0$,

ix) $T^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} M_\beta(\beta_0,\delta_0)^{-1} z$

x) and for $\{\beta^* : f_T(\beta^*,\delta_0)\}$, $T^{1/2}(\hat{\beta} - \beta^*) \xrightarrow{a.s.} 0$.

The most familiar application of this lemma is the feasible GLS estimator, which has the same distribution as the GLS estimator which uses the actual covariance matrix. Result (vii), that the normal equations f_T are unaffected by the value of the nuisance parameter δ on average, drives the

remaining results, beginning with the fact that the $T^{\frac{1}{2}}f_T$ has the same asymptotic distribution regardless of whether δ is set to its true value or to a consistent estimator. Note that both this lemma and Lemma 1 say nothing about the efficiency of the estimators that they discuss. It is well-known, for example, that if there are parametric restrictions among the covariance parameters and the regression coefficients, the feasible GLS estimator will generally be inefficient. This result is perfectly consistent with Lemma 2, which only states an equivalence and nothing about the properties of the infeasible estimator.

Finally, note that Lemmas 1 and 2 can easily be combined to yield an equivalence between the feasible linearized one-step estimator and its infeasible counterpart. We will make extensive use of such estimators. We have discussed the two methods separately to clearly distinguish familiar components of the structure of our estimators.

3.3 Asymptotic Properties of the GCM Estimator

Given the consistency of $\hat{\beta}^{OP}$, the consistency of $\hat{\beta}$ is neither surprising nor difficult to prove. The implicit function for our estimator, as given in equation (3.3) after normalization by T^{-1}, can be more completely written as

$$(3.11) \qquad f_T(\beta,\beta_0,\gamma_0) \equiv T^{-1} X'A_n(\beta_0,\gamma_0)\,\hat{u}_n(\beta,\gamma_0) = 0 .$$

For the duration of this section, we will continue to ommit the subscript for conditioning level, n, because it will remain fixed. We explicitly denote the matrix A in (3.11) as a function of both β and γ, as (3.5) implies. Since β_0 is unknown, the feasible counter-part derived from

$$(3.11') \qquad f_T(\beta,\hat{\beta}^{OP},\gamma_0) \equiv T^{-1} X'A_n(\hat{\beta}^{OP},\gamma_0)\,\hat{u}_n(\beta,\gamma_0) = 0 ,$$

will actually be used. Taking (3.11') as the definition of f_T ,

$$(3.12) \qquad \hat{\beta}_n = \hat{\beta}^{OP} - \left[E[\partial f_T(\hat{\beta}^{OP},\hat{\beta}^{OP},\gamma_0)/\partial\beta]\right]^{-1} f_T(\hat{\beta}^{OP},\hat{\beta}^{OP},\gamma_0) .$$

is our feasible one-step GCM estimator, where the weighting matrix in brackets is given in (3.6). Because $\hat{\beta}^{OP} \longrightarrow \beta_0$, the consistency of $\hat{\beta}$ must follow from the convergence of the second term to zero.

In order to obtain such asymptotic results, we must specialize the information sets $\mathcal{I}_{ni}$. As we have already indicated, it seems clear that we cannot have all $\binom{T}{n}$ different information sets because for $n << T$ this number can be extremely large and is not of order $O(T)$. In order to reduce the number of information sets, it seems intuitively sensible to group observations that are adjacent in the time-series. Therefore, we restrict $\mathcal{I}_{ni}$ to

$$(3.13) \qquad \mathcal{I}_{nt} \equiv \{t,\ t+1,\ \ldots,\ t+n-1\} \qquad (t=1,\ldots,T-n+1)$$

where the t subscript has been reintroduced because it has a natural time-series interpretation. An important implication of this choice of information sets is that the λ_{nt} defined in (3.2) now form an α-mixing process of the same order as the underlying u_t (see Lemma A.2). As a result, we will be able to extend the asymptotic distribution theory of the OP estimator to the case of the GCM estimator.

Before obtaining any asymptotic results, we must also ensure that our esitmator (3.12) is well-defined by adding the following assumption:

Assumption 5: $det\left[Var(\hat{u}_n)\right] > \Delta^T$, *for some* $\Delta > 0$.

This assumption guarantees the definiteness, and invertability, of $Var(\hat{u}_n)$. Given our previous assumptions, this assumption seems quite mild. We conjecture that it can actually be deduced from the previous ones.

We can now state our consistency result:

Theorem 3: Under assumptions 1(a,b,c), 2(a,b), 3(a), 4, and 5, the GCM estimator $\hat{\beta}_n$ in (3.12) is a strongly consistent estimator of β_0 as $T \longrightarrow \infty$.

In order to extend the Central Limit theory applied to the OP estimator, we must extend Assumptions 3(c,d).

Assumption 3':

a) The empirical distribution of X_{nt} $(t=a+1,\ldots,a+T-n)$ converges to a limit distribution H_{n0} , uniformly in a .

b) The empirical distribution of $(X_{nt},X_{n,t+m})$ $(t=a+1,\ldots,a+T-n-m)$ converges to a limit distribution H_{nm} , uniformly in a .

These assumptions are stronger than 3(cd) because they involve a multivariate distribution of dimension $2Kn$, in place of one of dimension $2K$. Note that Assumption 3'(b) implies Assumption 3'(a) if x_t contains a constant. The homogeneity in the X_{nt} that Assumptions 3'(a,b) requires allows us to prove

that the GCM estimator has an approximately normal distribution when T is large.

Theorem 4: Under Assumptions 1(a,b,c), 2(a,b), 3(a), 3'(a,b), 4, and 5, $T^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{D} N(0,\Omega_n)$ as $T \longrightarrow \infty$, where

$$\Omega_n = plim\ T^{-1}\ E[\partial u_n'/\partial\beta]\ V(u_n)^{-1}\ E[\partial u_n/\partial\beta'] .$$

4. The Efficiency of CMD Estimators

Given our parallels with the GLS estimation method, it is not surprising that for a particular $\hat{u}_n$, the matrix $V[\hat{u}_n]^{-1}$ is optimal in the sense that any other matrix metric yields a less efficient estimator. This result follows from an asymptotic version of the Gauss-Markov Theorem, which closely resembles Hansen (1982, Lemma 4.3):

Lemma 3: *Define the family of estimators* $\{\hat{\beta} : Z_T'\hat{u}_n(\hat{\beta}) = 0\}$ *given by 3.12). If such a* $\hat{\beta}$ *is consistent and asymptotically normal (CAN), the GCM estimator* $\hat{\beta}_n$ *that sets* $Z_T = E[\partial\hat{u}_n'/\partial\beta]\ V[\hat{u}]^{-1}$ *is best in the sense that the limiting covariance matrix of* $\hat{\beta}_n$ *is smallest (in the usual matrix sense).*

Proof: We follow Hansen's proof with the exception that the number of moment conditions, and hence some of our matrices, are growing with sample size T. Since we have already shown in the previous section, that the covariance matrices of the GCM estimators do converge, it suffices to show that the matrix inequalities hold point-wise in a sequence of consistent estimators for

the covariance matrices. This follows directly from Hansen's Lemma 4.3.

Q.E.D.

Having established that the CMD estimator is best among estimators that use a particular set of "residuals" $\hat{u}_n$, we turn to the choice of residuals or possible combinations of $\hat{u}_n$ by varying n, the level of conditioning. In the following lemma and theorem, we show that as the level of conditioning is increased, the GCM estimator becomes relatively more efficient. As a result, GCM estimators based on linear combinations of $\hat{u}_n$ and $\hat{u}_m$ for $m < n$ are identical to $\hat{\beta}_n$ based on $\hat{u}_n$ alone. Thus, a simple trade-off between computation and efficiency among GCM estimators holds.

