NBER WORKING PAPER SERIES

PANEL DATA

Gary Chamberlain

working Paper NO. 913

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

June 1982

NBER Working Paper #913
June 1982

Panel Data

ABSTRACT

We consider linear predictor definitions of noncausality or strict exogeneity and show that it is restrictive to assert that there exists a time-invariant latent variable c such that x is strictly exogenous conditional on c. A restriction of this sort is necessary to justify standard techniques for controlling for unobserved individual effects. There is a parallel analysis for multivariate probit models, but now the distributional assumption for the individual effects is restrictive. This restriction can be avoided by using a conditional likelihood analysis in a logit model. Some of these ideas are illustrated by estimating union wage effects for a sample of Young Men in the National Longitudinal Survey. The results indicate that the lags and leads could have been generated just by an unobserved individual effect, which gives some support for analysis of covariance-type estimates. These estimates indicate a substantial omitted variable bias. We also present estimates of a model of female labor force participation, focusing on the relationship between participation and fertility. Unlike the wage example, there is evidence against conditional strict exogeneity; if we ignore this evidence, the probit and logit approaches give conflicting results.

Gary Chamberlain
Department of Economics
University of Wisconsin
Madison, Wisconsin 53706
(608) 262-7789

PANEL DATA

TABLE OF CONTENTS

1. INTRODUCTION AND SUMMARY

The paper has four parts: the specification of linear models; the specification of nonlinear models; statistical inference; and empirical applications. The choice of topics is highly selective. We shall focus on a few problems and try to develop solutions in some detail.

The discussion of linear models begins with the following specification:

$$(1.1) \qquad y_{it} = \beta x_{it} + c_i \cdot u_{it} ,$$

$$(1.2) \qquad E(u_{it} | x_{i1}, \ldots . x_{iT}, c_i) = 0 \qquad (i=1, \ldots , N; \; t=1, \ldots , T) .$$

For example, in a panel of farms observed over several years, suppose that y_{it} is a measure of the output of the i^{th} farm in the t^{th} season, x_{it} is a measured input that varies over time, c_i is an unmeasured, fixed input reflecting soil quality and other characteristics of the farm's location, and u_{it} reflects unmeasured inputs that vary over time such as rainfall.

Suppose that data is available on $(x_{i1}, \ldots . x_{iT}, y_{i1}, \ldots . y_{iT})$ for each of a large number of units, but c_i is not observed. A cross-section regression of y_{i1} on x_{i1} will give a biased estimate of β if c is correlated with x, as we would expect it to be in the production function example. Furthermore, with a single cross section, there may be no internal evidence of this bias. If $T > 1$, we can solve this problem given the assumption in (1.2). The change in y satisfies

$$E(y_{i2} - y_{i1} | x_{i2} - x_{i1}) = \beta(x_{i2} - x_{i1}) ,$$

and the least squares regression of $y_{i2} - y_{i1}$ on $x_{i2} - x_{i1}$ provides a

consistent estimator of β (as $N \to \infty$) if the change in x has sufficient variation. A generalization of this estimator when $T > 2$ can be obtained from a least squares regression with individual specific intercepts.

The restriction in (1.2) is necessary for this result. For example, consider the following autoregressive specification:

$$y_{it} = \beta y_{i,t-1} + c_i + u_{it},$$

$$E(u_{it} | y_{i,t-1}, c_i) = 0.$$

It is clear that a regression of $y_{it} - y_{i,t-1}$ on $y_{i,t-1} - y_{i,t-2}$ will not provide a consistent estimator of β, since $u_{it} - u_{i,t-1}$ is correlated with $y_{i,t-1} - y_{i,t-2}$. Hence it is not sufficient to assume that

$$E(u_{it} | x_{it}, c_i) = 0.$$

Much of our discussion will be directed at testing the stronger restriction in (1.2).

Consider the (minimum mean-square error) linear predictor of c_i conditional on $x_{i1}, \ldots, x_{iT}$:

$$E^*(c_i | x_{i1}, \ldots, x_{iT}) = \eta + \lambda_1 x_{i1} + \ldots + \lambda_T x_{iT}. \tag{1.3}$$

Given the assumptions that variances are finite and that the distribution of $(x_{i1}, \ldots, x_{iT}, c_i)$ does not depend upon i, there are no additional restrictions in (1.3); it is simply notation for the linear predictor. Now consider the linear predictor of y_{it} given $x_{i1}, \ldots, x_{iT}$:

$$E^*(y_{it} | x_{i1}, \ldots, x_{iT}) = \zeta_t + \pi_{t1} x_{i1} + \ldots + \pi_{tT} x_{iT}.$$

Form the $T \times T$ matrix $\underset{\sim}{\Pi}$ with π_{ts} as the (t,s) element. Then the restriction in (1.2) implies that $\underset{\sim}{\Pi}$ has a distinctive structure:

$$\underset{\sim}{\Pi} = \beta \underset{\sim}{I} + \underset{\sim}{\ell} \underset{\sim}{\lambda}',$$

where I is the $T \times T$ identity matrix, ℓ is a $T \times 1$ vector of ones, and $\underset{\sim}{A}' = (\lambda_1, \ldots, \lambda_T)$. A test for this structure could usefully accompany estimators of β based on change regressions or on regressions with individual specific intercepts. Moreover, this formulation suggests an alternative estimator for β, which is developed in the inference section.

This test is an **exogeneity** test and it is useful to relate it to **Granger** (1969) and Sims (1972) causality. The novel feature is that we are testing for noncausality conditional on a latent variable. Suppose that $t=1$ is the first period of the individual's (economic) life. Within the linear predictor context, a Granger definition of "y does not cause x conditional on a latent variable c" is

$$(1.9) \quad E^*(x_{i,t+1} | x_{i1}, \cdots, x_{it}, y_{i1}, \ldots, y_{it}, c_i) = E^*(x_{i,t+1} | x_{i1}, \ldots, x_{it}, c_i) \qquad (t=1,2,\ldots).$$

A Sims definition is

$$E^*(y_{it} | x_{i1}, x_{i2}, \cdots, c_i) = E^*(y_{it} | x_{i1}, \cdots, x_{it}, c_i) \quad (t=1,2,\ldots).$$

In fact, these two definitions imply identical restrictions on the **covariance** matrix of $(x_{i1}, \ldots, x_{iT}, y_{i1}, \ldots, y_{iT})$. The **Sims** form fits directly into the $\underset{\sim}{\Pi}$ matrix framework and implies the following restrictions:

$$\underset{\sim}{\Pi} = \underset{\sim}{B} + \underset{\sim}{\gamma} \underset{\sim}{\lambda}',$$

where B is a lower triangular matrix and $\underset{\sim}{\gamma}$ is a $T \times 1$ vector.

We show how these nonlinear restrictions can be transformed into linear restrictions on a standard simultaneous equations model. We show also how a $\underset{\sim}{\gamma}\,\underset{\sim}{\lambda}'$ term can arise in an autoregressive model from the projection of an initial condition onto the x's.

In Section 3 we use a multivariate probit model to illustrate the new issues that arise in models that are nonlinear in the variables. Consider the following specification:

$$\tilde{y}_{it} = \beta\, x_{it} + c_i + u_{it},$$

$$y_{it} = 1 \text{ if } \tilde{y}_{it} \geq 0,$$

$$= 0 \text{ otherwise} \quad (i=1, \ldots, N;\ t=1, \ldots, T),$$

where, conditional on $x_{i1}, \ldots, x_{iT}, c_i$, the distribution of $(u_{i1}, \ldots, u_{iT})$ is multivariate normal $(N(\underset{\sim}{0}, \underset{\sim}{\Sigma}))$ with mean $\underset{\sim}{0}$ and covariance matrix $\underset{\sim}{\Sigma} = (\sigma_{jk})$. We observe $(x_{i1}, \ldots, x_{iT}, y_{i1}, \ldots, y_{iT})$ for a large number of individuals, but we do not observe c_i. For example, in the reduced form of a labor force participation model, y_{it} can indicate whether or not the i^{th} individual worked during period t, x_{it} can be a measure of the presence of young children, and c_i can capture unmeasured characteristics of the individual that are stable at least over the sample period. In the certainty model of Heckman and MaCurdy (1980), c_i is generated by the single life-time budget constraint.

If we treat the c_i as parameters to be estimated, then there is a severe incidental parameter problem. The consistency of the maximum

likelihood estimator requires that $T \to \infty$, but we want to do asymptotic inference with $N \to \infty$ for fixed T, which reflects the sample sizes in the panel data sets we are most interested in. So we consider a random effects estimator, which is based on the following specification for the distribution of c conditional on x:

$$(1.4) \quad c_i = \eta + \lambda_1 x_{i1} + \ldots + \lambda_T x_{iT} + v_i,$$

where the distribution of v_i conditional on $x_{i1}, \ldots, x_{iT}$ is $N(0, \sigma_v^2)$. This is similar to our specification in (1.3) for the linear model, but there is an important difference; (1.3) was just notation for the linear predictor, whereas (1.4) embodies substantive restrictions. We are assuming that the regression function of c on the x's is linear and that the residual variation is homoskedastic and normal. Given these assumptions, our analysis runs parallel to the linear case. There is a matrix $\underset{\sim}{\Pi}$ of multivariate probit coefficients which has the following structure:

$$\underset{\sim}{\Pi} = \text{diag}\{\alpha_1, \ldots, \alpha_T\}[\beta \underset{\sim}{I} + \underset{\sim}{\ell} \underset{\sim}{\lambda}'],$$

where $\text{diag}\{\alpha_1, \ldots, \alpha_T\}$ is a diagonal matrix of normalization factors with $\alpha_t = (\sigma_{tt} + \sigma_v^2)^{-\frac{1}{2}}$. We can impose these restrictions to obtain an estimator of $\alpha_t \beta$ which is consistent as $N \to \infty$ for fixed T. We can also test whether $\underset{\sim}{\Pi}$ in fact has this structure.

A quite different treatment of the incidental parameter problem is possible with a logit functional form for $P(y_{it} = 1 | x_{it}, c_i)$. The sum $\Sigma_{t=1}^{T} y_{it}$ provides a sufficient statistic for c_i. Hence we can use the distribution of $y_{i1}, \ldots, y_{iT}$ conditional on $x_{i1}, \ldots, x_{iT}, \Sigma_t y_{it}$

to obtain a conditional likelihood function that does not depend upon c_i. Maximizing it with respect to β provides an estimator that is consistent as $N \to \infty$ for fixed T, and the other standard properties for maximum likelihood hold as well. The power of the procedure is that it places no restrictions on the conditional distribution of c given x. It is perhaps the closest analog to the change regression in the linear model. A shortcoming is that the residual covariance matrix is constrained to be equicorrelated. Just as in the probit model, a key assumption is

$$(1.5) \qquad P(y_{it} = 1 | x_{i1}, \ldots, x_{iT}, c_i) = P(y_{it} = 1 | x_{it}, c_i),$$

and we discuss how it can be tested.

It is natural to ask whether (1.5) is testable without imposing the various functional form restrictions that underlie our tests in the probit and logit cases. First, some definitions. Suppose that $t = 1$ is the initial period of the individual's (economic) life; an extension of Sims' condition for x to be strictly exogenous is that y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_1, \ldots, x_t$. An extension of Granger's condition for "y does not cause x" is that x_{t+1} is independent of $y1, \ldots, y_t$ conditional on $x_1, \ldots, x_t$. Unlike the linear predictor case, now strict exogeneity is weaker than noncausality. Noncausality requires that y_t be independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_1, \ldots, x_t$ and on $y_1, \ldots, y_{t-1}$. If x is strictly exogenous and in addition y_t is independent of $x_1, \ldots, x_{t-1}$ conditional on x_t, then we shall say that the relationship of x to y is static.

Then our question is whether it is restrictive to assert that there exists a latent variable c such that the relationship of x to y is static conditional on c. We know that this is restrictive in the linear predictor case, since the weaker condition that x be strictly exogenous conditional on c is restrictive. Unfortunately, there are no restrictions when we replace zero partial correlation by conditional independence. It follows that conditional strict exogeneity is restrictive only when combined with specific functional forms -- a truly nonparametric test cannot exist.

Section 4 presents our framework for inference. Let $\underset{\sim}{r}_i' = (1, x_{i1}, \ldots, x_{iT}, y_{i1}, \ldots, y_{iT})$ and assume that $\underset{\sim}{r}_i$ is independent and identically distributed (i.i.d.) for i = 1,2,... . Let $\underset{\sim}{w}_i$ be the vector formed from the squares and cross-products of the elements in $\underset{\sim}{r}_i$. Our framework is based on a simple observation: the matrix $\underset{\sim}{\Pi}$ of linear predictor coefficients is a function of $E(\underset{\sim}{w}_i)$; if $\underset{\sim}{r}_i$ is i.i.d. then so is $\underset{\sim}{w}_i$; hence our problem is to make inferences about a function of a population mean under random sampling. This is straightforward and provides an asymptotic distribution theory for least squares that does not require a linear regression function or homoskedasticity.

Stack the columns of $\underset{\sim}{\Pi}'$ into a vector π and let $\underset{\sim}{\pi} = \underset{\sim}{h}(\underset{\sim}{\mu})$, where $\underset{\sim}{\mu} = E(\underset{\sim}{w}_i)$. Then the limiting distribution for least squares is normal with covariance matrix

$$\underset{\sim}{\Omega} = \frac{\partial \underset{\sim}{h}}{\partial \underset{\sim}{\mu}'} V(\underset{\sim}{w}_i) \frac{\partial \underset{\sim}{h}'}{\partial \underset{\sim}{\mu}} .$$

We impose restrictions on $\underset{\sim}{\Pi}$ by using a minimum distance estimator. The restrictions can be expressed es $\underset{\sim}{\mu} = \underset{\sim}{g}(\underset{\sim}{\theta})$, where $\underset{\sim}{\theta}$ is free to vary

within some set T. Given the sample mean $\underset{\sim}{\bar{w}} = \Sigma_{i=1}^{N} \underset{\sim}{w}_i/N$, we choose $\hat{\underset{\sim}{\theta}}$ to minimize the distance between $\bar{\underset{\sim}{w}}$ and $\underset{\sim}{g}(\underset{\sim}{\theta})$, using the following distance function:

$$\min_{\underset{\sim}{\theta} \in \Gamma} [\bar{\underset{\sim}{w}} - \underset{\sim}{g}(\underset{\sim}{\theta})]' \hat{V}^{-1}(\underset{\sim}{w}_i) [\bar{\underset{\sim}{w}} - \underset{\sim}{g}(\underset{\sim}{\theta})] ,$$

where $\hat{V}(\underset{\sim}{w}_i)$ is a consistent estimator of $V(\underset{\sim}{w}_i)$. This is a generalized least squares estimator for a multivariate regression model with nonlinear restrictions on the parameters; the only explanatory variable is a constant term. The limiting distribution of $\hat{\underset{\sim}{\theta}}$ is normal with covariance matrix

$$\left[\frac{\partial \underset{\sim}{g}'}{\partial \underset{\sim}{\theta}} V^{-1}(\underset{\sim}{w}_i) \frac{\partial \underset{\sim}{g}}{\partial \underset{\sim}{\theta}'} \right]^{-1}$$

An asymptotic distribution theory is also available when we use some matrix other than $\hat{V}^{-1}(\underset{\sim}{w}_i)$ in the distance function. This theory shows that $\hat{V}^{-1}(\underset{\sim}{w}_i)$ is the optimal choice. However, by using suboptimal norms, we can place a number of commonly used estimators within this framework.

The results on efficient estimation have some surprising consequences. The simplest example is a univariate linear predictor: $E^*(y_i | x_{i1}, x_{i2}) = \pi_0 + \pi_1 x_{i1} + \pi_2 x_{i2}$. Consider imposing the restriction that $\pi_2 = 0$; we do not want to maintain any other restrictions, such as linear regression, homoskedasticity, or normality. How shall we estimate π_1? Let $\hat{\underset{\sim}{\pi}}' = (\hat{\pi}_1, \hat{\pi}_2)$ be the estimator obtained from the least squares regression of y on x_1, x_2. We want to find a vector of the form $(\theta, 0)$ as close as possible to $(\hat{\pi}_1, \hat{\pi}_2)$, using $\hat{V}^{-1}(\hat{\underset{\sim}{\pi}})$ in the distance function. Since we are not using the conventional estimator of $V(\hat{\underset{\sim}{\pi}})$, the answer to this

minimization problem is not, in general, to set $\hat{\theta} = b_{yx_1}$, the estimator obtained from the least squares regression of y on x_1. We can do better by using $b_{yx_1} + \tau\hat{\pi}_2$; the asymptotic mean of $\hat{\pi}_2$ is zero if $\pi_2 = 0$, and if b_{yx_1} and $\hat{\pi}_2$ are correlated, then we can choose τ to reduce the asymptotic variance below that of b_{yx_1}.

This point has a direct counterpart in the estimation of simultaneous equations. The restrictions on the reduced form can be imposed using a minimum distance estimator. This is more efficient than conventional estimators since it is using the optimal norm. In addition, there are generalizations of two- and three-stage least squares that achieve this efficiency gain at lower computational cost.

A related application is to the estimation of restricted covariance matrices. Here the assumption to be relaxed is multivariate normality. We show that the conventional maximum likelihood estimator, which assumes normality, is asymptotically equivalent to a minimum distance estimator. But that minimum distance estimator is not. in general, using the optimal norm. Hence there is a feasible minimum distance estimator that is as least es good as the maximum likelihood estimator; it is strictly better in general for nonnormal distributions.

The minimum distance approach has an application to the multivariate probit model of Section 3. We begin by estimating T separate probit specifications in which all leads and lags of x are included in the specification for each y_{it}:

$$P(y_{it} = 1 | x_{i1}, \ldots, x_{iT}) = F(\pi_{to} + \pi_{t1}x_{i1} + \ldots + \pi_{tT}x_{iT}),$$

where F is the standard normal distribution function. Each of the T probit specifications is estimated using a maximum likelihood program for univariate probit analysis. There is some sacrifice of efficiency here, but it may be outweighed by the advantage of avoiding numerical integration. Given the estimator for $\underset{\sim}{\Pi}$, we derive its asymptotic covariance matrix and then impose and test restrictions by using the minimum distance estimator.