Lemma 4: *Let $\hat{\beta}_m$ and $\hat{\beta}_n$ be GCM estimators (3.12) for conditioning levels m and n respectively, where $m < n$. If and only if,*

$$E\left[\frac{\partial \hat{u}_m}{\partial \beta'}\right] = \Lambda_{mn}\Lambda_{nn}^{-1} E\left[\frac{\partial \hat{u}_n}{\partial \beta'}\right] , \tag{4.1}$$

where

$$\Lambda_{ij} \equiv E[\hat{u}_i \hat{u}_j'] , \qquad (i,j = m,n),$$

then $\hat{\beta}_n$ is efficient relative to $\hat{\beta}_m$.

Proof: Consider the estimator $\hat{\beta}_*$ based on $\hat{u}_* = (\hat{u}_m', \hat{u}_n')'$. According to Lemma 3, the optimal estimator will solve

$$E\left[\frac{\partial \hat{u}_*'}{\partial \beta}\right] \mathrm{Var}[\hat{u}_*]^{-1} \hat{u}_*(\hat{\beta}_*) = 0 \tag{4.2}$$

or

$$(4.3) \qquad \left[E\left[\frac{\partial \hat{u}_m'}{\partial \beta} \quad \frac{\partial \hat{u}_n'}{\partial \beta}\right]\right] \begin{bmatrix} \Lambda_{mm} & \Lambda_{mn} \\ \Lambda_{nm} & \Lambda_{nn} \end{bmatrix} \begin{bmatrix} \hat{u}_m(\hat{\beta}_*) \\ \hat{u}_n(\hat{\beta}_*) \end{bmatrix} = 0$$

Using a partitioned inverse, (4.3) can be rewritten as

$$(4.4) \qquad \left[F\ G \quad - F\ G\ \Lambda_{mn}\Lambda_{nn}^{-1} + E\left[\frac{\partial \hat{u}_n'}{\partial \beta}\right] \Lambda_{nn}^{-1}\right] \begin{bmatrix} \hat{u}_m(\hat{\beta}_*) \\ \hat{u}_n(\hat{\beta}_*) \end{bmatrix} = 0$$

where

$$(4.5) \qquad F \equiv E\left[\frac{\partial \hat{u}_n}{\partial \beta'} - \Lambda_{mn}\Lambda_{nn}^{-1}\frac{\partial \hat{u}_m}{\partial \beta'}\right]'$$

$$G \equiv \left[\Lambda_{mm} - \Lambda_{mn}\Lambda_{nn}^{-1}\Lambda_{nm}\right]^{-1} .$$

Because the random variables $\hat{u}_m$ and $\hat{u}_n$ are not linearly dependent (almost surely), (4.4) simplifies to the normal equation for $\hat{\beta}_n$ if and only if $F = 0$. In other words, $\hat{\beta}_* = \hat{\beta}_n$ if and only if $F = 0$. *Q.E.D.*

Theorem 5: $\hat{\beta}_n$ is more efficient than $\hat{\beta}_m$.

Proof: First, note that our selection of information sets in (3.13) has the property that for any m, n, i such that $m < n$, it follows that there is a j such that $\mathcal{I}_{mi} \subset \mathcal{I}_{nj}$. Now suppose that we repeat the experimental design

using X_{nj} . In a repeated design, the GCM estimator for the observations on $\hat{u}_n = \lambda_{nj}$ is the MLE corresponding to the sample indexed by $\mathfrak{s}_{nj}$, as explained in Sections 2 and 3, and is therefore relatively efficient compared to the GCM estimator for the observations on $\hat{u}_m = \lambda_{mi}$. Although the estimators will only be consistent for a linear function of β_0 if $n < K$, this relative efficiency will stand in the sense that the (possibly singular) asymptotic covariance matrices will have a positive semi-definite difference. This particular relative efficiency implies that the GCM estimator for $\hat{u}_n = [\lambda_{ni}]$ is also relatively efficient compared to the GCM estimator for $\hat{u}_m = \lambda_{mi}$. According to Lemma 4, this implies that

$$E\left[\frac{\partial \lambda_{mi}}{\partial \beta'}\right] = E[\lambda_{mi}\hat{u}_n'] \, \Lambda_{nn}^{-1} \, E\left[\frac{\partial \hat{u}_n}{\partial \beta'}\right] , \tag{4.6}$$

for any T , which implies in turn

$$E\left[\frac{\partial \hat{u}_m}{\partial \beta'}\right] = \Lambda_{mn}\Lambda_{nn}^{-1} \, E\left[\frac{\partial \hat{u}_n}{\partial \beta'}\right] . \tag{4.7}$$

Applying Lemma 4 again leads to the result of the Theorem. *Q.E.D.*

The proof of this theorem brings out an important point: for a conditioning level n , the CMD estimator uses the orthogonality conditions of the marginal likelihood functions of sub-samples of size n to form its own orthogonality condtions. In view of Lemma 3, the GCM estimator takes these marginal likelihood normal equations and combines them in an optimal (linear) fashion. Thus, an alternative motivation of GCM estimation begins with the

marginal log-likelihood functions for sub-samples and asks how to best combine the information in these functions to obtain an estimate of β_0 . However, the GCM estimator cannot be viewed as an optimal linear combination of the MLE's for each sub-sample. This follows from recognizing that, asymptotically, the MLE's are particular linear combinations of their respective scores whereas the GCM estimators are linear combinations of the "residuals" in those score functions. Nevertheless, our result of relative efficiency follows intuitively from the fact that for each MLE of level m there is another MLE of level n that is more efficient. The GCM estimators are in turn improvements upon the MLE's of a particular conditioning level. The rather obvious ordering of such underlying MLE's is extended to an ordering of GCM estimators.

5. Estimation with Unknown Covariance Matrix

Generally the covariance paramters γ_0 are unknown to the researcher and they must be estimated along with β_0 . In this section, we will discuss the extension of the results given above for known γ_0 to cover the estimation of β and γ together. This extension can be accomplished straight-forwardly by adding the orthogonality conditions for γ implied by the maximum likelihood normal equations to those already used by the GCM estimator for β . The lemmas and theorem given in the previous section apply directly to this extended GCM estimator, provided that the asymptotic results of Section 3 can also be extended.

Fortunately, the additional parameters require the evaluation of integrals of the same order as those in the previous situation. In this sense, the compuational burden is not greatly increased. The covariance parameters do, however, add significantly to the non-linearity of the

estimation problem, and the one-step estimator will be even more attractive, compared to actual solution of the conditional moment equations.

The normal equations for the τ vector for the MLE are

$$(5.1)\qquad \frac{\partial\, ln\, L(\beta,\tau;y)}{\partial\, \tau} \;=\; \tfrac{1}{2}\,\frac{\partial\, vec[\Gamma(\tau)^{-1}]'}{\partial\, \tau}\, vec[\Gamma(\tau) - E(u\,u'|y;\beta,\tau)] \;=\; 0$$

where L is defined in (2.8) and (2.9). From (5.1) we see that the relevant conditional moments are $vec[\Gamma(\tau) - E(u\,u'|y;\beta,\tau)] = 0$. Taking such functions to the intermediate conditioning levels of GCM estimators we form

$$(5.2)\qquad \omega_{nt}(\beta,\tau;y_{nt}) \;\equiv\; vec[E(u_{nt}u_{nt}') - E(u_{nt}u_{nt}'|y_{nt};\beta,\tau)] \qquad (t=1,\ldots,T-n),$$

$$(5.3)\qquad \hat{u}_n(\theta) \;\equiv\; \begin{bmatrix} \lambda_{nt}(\theta) \\ \omega_{nt}(\theta) \end{bmatrix}.$$

Our GCM estimator for $\theta \equiv (\beta,\tau)$ would then be formed from

$$(5.5)\qquad f_T(\theta) \;\equiv\; T^{-1}\, E\left[\frac{\partial\, \hat{u}_n'(\hat{\theta}^{OP})}{\partial\, \theta}\right] V(\hat{u}_n;\hat{\theta}^{OP})^{-1}\, \hat{u}_n(\theta)$$

as the one-step estimator

$$(5.6)\qquad \hat{\theta}_n \;=\; \hat{\theta}^{OP} \;-\; \{E[\partial f_T(\hat{\theta}^{OP})/\partial\theta]\}^{-1}\, f_T(\hat{\theta}^{OP})$$

just as in (3.12). An initial estimate of τ for the one-step estimator would be based on the estimator $\hat{\beta}^{OP}$ and a combination of the ω_{nt} in (5.2).

The extended $\hat{u}_n$ are formally identical to those in Sections 3 and 4 with respect to the characteristics that enable us to prove the preceding lemmas and theorems. The elements in (5.3) still form an α-mixing process of order $O(m^{-c})$ and their moments are appropriately bounded and differentiable. Without further complications, we can assert that Theorems 3, 4, and 5 continue to hold if one replaces $\hat{\beta}_n$ with $\hat{\theta}_n$ as defined in (5.5) and (5.6) and adds Assumption 2(c) wherever Assumption 2(b) appears.

6. AR(1) Probit

The simplest example of the GCM estimator is its application to a first-order autoregressive error term:

$$u_t = \rho u_{t-1} + \epsilon_t \tag{6.1}$$

where $E[\epsilon_t] = 0$, $V[\epsilon_t] = 1 - \rho^2$. One can obtain consistent estimates of β and ρ easily. As already pointed out, an estimate of β is $\hat{\beta}^{OP}$. One can derive a simple estimator for ρ using the OP residuals by finding a root to the equation

$$T^{-1} \sum_{t=2}^{T} \lambda_{1t}\lambda_{1,t-1} - E\left[\lambda_{1t}\lambda_{1,t-1}\right] = 0 \quad \text{where} \tag{6.2}$$

$$E\left[\lambda_{1t}\lambda_{1,t+m}\right] = \frac{\phi_t\phi_{t+m}(\Phi_{2,t,t+m} - \Phi_t\Phi_{t+m})}{\Phi_t\Phi_{t+m}(1-\Phi_t)(1-\Phi_{t+m})} \tag{6.3}$$

$$(m=\ldots,-1,0,1,\ldots),$$

$\lambda_{1t} \equiv (y_t - \Phi_t)/[\Phi_t(1-\Phi_t)]$, $\phi_t \equiv \phi(x_t'\hat{\beta}^{OP})$, $\Phi_t \equiv \Phi(x_t'\hat{\beta}^{OP})$, and $\Phi_{2,t,t+m} \equiv$

$\Phi_2(x_t'\hat{\beta}^{OP}, x_{t+m}'\hat{\beta}^{OP}, \rho^m)$. Experience shows that this estimator is also easy to compute.[6]

Given consistent estimates for β and ρ , let us proceed to the GCM estimator for a conditioning level of one ($n=1$). We use therefore the same residuals as ordinary probit, but combine them in a generalized fashion. We require two terms to compute the weighting matrix in (3.12) given by (3.6):

$$E[\partial\lambda_{1t}/\partial\beta] = \frac{\phi_t^2}{\Phi_t(1-\Phi_t)} x_t \quad \text{and}$$

$$Cov[\lambda_{1t}\lambda_{1,t+m};\beta,\rho] = E[\lambda_{1t}\lambda_{1,t+m};\beta,\rho]$$

which is given in (6.3). Note that in the case we are considering, in which ρ is estimated, ρ is not required for the construction of the "residuals" and hence its asymptotic distribution is not required, nor do we have to improve our estimate of ρ to get a more efficient estimator than $\hat{\beta}^{OP}$ from the GCM estimator.

7. Conclusion

In this paper, a general class of estimators called generalized conditional moment estimators has been introduced. For such limited dependent variable models with serial correlation as general probit, these estimators provide computationally feasible alternatives to ML estimation. In addition,

---

6 Dickens (1980) used this method for a panel data problem and found that quadratic approximations solved the implicit function rapidly and that the root was always unique. His Monte Carlo experiments suggest that this estimator can have reasonably small finite sample variance.

there is an efficiency ranking among the members of the GCM family where the efficiency grows with increasing computational difficulty.

General probit estimation was explained in detail. Other limited dependent variable models, like Tobit, can be estimated by the same methods, although we have not provided the asymptotic distribution theory to substantiate this claim here. Panel data problems, with relatively long time series, are an interesting special case because they provide a natural division for sub-sample conditioning sets, namely the cross-sections in each time period.

We have also provided asymptotic distribution theory for the ordinary probit and GCM estimators. The theory provides a direct route to the consistency and asymptotic normality of quasi maximimum likelihood estimators for discrete dependent data based on mixing conditions without reference to the marginal likelihood function. We have not shown the consistency or the asymptotic normality of the MLE and see this as an interesting topic for future research.

Further work might also investigate the uniqueness of the MLE and the roots of the GCM conditions. It is well-known that ordinary probit has a unique maximum likelihood estimator. Similarly, GCM estimators are also probably unique if one solves their normal equations, rather than employing the one-step version.

APPENDICES

A.1 Proof of the Consistency of $\hat{\beta}^{OP}$ (Theorem 1)

We are interested in dependent sequences of random variables, and like many others, we will assume that such sequences are α-mixing following the definitions given by White [1984a, pp. 44, 45]. We will make particular use of the mixing coefficient $\alpha(m)$ which is a measure of the dependence in a stochastic sequence [White, (1984a, p. 45)]:

Definition A.1: For a sequence of random vectors $\{Z_t\}$ with $B_{-\infty}^{n} \equiv \sigma(\ldots, Z_n)$ and $B_{n+m}^{\infty} \equiv \sigma(Z_{n+m}, \ldots)$, the mixing coefficient $\alpha(m)$ is defined to be

$$\alpha(m) \equiv \sup_t \sup_{\{A_1 \in B_{-\infty}^{t}, A_2 \in B_{t+m}^{\infty}\}} |P(A_1 \cap A_2) - P(A_1)P(A_2)| .$$

The sets B_{∞}^{t} and B_{t+m}^{∞} [which are formally defined by White (1984a, p. 44, Definition 3.40)] represent all the information contained in the respective sequences. The coefficient $\alpha(m)$ measures the dependence between two events separated by at least m time periods in terms of the difference between the joint probability of the events and the product of their marginal probabilities. The mixing coefficient measures the dependence in the entire sequence by taking the maximum coefficient over all possible events and all time.

Definition A.2: If $\alpha(m) \longrightarrow 0$ as $m \longrightarrow \infty$, then $\{Z_t\}$ is α-mixing.

In our proofs of the consistency of estimators, we will use two general results about random sequences that are weak mixing. The law of large numbers that we will use is taken from White (1984a, p. 47, Corollary 3.48):

Lemma A.1 (Law of Large Numbers): *Let* $\{Z_t\}$ *be a sequence with* $\alpha(m) = O(m^{-\lambda})$ $\lambda > r/(r-1)$, $r > 1$, *such that* $E(Z_t) = \mu_t < \infty$ *and* $E|Z_t|^{r+\delta} < \Delta < \infty$ *for some* $\delta > 0$, *and all* t . *Then* $\bar{Z}_T - \bar{\mu}_T \xrightarrow{a.s.} 0$, *where* $\bar{Z}_T \equiv T^{-1} \sum_{t=1}^{T} Z_t$ *and* $\bar{\mu}_T \equiv T^{-1} \sum_{t=1}^{T} \mu_t$.