Section 5 presents two empirical applications, which implement the, specifications discussed in Sections 2 and 3 using the inference procedures from Section 4. The linear example is based on the panel of Young Men in the National Longitudinal Survey (Parnes); y_t is the logarithm of the individual's hourly wage and x_t includes variables to indicate whether or not the individual's wage is set by collective bargaining; whether or not he lives in an SMSA; and whether or not he lives in the South. We present unrestricted least squares regressions of y_t on x_1, ... xT, and we examine the form of the $\underset{\sim}{\Pi}$ matrix. There are significant leads and lags, but there is evidence in favor of a static relationship conditional on a latent variable; the leads and lags could be interpreted as just due to c, with $E(y_t|x_1, \ldots, xT, c) = \beta x_t + c$. The estimates of β that control for c are smaller in absolute value than the cross-section estimates. The union coefficient declines by 402, with somewhat larger declines for the SMSA and region coefficients.

The second application presents estimates of a model of labor force participation. It is based on a sample of married women in the Michigan Panel Study of Income Dynamics. We focus on the relationship between

participation and the presence of young children. The unrestricted $\underset{\sim}{\Pi}$ matrix for the **probit** specification has significant leads and lags; but, unlike the wage example, there is evidence here that the leads and lags are not generated just by a latent variable. If we do impose this restriction, then the resulting estimator of β indicates that the **cross-**section estimates overstate the negative effect of young children on the woman's participation probability.

The estimates for the **logit** functional form present some interesting contrasts to the **probit** results. The cross-section estimates, as usual, are in close agreement with the **probit** estimates. But when we use the conditional maximum likelihood estimator to control for c, the effect of an additional young child on participation becomes substantially more negative than in the cross-section estimates; so the estimated sign of the bias is **opposite** to that of the **probit** results. Here the estimation method is having a first order effect on the results. There are a variety of possible explanations. It may be that the unrestricted distribution for c in the **logit** form is the key. Or, since there is evidence against the restriction that

$$P(y_{it}|x_{i1}, \ldots . x_{iT}, c_i) = P(y_{it}|x_{it}, c_i),$$

perhaps we are finding that imposing this restriction simply leads to different biases in the **probit** and **logit** estimates.

2. SPECIFICATION AND IDENTIFICATION: LINEAR MODELS

2.1. *A Production Function Example*

We shall begin with a production function example, due to Mundlak (1961).[2] Suppose that a farmer is producing a product with a Cobb-Douglas technology:

$$y_{it} = \beta x_{it} + c_i + u_{it} \qquad (0 < \beta < 1; \quad i=1, \ldots, N; \quad t=1, \ldots, T),$$

where y_{it} is the logarithm of output on the i^{th} farm in season t, x_{it} is the logarithm of a variable input (labor), c_i represents an input that is fixed over time (soil quality), and u_{it} represents a stochastic input (rainfall), which is not under the farmer's control. We shall assume that the farmer knows the product price (P) and the input price (W), which do not depend on his decisions, and that ha knows ci. The factor input decision, however, is made before knowing u_{it}, and we shall assume that x_{it} is chosen to maximize expected profits. Then the factor demand equation is

$$(2.1) \qquad x_t = \{\ell n\, \beta + \ell n\, [E(e^{u_t}|J_t)] + \ell n\, (P_t/W_t) + c\}/(1-\beta),$$

where J_t is the information set available to the farmer when he chooses x_t, and we have suppressed the i subscript.

Assume first that u_t is independent of Jt, so that the farmer cannot do better than using the unconditional mean. In that case we have

$$E(y_t|x_1, \ldots, x_T, c) = \beta x_t + c.$$

So if c is observed, only one period of data is needed: the least squares regression of Y_1 on x_1, c provides a consistent estimator of β as $N \to \infty$.

Now suppose that c is not observed by the econometrician, although it is known to the farmer. Consider the least squares regression of y_1 on x_1, using just a single cross-section of the data. The population counterpart is

$$E^*(y_1|x_1) = \pi_o + \pi x_1,$$

where E^* is the minimum mean-square error linear predictor (the wide-sense regression function):

$$\pi = Cov(y_1, x_1)/V(x_1), \ \pi_o = E(y_1) - \pi E(x_1).$$

We see from (2.1) that c and x_1 are correlated; hence $\pi \neq \beta$ and the least squares estimator of β does not converge to β as $N \to \infty$. Furthermore, with a single cross section, there may be no internal evidence of this omitted-variable bias.

Now the panel can help to solve this problem. Mundlak's solution was to include farm specific indicator variables: a least squares regression of y_{it} on x_{it}, $\underset{\sim}{d}_{it}$ (i=1, . . . , N; t=1, ... T), where $\underset{\sim}{d}_{it}$ is a $N \times 1$ vector of zeros except for a one in the i^{th} position. So this solution treats the c_i as a set of parameters to be estimated. It is a "fixed effects" solution, which we shall contrast with "random effects." The distinction is that under a fixed effects approach, we condition on the c_i, so that their distribution plays no role. A random effects approach invokes a distribution for c. In a Bayesian framework, β and the c_i would be treated symmetrically, with a prior distribution for both. Since I am only going to

use asymptotic results on inference, however, a "gentle" prior distribution for β will be dominated. That this need not be true for the c_i is one of the interesting aspects of our problem.

We shall do asymptotic inference as N tends to infinity for fixed T. Since the number of parameters (c_i) is increasing with sample size, there is a potential "incidental parameters" problem in the fixed effects approach. This does not, however, pose a deep problem in our example. The least squares regression with the indicator variables is algebraically equivalent to the least squares regression of $y_{it} - \bar{y}_i$ on $x_{it} - \bar{x}_i$ (i=1, ... N; t=1, ... T), where $\bar{y}_i = \Sigma_{t=1}^{T} y_{it}/T$, $\bar{x}_i = \Sigma_{t=1}^{T} x_{it}/T$. If T = 2, this reduces to a least squares regression of $y_{i2} - y_{i1}$ on xi2 - x_{i1}. Since

$$E(y_{i2} - y_{i1} | x_{i2} - x_{i1}) = \beta(x_{i2} - x_{i1}),$$

the least squares regression will provide a consistent estimator of β if there is sufficient variation in $x_{i2} - x_{i1}$.[3]

2.2 *Fixed Effects and Incidental Parameters*

The incidental parameters can create real difficulties. Suppose that u_{it} is independently and identically distributed (i.i.d.) across farms and periods with $V(u_{it}) = \sigma^2$. Then under a normality assumption, the maximum likelihood estimator of σ^2 converges (almost surely) to $\sigma^2(T-1)/T$ as $N \to \infty$ with T fixed.[4] The failure to correct for degrees of freedom leads to a serious inconsistency when T is small. For another example, consider the following autoregression:

$$y_{i1} = a\, y_{io} + c_i + u_{i1},$$

$$y_{i2} = \beta\, y_{i1} + c_i + u_{i2}.$$

Assume that u_{i1} and u_{i2} are i.i.d. conditional on y_{io} and ci, and that they follow a normal distribution ($N(0, \sigma^2)$). Consider the likelihood function corresponding to the distribution of (y_{i1}, y_{i2}) conditional on y_{io} and c_i. The log-likelihood function is quadratic in β, cl, ... c_N (given σ^2), and the maximum likelihood estimator of β is obtained from the least squares regression of $y_{i2} - y_{i1}$ on $y_{i1} - y_{io}$ (i-1, ... N). Since u_{i1} is correlated with y_{i1}, and

$$y_{i2} - y_{i1} = \beta(y_{i1} - y_{io}) + u_{i2} - u_{i1},$$

it is clear that

$$E(y_{i2} - y_{i1} | y_{i1} - y_{io}) \neq \beta(y_{i1} \quad y_{io}), \tag{2.9}$$

and the maximum likelihood estimator of β is not consistent. If the distribution of yio conditional on c_i does not depend on β or c_i, then the likelihood function based on the distribution of $(y_{io}$, yil, $y_{i2})$ conditional on c_i gives the same inconsistent maximum likelihood estimator of β. If the distribution of (yio, y_{i1}, y_{i2}) is stationary, then the estimator obtained from the least squares regression of $y_{i2} - y_{i1}$ on $y_{i1} - y_{io}$ *converges, as* $N \to \infty$, *to* $(\beta-1)/2$.[5]

2.3. *Random Effects and Specification Analysis*

We have seen that the success of the fixed effects estimator in the production function example must be viewed with some caution. The inci-

dental parameter problem will be even more serious when we consider nonlinear models. So we shall consider next a random effects treatment of the production function example; this will also provide a convenient framework for specification analysis.[6]

Assume that there is some joint distribution for $(x_{i1}, \ldots, x_{iT}, c_i)$, which does not depend upon i, and consider the regression function that does not condition on c:

$$E(y_{it}|x_{i1}, \ldots, x_{iT}) = \beta x_{it} + E(c_i|x_{i1}, \ldots, x_{iT}).$$

The regression function for c_i given $\underset{\sim}{x}_i \equiv (x_{i1}, \ldots, x_{iT})$ will generally be some nonlinear function. But we can specify a minimum mean-square error linear predictor:'

$$(2.2) \qquad E^*(c_i|x_{i1}, \ldots, x_{iT}) = \psi + \lambda_1 x_{i1} + \ldots + \lambda_T x_{iT} = \psi + \underset{\sim}{\lambda}' \underset{\sim}{x}_i,$$

where $\lambda = V^{-1}(\underset{\sim}{x}_i)\ \mathrm{Cov}(\underset{\sim}{x}_i, c_i)$. No restrictions are being imposed here -- (2.2) is simply giving our notation for the linear predictor.

Now we have

$$E^*(y_{it}|\underset{\sim}{x}_i) = \psi + \beta x_{it} + \underset{\sim}{\lambda}' \underset{\sim}{x}_i.$$

Combining these linear predictors for the T periods gives the following multivariate linear predictor:[8]

$$(2.3) \qquad E^*(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \underset{\sim}{\pi}_o + \underset{\sim}{\Pi}\ \underset{\sim}{x}_i,$$

$$\underset{\sim}{\Pi} = \mathrm{Cov}(\underset{\sim}{y}_i, \underset{\sim}{x}_i')\ V^{-1}(\underset{\sim}{x}_i) = \beta\, \underset{\sim}{I} + \underset{\sim}{\ell}\underset{\sim}{\lambda}',$$

where $\underset{\sim}{y}_i' = (y_{i1}, \ldots, y_{iT})$, $\underset{\sim}{I}$ is the $T \times T$ identity matrix, and $\underset{\sim}{\ell}$ is a $T \times 1$ vector of ones.

The $\underset{\sim}{\Pi}$ matrix is a useful tool for analyzing this model. Consider first the estimation of δ: if $T = 2$ we have

$$\underset{\sim}{\Pi} = (\pi_{jk}) = \begin{pmatrix} \beta + \lambda_1 & \lambda_2 \\ \lambda_1 & \beta + \lambda_2 \end{pmatrix}.$$

Hence

$$\beta = \pi_{11} - \pi_{21} = \pi_{22} - \pi_{12}.$$

So given a consistent estimator for $\underset{\sim}{\Pi}$, we can obtain a consistent estimator for δ. The estimation of $\underset{\sim}{\Pi}$ is almost a standard problem in multivariate regression; but, due to the nonlinearity in $E(c_i|\underset{\sim}{x}_i)$, we are estimating only a wide-sense regression function, and some care is needed. It turns out that there is a way of looking at the problem which allows a straightforward treatment, under very weak assumptions. We shall develop this in the section on inference.

We see in (2.3) that there are restrictions on the $\underset{\sim}{\Pi}$ matrix. The off-diagonal elements within the same column of $\underset{\sim}{\Pi}$ are all equal. The T^2 elements of $\underset{\sim}{\Pi}$ are functions of the $T+1$ parameters $\beta, \lambda_1, \ldots, \lambda_T$. This suggests an obvious specification test. Or, backing up a bit, we could begin with the specification that $\underset{\sim}{\Pi} = \beta \underset{\sim}{I}$. Then passing to (2.3) would be a test for whether there is a time-invariant omitted variable that is correlated with the x's. The test of $\Pi = \beta \underset{\sim}{I} + \underset{\sim}{\ell}\underset{\sim}{\lambda}'$ against an unrestricted $\underset{\sim}{\Pi}$ would be an omnibus test of a variety of misspecifications, some of which will be considered next.[9]

Suppose that there is serial correlation in u, with $u_t = \rho u_{t-1} + w_t$, where w_t is independent of J_t and we have suppressed the i subscripts. Now we have

$$E(e^{u_t}|J_t) = e^{\rho u_{t-1}} E(e^{w_t}).$$

So the factor demand equation becomes

$$x_t = \{\ell n\ \beta + \ell n\ [E(e^{u_t})] + \ell n\ (P_t/W_t) + \rho u_{t-1} + c\}/(1-\beta)$$

Suppose that there is no variation in prices across the farms, so that the P_t/W_t term is captured in period specific intercepts, which we shall suppress. Then we have

$$E^*(y_t|x_1, \ldots . x_T) = \beta x_t + (1-\rho^{-1})(\lambda_1 x_1 + \ldots + \lambda_T x_T) + \varphi x_{t+1},$$

where $\varphi = \rho^{-1}(1-\beta)$.

So the $\underset{\sim}{\Pi}$ matrix would indicate a distributed lead, even after controlling for c. If instead there is a first order moving average, $u_t = w_t + \rho w_{t-1}$, then

$$E(e^{u_t}|J_t) = e^{\rho w_{t-1}} E(e^{w_t}),$$

and a bit of algebra gives

$$E(y_t|x_1, \ldots, x_T) = x_t - \rho^{-1}(\lambda_1 x_1 + \ldots + \lambda_T x_T) + \varphi x_{t+1}.$$

Once again there is a distributed lead, but now β is not identified from the Π matrix.

2.4. *A Consumer Demand Example*

2.4.a. *Certainty*

We shall follow Ghez and Becker **(1975), Heckman** and **MaCurdy (1980),** and **MaCurdy (1981)** in presenting a life-cycle model under certainty. suppose that the consumer is maximizing

$$V = \sum_{t-1}^{\tau} \rho^{(t-1)} U_t(C_t)$$

subject to

$$\sum_{t=1}^{\tau} \gamma^{-(t-1)} P_t C_t \leq B, C_t \geq 0 \ (t=1, \ldots, \tau),$$

where $\rho^{-1} - 1$ is the rate of time preference, Y-1 is the (nominal) interest rate, C_t is **consumption** in period t, P_t is the price of the consumption good in period t, and B is the present value in the initial period of lifetime income. In this certainty model, the **consumer faces** a single lifetime budget constraint.

If the optimal consumption is positive in every period, then

$$U_t'(C_t) = (\gamma\rho)^{-(t-1)}(P_t/P_1)U_1'(C_1).$$

A convenient functional **form** is $U_t(C) = A_t C^{\delta}/\delta$ $(A_t > 0, \ \delta < 1)$; then we have

$$(2.4) \qquad y_t = \beta x_t + \varphi(t-1) + c + u_t ,$$

where yt = ℓn Ct, $x_t = \ell n$ Pt, $c = (\delta-1)^{-1} \ell n [U_1'(C_1)/P_1]$,

$u_t = (1-\delta)^{-1} \ell n$ At, $\beta = (\delta-1)^{-1}$, and $\varphi = (1-\delta)^{-1} \ell n (\gamma\rho)$. Note that c is determined by the marginal utility of initial wealth: $U_1'(C_1)/P_1 = \partial V/\partial B$.

We shall assume that At is not observed by the econometrician, and that it is independent of the P's. Then the model is similar to the production function example if there is price variation across consumers as well as over time. There will generally be correlation between c and $(x_1, \ldots . x_T)$. As before we have the prediction that $\underset{\sim}{\Pi} = \beta \, \underset{\sim}{\mathrm{I}} + \underset{\sim}{\ell} \, \underset{\sim}{\lambda}'$, which is testable. A consistent estimator of β can be obtained with only two periods of data since

$$(2.25) \qquad y_t - y_{t-1} = \beta (x_t - x_{t-1}) + \varphi + u_t - u_{t-1}.$$

We shall see next how these results are affected when we allow for some uncertainty.

2.4.b. Uncertainty

We shall present a highly simplified model in order to obtain some explicit results in the uncertainty case. The consumer is maximizing

$$E[\sum_{t=1}^{\tau} P^{t-1} U_t(C_t)]$$

subject to

$$P_1 C_1 + S_1 \leq B,$$

$$P_t C_t + S_t \leq \gamma S_{t-1}, \; C_t \geq 0, \; S_t \geq 0 \; (t=1, \ldots . \tau).$$

The only source of uncertainty is the future prices. The consumer is allowed to borrow against his future income, which has a present value of B in the initial period. The consumption plan must have C_t a function only of information available at date t.

It is convenient to set $\tau = \infty$ and to assume that P_{t+1}/P_t is i.i.d. $(t=1,2, \ldots)$. If $U_t(C) = A_t C^{\delta}/\delta$, then we have the following optimal plan:[10]

$$(2.5) \qquad C_1 = d_1 B/P_1, \; S_1 = (1-d_1)\, B,$$

$$C_t = d_t \gamma S_{t-1}/P_t, \; S_t = (1-d_t)\, \gamma S_{t-1} \qquad (t=2,3, \ldots).$$

where

$$d_t = [1 + f_{t+1} + (f_{t+1} f_{t+2}) + \ldots]^{-1},$$

$$f_t = (\rho \kappa\, A_t/A_{t-1})^{[1/(1-\delta)]}, \quad \kappa = \gamma^{\delta} E[(P_{t-1}/P_t)^{\delta}].$$

It follows that

$$y_t - y_{t-1} = (-1)(x_t - x_{t-1}) + \zeta + u_t - u_{t-1},$$

where y, x, u are defined as in (2.4) and $\zeta = (1-\delta)^{-1} \ell n\, (\rho\kappa) + \ell n\, \gamma$.