In addition, we will make use of the fact that measurable functions of mixing processes are mixing of the same order [see White (1984a, p. 47, Theorem 3.49)]:

Lemma A.2: *Let* g *be a measurable function onto* $\mathbb{R}^k$ *and define* $Y_t \equiv g(Z_t, \ldots, Z_{t+\tau})$, *where* τ *is finite. If the sequence of random vectors* $\{Z_t\}$ *is weak mixing such that* $\alpha(m) = O(m^{-\lambda})$ *for some* $\lambda > 0$, *then* $\{Y_t\}$ *is also weak mixing such that* $\alpha_Y(m) = O(m^{-\lambda})$.

For many proofs, we will require a stronger result than Lemma A.1. That is, we will appeal to a uniform law of large numbers. We will use a convenient one recently given by Andrews (1985):

Lemma A.3 (Uniform Law of Large Numbers): *Let* $\{Z_t(\theta)\}$ *be a sequence of random variables such that* $E[Z_t(\theta)] = \mu_t(\theta)$ $(t=1,2,\ldots)$. *Let* $Z_t^*(\hat{\theta},\rho) \equiv \sup_{\|\theta - \hat{\theta}\| < \rho} Z_t(\theta)$ *and* $Z_{*t}(\hat{\theta},\rho) \equiv \inf_{\|\theta - \hat{\theta}\| < \rho} Z_t(\theta)$;[7]

---

[7] $\|\cdot\|$ denotes the Euclidean length of a vector.

$\mu_t^*(\theta,\rho) \equiv E[Z_t^*(\theta,\rho)]$ *and* $\mu_{*t}(\theta,\rho) \equiv E[Z_{*t}(\theta,\rho)]$. *Let* $\bar{Z}_T(\theta) \equiv T^{-1}\Sigma_t^T Z_t(\theta)$, $\bar{Z}_T^*(\theta,\rho) \equiv T^{-1}\Sigma_t^T Z_t^*(\theta,\rho)$, *and so on. If*

i) $\bar{Z}_T^*(\theta,\rho) - \mu_T^*(\theta,\rho)$ *and* $\bar{Z}_{*T}(\theta,\rho) - \mu_{*T}(\theta,\rho)$ *converge a.s. to zero as* $T \to \infty$,

ii) $\theta \in \Theta$, *a compact, convex subset of* $\mathbb{R}^K$,

iii) $\exists\ \rho > 0$ *such that* $\overline{lim}_{T\to\infty}\ T^{-1}\ \Sigma_t^T\ E\left[sup_{\|\hat{\theta}\ -\ \theta\| \le \rho}\ \|\partial Z_t(\hat{\theta})/\partial\theta\|\right] < \infty$, $\forall\ \theta \in \Theta$,

then (a) $\bar{Z}_T(\theta) - \bar{\mu}_T(\theta) \xrightarrow{a.s.} 0$ *uniformly in* θ *and (b)* $\bar{\mu}_T(\theta)$ *is continuous on* Θ , *uniformly over* $T \ge 1$.

Now consider the binomial dependent data y_t and take

$$(A.1) \qquad Z_t(x_t'\beta) = log\left[\Pr(y_t,x_t'\beta)/\Pr(y_t,x_t'\beta_0)\right]$$

where

$$(A.2) \qquad \Pr(y_t,x_t'\beta) \equiv \left\{ \begin{array}{ll} \Phi(-x_t'\beta) , & \text{if } y_t = 0 \\ 1 - \Phi(-x_t'\beta) , & \text{if } y_t = 1 \end{array} \right\}$$

for a fixed $\beta \in B$. This quantity is of interest because of the fundamental information inequality

$$(A.3) \qquad E_{\beta_0}\left[log\left[\Pr(y_t,x_t'\beta)/\Pr(y_t,x_t'\beta_0)\right]\right] = E_{\beta_0}\left[Z_t(x_t'\beta)\right] < 0 ,$$
$$\forall\ x_t'\beta \neq x_t'\beta_0 ,$$

which states that on average $Z_t(x_t'\beta)$ is maximized at $x_t'\beta_0$, where β_0 is the true parameter value [see Lehman (1983, pp. 409 ff.)]. This suggests, in

turn, that the sequence of maximizers $\{\hat{\beta}_T\}$ of the sample quasi log-likelihood function

$$(A.4) \quad \sum_{t=1}^{T} log\left[\Pr(y_t, x_t'\beta)\right] = \sum_{t=1}^{T} Z_t(x_t'\beta) + \sum_{t=1}^{T} log\left[\Pr(y_t, x_t'\beta_0)\right]$$

will be, on average, equal to $x_t'\beta_0$, and converge asymptotically to $x_t'\beta_0$. This is, in fact, the case.

Follow this simple argument:

(1) Because B is compact and x_t is bounded $x_t'\beta$ is bounded. This implies bounds on Z_t and all of its moments.

(2) $\{y_t\} = \{1(x_t'\beta + u_t \geq 0)\}$ is α-mixing because (i) x_t is non-stochastic and therefore α-mixing, (ii) u_t is α-mixing by Assumption 1, and (iii) Lemma A.2. $Z_t(x_t'\beta)$ is also clearly α-mixing of the same order as u_t.

(3) The conditions of the law of large numbers are met for $\{Z_t(x_t'\beta)\}$ and $n^{-1}\sum_{t=1}^{n}\left[Z_t(x_t'\beta) - E\{Z_t(x_t'\beta)\}\right] = \bar{Z}_T(\beta) - \bar{\mu}_T(\beta) \xrightarrow{a.s.} 0$, for each $\beta \in B$. Furthermore, this convergence is uniform in $\beta \in B$ by Lemma A.3.

(4) Finally, $\bar{\mu}_T(\beta) \equiv \int E[Z_t(x_t'\beta)]\, dH_T$ converges to $\bar{\mu}(\beta) \equiv \int E[Z_t(x_t'\beta)]\, dH$, according to Theorem 1 of Jennrich [1969, p. 635], using the stationarity of u_t and the convergence of $\{H_n\}$ to H. As we noted above, $\bar{\mu}(\beta)$ has a unique maximum at β_0. Therefore, Lemma 3 of Amemiya [1973, p. 1002] implies that the sequence $\{\hat{\beta}_T\}$ converges a.s. to β_0.

Note the following:

(1) This proof does not depend on the assumption that the u_t are a Gaussian ARMA process. Nor has any restriction on the nature of dependence, other than α-mixing, been imposed: it is necessary that $\alpha(m) \longrightarrow 0$ as $m \longrightarrow 0$, but not at any particular rate.

(2) The method of proof extends immediately to other quasi log-likelihood functions each term of which involves a subset of the y_t's and for which the maximum distance τ between subscripts is bounded by a fixed finite number. Such extensions are used in Appendix 3.

(3) A corollary to observation (2) is that the consistency of the MLE for this data cannot be established in this way, for in that case τ is the sample size and not finite.

A.2 Proof of the Asymptotic Normality of $\hat{\beta}^{OP}$ (Theorem 2)

In oder to prove the asymptotic normality of $\hat{\beta}^{OP}$, we will use the following central limit theorem presented in White (1984a, p. 124, Thm. 5.19):

Lemma A.4 (Central Limit Theorem): *Let* $\{Z_t\}$ *be a sequence of* α*-mixing random scalars such that* $\alpha(m) = O(m^{-\lambda})$, $\lambda > r/(r-1)$, $r > 1$, *with* $E(Z_t) = \mu_t$ *and* $Var(Z_t) = \sigma_t^2$, $\sigma_t^2 \neq 0$, *and* $E|Z_t|^{2r} < \Delta < \infty$ *for all* t. *Define* $\bar{\sigma}^2_{a,T} \equiv V\left[T^{-1/2} \sum_{t=a+1}^{a+T} Z_t\right]$. *If there exists* $\bar{\sigma}^2$, $0 < \bar{\sigma}^2 < \infty$, *such that* $\bar{\sigma}^2_{a,T} \longrightarrow \bar{\sigma}^2$ *as* $T \longrightarrow \infty$ *uniformly in* a, *then* $T^{1/2}(\bar{Z}_n - \bar{\mu}_T)/\bar{\sigma}$ *converges in distribution to a* $N(0,1)$ *random variable.*

The probit quasi-MLE $\hat{\beta}^{OP}$ is the implicit solution to the vector equation

$$\sum_{t=1}^{T} \hat{u}_t^{OP} x_t = 0 \tag{A.5}$$

where $\hat{u}_t^{OP} = (y_t - \hat{\Phi}_t)\cdot\hat{\phi}_t/[\hat{\Phi}_t(1 - \hat{\Phi}_t)]$ is a scalar, $\hat{\Phi}_t \equiv \Phi(-x_t'\hat{\beta}^{OP})$, and $\hat{\phi}_t \equiv \phi(-x_t'\hat{\beta}^{OP})$. Given the consistency of $\hat{\beta}^{OP}$ shown in Appendix 1, the critical part of the asymptotic normality argument is whether (A.5) evaluated at β_0 and standardized by $T^{-1/2}$ is asymptotically normal; that is, does

$$T^{-1/2} \sum_{t=1}^{T} u_t x_t \xrightarrow{D} N(0,V), \tag{A.6}$$

where $u_t = (y_t - \Phi_t)\cdot\phi_t/[\Phi_t(1 - \Phi_t)]$ is a scalar, $\Phi_t \equiv \Phi(-x_t'\beta)$, and $\phi_t \equiv \phi(-x_t'\beta)$, and V is a positive definite matrix? The individual terms in (A.6) all have expectation zero, and as a sequence they are α-mixing by Lemma A.2. Note once again that the compactness of $\mathfrak{X}$ and B guarantees that all the moments of these elements are bounded.

The strategy we take in establishing (A.6) is to establish via Lemma A.3 the asymptotic normality of $T^{1/2}(\bar{Z}_T - \bar{\mu}_T)/\bar{\sigma}$, where

$$Z_t = \xi'(u_t x_t) = u_t(\xi' x_t) \tag{A.7}$$

for an arbitrary $K \times 1$ vector ξ of constants. Given that $T^{1/2}(\bar{Z}_T - \bar{\mu}_T)/\bar{\sigma}$ has an asymptotic normal distribution for arbitrary ξ, it then follows that (A.6) must hold. In order to invoke Lemma A.3, it is necessary to demonstrate the existence of $\bar{\sigma}^2$ and the uniform convergence of $\bar{\sigma}^2_{a,T}$ to $\bar{\sigma}^2$. We will make extensive use of the following lemma to show such results.

Lemma A.5: *Consider the sequence* $\{(x_t, z_t)\}$ *where*

(i) $x_t \in \mathfrak{X}$, *a compact subset of* $\mathbb{R}^K$,

(ii) *the empirical distribution of* x_t $(t=a+1,\ldots,a+T)$ *converges to a limit distribution* H_0, *uniformly in* a,

(iii) *the empirical distribution of* (x_t, x_{t+m}) $(t=a+1,a+2,\ldots,a+T)$ *converges to a limit distribution* H_m, *uniformly in* a, $m=1,2,\ldots$,

(iv) z_t $(t=a+1,a+2,\ldots,T)$ *is a sequence of vectors of length* L *containing Bernoulli random variables that is* α*-mixing of order* $O(m^{-c})$, $E(z_t) = P(x_t)$, $E(z_t z_{t+m}) = C(x_t, x_{t+m})$,

(v) the vector-valued function $P:\mathbb{S}\rightarrow\mathbb{R}^{L}$ *, and the matrix-valued functions* $f:\mathbb{S}\rightarrow\mathbb{R}^{L}\times\mathbb{R}^{N_1}$ *,* $g:\mathbb{S}\rightarrow\mathbb{R}^{L}\times\mathbb{R}^{N_2}$ *, and* $C:\mathbb{S}\times\mathbb{S}\rightarrow\mathbb{R}^{L}\times\mathbb{R}^{L}$ *are absolutely and uniformly integrable with respect to* H_m $(m=0,1,2,\ldots)$ *,*

then

$$(A.8)\qquad T^{-1}\sum_{t=a+1}^{a+T}\sum_{v=a+1}^{a+T} f(x_t)'Cov(z_t,z_v)\,g(x_v)$$

converges to a matrix constant as $T\rightarrow\infty$ *, uniformly in* a *.*

Proof: Expression (A.8) can be rewritten as the sum of three terms:

$$(A.9)\qquad T^{-1}\sum_{t=a+1}^{a+T} f(x_t)'Var(z_t)\,g(x_t) \;+$$

$$(A.10)\qquad \sum_{m=1}^{T-1} T^{-1}\sum_{t=a+1}^{a+T-m} f(x_t)'Cov(z_t,z_{t+m})\,g(x_{t+m}) \;+$$

$$(A.11)\qquad \sum_{m=1}^{T-1} T^{-1}\sum_{t=a+1}^{a+T-m} f(x_{t+m})'Cov(z_{t+m},z_t)\,g(x_t) \;.$$

Expression (A.9) converges by conditions (i), (ii), and (v), by application of the Helly-Bray Theorem. By the same logic, each of the expressions inside the first sum in (A.10) and (A.11) converges given condition (iii). It remains to show that the sums over the index m converge. If the covariance terms die out fast enough as m grows, the double sums in (A.10) and (A.11) will converge.

We will prove the stronger result that (A.10) and (A.11) are absolutely summable. The key to our proof is that the mixing condition (iv) places bounds on the covariance terms. Consider the $(i,j)^{th}$ element of $Cov(z_t, z_{t+m})$:

$$[Cov(z_t, z_{t+m})]_{i,j} = Pr(z_{t,i} = 1,\ z_{t+m,i} = 1) - Pr(z_{t,i} = 1) \cdot Pr(z_{t+m,i} = 1)$$

$$\Rightarrow \quad |[Cov(z_t, z_{t+m})]_{i,j}| \leq \alpha(m)$$

by Definition A.1 and condition (iv), where the (i,j) subscript in the last expression labels matrix elements. Therefore, for any fixed $\xi_1 \in \mathbb{R}^{N_1}$ and $\xi_2 \in \mathbb{R}^{N_2}$, a bilinear form in (A.10) satisfies the inequalities

$$\sum_{m=1}^{T-1} T^{-1} \sum_{t=a+1}^{a+T-m} |\xi_1' f_t' Cov_{t,t+m}\, g_{t+m}\, \xi_2| \leq$$

$$\sum_{m=1}^{T-1} T^{-1} \sum_{t=a+1}^{a+T-m} |\xi_1' f_t' Cov_{t,t+m}\, f_t\, \xi_1|^{1/2} \cdot |\xi_2' g_{t+m}' Cov_{t,t+m}\, g_{t+m}\, \xi_2|^{1/2} \leq$$

$$(A.12) \qquad (N_1 \cdot N_2)^{1/2} \sum_{m=1}^{T-1} \alpha(m)\, T^{-1} \sum_{t=a+1}^{a+T-m} \|f_t\, \xi_1\| \cdot \|g_{t+m}\, \xi_2\| \cdot$$

using the Cauchy-Schwarz inequality, where subscripts have been used to abbreviate arguments indexed by time. Again, conditions (i), (ii), (iii), and (v) assure the convergence of each sum inside the first summation, and furthermore, each limit is bounded uniformly in m . By condition (iv), the

series $\sum_{m=1}^{\infty} \alpha(m)$ converges [see Rudin (1976, p. 62, Theorem 3.28)]. Combining these results, it follows that (A.12) is absolutely summable, so that (A.10) converges to a constant as required. In a similar fashion, (A.11) converges to a constant. Conditions (ii) and (iii) guarantee that this convergence is uniform in a. *Q.E.D*