We see that, in this particular example, the appropriate interpretation of the change regression is very sensitive to the amount of information available to the consumer. In the uncertainty case, a regression of $(\ell n\, C_t - \ell n\, C_{t-1})$ on $(\ell n\, P_t - \ell n\, P_{t-1})$ does not provide a consistent estimator of $(\delta-1)^{-1}$; in fact, the estimator converges to -1, with the implied estimator of δ converging to 0.

2.a.c. *Labor Supply*

We shall consider a certainty model in which the consumer is maximizing

$$(2.6) \qquad v = \sum_{t=1}^{\tau} \rho^{(t-1)} U_t(C_t, L_t)$$

subject to

$$\sum_{t=1}^{\tau} \gamma^{-(t-1)} (P_t C_t + W_t L_t) \leq B + \sum_{t-1}^{\tau} \gamma^{-(t-1)} W_t \bar{L},$$

$$C_t \geq 0, \; 0 \leq L_t \leq \bar{L} \quad (t=1, \ldots . \tau),$$

where L_t is leisure, W_t is the wage rate, B is the present value in the initial period of nonlabor income, and $\bar{L}$ is the time endowment. We shall assume that the inequality constraints on L are not binding; the participation decision will be discussed in tbe section on nonlinear models. If U_t is additively separable,

$$U_t(C, L) = U_t^*(C) + \tilde{U}_t(L),$$

and if $\tilde{U}_t(L) = A_t L^{\delta}/\delta$, then we have

$$(2.7) \qquad y_t = \beta x_t + \varphi(t-1) + c + u_t,$$

where $y_t = \ell n \, L_t$, $x_t = \ell n \, W_t$, $c = (\delta-1)^{-1} \ell n \, [\tilde{U}_1'(L_1)/W_1]$, $u_t = (1-\delta)^{-1} \ell n \, A_t$, $\beta = (\delta-1)^{-1}$, and $\varphi = (1-6)^{-1} \ell n \, (\gamma\rho)$. Once again c is determined by the marginal utility of initial wealth: $\tilde{U}_1'(L_1)/W_1 = \partial V/\partial B$.

We shall assume that A_t is not observed by the econometrician. There will generally be a correlation between c and $(x_1, \ldots . x_T)$, since L_1 depends

hypothesis in (2.8) implies that if $T \geq 4$, there are $(T-3)(T-2)/2$ over-identifying restrictions.

Consider next a Granger definition of "y does not cause x conditional on c":

$$E^*(x_{t+1}|x_1, \ldots, x_t, y_1, \ldots, y_t, c) = E^*(x_{t+1}|x_1, \ldots, x_t, c) \quad (t=1, \ldots, T-1). \tag{2.10}$$

Define the following linear predictors:

$$x_{t+1} = \psi_{t1}x_1 + \ldots + \psi_{tt}x_t + \varphi_{t1}y_1 + \ldots + \varphi_{tt}y_t + \zeta_{t+1}c + v_{t+1},$$

$$E^*(v_{t+1}|x_1, \ldots, x_t, y_1, \ldots, y_t, c) = 0 \quad (t=1, \ldots, T-1).$$

Then (2.10) is equivalent to $\varphi_{ts} = 0$. We can rewrite the system, imposing $\varphi_{ts} = 0$, as follows:

$$x_{t+1} = \tilde{\psi}_{t1}x_1 + \ldots + \tilde{\psi}_{t,t-1}x_{t-1} + \tau_t x_t + v_{t+1}, \tag{2.11}$$

$$\tilde{\psi}_{ts} = \psi_{ts} - (\zeta_{t+1}/\zeta_t)\psi_{t-1,s}, \quad \tau_t = \psi_{tt} + (\zeta_{t+1}/\zeta_t),$$

$$v_{t+1} = v_{t+1} - (\zeta_{t+1}/\zeta_t)v_t, \quad E(x_s \tilde{v}_{t+1}) = E(y_s \tilde{v}_{t+1}) = 0$$

$$(s < t-1; \; t-2, \ldots, T-1).$$

In the equation for x_{t+1}, there are t unknown parameters, $\psi_{t1}, \ldots, \psi_{t,t-1}, \tau_t$, and $2(t-1)$ orthogonality conditions. Hence there are t-2 restrictions $(3 \leq t \leq T-1)$.

It follows that the Granger condition for "y does not cause x conditional on c" implies $(T-3)(T-2)/2$ restrictions, which is the same number of restrictions implied by the Sims condition. In fact, it is a consequence of Sims' (1972) theorem, as extended by Hosoya (1977), that the two sets of

restrictions are equivalent; this is not immediately obvious from a direct comparison of (2.9) and (2.11).

In terms of the Π matrix, conditional strict exogeneity implies that

$$\underset{\sim}{\Pi} = \underset{\sim}{B} + \underset{\sim}{\gamma}\,\underset{\sim}{\lambda}',$$

$$\underset{\sim}{B} = \begin{bmatrix} \beta_{11} & 0 & \cdots & 0 \\ \beta_{21} & \beta_{22} & 0 & \cdots & 0 \\ \vdots & & & & \\ \beta_{T1} & \beta_{T2} & \cdots & & \beta_{TT} \end{bmatrix}, \quad \underset{\sim}{\gamma} = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_T \end{pmatrix}, \quad \underset{\sim}{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_T \end{pmatrix}.$$

These nonlinear restrictions can be imposed and tested using the minimum distance estimator to be developed in the inference section. Alternatively, we can use the transformations in (2.9) or in (2.11). These transformations give us "simultaneous equations" systems with linear restrictions; (2.9) can be estimated using three-stage least squares. A generalization of three-stage least squares, which does not require homoskedasticity assumptions, is developed in the inference section. It is asymptotically equivalent to imposing the nonlinear restrictions directly on $\underset{\sim}{\Pi}$, using the minimum distance estimator.

2.6. *Lagged Dependent Variables*

For a specific example, write the labor supply model in (2.7) as follows:

$$(2.12) \qquad y_t = \delta_1 x_t + \delta_2 x_{t-1} + \delta_3 y_{t-1} + v_t,$$

$$E^*(v_t | x_1, \ldots, x_T) = 0 \quad (t=1, \ldots, T);$$

this reduces to (2.7) if $\delta_2 = -\delta_1$ and $\delta_3 = 1$. If we assume that $v_t = w + e_t$,

upon wages in all periods. If At is independent of the W's, then we have the prediction that $\underset{\sim}{\Pi} = \beta \underset{\sim}{I} + \underset{\sim}{\ell}\underset{\sim}{\lambda}'$. If, however, wages are partly determined by the quantity of previous work experience, then there will be lags and leads in addition to those generated by c, and $\underset{\sim}{\Pi}$ will not have this simple structure[11].

It would be useful at this point to extend the uncertainty model to incorporate uncertainty about future wages. Unfortunately, a comparably simple explicit solution is not available. But we may conjecture that the correct interpretation of a regression of $(\ell n\ L_t - \ell n\ L_{t-1})$ on $(\ell n\ W_t - \ell n\ W_{t-1})$ is also sensitive to the amount of Information available to the consumer.

2.5. *Strict Exogeneity Conditional on a Latent Variable*

We shall relate the specification analysis of $\underset{\sim}{\Pi}$ to the causality definitions of Granger (1969) and Sims (1972). Consider a sample in which t=1 is the first period of the individual's (economic) life.[12] A Sims definition of "x is strictly exogenous" is

$$E^*(y_t|x_1,x_2,\ldots) = E^*(y_t|x_1,\ldots,x_t) \qquad (t=1,2,\ldots).$$

In this case $\underset{\sim}{\Pi}$ is lower triangular: he elements above the main diagonal are all zero. This fails to hold in the models we have been considering, due to the omitted variable c. But, in some cases, we do have the following property:

$$(2.8) \qquad E^*(y_t|x_1,x_2,\ldots,c) = E^*(y_t|x_1,\ldots,x_t,c) \quad (t=1,2,\ldots)$$

It was stressed by Granger (1969) that the assessment of noncausality depends crucially on what other variables are being conditioned on. The

novel feature of (2.8) is that we are asking whether there exists some latent variable (c) such that x is strictly exogenous conditional on c. The question is not vacuous since c is restricted to be time invariant.

Let us examine what restrictions are implied by (2.8). Define the following linear predictors:[13]

$$y_t = \beta_{t1} x_1 + \ldots + \beta_{tT} x_T + \gamma_t c + u_t,$$

$$E^*(u_t | x_1, \ldots, x_T, c) = 0 \qquad (t=1, \ldots, T).$$

Then (2.8) is equivalent to $\beta_{ts} = 0$ for $s > t$. If $\gamma_1 \neq 0$, we can choose a scale normalization for c such that $\gamma_1 = 1$. Then we can rewrite the system with $\beta_{ts} = 0$ $(s > t)$ as follows:

$$(2.9) \qquad y_t = \tilde{\beta}_{t1} x_1 + \beta_{t2} x_2 + \ldots + \beta_{tt} x_t + \gamma_t y_1 + \tilde{u}_t,$$

$$\tilde{\beta}_{t1} = \beta_{t1} - \gamma_t \beta_{11}, \quad \tilde{u}_t = u_t - \gamma_t u_1,$$

$$E(x_s u_t) = 0 \quad (s=1, \ldots, T; \; t=2, \ldots, T)$$

Consider the "instrumental variable" orthogonality conditions implied by $E(x_s \tilde{u}_t) = 0$. In the y_T equation, we have T+1 unknown coefficients: $\tilde{\beta}_{T1}, \beta_{T2}, \ldots, \beta_{TT}, \gamma_T$, and T orthogonality conditions. So these coefficients are not identified. In the y_{T-1} equation, however, we have just enough orthogonality conditions; and in the y_{T-j} equation ($j \leq T-2$), we have j-1 more than we need since there are T-j+1 unknown coefficients: $\tilde{\beta}_{T-j,1}, \beta_{T-j,2}, \ldots, \beta_{T-j,T-j}, \gamma_{T-j}$, and T orthogonality conditions: $E(x_s \tilde{u}_{T-j}) = 0$ $(s=1, \ldots, T)$. It follows that, subject to a rank condition, we can identify β_{ts}, γ_t, and $\tilde{\beta}_{t1}$ for $2 \leq s \leq t \leq T-1$. In addition the

where w is uncorrelated with the x's and e_t is i.i.d. and uncorrelated with the x's and w, then we have the autoregressive, variance-components model of Balestra and Nerlove (1966).[14] In keeping with our general approach, we shall avoid placing restrictions on the serial correlation structure of v_t, our inference procedures will be based on the strict exogeneity condition that $E^*(v_t|x_1, \ldots, x_T) = 0$.

We can fit this model into the $\underset{\sim}{\Pi}$ matrix framework by using recursive substitution to obtain the reduced form:

$$y_t = \beta_{t1}x_1 + \ldots + \beta_{tt}x_t + \gamma_t c + u_t ,$$

$$E^*(u_t|x_1,\ldots,x_T) = 0,$$

where

$$\beta_{ts} = (\delta_2 + \delta_3\delta_1)\delta_3^{t-s-1}, \quad \beta_{tt} = \delta_1, \quad \gamma_t = \delta_3^{t-1},$$

$$c = \delta_2 x_o + \delta_3 y_o, \quad u_t = v_t + \delta_3 v_{t-1} + \ldots + \delta_3^{t-1} v_1$$

$$(1 \leq s \leq t\text{-}1, \; t=1, \ldots, T).$$

(We are assuming that (2.12) holds for $t \geq 1$, but data on (x_o, y_o) are not available.)

Hence this model satisfies the conditional strict exogeneity restrictions,

$$\underset{\sim}{\Pi} = \underset{\sim}{B} + \underset{\sim}{\gamma}\,\underset{\sim}{\lambda}',$$

where B is lower triangular. The $\underset{\sim}{\gamma}\,\underset{\sim}{\lambda}'$ term is generated by the projection of the initial condition $(\delta_2 x_o + \delta_3 y_o)$ on xl, ... x_T.[15]

Estimation can proceed by using the minimum distance procedure to impose the nonlinear restrictions on $\underset{\sim}{\Pi}$. Alternatively, we can complete the system

in (2.12) with

$$y_1 = \varphi_1 x_1 + \ldots + \varphi_T x_T + v_1;$$

this is just notation for the identity

$$y_1 = E^*(y_1 | x_1, \ldots, x_T) + [y_1 - E^*(y_1 | x_1, \ldots, x_T)].$$

Then we can apply the generalized three-stage least squares estimator to be developed in the inference section. It achieves the same limiting distribution at lower computational coat, since the restrictions in this form are linear and can be imposed without requiring iterative optimization techniques.

Now consider a second order autoregression:

$$y_t = \delta_1 x_t + \delta_2 x_{t-1} + \delta_3 y_{t-1} + \delta_4 y_{t-2} + v_t,$$

$$E^*(v_t | x_1, \ldots, x_T) = 0 \qquad (t=1, \ldots, T).$$

Recursive substitution gives

$$y_t = \beta_{t1} x_1 + \ldots + \beta_{tt} x_t + \gamma_{t1} c_1 + \gamma_{t2} c_2 + u_t,$$

$$E^*(u_t | x_1, \ldots, x_T) = 0 \qquad (t=1, \ldots, T),$$

where

$$c_1 = \delta_2 x_o + \delta_3 y_o + \delta_4 y_{-1}, \quad c_2 = y_o,$$

and there are nonlinear restrictions on the parameters. . The $\underset{\sim}{\Pi}$ matrix has the following form:

$$\underset{\sim}{\Pi} = \underset{\sim}{B} + \underset{\sim}{\gamma}_1\underset{\sim}{\lambda}_1' + \underset{\sim}{\gamma}_2\underset{\sim}{\lambda}_2' ,$$

where B is lower triangular, $\underset{\sim}{\gamma}_j' = (\gamma_{1j}, \ldots, \gamma_{Tj})$, and $E^*(c_j|\underset{\sim}{x}) = \underset{\sim}{\lambda}_j'\underset{\sim}{x}$ (j=1,2)

This specification suggests a natural extension of the conditional strict **exogeneity** idea, with the conditioning set indexed by the number of latent variables. We shall say that "**x** is strictly exogenous conditional on c_1, c_2" if

$$E^*(y_t|\ldots, x_{t-1}, x_t, x_{t+1}, \ldots . c_1, c_2) = E^*(y_t|x_t, x_{t-1}, \ldots, c_1, c_2).$$

We can also introduce a **Granger** version of this condition and generalize the analysis in Section 2.5.

Serial Correlation or *Partial Adjustment?*

Griliches' (1967) considered the problem of distinguishing between the following two **models:** a partial adjustment model,[16]

$$(2.13) \qquad y_t = \beta x_t + \gamma y_{t-1} + v_t ,$$

and a model with no structural lagged **dependent variable** but with a residual following a first order **Markov** process:

$$(2.14) \qquad y_t = \beta x_t + u_t,$$

$$u_t = \rho u_{t-1} + e_t, \; e_t \text{ i.i.d.};$$

in both **cases x** is strictly exogenous:

$$E^*(v_t|x_1, \ldots . x_T) = E^*(u_t|x_1, \ldots . x_T) = 0 \quad (t=1, \ldots . T).$$

In the serial correlation case, we have

$$y_t = \beta x_t - \rho\beta x_{t-1} + \rho y_{t-1} + e_t ;$$

as Griliches observed, the least squares regression will have a distinctive pattern -- the coefficient on lagged x equals (as $N \to \infty$) minus the product of the coefficients on current x and lagged y.

I want to point out that this prediction does not rest on the serial correlation structure of u. It is a direct implication of the assumption that u is uncorrelated with $x_1, \ldots, x_T$:

$$\begin{aligned} E^*(y_t|x_t, x_{t-1}, y_{t-1}) &= \beta x_t + E^*(u_t|x_t, x_{t-1}, y_{t-1}) \\ &= \beta x_t + E^*(u_t|u_{t-1}) \\ &= \beta x_t + \varphi_t u_{t-1} \\ &= \beta x_t - \varphi_t\beta x_{t-1} + \varphi_t y_{t-1}. \end{aligned}$$

Here $\varphi_t u_{t-1}$ is simply notation for the linear predictor. In general u_t is not a first order process ($E^*(u_t|u_{t-1}, u_{t-2}) \neq E^*(u_t|u_{t-1})$), but this does not affect our argument.

Within the $\underset{\sim}{\Pi}$ matrix framework, the distinction between the two models is that (2.14) implies a diagonal $\underset{\sim}{\Pi}$ matrix, with no distributed lag, whereas the partial adjustment specification in (2.13) implies that $\underset{\sim}{\Pi} = \underset{\sim}{B} + \underset{\sim}{Y}\underset{\sim}{\lambda}'$, with a distributed lag in the lower triangular $\underset{\sim}{B}$ matrix and a rank one set of lags and leads in $\underset{\sim}{Y}\underset{\sim}{\lambda}'$.

We can generalize the serial correlation model to allow for an individual specific effect that may be correlated with x:

$$y_t = \beta x_t + c + u_t,$$

$$E^*(u_t | x_1, \ldots . x_T) = 0.$$

Now both the serial correlation and the partial adjustment models have a rank one set of lags and leads in $\underset{\sim}{\Pi}$, but we can distinguish between them because only the partial adjustment model has a distributed lag in the $\underset{\sim}{B}$ matrix. So the absence of structural lagged dependent variables is signalled by the following special case of conditional strict exogeneity:

$$E^*(y_t | x_1, \ldots, x_T, c) = E^*(y_t | x_t, c).$$

In this case the relationship of x to y is "static" conditional on c. We shall pursue this distinction in nonlinear models in Section 3.3.

2.7. *Residual Covariances: Heteroskedasticity and Serial Correlation*

2.7.a. *Heteroskedasticity*

If $E(c_i | \underset{\sim}{x}_i) \neq E^*(c_i | \underset{\sim}{x}_i)$, then there will be heteroskedasticity, since the residual will contain $E(c_i | \underset{\sim}{x}_i) - E^*(c_i | \underset{\sim}{x}_i)$. Another source of heteroskedasticity is random coefficients:

$$y_{it} = b_i x_{it} + c_i + u_{it},$$

$$b_i = \beta + w_i, \; E(w_i) = 0,$$

$$y_{it} = \beta x_{it} + c_i + (w_i x_{it} + u_{it}).$$

If w is independent of x, then $\underset{\sim}{\Pi} = \beta \underset{\sim}{I} + \underset{\sim}{\ell} \underset{\sim}{\lambda}'$, and our previous discussion, is relevant for the estimation of β. We shall handle the heteroskedasticity

problem in the inference section by allowing, $E[(\underset{\sim}{y}_i - \underset{\sim}{\Pi}\,\underset{\sim}{x}_i)(\underset{\sim}{y}_i - \underset{\sim}{\Pi}\,\underset{\sim}{x}_i)'|\underset{\sim}{x}_i]$ to be an arbitrary function of $\underset{\sim}{x}_i$.[17]

2.7.b. *Serial Correlation*

It may be of interest to impose restrictions on the residual covariances, such as a variance-components structure together with an autoregressive-moving average scheme.[18] Consider the homoskedastic case in which

$$\underset{\sim}{\Omega} = E[(\underset{\sim}{y}_i - \underset{\sim}{\Pi}\,\underset{\sim}{x}_i)(\underset{\sim}{y}_i - \underset{\sim}{\Pi}\,\underset{\sim}{x}_i)'|\underset{\sim}{x}_i]$$

does not depend upon $\underset{\sim}{x}_i$. Then the restrictions can be expressed as $\Omega_{jk} = g_{jk}(\underset{\sim}{\theta})$, where the g's are known functions and $\underset{\sim}{\theta}$ is an unrestricted parameter vector. We shall discuss a minimum distance procedure for imposing such restrictions in the inference section.

3. SPECIFICATION AND IDENTIFICATION: NONLINEAR MODELS

3.1. *A Random Effects Probit Model*

Our treatment of individual effects carries over with some important qualifications to nonlinear models. We shall illustrate with a labor force participation example. If the upper bound on leisure is binding in (2.6), then

$$\rho^{(t-1)}\tilde{U}_t'(\bar{L}) > m\gamma^{-(t-1)}w_t,$$

where m is the Lagrange multiplier corresponding to the lifetime budget constraint (the marginal utility of initial wealth). Let $y_{it} = 1$ if individual i works in period t, $y_{it} = 0$ otherwise. Let

$$\ell n\ W_{it} = \varphi_1 x_{it} + e_{1it},$$

$$\ell n\ A_{it} = \varphi_2 x_{it} + e_{2it},$$

where x_{it} contains measured variables that predict wages and tastes for leisure. We shall simplify the notation by supposing that x_{it} consists of a single variable. Then $y_{it} = 1$ if

$$(\varphi_1 - \varphi_2) x_{it} - (t-1)\ \ell n\ (\gamma\rho) + \text{en}\ m_i + (1-6)\ \text{en}\ \bar{L} + e_{1it} - e_{2it} \geq 0,$$

which we shall write as

$$\beta\, x_{it} + \varphi(t-1) + c_i + u_{it} \geq 0. \tag{3.1}$$

Now we need a distributional assumption for the **u's.** We shall assume that $(u_1, \ldots, u_T)$ is independent of c and the x's, with **a multivariate** normal distribution $(N(0, \Sigma))$. So we have a **probit** model (suppressing the **i** subscripts and period-specific intercepts):

$$P(y_t = 1 | x_1, \ldots, x_T, c) = F[\sigma_{tt}^{-\frac{1}{2}}(\beta x_t + c)],$$

where $F(\ )$ is the standard **normal** distribution function and σ_{tt} is the $t^{\underline{th}}$ diagonal element of Σ.

Next we shall specify a distribution for c conditional on $\mathbf{x} = (x_1, \ldots, x_T)$:

$$c = \psi + \lambda_1 x_1 + \ldots + \lambda_T x_T + v,$$

where **v** is independent of the x's and has a normal distribution $(N(0, \sigma_v^2))$.

There is a very important difference in this step compared with the linear case. In the linear case it was not restrictive to decompose $\mathbf{c}$ into its linear projection on $\mathbf{x}$ and an orthogonal residual. Now, however, we are assuming that the regression function $E(\mathbf{c}|\mathbf{x})$ is actually linear, that $\mathbf{v}$ is independent of $\mathbf{x}$, and that $\mathbf{v}$ has a normal distribution. These are restrictive assumptions and there may be a payoff to relaxing them.

Given these assumptions, the distribution for y_t conditional on $x_1, \ldots . x_T$ but marginal on $\mathbf{c}$ also has a **probit** form:

$$P(y_t = 1 | x_1, \ldots . x_T) = F[\alpha_t(\beta x_t + \lambda_1 x_1 + \ldots + \lambda_T x_T)],$$

$$\alpha_t = (\sigma_{tt} + \sigma_v^2)^{-\frac{1}{2}} .$$

Combining these T specifications gives the following matrix of coefficients:[19]

$$\Pi = \text{diag}\{\alpha_1, \ldots . \alpha_T\} [\beta \underset{\sim}{I}_T + \underset{\sim}{\ell} \underset{\sim}{\lambda}'] . \tag{3.2}$$

This differs from the linear case only in the diagonal matrix of normalization factors α_t. There are now nonlinear restrictions on $\underset{\sim}{\Pi}$, but the identification analysis is still straightforward. We have

$$\alpha_t \beta = \frac{\alpha_t}{\alpha_1} \pi_{11} - \pi_{t1} = \pi_{tt} - \frac{\alpha_t}{\alpha_1} \pi_{1t} ,$$

$$\frac{\alpha_t}{\alpha_1} = (\pi_{tt} + \pi_{t1})/(\pi_{11} + \pi_{1t}) \qquad (t=2, \ldots . T),$$

if $\beta + \lambda_1 + A, \neq 0$. Then, as in the linear case, we can solve for $\alpha_1 \beta$ and $\alpha_1 \underset{\sim}{\lambda}$. Only ratios of coefficients are identified, and so we can use a scale normalization such as $\alpha_1 \equiv 1$.

As for inference, a computationally simple approach is to estimate T cross-sectional probit specifications by maximum likelihood, where $x_1, \ldots, x_T$ are included in each of the T specifications. This gives $\hat{\pi}_t$ $(t=1, \ldots . T)$ and we can use a Taylor expansion to derive the covariance matrix of the asymptotic normal distribution for $(\hat{\pi}_1, \ldots . \hat{\pi}_T)$. Then restrictions can be imposed on Π using a minimum distance estimator, just as in the linear case.

We shall conclude our discussion of this model by considering the interpretation of the coefficients. We began with the probit specification that

$$P(y_t = 1 | x_1, \ldots . XT, c) = F[\sigma_{tt}^{-\frac{1}{2}} (\beta x_t + c)].$$

So one might argue that the correct measure of the effect of x_t is based on $\sigma_{tt}^{-\frac{1}{2}} \beta$, whereas we have obtained $(\sigma_{tt} + \sigma_v^2)^{-\frac{1}{2}} \beta$, which is then an underestimate. But there is something curious about this argument, since the "omitted variable" v is independent of $x_1, \ldots . x_T$. Suppose that we decompose u_t in (3.1) into $u_{1t} + u_{2t}$ and that measurements on u_{1t} become available. Then this argument implies that the correct measure of the effect of x_t is based on $[V(u_{2t})]^{-\frac{1}{2}} \beta$. As the data collection becomes increasingly successful. there is less and lass variance left in the residual u_{2t}, and $[V(u_{2t})]^{-\frac{1}{2}}$ becomes arbitrarily large.

The resolution of this puzzle is that the effect of x_t depends upon the value of c, and the effect evaluated at the average value for c is not equal to the average of the effects, averaging over the distribution for c. Consider the effect on the probability that y_t = 1 of increasing x_t from x' to x"; using the average value for c gives

$$F[\sigma_{tt}^{-\frac{1}{2}}(\beta x'' + E(c))] - F[\sigma_{tt}^{-\frac{1}{2}}(\beta x' + E(c))].$$

The problem with this measure is that it may be relevant for Only a small fraction of the population. I think that a more appropriate measure is the mean effect for a randomly dram individual:

$$\int [P(y_t = 1|x_t = x'', c) - P(y_t = 1|x_t = x', c)]\mu(dc),$$

where $\mu(dc)$ gives the population probability measure for c.

We shall see how to recover this measure within our framework. Let $z = \lambda_1 x_1 + \ldots + \lambda_T x_T$; let $\mu(dz)$ and $\mu(dv)$ give the population probability measures for the independent random variables z and v. Then

$$P(y_t - 1|x_t, c) = P(y_t = 1|x_1, \ldots, x_T, c)$$

$$= P(y_t = 1|x_t, z, v);$$

$$\int P(y_t = 1|x_t, z, v)\mu(dz)\mu(dv)$$

$$= \int P(y_t = 1|x_t, z, v)\mu(dv|x_t, z)\mu(dz)$$

$$= \int P(y_t = 1|x_t, z)\mu(dz),$$

where $\mu(dv|x_t, z)$ is the conditional probability measure, which equals the unconditional measure since v is independent of x_t and z. (It is important to note that the last integral does <u>not</u>, in general, equal $P(y_t = 1|x_t)$. For if x_t end z are correlated, as they are in our case, then

$$P(y_t = 1|x_t) = \int P(y_t = 1|x_t, z)\mu(dz|x_t)$$

$$\neq \int P(y_t = 1|x_t, z)\mu(dz).)$$

We have shown that

$$(3.3)\qquad \int [P(y_t = 1|x_t = x'', c) - P(y_t = 1|x_t = x', c)]\mu(dc)$$

$$= \int [P(y_t = 1|x_t = x'', z) - P(y_t = 1|x_t = x', z)]\mu(dz).$$

The integration with respect to the marginal distribution for z can be done using the empirical distribution function, which gives the following consistent (as $N \to \infty$) estimator of (3.15):

$$(3.4)\qquad \frac{1}{N}\sum_{i=1}^{N} \{F[\alpha_t(\beta x'' + \lambda_1 x_{i1} + \ldots + \lambda_T x_{iT})]$$

$$- F[\alpha_t(\beta x' + \lambda_1 x_{i1} + \ldots + \lambda_T x_{iT})]\}.$$

3.2 A Fixed Effects Logit Model: Conditional Likelihood

A weakness in the probit model was the specification of a distribution for c conditional on x. A convenient form was chosen, but it was only an approximation, perhaps a poor one. We shall discuss a technique that does not require us to specify a particular distribution for c conditional on x; it will, however, have its own weaknesses.

Consider the following specification:

$$(3.5)\qquad P(y_t = 1|x_1, \ldots x_T, c) = G(\beta x_t + c),\ G(z) = e^z/(1 + e^z),$$

where $y_1, \ldots y_T$ are independent conditional on $x_1, \ldots x_T$, c.

Suppose that T=2 and compute the probability that $y_2 = 1$ conditional, on $y_1 + y_2 = 1$:

$$(3.6)\qquad P(y_2 = 1|x_1, x_2, c, y_1 + y_2 = 1) = G[\beta(x_2 - x_1)],$$

which does not depend upon c. Given a random sample of individuals, the conditional log-likelihood function is

$$L = \sum_{i \in B} \{w_i \ln G[\beta(x_{i2} - x_{i1})] + (1-w_i) \ln G[-\beta(x_{i2}-x_{i1})]\},$$

where

$$w_i = \begin{cases} 1 \text{ if } (y_{i1}, y_{i2}) = (0, 1) \\ 0 \text{ if } (y_{i1}, y_{i2}) = (1, 0), \end{cases}$$

$$B = \{i | y_{i1} + y_{i2} = 1\}.$$

This conditional likelihood function does not depend upon the incidental parameters. It is in the form of a binary logit likelihood function in which the two outcomes are (0.1) and (1,0) with explanatory variables $x_2 - x_1$. This is the analog of differencing in the two period linear model. The conditional maximum likelihood (ML) estimate of β can be obtained simply from a ML binary logit program. This conditional likelihood approach was used by Rasch (1960, 1961) in his model for intelligence tests.20

The conditional ML estimator of β is consistent provided that the conditional likelihood function satisfies regularity conditions, which impose mild restrictions on the c_i. These restrictions, which are satisfied if the ci are a random sample from some distribution, are discussed in Andersen (1970). Furthermore, the inverse of the information' matrix based on the conditional likelihood function provides a covariance matrix for the asymptotic ($N \to \infty$) normal distribution of the conditional ML estimator of β.

These results should be contrasted with the inconsistency of the standard fixed effects ML estimator, in which the likelihood function is

based on the distribution of $y_1, \ldots, y_T$ conditional on $x_1, \ldots, x_T$, c. For example, suppose that T = 2, $x_{i1} = 0$, $x_{i2} = 1$ (i=1, ... N). The following limits exist with probability one if the c_i are a random sample from some distribution:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[y_{i1}(1-y_{i2})|c_i] = \varphi_1 ,$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} E[(1-y_{i1})y_{i2}|c_i] = \varphi_2 ,$$

where

$$E[y_{i1}(1 - y_{i2})|c_i] = G(c_i)G(-\beta - c_i),$$

$$E[(1-y_{i1})\, y_{i2}|c_i] = G(-c_i)G(\beta + c_i).$$

Andersen (1973, p. 66) shows that the ML estimator of β converges with probability one to 2β as $N \to \infty$. A simple extension of his argument shows that if G is replaced by any distribution function ($\tilde{G}$) corresponding to a symmetric, continuous, nonzero probability density, then the ML estimator of β converges with probability one to

$$2\tilde{G}^{-1}\left(\frac{\varphi_2}{\varphi_1+\varphi_2}\right) .$$

The logit case is special in that $\varphi_2/\varphi_1 = e^{\beta}$ for any distribution for c. In general the limit depends on this distribution; but if all of the $c_i = 0$, then once again we obtain convergence to 26 as $N \to \infty$.

For general T, conditioning on $\Sigma_t y_{it}$ (i=1, ... N) gives the following conditional log-likelihood function:

$$L = \sum_{i=1}^{N} \ell n \left[\exp\left(\beta \sum_{t=1}^{T} x_{it} y_{it}\right) \Big/ \sum_{d \in B_i} \exp\left(\beta \sum_{t=1}^{T} x_{it} d_t\right)\right],$$

$$B_i = \left\{d = (d_1, \ldots, d_T) \,\middle|\, d_t = 0 \text{ or } 1 \text{ and } \sum_{t=1}^{T} d_t = \sum_{t=1}^{T} y_{it}\right\}.$$

L is in the conditional logit form considered by McFadden (1974), with the alternative set (B_i) varying across the observations. Hence it can be maximized by standard programs. There are T+1 distinct alternative sets corresponding to $\Sigma_t y_{it} = 0,1, \ldots, T$. Groups for which $\Sigma_t y_{it} = 0$ or T contribute zero to L, however, and so only T-1 alternative sets are relevant. The alternative set for the group with $\Sigma_t y_{it} = s$ has $\binom{T}{s}$ elements, corresponding to the distinct sequences of T trials with s successes. For example, with T=3 and s=1 there are three alternatives with the following conditional probabilities:

$$P(1,0,0 \mid \underset{\sim}{x}_i, c_i, \textstyle\sum_t y_{it} = 1) = \exp[\beta(x_{i1} - x_{i3})]/D,$$

$$P(0,1,0 \mid \underset{\sim}{x}_i, c_i, \textstyle\sum_t y_{it} = 1) = \exp[\beta(x_{i2} - x_{i3})]/D,$$

$$P(0,0,1 \mid \underset{\sim}{x}_i, c_i, \textstyle\sum_t y_{it} = 1) = 1/D,$$

$$D = \exp[\beta(x_{i1}-x_{i3})] + \exp[\beta(x_{i2}-x_{i3})] + 1.$$

A weakness in this approach is that it relies on the assumption that the Yt are independent conditional on x, c, with an identical form for the conditional probability each period: $P(y_t = 1|\underset{\sim}{x}, c) = G(\beta x_t + c)$. In the probit framework, these assumptions translate into $\Sigma = \sigma^2 \underset{\sim}{I}$, so that $v + u_t$ generates an equicorrelated matrix: $\sigma_v^2 \underset{\sim}{\ell}\,\underset{\sim}{\ell}' + \sigma^2 I$. We have seen that it is straightforward to allow Σ to be unrestricted in the probit framework; that is not true here.

An additional weakness is that we are limited in the sorts of probability statements that can be made. We obtain a clean estimate of the effect of x_t on the log odds:

$$\ell n \left[\frac{P(y_t=1|x_t=x'', c)}{P(y_t=0|x_t=x'', c)} \Big/ \frac{P(y_t=1|x_t=x', c)}{P(y_t=0|x_t=x', c)} \right] = \beta(x'' - x');$$

the special feature of the logistic functional form is that this function of the probabilities does not depend upon c; so the problem of integrating over the marginal distribution of c (instead of the conditional distribution of c given x) does not arise. But this is not the only function of the probabilities that one might want to know. In the probit section we considered

$$P(y_t = 1|x_t = x'', c) - P(y_t = 1|x_t = x', c),$$

which depends upon c for probit or logit, and we averaged over the marginal distribution for c:

$$\int [P(y_t = 1|x_t = x'', c) - P(y_t = 1|x_t = x', c)]\mu(dc). \tag{3.7}$$

This requires us to specify a marginal distribution for c, which is what the conditioning argument trys to avoid. We cannot estimate (3.7) if all we have is the conditional ML estimate of β.