Returning to the asymptotic behavior of the OP estimator, consider

$$\bar{\sigma}^2_{a,T} \equiv Var\left[T^{-\frac{1}{2}} \sum_{t=a+1}^{a+T} Z_t\right] = T^{-1} \sum_{t=a+1}^{a+T} \sum_{v=a+1}^{a+T} (\xi' x_t x_v' \xi)\, Cov[u_t, u_{t+m}] .$$

Assumptions 3(acd) coincide with conditions (i), (ii), and (iii) of Lemma A.5. Defining $z_t \equiv (y_t, 1)'$ $(L=2)$, condition (iv) is satisfied according to Assumptions 1(a,b). This leaves the definitions of the functions in condition (v), which are

$$f(x_t) = g(x_t) \equiv [\phi_t/\Phi_t(1-\Phi_t) \quad -\phi_t/(1-\Phi_t)]\, x_t$$

$$P(x_t) \equiv (\Phi_t, 1)'$$

$$C(x_t, x_{t+m}) \equiv \begin{bmatrix} (\Phi_2(x_t'\beta_0, x_{t+m}'\beta_0; \rho_m) - \Phi_t \cdot \Phi_{t+m}) & 0 \\ 0 & 0 \end{bmatrix} .$$

where Φ_2 is defined in (2.7) and $N_1 = N_2 = K$. These functions satisfy (v) because they are continous and infinitely differentiable. Lemma A.5 implies, therefore, that $\bar{\sigma}^2_{a,T}$ does converge uniformly to a constant $\bar{\sigma}^2 = \xi' V \xi$ and by straight-forward algebra

$$V = \Delta_0 + 2 \sum_1^{\infty} \Delta_m$$

where Δ_m is defined in (2.6). Finally, we conclude that (A.6) holds by application of Lemma A.4.

Given the asymptotic normality of the score vector in (A.6), proof of the asymptotic normality of $\hat{\beta}^{OP}$ is close at hand. Consider a Taylor Series expansion of (A.5) around β_0 standardized by $T^{-\frac{1}{2}}$:

$$\text{(A.13)} \quad T^{-\frac{1}{2}} \sum_{t=1}^{T} \hat{u}_t^{OP} x_t = T^{-\frac{1}{2}} \sum_{t=1}^{T} u_t x_t + \left[T^{-1} \sum_{t=1}^{T} \bar{\Lambda}_t \right] T^{\frac{1}{2}} (\hat{\beta}^{OP} - \beta_0) = 0$$

where

$$\text{(A.14)} \quad \bar{\Lambda}_t \equiv - \bar{\phi}_t \left[\frac{(y_t - \bar{\Phi}_t)[x_t'\bar{\beta} \cdot \bar{\Phi}_t(1-\bar{\Phi}_t) + \bar{\phi}_t - 2\bar{\Phi}_t\bar{\phi}_t] - \bar{\phi}_t(1-\bar{\Phi}_t)}{\bar{\Phi}_t^2 \cdot (1-\bar{\Phi}_t)^2} \right] x_t x_t' ,$$

$\bar{\Phi}_t \equiv \Phi(x_t'\bar{\beta})$, $\bar{\phi}_t \equiv \phi(x_t'\bar{\beta})$, and $\bar{\beta}$ lies on the line segment joining $\hat{\beta}^{OP}$ and β_0 . Then (A.13) implies

$$\text{(A.15)} \quad T^{\frac{1}{2}} (\hat{\beta}^{OP} - \beta_0) = \left[T^{-1} \sum_{t=1}^{T} \bar{\Lambda}_t \right]^{-1} \left[T^{-\frac{1}{2}} \sum_{t=1}^{T} u_t x_t \right] .$$

As $T \longrightarrow \infty$, the bracketed inverse in (A.15) converges to Δ_0^{-1} , where Δ_0 is defined by (2.6), by application of Lemma 4 of Amemiya (1973), using $\bar{\beta} \xrightarrow{a.s.} \beta_0$ (because $\hat{\beta}^{OP} \xrightarrow{a.s.} \beta_0$) and Assumption 2(a,c,d). Since the asymptotic distribution of $T^{-\frac{1}{2}} \Sigma_t u_t x_t$ has already been derived in (A.6), the conclusion of Theorem 2 follows immediately.

A.3 Asymptotic Properties of GCM Estimators

Proof of Lemma 2: The first-order Taylor series expansion of (3.10) evaluated at $(\hat{\beta},\hat{\delta})$ around (β_0,δ_0) is

$$f_T(\hat{\beta},\hat{\delta}) = 0 = f_T(\beta_0,\delta_0) + [\partial f_T(\bar{\beta},\bar{\delta})/\partial\beta' \;\; \partial f_T(\bar{\beta},\bar{\delta})/\partial\delta'] \begin{bmatrix} \hat{\beta} - \beta_0 \\ \hat{\delta} - \delta_0 \end{bmatrix}$$

$$\text{(A.16)} \qquad \Rightarrow \hat{\beta} - \beta_0 = -[\partial f_T(\bar{\beta},\bar{\delta})/\partial\beta']^{-1} [f_T(\beta_0,\delta_0) - (\partial f_T(\bar{\beta},\bar{\delta})/\partial\delta')(\hat{\delta} - \delta_0)] ,$$

by Lemma 2, (iii), where $(\bar{\beta},\bar{\delta})$ lies on the line segment between $(\hat{\beta},\hat{\delta})$ and (β_0,δ_0). Conditions (i) and (ii) require both $f_T(\beta_0,\delta_0)$ and $\hat{\delta} - \delta_0$ to converge a.s. to zero. Condition (iii) guarantees that these zeroes imply the right-hand side converges to zero, so that the first part of Lemma 2 is proved.

Renormalizing (A.16) by $T^{1/2}$, we have

$$T^{1/2}(\hat{\beta} - \beta_0) = -[\partial f_T(\bar{\beta},\bar{\delta})/\partial\beta']^{-1} [T^{1/2}\cdot f_T(\beta_0,\delta_0) - (\partial f_T(\bar{\beta},\bar{\delta})/\partial\delta') \; T^{1/2}\cdot(\hat{\delta} - \delta_0)] .$$

Now $T^{1/2}\cdot f_T(\beta_0,\delta_0)$ and $T^{1/2}\cdot(\hat{\delta} - \delta_0)$ have limiting distributions, according to Lemma 2, (iv) and (v). The leading matrix inverse converges a.s. to $-[E(\partial f_T(\beta_0,\delta_0)/\partial\beta')]^{-1}$ by (vi). Using (3.10),

$$\partial f_T(\beta,\delta)/\partial\delta' = [\hat{u}(\beta)' \otimes I] \; \partial vec[X(\delta)]/\partial\delta$$

which has zero expectation at $\beta = \beta_0$ because $E[\hat{u}(\beta_0)|X(\delta)] = 0$. Again, (vi) implies that $\partial f_T(\bar{\beta},\bar{\delta})/\partial\delta'$ will converge to zero, so that

$$T^{1/2}(\hat{\beta} - \beta_0) - [-E(\partial f_T(\beta_0,\delta_0)/\partial\beta')]^{-1} T^{1/2} f_T(\beta_0,\delta_0) \xrightarrow{a.s.} 0 .$$

This expression is equivalent to the second result of Lemma 2.

Proof of Theorem 3: First, we apply Lemma 2 to show that $f_T(\beta)$ given by (3.11') has a strongly consistent root when β_0 is replaced by $\hat{\beta}^{OP}$ in the matrix A_n. Then we will apply Lemma 1 to show that a linear approximation to that root is also strongly consistent.

Condition (i) of Lemma 2 is the result of Theorem 1 which gives the strong consistency of $\hat{\beta}^{OP}$. A proof of condition (ii) has the same elements as the proof of Theorem 1. First note that $\lambda_{nt}(\beta_0)$ is α-mixing because (1) x_t is non-stochastic, (2) u_t is α-mixing, and (3) Lemma A.2. Again because x_t is non-stochastic, the elements in the sums of $T^{-1} X'A(\beta_0)\hat{u}(\beta_0)$ are α-mixing. We can therefore apply Lemma A.1 to obtain condition (ii):

$$f_T(\beta,\beta_0,\gamma_0) \longrightarrow 0 \quad a.s. \quad as \quad T \longrightarrow \infty$$

for f_T defined by (3.11). Condition (iii) follows from the compactness of $\mathfrak{B}$ and B, and from the differentiability of A_n and $\hat{u}_n$.

Having demonstrated the consistency for known $A(\beta_0)$, we move on to $A(\hat{\beta}^{OP})$ using Andrew's uniform LLN, Lemma A.3. That is, we wish to show that

$$f_T(\beta,\hat{\beta}^{OP},\gamma_0) \longrightarrow 0 \quad a.s. \quad as \quad T \longrightarrow \infty ,$$

to meet condition (ii) of Lemma 1. Condition (iii) of Lemma A.3 follows from the facts that $A(\beta)$ is continuous and infinitely differentiable on the parameter space B, that $\mathfrak{S}$ is compact, and Assumption 5. Therefore, the function (3.11) satisfies $f_T(\beta,\hat{\beta}^{OP},\gamma_0) - f_T(\beta,\beta_0,\gamma_0) \longrightarrow 0$ a.s. and the (i)-(iii) of Lemma 1 are satisfied, thereby proving the consistency of the GCM estimator in (3.12).

Proof of Theorem 4: As in the proof of Theorem 2, the key to proving the asymptotic normality of the GCM estimator rests in showing that the relevant varaince converges uniformly in the starting point of a time sequence. In the present case, we wish to show that

$$T^{-1}\ \xi' E[\partial \hat{u}_n'/\partial\beta]\ V(\hat{u}_n)^{-1}\ E[\partial \hat{u}_n/\partial\beta']\ \xi$$

converges appropriately, for any fixed $\xi \in \mathbb{R}^K$. We accomplish this by first proving

Lemma A.6: *Given Assumptions 1(a,b,c), 2(a), 3(a), 3'(a,b), 4, and 5, $T^{-1}\ \xi' E[\partial \hat{u}_n'/\partial\beta]\ V(\hat{u}_n)\ E[\partial \hat{u}_n/\partial\beta']\ \xi$ converges uniformly in a to a constant matrix, for any fixed vector $\xi \in \mathbb{R}^K$, where $\hat{u}_n \equiv [\lambda_{nt}; t=a+1,\ldots,a+T]$, λ_{nt} is given by (3.2), and η_{nt} is given by (3.13).*

Proof: This follows from Lemma A.4 where $x_t \equiv X_{nt}$ and z_t is a vector consisting of a one and n^2 indicator variables for the n^2 possible outcomes of y_{nt}. We construct z_t from a base 2 expansion as

$$z_{ti} = \left\{ \begin{array}{ll} 1 & \text{if } i-1 = \prod_{j=1}^{n} 2^{j \cdot y_{ntj}} \\ 0 & \text{if otherwise} \end{array} \right\} \qquad (i=2,\ldots,n^2+1)$$

where y_{ntj} is the j^{th} element of y_{nt} and z_{ti} is the i^{th} element of z_t. The required functions are

$$f(X_{nt}) = g(X_{nt}) \equiv M(X_{nt})\, E[\partial \lambda_{nt}/\partial \beta']$$

where $M(X_{nt}) = [E(u_{nt}|z_{ti}=1)';i=2,\ldots,n^2+1]$, so that $M(X_{nt})'z_t \equiv \lambda_{nt}$ and $[M(X_{nt})'z_t] \equiv \hat{u}_n$. *Q.E.D.*

Lemma A.7: *Given Assumptions 1(a,b,c), 2(a), 3(a), 3'(a,b), 4, and 5, $T^{-1}\, \xi' E[\partial \hat{u}_n'/\partial \beta]\, V(\hat{u}_n)^{-1}\, E[\partial \hat{u}_n/\partial \beta']\, \xi$ converges uniformly in a to a constant matrix, for any fixed vector $\xi \in \mathbb{R}^K$, where $\hat{u}_n \equiv [\lambda_{nt};t=a+1,\ldots,a+T]$, λ_{nt} is given by (3.2), and $\hat{\beta}_{nt}$ is given by (3.13).*

Proof: First, we show that if $\sum_{t=1}^{T} |a_{Tt} w_{Tt}|$ has a limit as $T \longrightarrow \infty$ and $w_{Ti} > \bar{w} > 0$, $\forall\, i$, then $\sum_{t=1}^{T} |a_{Tt}/w_{Tt}|$ also has a limit. According to Cauchy's Theorem, the condition implies

$$\forall\, \epsilon > 0,\ \exists\, T_0 :\ \forall\, T_2 > T_1 > T_0,\ \sum_{t=T_1}^{T_2} |a_{Tt} w_{Tt}| < \epsilon$$

$$\Rightarrow \sum_{t=T_1}^{T_2} |a_{Tt}| < \epsilon/\bar{w},$$

$$\Rightarrow \sum_{t=T_1}^{T_2} |a_{Tt}/w_{Tt}| \leq \sum_{t=T_1}^{T_2} |a_{Tt}|/\bar{w} \leq \epsilon/\bar{w} ,$$

so that the implication follows from a second application of Cauchy's theorem. *Q.E.D.*

The result of the lemma follows from what we have just shown and the result of Lemma A.6. If we decompose $V(\hat{u}_n)$ into to its spectral decomposition $Z_n W_n Z_n'$, where W_n is a diagonal matrix of the eigenvalues of $V(\hat{u}_n)$ and Z_n is a matrix of corresponding eigenvectors, each element of the convergent matrix in Lemma A.6 can be written as

$$\sum_{t=a+1}^{n(a+T-n)} z_t^2 \, w_{Tt}$$

where z_t is the t^{th} element of $Z_n' E[\partial \hat{u}_n/\partial \beta'] \xi$, so that

$$\sum_{t=a+1}^{n(a+T-n)} z_t^2/w_{Tt} = \xi' E[\partial \hat{u}_n'/\partial \beta] \, Z_n A_n^{-1} Z_n' E[\partial \hat{u}_n/\partial \beta'] \, \xi$$

$$= T^{-1} \, \xi' \, E[\partial \hat{u}_n'/\partial \beta] \, V(\hat{u}_n)^{-1} \, E[\partial \hat{u}_n/\partial \beta'] \, \xi$$

is also a convergent series. The convergence remains uniform in a. *Q.E.D.*

Lemma A.8: *Given Assumptions 1(a,b,c), 2(a), 3(a), 3'(a,b), 4, and 5, $\partial f_T(\beta_1,\beta_2,\gamma_0)/\partial \beta_1'$ and $\partial f_T(\beta_1,\beta_2,\gamma_0)/\partial \beta_2'$ converge uniformly in (β_1,β_2) to their expectations, where f_T is defined in (3.11).*