Our specification in (3.5) asserts that y_t is independent of $x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T$ conditional on x_t, c. This can be relaxed somewhat, but the conditional likelihood argument certainly requires more than

$$P(y_t = 1|x_t, c) = G(\beta x_t + c);$$

to see this, try to derive (3.6) with x2 = y_1. We can, however, implement the following specification (with $\underset{\sim}{x}' = (x_1, \ldots, x_T)$):

$$(3.8) \qquad P(y_t = 1|\underset{\sim}{x}, c) = G(\beta_{to} + \beta_{t1}x_1 + \ldots + \beta_{tt}x_t + c),$$

where $y_1, \ldots, y_T$ are independent conditional on $\underset{\sim}{x}$, c. This corresponds to our specification of "x is strictly exogenous conditional on c" in Section 2.5, except that $\gamma_t = 1$ in the term $\gamma_t c$ -- it is not straightforward to allow a time-varying coefficient on c in the conditional likelihood approach. The extension of (3.6) is

$$(3.9) \qquad P(y_t = 1|\underset{\sim}{x}, c, y_1 + y_t = 1) = G(\tilde{\beta}_{to} + \tilde{\beta}_{t1}x_1 + \beta_{t2}x_2 + \ldots + \beta_{tt}x_t)$$

$$(t=2, \ldots, T),$$

where $\beta_{tj} = \beta_{tj} - \beta_{1j}$ $(j=0,1)$. So if x has sufficient variation, we can obtain consistent estimates of $\tilde{\beta}_{to}$, $\tilde{\beta}_{t1}$, and β_{ts} $(s=2, \ldots, t)$.

Only these parameters are identified, since we can transform the model replacing c by $c = \beta_{10} + \beta_{11}x_1 + c$ without violating any restrictions.

The restrictions in (3.5) or in (3.8) can be tested against the following alternative:

$$(3.10) \quad N_Y, = 1|\underset{\sim}{x}, c) = G(\pi_{to} + \pi_{t1}x_1 + \ldots + \pi_{tT}x_T + c).$$

We can identify only $\pi_{tj} - \pi_{1j}$ and so we can normalize $\pi_{1j} = 0$ (j=0, ..., T; t=2, ... T). The maximized values of the conditional log likelihoods can be used to form χ^2 statistics. [21] There are (T-2)(T-1)/2 restrictions in passing from (3.10) to (3.8), and (3.5) imposes an additional (T-1)(T+4)/2 - 1 restrictions.

3.3. *Serial Correlation And Lugged Dependent Variables*

Consider the following two models:

$$(3.11a) \quad y_t = \begin{cases} 1 \text{ if } y_t^* = \gamma y_{t-1} + e_t \geq 0 \\ 0 \text{ otherwise} \end{cases},$$

$$(3.11b) \quad y_t = \begin{cases} 1 \text{ if } y_t^* = u_t \geq 0 \\ 0 \text{ otherwise; } u_t = \rho u_{t-1} + e_t \end{cases};$$

in both cases et is i.i.d. $N(0, \sigma^2)$. Heckman (1978) observed that we can distinguish between these two models.[22] In the first model,

$$P(y_t = 1|y_{t-1}, y_{t-2}, \ldots) = P(y_t = 1|y_{t-1}) = F(\gamma y_{t-1}/\sigma),$$

where $F(\ )$ is the standard normal distribution function. In the second model, however, $P(y_t = 1|y_{t-1}, y_{t-2}, \ldots)$ depends upon the entire history of the process. If we observed u_{t-1}, then previous outcomes would be irrelevant. In fact, we observe only whether $u_{t-1} \geq 0$; hence conditioning in addition on whether $u_{t-2} \geq 0$ affects the distribution of u_{t-1} and y_t. So the lagged y implies a Markov chain whereas the Markov assumption for the probit residual does not imply a Markov chain for the binary sequence that it generates.

There is an analogy with the following linear models:

(3.12a) $$y_t = \gamma y_{t-1} + e_t ,$$

(3.12b) $$y_t = u_t, \; u_t = e_t + \rho e_{t-1},$$

where e_t is i.i.d. $N(0, \sigma^2)$. We know that if $u_t = \rho u_{t-1} + e_t$, then no distinction would be possible, without introducing more structure, since both models imply a linear Markov process. With the moving average residual, however, the serial correlation model implies that the entire past history is relevant for predicting y. So the distinction between the two models rests on the order of the dependence on previous realizations of y_t.

We can still distinguish between the two models even when $(u_1, \ldots, u_T)$ has a general multivariate normal distribution $(N(\underset{\sim}{\mu}, \underset{\sim}{\Sigma}))$. Given nor-

malizations such as $V(u_t) = 1$ (e-1, ... T), the serial correlation model has $T(T+1)/2$ free parameters. Hence if $T \geq 3$, there are restrictions on the $2^T - 1$ parameters of the multinomial distribution for $(y_1, \ldots, y_T)$. In particular, the most general multivariate probit model cannot generate a Markov chain. So we can add a lagged dependent variable and identify Y.

This result relies heavily on the restrictive nature of the multivariate probit functional form. A more robust distinction between the two models is possible when there is variation over time in x_t. We shall pursue this after first presenting a generalization of strict exogeneity and noncausality for nonlinear models.

Let t=1 be the first period of the individual's (economic) life. An extension of Granger's definition of "y does not cause x" is that x_{t+1} is independent of $y_1, \ldots, y_t$ conditional on $x_1, \ldots, x_t$. An extension of Sims' strict exogeneity condition is thst y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_1, \ldots, x_t$. In contrast to the linear predictor case, these two definitions are no longer equivalent.[23] For consider the following counterexample: let y_1, y_2 be independent Bernoulli random variables with $P(y_t = 1) = P(y_t = -1) = 1/2$ (t=1,2). Let x3 = $y_1 y_2$. Then y_1 is independent of x3 and y_2 is independent of x_3. Let all of the other random variables be degenerate (equal to zero, say). Then x is strictly exogenous but x3 is clearly not independent of y_1, y_2 conditional on xl, x_2. The counterexample works for the following reason: if a random variable is uncorrelated with each of two other random variables, then it is uncorrelated with every linear combination of them; but If it is independent of each of the other random variables, it need not be independent of every function of them.

Consider the following modification of Sims' condition: y_t is independent of $x_{t+1}, x_{t+2}, \ldots$ conditional on $x_1, \ldots, x_t, y_1, \ldots, y_{t-1}$ $(t=1,2,\ldots)$. Chamberlain (1982) shows that, subject to a regularity condition, this is equivalent to our extended definition of Granger non-causality. The regularity condition is trivially satisfied whenever y_t has a degenerate distribution prior to some point. So it is satisfied in our case since $y_0, y_{-1}, \ldots$ have degenerate distributions.

It is straightforward to introduce a time-invariant latent variable into these definitions. We shall say that "*y does not cause x conditional on a latent variable c*" if either

$$x_{t+1} \text{ is independent of } y_1, \ldots, y_t \text{ conditional on } x_1, \ldots, x_t, c \quad (t=1,2,\ldots I,$$

or

$$y_t \text{ is independent of } x_{t+1}, x_{t+2}, \ldots \text{ conditional on } x_1, \ldots, x_t, y_1, \ldots, y_{t-1}, c \quad (t=1,2,\ldots);$$

they are equivalent. We shall say that "*x is strictly exogenous conditional on a latent variable c*" if

$$y_t \text{ is independent of } x_{t+1}, x_{t+2}, \ldots \text{ conditional on } x_1, \ldots, x_t, c \ (t=1,2,\ldots).$$

Now let us return to the problem of distinguishing between serial correlation and structural lagged dependent variables. Assume throughout the discussion that x_t and y_t are not independent. We shall say that the relationship of x to y is *static* if

$$x \text{ is strictly exogenous and } y_t \text{ is independent of } x_1, \ldots, x_{t-1} \text{ conditional on } x_t.$$

Then I propose the following distinctions:

> *There is residual serial correlation* if y_t *is not independent of* $y_1, \ldots, y_{t-1}$ *conditional on* $x_1, \ldots, x_t$;
>
> If *the* relationship of x *to* y *is static, then there are no structural lagged dependent variables.*

Suppose that y_t and x_t are binary and consider the probability that $y_2 = 1$ conditional on $(x_1, x_2) = (0, 0)$ and conditional on $(x1, x_2) = (1, 0)$. Since y_t and x_t are assumed to be dependent, the distribution of y_1 is generally different in the two cases. If y_1 has a structural effect on y_2, then the conditional probability of $y_2 = 1$ should differ in the two cases, so that y_2 is not independent of x_1 conditional on x_2.

Note that this condition is one-sided: I am only offering a condition for there to be no structural effect of y_{t-1} on y_t. There can be distributed lag relationships in which we would not want to say that y_{t-1} has a structural effect on y_t. Consider the production function example with serial correlation in rainfall; assume for the moment that there is no variation in c. If the serial correlation in rainfall is not incorporated in the farmer's information set, then our definitions assert that there is residual serial correlation but no structural lagged dependent variables, since the relationship of x to y is static. Now suppose that the farmer does use previous rainfall to predict future rainfall. Then the relationship of x to y is not static since x is not strictly exogenous. But we my not want to say that the relationship between y_{t-1} and y_t is structural, since the technology does not depend upon y_{t-1}.

How are these distinctions affected by latent variables? It should be clear that a time-invariant latent variable can produce residual serial correlation. A major theme of the paper has been that such a latent variable can also produce a failure of strict exogeneity. So consider conditional versions of these properties:

> *There is residual serial correlation conditional on a latent variable c if* y_t *is not independent of* $y_1, \ldots, y_{t-1}$ *conditional on* $x_1, \ldots, x_t$, c;
>
> *The relationship of* x *to* y *is static conditional on a latent variable c* if x is *strictly exogenous conditional on c and* **if** y_t *is independent* **of** $x_1, \ldots, x_{t-1}$ *conditional on* x_t, **c;**
>
> *If the relationship of* c *to* y *is static conditional on a latent variable c, then there are no structural lagged dependent variables.*

A surprising feature of the linear predictor definition of strict **exogeneity** is that it is restrictive to assert that there exists **some** time-invariant latent variable c such that **x** is strictly exogenous conditional on c. This is no longer true when we **use** conditional independence to define strict exogeneity. For a counterexample, suppose that $\mathbf{x_t}$ is a binary variable and consider the conditional strict exogeneity question, "Does there exist a time-invariant random variable **c** such that $\mathbf{y_t}$ is independent of $\mathbf{x_1}, \ldots$ xT conditional on $\mathbf{x_1}, \ldots \mathbf{x_t}$, **c?"** The answer is "yes" since we can order the 2^T possible outcomes of the binary sequence $(\mathbf{x_1}, \ldots \mathbf{x_T})$ and set **c=j** if the $j^{\underline{th}}$ outcome occurs $(j=1, \ldots, 2^T)$. Now $\mathbf{y_t}$ is independent of $\mathbf{x_1}, \ldots, \mathbf{x_T}$ conditional on c!

For a **nondegenerate** counterexample, let **y** and **x** be binary random variables with

$$P(y = \alpha_j,\ x = \alpha_k) = \tau_{jk} > 0, \quad \sum_{j,k=1}^{2} \tau_{jk} = 1,$$

where $\alpha_1 = 1$, $\alpha_2 = 0$. Let $\gamma' = (\tau_{11}, \tau_{12}, \tau_{21}, \tau_{22})$. Then we can set

$$\underset{\sim}{\gamma} = \sum_{m=1}^{4} \gamma_m \underset{\sim}{e}_m,\ \gamma_m > 0, \quad \sum_{m=1}^{4} \gamma_m = 1,$$

where $\underset{\sim}{e}_m$ is a vector of zeros except for a one in the m^{th} component. Hence $\underset{\sim}{\gamma}$ is in the interior of the convex hull of $\{\underset{\sim}{e}_m,\ m=1, \ldots, 4\}$. Now consider the vector

$$\underset{\sim}{y}(\delta,\lambda) = \begin{vmatrix} \delta\lambda \\ \delta(1-\lambda) \\ (1-\delta)\lambda \\ (1-\delta)(1-\lambda) \end{vmatrix} .$$

The components of $\underset{\sim}{y}(\delta,\lambda)$ give the probabilities $P(y = \alpha_j,\ x = \alpha_k)$ when y and x are independent with $P(y = 1) = \delta$, $P(x = 1) = \lambda$. Set $\underset{\sim}{e}^*_m = \underset{\sim}{y}(\delta_m, \lambda_m)$ with $0 < \delta_m < 1$, $0 < \lambda_m < 1$. Then $\underset{\sim}{\gamma}$ will be in the interior of the convex hull of $\{\underset{\sim}{e}^*_m,\ m=1,\ldots,4\}$ if we choose δ_m, λ_m so that $\underset{\sim}{e}^*_m$ is sufficiently close to $\underset{\sim}{e}_m$. Hence

$$\underset{\sim}{\gamma} = \sum_{m=1}^{4} \gamma^*_m \underset{\sim}{e}^*_m\ ,\ \gamma^*_m > 0, \quad \sum_{m=1}^{4} \gamma^*_m = 1$$

Let the components of $\underset{\sim}{e}^*_m$ be $(\tau^m_{11}, \tau^m_{12}, \tau^m_{21}, \tau^m_{22})$. Let c be a random variable with $P(c = m) = \gamma^*_m$ $(m=1, \ldots, 4)$, and set

$$P(y = \alpha_j,\ x = \alpha_k \mid c = m) = \tau^m_{jk} .$$

Now y is independent of **x** conditional on c, and the conditional distributions are nondegenerate.

If $(x_1, \ldots, x_T, y_1, \ldots, y_T)$ has a general **multinomial** distribution, then a straightforward extension of this argument shows that there exists a random variable c such that $(y_1, \ldots, y_T)$ is independent of $(x_1, \ldots, x_T)$ conditional on **c,** and the conditional distributions are nondegenerate.

A similar point applies to factor analysis. Consider a linear **one-**factor model. The specification is that there exists a latent variable c such that the partial correlations between $y_1, \ldots, y_T$ are zero given c. This is restrictive if $T > 3$. But we now **know** that it is not restrictive to assert that there exists a **latent** variable c such that $y_1, \ldots, y_T$ are independent conditional on c.

It follows that we cannot test for conditional strict **exogeneity** without imposing functional form restrictions; **nor** can we test for a conditionally static relationship without restricting the functional forms.

This point is intimately related to the fundamental difficulties created by incidental parameters in nonlinear models. **The** labor force participation example is assumed to be static conditional on c. We shall present some tests of this in Section 5, but we **shall** be jointly testing that proposition and the functional forms — a truly nonparametric **test** cannot exist. We stressed in the **probit** model that the specification for the distribution of c conditional on **x** is restrictive; we avoided such a restrictive specification in the **logit** model but only by imposing a restrictive functional **form** on the distribution of y conditional on **x,** c.

3.4. *Duration Models*

In many problems the basic data is the amount of tine spent in a state. For example, a complete description of an individual's labor force participation history is the duration of the first spell of participation and the date it began, the duration of the following spell of nonparticipation, and so on. This complete history will generate a binary sequence when it is cut up into fixed length periods, but these periods may have little to do with the underlying process.[24]

In particular, the measurement of serial correlation depends upon the period of observation. As the period becomes shorter, the probability that a person who worked last period will work this period approaches one. So finding significant serial correlation may say very little about the underlying process. Or consider a spell that begins near the end of a period; then it is likely to overlap into the next period, so that previous employment raises the probability of current employment.

Consider the underlying process of time spent in one state followed by time spent in the other state. If the individual's history does not help to predict his future given his current state, then this is a Markov process. Whereas serial independence in continuous time has the absurd Implication that mean duration of a spell is zero, the Markov property does provide a fruitful starting point. It has two Implications: the individual's history prior to the current spell should not affect the distribution of the length of the current spell; and the amount of time spent in the current state should not affect the distribution of remaining time in that state.

So the first requirement of the Markov property is that durations of the spells be independent of each other. Assuming stationarity, this

implies an alternating renewal process. The second requirement is that the distribution of duration be exponential, so that we have an alternating Poisson process. We shall refer to departures from this model as duration dependence.

A test of this Markov property using binary sequences will depend upon what sampling scheme is being used. The simplest case is point sampling, where each period we determine the individual's state at a particular point in time, such as July 1 of each year. Then if an individual is following an alternating Poisson process, her history prior to that point is irrelevant in predicting her state at the next interview. So the binary sequence generated by point sampling should be a Markov chain.

It is possible to test this in a fixed effects model that allows each individual to have her own two exponential rate parameters (c_{i1}, c_{i2}) in the alternating Poisson process. The idea is related to the conditional likelihood approach in the fixed effects logit model. Let s_{ijk} be the number of times that individual i is observed making a transition from state j to state k (j, k = 1,2). Then the initial state and these four transition counts are sufficient statistics for the Markov chain. Sequences with the same initial state and the same transition counts should be equally likely. This is the Markov form of de Finetti's (1975) partial exchangeability.[25] So we can test whether the Markov property holds conditional on c_{i1}, c_{i2} by testing whether there is significant variation in the sample frequencies of sequences with the same transition counts.

This analysis is relevant if, for example, each year the survey question is "Did you have a job on July 1?" In the Michigan Panel Study of Income Dynamics, however, the most commonly used question for generating participation sequences is "Did your wife do any work for money last year?" This interval sampling leads to a more complex analysis, since even if the individual is following an alternating Poisson process, the binary sequence

generated by this sampling scheme is not a Markov chain. suppose that $y_{t-1} = 1$, so that we know that the individual worked at some point during the previous period. What is relevant, however, is the individual's state at the end of the period, and y_{t-2} will affect the probability that the spell of work occurred early in period t-1 instead of late in the period.

Nevertheless, it is possible to test whether the underlying process is alternating Poisson. The reason is that if $y_{t-1} = 0$, we know that the individual never worked during period t-1, and so we know the state at the end of that period; hence $y_{t-2}, y_{t-3}, \ldots$ are irrelevant. So we have

$$P(y_t = 1 | c_1, c_2, y_{t-1}, y_{t-2}, \ldots)$$

$$= NY, = 1 | c_1, c_2, y_{t-1} = \cdots = y_{t-d} = 1, y_{t-d-1} = 0)$$

$$= P(y_t = 1 | c_1, c_2, d),$$

where d is the number of consecutive preceding periods that the individual was in state 1.