Proof: (1) $\partial f_T(\beta_1,\beta_2,\gamma_0)/\partial\beta_1' = T^{-1} E[\partial\hat{u}_n'/\partial\beta] V(\hat{u}_n)^{-1} \partial\hat{u}_n/\partial\beta'$ which has a convergent expectation according to the previous lemma. That it converges to the limit of its expectation follows from Lemma A.1. That the convergence is uniform follows from Lemma A.3. (2) $\partial f_T(\beta_1,\beta_2,\gamma_0)/\partial\beta_2' = T^{-1} (\hat{u}_n' \otimes E[\partial\hat{u}_n'/\partial\beta]) \partial vec[V(\hat{u}_n)^{-1}]/\partial\beta$ which has expectation zero. As before, in the proofs of Theorems 1 and 3, this object is contains sums of α-mixing processes with bounded moments so that Lemma A.1 establishes a.s. convergence to zero. In addition, Lemma A.3 provides uniform convergence.

Lemma A.9: *Given Assumptions 1(a,b,c), 2(a), 3(a), 3'(a,b), 4, and 5,* $T^{1/2} f_T \xrightarrow{D} N(0,V)$ *as* $T \longrightarrow \infty$, *where*

$$V = plim\ T^{-1} E[\partial\hat{u}_n'/\partial\beta] V(\hat{u}_n)^{-1} E[\partial\hat{u}_n/\partial\beta'] .$$

and $f_T = f_T(\beta_0,\beta_0,\gamma_0)$ *is defined by (3.11).*

Proof: We will apply Lemma A.4. That the elements of f_T are α-mixing $O(m^{-c})$ follows from its construction. Their expectations are zero and higher moments are bounded by our assumptions. Lemma A.7 shows that the variance of $T^{1/2} \xi' f_T$ converges uniformly to a constant. Thus, the requirements of Lemma A.4 are satisfied. *Q.E.D*

It remains to put these elements together to finish the proof of Theorem 4. We can apply Lemma 2 to the consistent root of (3.11) because (1) condition (iv) is given by Theorem 2 for $\delta_0 = \beta_0$ and $\hat{\delta} = \hat{\beta}^{OP}$, (2) condition (v) is given by Lemma A.9, and (3) condition (vi) is given by Lemma A.8. Result (viii) of Lemma 2 also establishes that condition (iv) of Lemma 1 is met when f_T is defined by (3.11'). Condition (v) of Lemma 1 follows from Lemma A.8. Application of Lemma 1 to (3.11'), therefore yields the required asymptotic distribution for the GCM estimator in (3.12).

REFERENCES

Amemiya,T. (1973), "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, 41, 997.

Andrews, D.W.K. (1985), "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers," Discussion Paper, Cowles Foundation, Yale University.

Avery, R.B., L.P. Hansen, and V.J. Hotz (1983), "Multiperiod Probit Models and Orthogonality Condition Estimation," *International Economic Review*, 24, 21-35.

Dagenais, M.G. (1982), "The Tobit Model with Serial Correlation," *Economics Letters*, 10, 263-267.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the *EM* Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39(1), 1-38.

Dickens, W.T. (1980), "Union Representation Elections: Campaign and Vote," Unpublished Ph.D. Dissertation, Massachusetts Institute of Technology, Cambrdige, Massachusetts.

Fair, R.C. (1977), "A Note on the Computation of the Tobit Estimator," *Econometrica*, 40, 497.

Gourieroux, C., A. Monfort and A. Trognon (1984), "Estimation and Test in Probit Models with Serial Correlation," in J.P. Florens, M. Mouchart, J.P. Raoult and L. Simar, eds., *Alternative Approaches to Time Series Analysis*, Proceedings of the Third Franco-Belgian Meeting of Statisticians, November 1982 (Brussels: Publications des Facultés Universitaires Saint-Louis).

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029-1054.

Hartley, M.J. (1974), "On the Estimation of the Probit Model Via Maximum Likelihood Methods," Discussion Paper, Department of Economics, State University of New York at Buffalo.

Heckman, J.J. (1981a), "Statistical Models for Discrete Panel Data," in C.F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge: MIT Press), 114-178.

Heckman, J.J. (1981b), "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," in C.F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge: MIT Press), 178-195..

Jarque, C.M. and A.K. Bera (1982), "Efficient Specification Tests for Limited Dependent Variable Models," *Economics Letters*, 9, 153-160.

Jennrich, R.I. (1969), "Asymptotic Properties of Non-linear Least Squares," *Annals of Mathematical Statistics*, 40, 633-643.

Kiefer, N.M. (1980), "A Note on Switching Regressions and Logistic Discrimination," *Econometrica*, 48, 1065-1069.

Kiefer, N.M, (1982), "Testing for Dependence in Multivariate Probit Models," *Biometrika*, 69, 161-166.

Lehman, E.L. (1983), *Theory of Point Estimation*, New York: Wiley.

Levine, D. (1983), "A Remark on Serial Correlation in Maximum Likelihood," *Journal of Econometrics*, 23(3), 337-342.

Newey, W.K. (1983), "Maximum Likelihood Step Size and the One Step Theorem," Econometric Research Program Research Memorandum No. 308, Princeton University.

Ochi, Y. and R.L. Prentice (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, 71, 531-543.

Poirier, D.J. and P.A. Ruud (1980), "Specification Error in Probit Models," Paper Presented at the NBER-NSF Conference on Decision Rules and Uncertainty, June 1980, Cambridge, Massachusetts.

Poirier, D.J. and P.A. Ruud (1981a), "Limited Dependent Variable Models with Autocorrelation," Paper Presented at the NBER-NSF Conference on Decision Rules and Uncertainty, March 1981, Cambridge, Massachusetts.

Poirier, D.J. and P.A. Ruud (1981b), "Conditional Minimum Distance Estimation in Limited Dependent Variable Models," Paper Presented at the North American Meetings of the Econometric Society, December 1981, Washington, D.C.

Quandt, R.E. (1981), "Autocorrelated Errors in Simple Disequilibrium Models," *Economics Letters*, 7, 55-61.

Robinson, P.M. (1980), "Estimation and Forecasting for Time Series Containing Censored or Missing Observations," in O.D. Anderson, ed., *Time Series* (Amsterdam: North-Holland).

Robinson, P.M. (1982), "On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables," *Econometrica*, 50(1), 27-41.

Robinson, P.M., A.K. Bera, and C.M. Jarque (1985), "Tests for Serial Dependence in Limited Dependent Variable Models," *International Economic Review*, 26(3), 629-638.

Rothenberg, T.J. and C.T. Leenders (1964), "Efficient Estimation of Simultaneous Equation Systems," *Econometrica*, 32, 57-76.

Rudin, W. (1976), *Principles of Mathematical Analysis*, Third Ed., New York: McGraw-Hill.

Ruud, P.A. (1981), "Specification Errors in Limited Dependent Variable Models," Unpublished Ph.D. Dissertation, Massachusetts Institute of Technology, Cambrdige, Massachusetts.

Tallis, G.M. (1961), "The Moment Generating Function of the Truncated Multi-normal Distribution," *Journal of the Royal Statistical Society*, Ser. B, 23(1), 223-229.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1-26.

White, H. (1984a), *Asymptotic Theory for Econometricians*, New York: Academic Press.

White, H. (1984b), "Maximum Likelihood Estimation of Misspecified Dynamic Models," *Misspecification Analysis*, ed. T. Dijkstra, Berlin: Springer-Verlag.