Let s_{01} be the number of times in the sequence that 1 is preceded by 0; let s_{011} be the number of times that 1 is preceded by 0, 1; etc. Then sufficient statistics are $s_{01}, s_{011}, \ldots$, as well as the number of consecutive ones at the beginning (n_1) and at the end (n_T) of a sequence. [26] For an example with $T = 5$; let $n_1 = 0$, $n_5 = 0$, $s_{01} = 1$, $s_{011} = 1$, $s_{0111} = \ldots = 0$; then we have

$$P(0, 1, 1, 0, 0 | \underset{\sim}{c})$$

$$= NY_1 = 0 | \underset{\sim}{c}) P(1 | 0, \underset{\sim}{c}) P(1 | 0,1, \underset{\sim}{c}) P(0 | 0,1,1, \underset{\sim}{c}) P(0 | 0, \underset{\sim}{c});$$

$$P(0, 0, 1, 1, 0 | \underset{\sim}{c})$$

$$= P(y_1 = 0 | \underset{\sim}{c}) P(0 | 0, \underset{\sim}{c}) P(1 | 0, \underset{\sim}{c}) P(1 | 0,1, \underset{\sim}{c}) P(0 | 0,1,1, \underset{\sim}{c}),$$

where $c = (c_1, c_2)$. Thus these two sequences are equally likely conditional on $\underset{\sim}{c}$, and letting μ be the probability measure for c gives

$$P(0,1,1,0,0) = \int P(0,1,1,0,0|\underset{\sim}{c})\mu(d\underset{\sim}{c})$$

$$= \int P(0,0,1,1,0|\underset{\sim}{c})\mu(d\underset{\sim}{c}) = P(0,0,1,1,0)$$

So the alternating Poisson process implies restrictions on the multinomial distribution for the binary sequence.

These tests are indirect. The duration dependence question is clearly easier to answer using surveys that measure durations of spells. Such duration data raises a number of new econometric problems, but we shall not pursue them here.[27] I would simply like to make one connection with the methods that we have been discussing.

Let us simplify to a one state process; for example, y_{it} can be the duration of the time interval between the starting date of the $i^{\underline{th}}$ individual's $t^{\underline{th}}$ job and his $(t+1)^{\underline{th}}$ job. Suppose that we observe $T > 1$ jobs for each of the N individuals, a not innocuous assumption. Impose, the restriction that $y_{it} > 0$ by using the following specification:

$$y_{it} = \exp(\beta x_{it} + c_i + u_{it}),$$

$$E^*(u_{it}|\underset{\sim}{x}_i) = 0 \quad (t=1,\ldots,T),$$

where $\underset{\sim}{x}_i' = (x_{i1}, \ldots . x_{iT})$. Then

$$E^*(\ell n\, y_{it}|\underset{\sim}{x}_i) = \beta x_{it} + \underset{\sim}{\lambda}'\underset{\sim}{x}_i ,$$

and our Section 2 analysis applies. The strict exogeneity assumption has

a surprising implication in this context. Suppose that x_{it} is the individual's age at the beginning of the t^{th} job. Then $x_{it} - x_{i\,t-1} = y_{i,t-1}$ -- age is not strictly exogenous.[28]

4. INFERENCE

Consider a sample $\underset{\sim}{r}_i' = (\underset{\sim}{x}_i', \underset{\sim}{y}_i')$, i=1, ... N, where $\underset{\sim}{x}_i' = (x_{i1}, \ldots . x_{iK})$, $\underset{\sim}{y}_i' = (y_{i1}, \ldots . y_{iM})$. We shall assume that $\underset{\sim}{r}_i$ is independent and identically distributed (i.i.d.) according to some multivariate distribution with finite fourth moments and $E(\underset{\sim}{x}_i \underset{\sim}{x}_i')$ nonsingular. Consider the minimum mean-square error linear predictors,[29]

$$E^*(y_{im}|\underset{\sim}{x}_i) = \underset{\sim}{\pi}_m' \underset{\sim}{x}_i \qquad (m=1, \ldots . M),$$

which we can write as

$$E^*(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \underset{\sim}{\Pi}\, \underset{\sim}{x}_i, \qquad \underset{\sim}{\Pi} = E(\underset{\sim}{y}_i\underset{\sim}{x}_i')\,[E(\underset{\sim}{x}_i\underset{\sim}{x}_i')]^{-1}.$$

We want to estimate Π subject to restrictions and to test those restrictions. For example, we may want to test whether a submatrix of II has the form $\beta \underset{\sim}{I} + \underset{\sim}{\ell}\, \underset{\sim}{A}'$.

We shall not assume that the regression function $E(\underset{\sim}{y}_i|\underset{\sim}{x}_i)$ is linear. For although $E(\underset{\sim}{y}_i|\underset{\sim}{x}_i, c_i)$ may be linear (indeed, we hope that it is), there is generally no reason to insist that $E(c_i|\underset{\sim}{x}_i)$ is linear. So we shall present a theory of inference for linear predictors. Furthermore, even if the regression function is linear, there may be heterdskedasticity -- due to random coefficients, for example. So we shall allow $E[(\underset{\sim}{y}_i - \underset{\sim}{\Pi}\underset{\sim}{x}_i)(\underset{\sim}{y}_i - \underset{\sim}{\Pi}\underset{\sim}{x}_i)'|\underset{\sim}{x}_i]$ to be an arbitrary function of $\underset{\sim}{x}_i$

4.1 *The Estimation of Linear Predictors*

Let $\underset{\sim}{w}_i$ be the vector formed from the distinct elements of $\underset{\sim}{r}_i\underset{\sim}{r}_i'$ that have nonzero variance.[30] Since $\underset{\sim}{r}_i' = (\underset{\sim}{x}_i', y_i')$ is i.i.d., it follows that $\underset{\sim}{w}_i$ is i.i.d. This simple observation is the key to our results. Since $\underset{\sim}{\Pi}$ is a function of $E(\underset{\sim}{w}_i)$, our problem is to make inferences about a function of a population mean, under random sampling.

Let $\mu = E(\underset{\sim}{w}_i)$ and let π be the vector formed from the columns of $\underset{\sim}{\Pi}'$ ($a = \text{vec}(\underset{\sim}{\Pi}')$). Then $\underset{\sim}{\pi}$ is a function of $\underset{\sim}{\mu}$: $\underset{\sim}{\pi} = \underset{\sim}{h}(\underset{\sim}{\mu})$. Let $\bar{\underset{\sim}{w}} = \sum_{i=1}^{N} \underset{\sim}{w}_i/N$; then $\hat{\underset{\sim}{\pi}} = \underset{\sim}{h}(\bar{\underset{\sim}{w}})$ is the least squares estimator:

$$\hat{\pi} = \text{vec}\left[\left(\sum_{i=1}^{N} \underset{\sim}{x}_i\underset{\sim}{x}_i'\right)^{-1} \sum_{i=1}^{N} \underset{\sim}{x}_i y_i'\right].$$

By the strong law of large numbers, $\bar{\underset{\sim}{w}}$ converges almost surely to μ^0 as $N \to \infty$ ($\bar{\underset{\sim}{w}} \xrightarrow{a.s.} \mu^0$), where μ^0 is the true value of μ. Let $\pi^0 = \underset{\sim}{h}(\underset{\sim}{\mu}^0)$. Since $\underset{\sim}{h}(\underset{\sim}{\mu})$ is continuous at $\underset{\sim}{\mu} = \underset{\sim}{\mu}^0$, we have $\hat{\underset{\sim}{\pi}} \xrightarrow{a.s.} \underset{\sim}{\pi}^0$. The central limit theorem implies that

$$\sqrt{N}\,(\bar{\underset{\sim}{w}} - \underset{\sim}{\mu}^0) \xrightarrow{D} N(\underset{\sim}{0}, V(\underset{\sim}{w}_i)).$$

Since $\underset{\sim}{h}(\mu)$ is differentiable at $\underset{\sim}{\mu} = \underset{\sim}{\mu}^0$, the δ-method gives

$$\sqrt{N}\,(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}^0) \xrightarrow{D} N(\underset{\sim}{0}, \underset{\sim}{\Omega}),$$

where

$$\underset{\sim}{\Omega} = \frac{\partial \underset{\sim}{h}(\underset{\sim}{\mu}^0)}{\partial \mu'}\, V(\underset{\sim}{w}_i)\, \frac{\partial \underset{\sim}{h}'(\underset{\sim}{\mu}^0)}{\partial \mu}.$$ [31]

We have derived the limiting distribution of the least squares estimator. This approach was used by Cramer (1946) to obtain limiting normal distributions for sample correlation and regression coefficients (p. 367); he presents an

explicit formula for the variance of the limiting distribution of a sample correlation coefficient (p. 359). Kendall and Stuart (1961, p. 293) and Goldberger (1974) present the formula for the variance of the limiting distribution of a simple regression coefficient.

Evaluating the partial derivatives in the formula for Ω is tedious. That calculation can be simplified since $\hat{\pi}$ has a "ratio" form. In the case of simple regression with a zero intercept, we have $\pi = E(y_i x_i)/E(x_i^2)$ and

$$\sqrt{N}(\hat{\pi} - \pi^o) = \left(\sum_{i=1}^{N} y_i x_i - \pi^o \sum_{i-1}^{N} x_i^2\right)/\left[\sqrt{N}\left(\sum_{i=1}^{N} x_i^2/N\right)\right].$$

Since $\Sigma_{i=1}^{N} x_i^2/N \xrightarrow{a.s.} E(x_i^2)$, we obtain the same limiting distribution by working with

$$\sum_{i=1}^{N} [(y_i - \pi^o x_i)x_i]/[\sqrt{N}\, E(x_i^2)].$$

The definition of π^o gives $E[(y_i - \pi^o x_i)x_i] = 0$, and so the central limit theorem implies that

$$\sqrt{N}\,(\hat{\pi} - \pi^o) \xrightarrow{D} N\{0,\ E[(y_i - \pi^o x_i)^2 x_i^2]/[E(x_i^2)]^2\}.$$

This approach was used by White (1980) to obtain the limiting distribution for univariate regression coefficients.[32] In the Appendix (Proposition 7) we follow White's approach to obtain

$$(4.1) \qquad \underset{\sim}{\Omega} = E[(\underset{\sim}{y}_i - \underset{\sim}{\Pi}^o \underset{\sim}{x}_i)(\underset{\sim}{y}_i - \underset{\sim}{\Pi}^o \underset{\sim}{x}_i)' \otimes \underset{\sim}{\Phi}_x^{-1}(\underset{\sim}{x}_i \underset{\sim}{x}_i')\underset{\sim}{\Phi}_x^{-1}],$$

where $\underset{\sim}{\Phi}_x = E(\underset{\sim}{x}_i \underset{\sim}{x}_i')$. A consistent estimator of Ω is readily available from the corresponding sample moments:

$$(4.2)\qquad \hat{\underset{\sim}{\Omega}} = \frac{1}{N}\sum_{i=1}^{N}[(\underset{\sim}{y}_i - \hat{\underset{\sim}{\Pi}}\,\underset{\sim}{x}_i)(\underset{\sim}{y}_i - \hat{\underset{\sim}{\Pi}}\,\underset{\sim}{x}_i)' \otimes \underset{\sim}{S}_x^{-1}(\underset{\sim}{x}_i\underset{\sim}{x}_i')\,\underset{\sim}{S}_x^{-1}]$$

$$\xrightarrow{a.s.} \underset{\sim}{\Omega}\ ,$$

where $\underset{\sim}{S}_x = \Sigma_{i=1}^{N}\,\underset{\sim}{x}_i\underset{\sim}{x}_i'/N.$

If $E(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \underset{\sim}{\Pi}\,\underset{\sim}{x}_i$, so that the regression function is linear, than

$$\underset{\sim}{\Omega} = E[V(\underset{\sim}{y}_i|\underset{\sim}{x}_i) \otimes \underset{\sim}{\Phi}_x^{-1}(\underset{\sim}{x}_i\underset{\sim}{x}_i')\,\underset{\sim}{\Phi}_x^{-1}].$$

If $V(\underset{\sim}{y}_i|\underset{\sim}{x}_i)$ is uncorrelated with $\underset{\sim}{x}_i\underset{\sim}{x}_i'$, then

$$\underset{\sim}{\Omega} = E[V(\underset{\sim}{y}_i|\underset{\sim}{x}_i)] \otimes \underset{\sim}{\Phi}_x^{-1}\ .$$

If the conditional variance is homoskedastic, so that $V(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \underset{\sim}{\Sigma}$ does not depend on $\underset{\sim}{x}_i$, then

$$\underset{\sim}{\Omega} = \underset{\sim}{\Sigma} \otimes \underset{\sim}{\Phi}_x^{-1}.$$

4.2 Imposing Restrictions: The Minimum Distance Estimator

Since $\underset{\sim}{\Pi}$ is a function of $E(\underset{\sim}{w}_i)$, restrictions on $\underset{\sim}{\Pi}$ imply restrictions on $E(\underset{\sim}{w}_i)$. Let the dimension of $\underset{\sim}{\mu} = E(\underset{\sim}{w}_i)$ be q.[33] We shall specify the restrictions by the condition that μ depends only on a $p\times 1$ vector θ of unknown parameters: $\underset{\sim}{\mu} = \underset{\sim}{g}(\underset{\sim}{\theta})$, where $\underset{\sim}{g}$ is a known function and $p \leq q$. The domain of θ is T, a subset of p-dimensional Euclidean space (R^p) that contains the true value θ^o. So the restrictions Imply that $\underset{\sim}{\mu}^o = \underset{\sim}{g}(\underset{\sim}{\theta}^o)$ is confined to a certain subset of R^q.

We can impose the restrictions by using a minimum distance estimator: choose $\hat{\theta}$ to

$$\min_{\theta \in T} \sum_{i=1}^{N} [w_i - g(\theta)]' A_N [w_i - g(\theta)],$$

where $A_N \xrightarrow{a.s.} \Psi$ and Ψ is positive definite.[34] This minimization problem is equivalent to the following one: choose $\hat{\theta}$ to

$$\min_{\theta \in T} [\bar{w} - g(\theta)]' A_N [\bar{w} - g(\theta)].$$

The properties of $\hat{\theta}$ are developed, for example, in **Malinvaud** (1970, Chap. 9). Since g does not depend on any exogenous variables, the derivation of these properties can be simplified considerably, as in **Chiang** (1956: and **Ferguson** (1958).[35]

For completeness, we shall state a set of regularity conditions and the properties that they imply:

Assumption 1. $a_N \xrightarrow{a.s.} g(\theta^0)$; T is a compact subset of R^p that contains θ^0; g is continuous on T, and $g(\theta) = g(\theta^0)$ for $\theta \in T$ implies that $\theta = \theta^0$; $A_N \xrightarrow{a.s.} \Psi$, where Y is positive definite.

Assumption 2. $\sqrt{N}[a_N - g(\theta^0)] \xrightarrow{D} N(0, A)$; T contains a neighborhood T_0 of θ^0 in which g has continuous second partial derivatives; rank (G) = p, where $G = \partial g(\theta^0)/\partial \theta'$.

Choose $\hat{\theta}$ to

$$\min_{\theta \in T} [a_N - g(\theta)]' A_N [a_N - g(\theta)].$$

Proposition I. If Assumption 1 is satisfied, then $\hat{\underset{\sim}{\theta}} \xrightarrow{a.s.} \underset{\sim}{\theta}^{o}$.

Proposition 2. If assumptions 1 and 2 are satisfied, then $\sqrt{N}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}^{o}) \xrightarrow{D} N(\underset{\sim}{0}, \underset{\sim}{A})$, where

$$\underset{\sim}{A} = (\underset{\sim}{G}' \underset{\sim}{\Psi} \underset{\sim}{G})^{-1} \underset{\sim}{G}' \underset{\sim}{\Psi} \underset{\sim}{A} \underset{\sim}{\Psi} \underset{\sim}{G} (\underset{\sim}{G}' \underset{\sim}{\Psi} \underset{\sim}{G})^{-1}.$$

If A is positive definite, then $\underset{\sim}{A} = (\underset{\sim}{G}' \underset{\sim}{\Delta}^{-1} \underset{\sim}{G})^{-1}$ is positive semi-definite; hence an optimal choice for $\underset{\sim}{\Psi}$ is $\underset{\sim}{A}^{-1}$.

Proposition 3. If Assumptions 1 and 2 are satisfied, if $\underset{\sim}{\Delta}$ is a q×q positive-definite matrix, and if $\underset{\sim}{A}_N \xrightarrow{a.s.} \underset{\sim}{\Delta}^{-1}$, then

$$N[\underset{\sim}{a}_N - \underset{\sim}{g}(\hat{\underset{\sim}{\theta}})]' \underset{\sim}{A}_N [\underset{\sim}{a}_N - \underset{\sim}{g}(\hat{\underset{\sim}{\theta}})] \xrightarrow{D} \chi^2(q-p).$$

Now consider imposing additional restrictions, which are expressed by the condition that $\underset{\sim}{\theta} = \underset{\sim}{f}(\underset{\sim}{\alpha})$, where $\underset{\sim}{\alpha}$ is s×1 ($s \leq p$). The domain of $\underset{\sim}{\alpha}$ is T_1, a subset of R^s that contains the true value $\underset{\sim}{\alpha}^{0}$. so $\underset{\sim}{\theta}^{o} = \underset{\sim}{f}(\underset{\sim}{\alpha}^{o})$ is confined to a certain subset of R^p.

Assumption 2'. T_1 is a compact subset of R^s that contains $\underset{\sim}{\alpha}^{o}$; f is a continuous mapping from T_1 into T; $\underset{\sim}{f}(\underset{\sim}{\alpha}) = \underset{\sim}{\theta}^{o}$ for $\underset{\sim}{\alpha} \in T_1$ implies $\underset{\sim}{\alpha} = \underset{\sim}{\alpha}^{o}$; T_1 contains a neighborhood of $\underset{\sim}{\alpha}^{o}$ in which $\underset{\sim}{f}$ has continuous second partial derivatives; rank (F) = s, where $\underset{\sim}{F} = \partial \underset{\sim}{f}(\underset{\sim}{\alpha}^{o})/\partial \underset{\sim}{\alpha}'$.

Let $h(\underset{\sim}{\alpha}) = g[\underset{\sim}{f}(\underset{\sim}{\alpha})]$. Choose $\hat{\underset{\sim}{\alpha}}$ to

$$\min_{\underset{\sim}{\alpha} \in T_1} [\underset{\sim}{a}_N - \underset{\sim}{h}(\underset{\sim}{\alpha})]' \underset{\sim}{A}_N [\underset{\sim}{a}_N - \underset{\sim}{h}(\underset{\sim}{\alpha})].$$

Proposition 3'. If Assumptions 1, 2, and 2' are satisfied, if Δ ispositive definite, and if $\underset{\sim}{A}_N \xrightarrow{a.s.} \Delta^{-1}$, then $d_1 - d_2 \xrightarrow{D} \chi^2(p - s)$, where

$$d_1 = N[\underset{\sim}{a}_N - \underset{\sim}{h}(\hat{\underset{\sim}{\alpha}})]' \underset{\sim}{A}_N[\underset{\sim}{a}_N - \underset{\sim}{h}(\hat{\underset{\sim}{\alpha}})],$$

$$d_2 = N[\underset{\sim}{a}_N - \underset{\sim}{g}(\hat{\underset{\sim}{\theta}}) \text{ I } ' \underset{\sim}{A}_N[\underset{\sim}{a}_N - \underset{\sim}{g}(\hat{\underset{\sim}{\theta}})].$$

Furthermore, $d_1 - d_2$ is independent of d_2 in their limiting joint distribution.

Suppose that the restrictions involve only Π. We specify the restrictions by the condition that $\pi = \underset{\sim}{f}(\underset{\sim}{6})$, where $\underset{\sim}{\delta}$ is $s \times 1$ and the domain of 6 is T_1, a subset of R^s that includes the true value δ^o. Consider the following estimator of $\underset{\sim}{\delta}^o$: choose $\hat{\underset{\sim}{\delta}}$ to

$$\min_{\delta \in T_1} [\hat{\underset{\sim}{\pi}} - \underset{\sim}{f}(\underset{\sim}{\delta})]' \hat{\underset{\sim}{\Omega}}^{-1} [\hat{\underset{\sim}{\pi}} - \underset{\sim}{f}(\underset{\sim}{\delta})],$$

where $\hat{\underset{\sim}{\Omega}}$ is given in (4.2), and we assume that $\underset{\sim}{\Omega}$ in (4.1) is positive definite. If Tl and $\underset{\sim}{f}$ satisfy assumptions 1 and 2, then $\hat{\delta} \xrightarrow{a.s.} \underset{\sim}{\delta}^o$,

$$\sqrt{N}(\hat{\underset{\sim}{\delta}} - \underset{\sim}{\delta}^o) \xrightarrow{D} N(\underset{\sim}{0}, [\underset{\sim}{F}' \Omega^{-1} \underset{\sim}{F}]^{-1}),$$

and

$$N[\hat{\underset{\sim}{\pi}} - \underset{\sim}{f}(\hat{\underset{\sim}{\delta}})]' \hat{\underset{\sim}{\Omega}}^{-1}[\hat{\underset{\sim}{\pi}} - \underset{\sim}{f}(\hat{\underset{\sim}{\delta}})] \xrightarrow{D} \chi^2(MK-s),$$

where $F = \partial \underset{\sim}{f}(\underset{\sim}{\delta}^o)/\partial \underset{\sim}{\delta}'$.

We can also estimate $\underset{\sim}{\delta}^o$ by applying the minimum distance procedure to w instead of to $\hat{\pi}$. Suppose that the components of $\underset{\sim}{w}_i$ are arranged so that $\underset{\sim}{w}_i' = (\underset{\sim}{w}_{i1}', \underset{\sim}{w}_{i2}')$, where $\underset{\sim}{w}_{i1}$ contains the components of $\underset{\sim}{x}_i\underset{\sim}{y}_i'$. Partition $\underset{\sim}{\mu} = E(\underset{\sim}{w}_i)$ conformably: $\mu' = (\underset{\sim}{\mu}_1', \underset{\sim}{\mu}_2')$. Set $\underset{\sim}{\theta}' = (\underset{\sim}{\theta}_1', \underset{\sim}{\theta}_2') = (\underset{\sim}{\delta}', \underset{\sim}{\mu}_2')$.

Assume that $V(\underset{\sim}{w}_i)$ is positive definite. Now choose $\hat{\theta}$ to

$$\min_{\theta \in T} [\bar{\underset{\sim}{w}} - \underset{\sim}{g}(\underset{\sim}{\theta})]' \underset{\sim}{A}_N [\bar{\underset{\sim}{w}} - \underset{\sim}{g}(\underset{\sim}{\theta})],$$

where $\underset{\sim}{A}_N \xrightarrow{a.s.} V^{-1}(\underset{\sim}{w}_i)$,

$$\underset{\sim}{g}(\underset{\sim}{\theta}) = \begin{vmatrix} \bar{g}_1[\underset{\sim}{f}(\underset{\sim}{\delta}), \underset{\sim}{\mu}_2] \\ \\ \underset{\sim}{\mu}_2 \end{vmatrix},$$

and $\underset{\sim}{g}_1(\underset{\sim}{\pi}, \underset{\sim}{\mu}_2) = \underset{\sim}{\mu}_1$. Then $\hat{\underset{\sim}{\theta}}_1$ gives an estimator of $\underset{\sim}{\delta}^o$; it has the same limiting distribution as the estimator $\hat{\underset{\sim}{\delta}}$ that we obtained by applying the minimum distance procedure to $\hat{\underset{\sim}{\pi}}$.[36]

This framework leads to some surprising results on efficient estimation. For a simple example, we shall use a univariate linear predictor model,

$$E^*(y_i | x_{i1}, x_{i2}) = \pi_o + \pi_1 x_{i1} + \pi_2 x_{i2}.$$

Consider imposing the restriction $\pi_2 = 0$. Then the conventional estimator of π_1 is b_{yx_1}, the slope coefficient in the least squares regression of y on x_1. We shall show that this estimator is generally less efficient than the minimum distance estimator if the regression function is nonlinear or if there is hetaroskedasticity.

Let $\hat{\pi}_1$, $\hat{\pi}_2$ be the slope coefficients in the least squares multiple regression of y on x_1, x_2. The minimum distance estimator of π_1 under the restriction $\pi_2 = 0$ can be obtained as $\hat{\delta} = \hat{\pi}_1 + \tau \hat{\pi}_2$, where τ is chosen to minimize the (estimated) variance of the limiting distribution of $\hat{\delta}$; this gives

$$\hat{\delta} = \hat{\pi}_1 - \frac{\hat{\omega}_{12}}{\hat{\omega}_{22}} \hat{\pi}_2 ,$$

where $\hat{\omega}_{jk}$ is the estimated covariance between $\hat{\pi}_j$ and $\hat{\pi}_k$ in their Limiting distribution. Since $\hat{\pi}_1 = b_{yx_1} - \hat{\pi}_2 b_{x_2x_1}$, we have

$$\hat{\delta} = b_{yx_1} - (b_{x_2x_1} + \frac{\hat{\omega}_{12}}{\hat{\omega}_{22}}) \hat{\pi}_2 .$$

If $E(y_i|x_{i1}, x_{i2})$ is linear and if $V(y_i|x_{i1}, x_{i2}) = \sigma^2$, then $\omega_{12}/\omega_{22} = -\mathrm{Cov}(x_{i1}, x_{i2})/V(x_{i1})$ and $\hat{\delta} = b_{yx_1}$. But in general $\hat{\delta} \neq b_{yx_1}$ and $\hat{\delta}$ is more efficient than b_{yx_1}. The source of the efficiency gain is that the limiting distribution of $\hat{\pi}_2$ has a zero mean (if $\pi_2 = 0$), and so we can reduce variance without introducing any bias if $\hat{\pi}_2$ is correlated with b_{yx_1}. Under the assumptions of linear regression and homoskedasticity, b_{yx_1} and $\hat{\pi}_2$ are uncorrelated; but this need not be true in the more general framework that we are using.

4.3 Simultaneous Equations: A Generalization of Two-and Three-Stage Least Squares

Given the discussion on imposing restrictions, it is not surprising that two-stage least squares is not, in general, an efficient procedure for combining instrumental variables. I shall demonstrate this with a simple example. Assume that $(y_i, z_i, x_{i1}, x_{i2})$ is i.i.d. according to some distribution with finite fourth moments, and that

$$y_i = \delta z_i + v_i ,$$

where $E(v_i x_{i1}) = E(v_i x_{i2}) = 0$. Assume also that $E(z_i x_{i1}) \neq 0$, $E(z_i x_{i2}) \neq 0$.

Then there are two instrumental variable estimators that both converge a.s. to δ:

$$\hat{\delta}_j = \sum_{i=1}^{N} y_i x_{ij} / \sum_{i=1}^{N} z_i x_{ij} \qquad (j=1, 2),$$

$$\sqrt{N} \left\{ \begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix} - \begin{pmatrix} \delta \\ \delta \end{pmatrix} \right\} \xrightarrow{D} N(0, \underset{\sim}{\Lambda}) ,$$

where the j, k element of Λ is

$$\lambda_{jk} = \frac{E[(y_i - \delta z_i)^2 x_{ij} x_{ik}]}{E(z_i x_{ij})\ E(z_i x_{ik})} \qquad (j,k = 1,2).$$

The two-stage least squares estimator combines $\hat{\delta}_1$ and $\hat{\delta}_2$ by forming $\hat{z}_i = \hat{\pi}_1 x_{i1} + \hat{\pi}_2 x_{i2}$, based on the least squares regression of z on x_1, x2 (assume that $E[(x_{i1}, x_{i2})'(x_{i1}, x_{i2})]$ is nonsingular):

$$\hat{\delta}_{TSLS} = \sum_{i=1}^{N} y_i \hat{z}_i / \sum_{i=1}^{N} z_i \hat{z}_i = \hat{\alpha}\hat{\delta}_1 + (1-\hat{\alpha})\ \hat{\delta}_2 ,$$

where

$$\hat{\alpha} = \hat{\pi}_1 \sum_{i=1}^{N} z_i x_{i1} / (\hat{\pi}_1 \sum_{i=1}^{N} z_i x_{i1} + \hat{\pi}_2 \sum_{i=1}^{N} z_i x_{i2}).$$

Since $\hat{\alpha} \xrightarrow{a.s.} \alpha$, $\sqrt{N}(\hat{\delta}_{TSLS}$ - 6) has the same limiting distribution as

$$\sqrt{N}[\alpha(\hat{\delta}_1 - \delta) + (1-\alpha)(\hat{\delta}_2 - \delta)].$$

This suggests finding the τ that minimizes the variance of the limiting distribution of $\sqrt{N}[\tau(\hat{\delta}_1 - \delta) + (1-\tau)(\hat{\delta}_2 - \delta)]$. The answer leads to the minimum

distance estimator: choose $\hat{\theta}$ to

$$\min_{\theta} \left[\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix} - \begin{pmatrix} \theta \\ \theta \end{pmatrix} \right]' \underset{\sim}{A}^{-1} \left[\begin{pmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix} - \begin{pmatrix} \theta \\ \theta \end{pmatrix} \right]$$

$$\hat{\theta} = \tau \hat{\delta}_1 + (1-\tau) \hat{\delta}_2 ,$$

where

$$\tau = (\lambda^{11} + \lambda^{12})/(\lambda^{11} + 2\lambda^{12} + \lambda^{22}) ,$$

and λ^{jk} is the j,k element of A-1. The estimator obtained by using a consistent estimator of Λ has the same limiting distribution.

In general $\tau \neq \alpha$ since τ is a function of fourth moments and α is not. Suppose, for example, that $z_i = x_{i2}$. Then $\alpha = 0$ but $\tau \neq 0$ unless

$$E[(y_i - \delta z_i)^2 (\frac{x_{i2}^2}{E(x_{i2}^2)} - \frac{x_{i1}x_{i2}}{E(x_{i1}x_{i2})})] = 0.$$

If we add another equation, then we can consider the conventional three-stage least squares estimator. Its limiting distribution is derived in the Appendix (Proposition 7); however, viewed as a minimum distance estimator, it is using the wrong norm in general.

Consider the standard simultaneous equations model:

$$\underset{\sim}{y}_i = \underset{\sim}{\Pi} \underset{\sim}{x}_i + \underset{\sim}{u}_i, \qquad E(\underset{\sim}{u}_i \underset{\sim}{x}_i') = \underset{\sim}{0},$$

$$\underset{\sim}{\Gamma} \underset{\sim}{y}_i + \underset{\sim}{B} \underset{\sim}{x}_i = \underset{\sim}{v}_i ,$$

where $\underset{\sim}{\Gamma} \underset{\sim}{\Pi} + \underset{\sim}{B} = 0$ and $\underset{\sim}{\Gamma} \underset{\sim}{u}_i = \underset{\sim}{v}_i$. We are continuing to assume that $\underset{\sim}{y}_i$ is M×1, $\underset{\sim}{x}_i$ is K×1, $\underset{\sim}{r}_i' = (\underset{\sim}{x}_i', \underset{\sim}{y}_i')$ is i.i.d. according to a distribution with finite fourth moments (i=1, ... N), and that $E(\underset{\sim}{x}_i \underset{\sim}{x}_i')$ is nonsingular.

There are restrictions on Γ and B: $m(\Gamma, B) = 0$, where m is a known function. Assume that the implied restrictions on Π can be specified by the condition that $\pi = \text{vec}(\Pi') = f(\delta)$, where the domain of δ is T_1, a subset of R^s that includes the true value $\delta^o (s \leq MK)$. Assume that Tl and f satisfy assumptions 1 and 2; these properties could be derived from regularity conditions on m, as in Malinvaud (1970, proposition 2, p. 670).

Choose $\hat{\delta}$ to

$$\min_{\delta \in T_1} [\hat{\pi} - f(\delta)]' \hat{\Omega}^{-1} [\hat{\pi} - f(\delta)],$$

where $\hat{\Omega}$ is given in (4.2) and we assume that Ω in (4.1) is positive definite. Let $F = \partial f(\delta^o)/\partial \delta'$. Then we have $\sqrt{N}(\hat{\delta} - \delta^o) \xrightarrow{D} N(0, \Lambda)$, where $\Lambda = (F' \Omega^{-1} F)^{-1}$. This generalizes Malinvaud's minimum distance estimator (p. 676); it reduces to his estimator if $u_i^o u_i^{o'}$ is uncorrelated with $x_i x_i'$, so that $\Omega = E(u_i^o u_i^{o'}) \otimes [E(x_i x_i')]^{-1}$ $(u_i^o = y_i - \Pi^o x_i)$.

Now suppose that the only restrictions on Γ and B are that certain coefficients are zero, together with the normalization restrictions that the coefficient of y_{im} in the m^{th} structural equation is one. Then we can give a* explicit formula for A. Write the m^{th} structural equation as

$$y_{im} = \delta_m' z_{im} + v_{im},$$

where the components of z_{im} are the variables in y_i and x_i that appear in the m^{th} equation with unknown coefficients. Let there be M structural equations and assume that the true value Γ^o is nonsingular. Let $\delta' = (\delta_1', \ldots, \delta_M')$ be $s \times 1$, and let $\Gamma(\delta)$ and $B(\delta)$ be parametric representations of Γ and B that satisfy the zero restrictions and the normalization rule. We can choose a compact set $T_1 \subset R^s$ containing a neighborhood of the true value δ^o, such that $\Gamma(\delta)$ is nonsingular for $\delta \in T_1$. Then $\pi = f(\delta)$,

where $\underset{\sim}{f}(\underset{\sim}{\delta}) = \text{vec}\,[-\underset{\sim}{\Gamma}^{-1}(\underset{\sim}{\delta})\ B(\delta)]'$.

Assume that $\underset{\sim}{f}(\underset{\sim}{\delta}) = \underset{\sim}{\pi}^{o}$ implies that $\underset{\sim}{\delta} = \underset{\sim}{\delta}^{o}$, so that the structural parameters are identified. Then T_1 and f satisfy Assumptions 1 and 2, and $\sqrt{N}(\hat{\underset{\sim}{\delta}} - \underset{\sim}{\delta}^{o}) \xrightarrow{D} N(\underset{\sim}{0}, \underset{\sim}{\Lambda})$. The formula for $\partial\underset{\sim}{\pi}/\partial\underset{\sim}{\delta}'$ is given in Rothenberg (1973, p. 69):

$$\frac{\partial \underset{\sim}{\pi}}{\partial \underset{\sim}{\delta}'} = -(\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{I}_K)\,[\underset{\sim}{\Phi}_{zx}(\underset{\sim}{I}_M \otimes \underset{\sim}{\Phi}_x^{-1})]' \quad ,$$

where $\underset{\sim}{\Phi}_{zx}$ is block diagonal: $\underset{\sim}{\Phi}_{zx} = \text{diag}\{E(\underset{\sim}{z}_{i1}\,\underset{\sim}{x}_i'), \ldots, E(\underset{\sim}{z}_{iM}\,\underset{\sim}{x}_i')\}$, and $\underset{\sim}{\Phi}_x = E(\underset{\sim}{x}_i\underset{\sim}{x}_i')$. So we have

$$\underset{\sim}{\Lambda} = \{\underset{\sim}{\Phi}_{zx}[E(\underset{\sim}{v}_i^{o}\underset{\sim}{v}_i^{o\prime} \otimes \underset{\sim}{x}_i\underset{\sim}{x}_i')]^{-1}\,\underset{\sim}{\Phi}_{zx}'\}^{-1}, \tag{4.3}$$

where $\underset{\sim}{v}_i^{o} = \underset{\sim}{\Gamma}^{o}\underset{\sim}{y}_i + \underset{\sim}{B}^{o}\underset{\sim}{x}_i$. If $\underset{\sim}{u}_i^{o}\underset{\sim}{u}_i^{o\prime}$ is uncorrelated with $\underset{\sim}{x}_i\underset{\sim}{x}_i'$, then this reduces to

$$\underset{\sim}{\Lambda} = \{\underset{\sim}{\Phi}_{zx}\,[E^{-1}(\underset{\sim}{v}_i^{o}\underset{\sim}{v}_i^{o\prime}) \otimes \underset{\sim}{\Phi}_x^{-1}]\,\underset{\sim}{\Phi}_{zx}'\}^{-1} \quad ,$$

which is the conventional asymptotic covariance matrix for three-stage least squares (Zellner and Theil (1962)).

I shall present a generalization of three-stage least squares that has the same limiting distribution as the generalized minimum distance estimator. Let $\beta = \text{vec}(B')$ and note that $\pi = -(\Gamma^{-1} \otimes \underset{\sim}{I})\underset{\sim}{\beta}$. Then we have

$$[\hat{\underset{\sim}{\pi}} + (\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{I})\,\underset{\sim}{\beta}]'\,\underset{\sim}{\Omega}^{-1}[\hat{\underset{\sim}{\pi}} + (\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{I})\,\underset{\sim}{\beta}]$$

$$= [(\underset{\sim}{\Gamma} \otimes \underset{\sim}{I})\hat{\underset{\sim}{\pi}} + \underset{\sim}{\beta}]'\,\underset{\sim}{\Theta}^{-1}[(\underset{\sim}{\Gamma} \otimes \underset{\sim}{I})\hat{\underset{\sim}{\pi}} + \underset{\sim}{\beta}],$$

where

$$\underset{\sim}{\Theta} = (\underset{\sim}{I} \otimes \underset{\sim}{\Phi}_{x}^{-1})\, E(\underset{\sim}{\Gamma}\, \underset{\sim}{u}_{i}^{o}\underset{\sim}{u}_{i}^{\prime o}\, \underset{\sim}{\Gamma}' \otimes \underset{\sim}{x}_{i}\underset{\sim}{x}_{i}')\, (I \otimes \underset{\sim}{\Phi}_{x}^{-1}).$$

Let $\underset{\sim}{S}_{zx}$ be the following block-diagonal matrix:

$$\underset{\sim}{S}_{zx} = \operatorname{diag}\left\{\frac{1}{N}\sum_{i=1}^{N} z_{i1}\underset{\sim}{x}_{i}', \ldots, \frac{1}{N}\sum_{i=1}^{N} z_{iM}\underset{\sim}{x}_{i}'\right\},$$

and let

$$\underset{\sim}{S}_{x} = \frac{1}{N}\sum_{i=1}^{N} \underset{\sim}{x}_{i}\underset{\sim}{x}_{i}', \quad \underset{\sim}{s}_{xy} = \frac{1}{N}\sum_{i=1}^{N} y_{i} \otimes \underset{\sim}{x}_{i}.$$

Let

$$\underset{\sim}{\Psi} = E(\underset{\sim}{v}_{i}^{o}\underset{\sim}{v}_{i}^{\prime o} \otimes \underset{\sim}{x}_{i}\underset{\sim}{x}_{i}'), \quad \hat{\underset{\sim}{\Psi}} = \frac{1}{N}\sum_{i=1}^{N} (\hat{\underset{\sim}{v}}_{i}\hat{\underset{\sim}{v}}_{i}' \otimes \underset{\sim}{x}_{i}\underset{\sim}{x}_{i}'),$$

where

$$\hat{\underset{\sim}{v}}_{i} = \hat{\underset{\sim}{\Gamma}}\, \underset{\sim}{y}_{i} + \hat{\underset{\sim}{B}}\, \underset{\sim}{x}_{i} \text{ and } \hat{\underset{\sim}{\Gamma}} \xrightarrow{a.s.} \underset{\sim}{\Gamma}^{o}, \hat{\underset{\sim}{B}} \xrightarrow{a.s.} \underset{\sim}{B}^{o}.$$

Now replace Θ by

$$\hat{\underset{\sim}{\Theta}} = (\underset{\sim}{I} \otimes \underset{\sim}{S}_{x}^{-1})\, \hat{\underset{\sim}{\Psi}}(\underset{\sim}{I} \otimes \underset{\sim}{S}_{x}^{-1}),$$

and note that

$$(\underset{\sim}{I} \otimes \underset{\sim}{S}_{x})[(\underset{\sim}{\Gamma} \otimes \underset{\sim}{I})\, \hat{\underset{\sim}{\pi}} + \underset{\sim}{\beta}] = \underset{\sim}{s}_{xy} - \underset{\sim}{S}_{zx}'\, \underset{\sim}{\delta}.$$

Then we have the following distance function:

$$(\underset{\sim}{s}_{xy} - \underset{\sim}{S}_{zx}'\, \underset{\sim}{\delta})'\, \hat{\underset{\sim}{\Psi}}^{-1}\, (\underset{\sim}{s}_{xy} - \underset{\sim}{S}_{zx}'\, \underset{\sim}{\delta}).$$

This corresponds to **Basmann's** (1965) interpretation of three-stage least squares. [17]

Minimizing with respect to δ gives

$$\hat{\delta}_{G3} = (S_{zx}\hat{\Psi}^{-1}S'_{zx})^{-1}(S_{zx}\hat{\Psi}^{-1}s_{xy}).$$

The limiting distribution of this estimator is derived in the Appendix (Proposition 7). We record it es

Proposition 4. $\sqrt{N}(\hat{\delta}_{G3} - \delta^{o}) \xrightarrow{D} N(0, A)$, where A is given in (4.3). This generalized three-stage least squares estimator is asymptotically efficient within the class of minimum distance estimators.

Our derivation of the limiting distribution of $\hat{\delta}_{G3}$ relies on linearity. For a generalized nonlinear three-stage least squares estimator, see Hansen (1982).

Finally, we shall consider the generalization of two-stage least squares.[38] Suppose that

$$y_{i1} = \delta_1' z_{i1} + v_{i1},$$

where $E(x_i v_{i1}) = 0$, z_{i1} is $s_1 \times 1$, and rank $[E(x_i z_{i1}')] = s_1$. We complete the system by setting

$$y_{im} = \pi_m' x_i + u_{im},$$

where $E(x_i u_{im}) = 0$ $(m = 2, \ldots, M)$. So $z_{im} = x_i$ $(m = 2, \ldots, M)$, and

$$\Phi_{zx} = \text{diag}\{E(z_{i1}x_i'), I_{M-1} \otimes E(x_i x_i')\}.$$

Let $\delta' = (\delta_1', \pi_2', \ldots, \pi_M')$ and apply the minimum distance procedure to obtain $\hat{\delta}$; since we are ignoring any restrictions on π_m $(m = 2, \ldots, M)$,

$\hat{\delta}$ is a limited information minimum distance estimator.

We have $\sqrt{N}(\hat{\underset{\sim}{\delta}}_1 - \underset{\sim}{\delta}_1^o) \xrightarrow{D} N(\underset{\sim}{0}, \underset{\sim}{\Lambda}_{11})$, and evaluating the partitioned inverse gives

$$\underset{\sim}{\Lambda}_{11} = \{E(\underset{\sim}{z}_{i1}\,\underset{\sim}{x}_i')\,[E((v_{i1}^o)^2\,\underset{\sim}{x}_i\underset{\sim}{x}_i')]^{-1}\,E(\underset{\sim}{x}_i\underset{\sim}{z}_{i1}')\}^{-1}, \tag{4.4}$$

where $v_{i1}^o = y_{i1} - \underset{\sim}{\delta}_1'^o \underset{\sim}{z}_{i1}$.

We can obtain the same limiting distribution by using the following generalization of two-stage least squares: Let

$$\underset{\sim}{Z}_1' = (\underset{\sim}{z}_{11}, \ldots, \underset{\sim}{z}_{N1}),\ \underset{\sim}{X}' = (\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_N),$$

$$\underset{\sim}{y}_1' = (y_{11}, \ldots, y_{N1}), \text{ and}$$

$$\hat{\underset{\sim}{\Psi}}_{11} = \frac{1}{N}\sum_{i=1}^{N} (y_{i1} - \hat{\underset{\sim}{\delta}}_1'\,\underset{\sim}{z}_{i1})^2\,\underset{\sim}{x}_i\underset{\sim}{x}_i',$$

where $\hat{\underset{\sim}{\delta}}_1 \xrightarrow{a.s.} \underset{\sim}{\delta}_1^o$ (for example, $\hat{\underset{\sim}{\delta}}_1$ could be an instrumental variable estimator); then

$$\hat{\underset{\sim}{\delta}}_{1G2} = (\underset{\sim}{Z}_1'\,\underset{\sim}{X}\,\hat{\underset{\sim}{\Psi}}_{11}^{-1}\,\underset{\sim}{X}'\underset{\sim}{Z}_1)^{-1}\,(\underset{\sim}{Z}_1'\,\underset{\sim}{X}\,\hat{\underset{\sim}{\Psi}}_{11}^{-1}\,\underset{\sim}{X}'\,\underset{\sim}{y}_1).$$

This is the estimator of $\underset{\sim}{\delta}_1$ that we obtain by applying generalized three-stage least squares to the completed system, with no restrictions on $\underset{\sim}{\pi}_m$ ($m = 2, \ldots, M$). The limiting distribution of this estimator is derived in the Appendix (Proposition 7):

Proposition 5. $\sqrt{N}(\hat{\delta}_{1G2} - \delta_1^o) \xrightarrow{D} N(0, A_{11})$, where A_{11} is given in (4.4). This generalized two-stage least squares estimator is asymptotically efficient in the class of limited information minimum distance estimators.

4.4 *Asymptotic Efficiency: A Comparison with the Quasi-Maximum Likelihood Estimator*

Assume that r_i is i.i.d. (i=1,2,...) from a distribution with $E(r_i) = \tau$, $V(r_i) = \Sigma$, where Σ is a JxJ positive-definite matrix; the fourth moments are finite. Suppose that we wish to estimate functions of Σ subject to restrictions. Let $\sigma = \text{vec}(\Sigma)$ and express the restrictions by the condition that $\sigma = g(\theta)$, where g is a function from T into R^q with a domain $T \subset R^p$ that contains the true value $\theta^o (q = J^2;\ p \leq J(J+1)/2)$. Let

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})(r_i - \bar{r})',$$

and let $\bar{s} = \text{vec}(\bar{S})$.

If the distribution of r_i is multivariate normal, then the log-likelihood function is

$$L = \frac{N}{2} \ell n\, |\Sigma^{-1}| - \frac{N}{2} tr\{\Sigma^{-1}[\bar{S} + (\bar{r} - \tau)(\bar{r} - \tau)']\}.$$

If there are no restrictions on τ, then the maximum likelihood estimator of θ^o is a solution to the following problem: Choose $\hat{\theta}$ to solve

$$\frac{\partial g'(\theta)}{\partial \theta}\,[\Sigma^{-1}(\theta) \otimes \Sigma^{-1}(\theta)](\bar{s} - g(\theta)) = 0.$$

We shall derive the properties of this estimator when the distribution of

$\underset{\sim}{r}_i$ is not necessarily normal; in that case we shall refer to the estimator as a quasi-maximum likelihood estimator ($\hat{\underset{\sim}{\theta}}_{QML}$).[39]

MaCurdy (1979) considered a version of this problem and showed that, under suitable regularity conditions, $\sqrt{N}(\hat{\underset{\sim}{\theta}}_{QML} - \underset{\sim}{\theta}^o)$ has a limiting normal distribution; the covariance *matrix*, however, is not given by the standard information matrix formula. We would like to compare this distribution with the distribution of the minimum distance estimator.

This comparison can be readily made by using theorem 1 in Ferguson (1958). In our notation, Ferguson considers the following problem: Choose $\hat{\theta}$ to solve

$$\underset{\sim}{W}(\bar{\underset{\sim}{s}}, \underset{\sim}{\theta}) [\bar{\underset{\sim}{s}} - \underset{\sim}{g}(\underset{\sim}{\theta})] = \underset{\sim}{0}.$$

He derives the limiting distribution of $\sqrt{N}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}^o)$ under regularity conditions on the functions $\underset{\sim}{W}$ and $\underset{\sim}{g}$. These regularity conditions are particularly simple in our problem since W does not depend on $\bar{\underset{\sim}{s}}$. We can state them as follows:

Assumption 3. $\Xi_o \subset R^p$ is an open set containing $\underset{\sim}{\theta}^o$; g is a continuous, one-to-one mapping of Ξ_o into R^q with a continuous inverse; g has continuous second partial derivatives in Ξ_o; rank $[\partial \underset{\sim}{g}(\underset{\sim}{\theta})/\partial \underset{\sim}{\theta}'] = p$ for $\underset{\sim}{\theta} \epsilon \Xi_o$; $\underset{\sim}{\Sigma}(\underset{\sim}{\theta})$ is non-singular for $\underset{\sim}{\theta} \epsilon \Xi_o$.

In addition, we shall need $\bar{\underset{\sim}{s}} \xrightarrow{a.s.} \underset{\sim}{g}(\underset{\sim}{\theta}^o)$ and the central limit theorem result that $\sqrt{N}(\bar{\underset{\sim}{s}} - \underset{\sim}{g}(\underset{\sim}{\theta}^o)) \xrightarrow{D} \underset{\sim}{N}(\underset{\sim}{0}, \underset{\sim}{\Delta})$, where $A = V[(\underset{\sim}{r}_i - \underset{\sim}{\tau}^o) \otimes (\underset{\sim}{r}_i - \underset{\sim}{\tau}^o)]$.

Then Ferguson's theorem implies that the likelihood equations almost surely have a unique solution within Ξ_o for sufficiently large N, and $\sqrt{N}(\hat{\underset{\sim}{\theta}}_{QML} - \underset{\sim}{\theta}^o) \xrightarrow{D} N(0, \underset{\sim}{\Delta})$, where

$$\Lambda = (G'\Psi G)^{-1} G'\Psi A \Psi G(G'\Psi G)^{-1},$$

and $G = \partial g(\theta^o)/\partial\theta'$, $\Psi = (\Sigma^o \otimes \Sigma^o)^{-1}$. It will be convenient to rewrite this, imposing the symmetry restrictions on Σ. Let σ^* be the $J(J+1)/2 \times 1$ vector formed by stacking the columns of the lower triangle of Σ. We can define a $J^2 \times [J(J+1)/2]$ matrix T such that $\sigma = T\sigma^*$. The elements in each row of T are all zero except for a single element which is one; T has full column rank. Let $\bar{s} = T'\bar{s}^*$, $g(\theta) = T\, g^*(\theta)$, $G^* = \partial g^*(\theta^o)/\partial\theta'$, $\Psi^* = T'\Psi T$; then $\sqrt{N}[\bar{s}^* - g^*(\theta^o)] \xrightarrow{D} N(0, A^*)$, where A^* is the covariance matrix of the vector formed from the columns of the lower triangle of $(r_i - \tau^o)(r_i - \tau^o)'$. Now we can set

$$\Lambda = (G'^* \Psi^* G^*)^{-1}(G'^* \Psi^* A^* \Psi^* G^*)(G'^* \Psi^* G^*)^{-1}.$$

Consider the following minimum distance estimator: Choose $\hat{\theta}_{MD}$ to

$$\min_{\theta\in T} [\bar{s}^* - g^*(\theta)]' A_N [\bar{s}^* - g^*(\theta)],$$

where T is a compact subset of Ξ_o that contains a neighborhood of θ^o and $A_N \xrightarrow{a.s.} \Psi^*$. Then the following result is implied by Proposition 2.

Proposition 6. If Assumption 3 is satisfied, then $\sqrt{N}(\hat{\theta}_{QML} - \theta^o)$ has the same limiting distribution as $\sqrt{N}(\hat{\theta}_{MD} - \theta^o)$.

If A^* is nonsingular, an optimal minimum distance estimator has $A_N \xrightarrow{a.s.} \zeta\Delta^{*-1}$, where ζ is an arbitrary positive real number. If the distribution of r_i is normal, then $A^{*-1} = (1/2)\Psi^*$; but in general Δ^{*-1} is not proportional to Ψ^*, since A^* depends on fourth moments and Ψ^* is a function of second moments.

So in general $\hat{\underset{\sim}{\beta}}_{QML}$ is less efficient than the optimal minimum distance estimator that uses

$$(4.5) \qquad \underset{\sim}{A}_N = \left[\frac{1}{N}\sum_{i=1}^{N}(\underset{\sim}{s}_i^* - \bar{\underset{\sim}{s}}^*)(\underset{\sim}{s}_i^* - \bar{\underset{\sim}{s}}^*)'\right]^{-1},$$

where $\underset{\sim}{s}_i^*$ is the vector formed from the lower triangle of $(\underset{\sim}{r}_i - \bar{\underset{\sim}{r}})(\underset{\sim}{r}_i - \bar{\underset{\sim}{r}})'$.

More generally, we can consider the class of consistent estimators that are continuously differentiable functions of $\bar{\underset{\sim}{s}}^*$: $\hat{\underset{\sim}{\theta}} = \hat{\underset{\sim}{\theta}}(\bar{\underset{\sim}{s}}^*)$. Chiang (1956) shows that the minimum distance estimator based on $\underset{\sim}{\Delta}^{*-1}$ has the minimal asymptotic covariance matrix within this class. The minimum distance estimator based on $\underset{\sim}{A}_N$ in (4.5) attains this lower bound.

4.5. *Multivariate Probit Models*

Suppose that

$$y_{im} = 1 \text{ if } \underset{\sim}{\pi}_m' \underset{\sim}{x}_i + u_{im} \geq 0,$$

$$= 0 \text{ otherwise } (i=1, \ldots, N;\ m=1, \ldots, M),$$

where the distribution of $\underset{\sim}{u}_i' = (u_{i1}, \ldots, u_{iM})$ conditional on $\underset{\sim}{x}_i$ is multivariate normal, $N(\underset{\sim}{0}, \underset{\sim}{\Sigma})$. There may be restrictions on $\underset{\sim}{\pi}' = (\underset{\sim}{\pi}_1', \ldots, \underset{\sim}{\pi}_M')$, but we want to allow $\underset{\sim}{\Sigma}$ to be unrestricted, except for the scale normalization that the diagonal elements of $\underset{\sim}{\Sigma}$ are equal to one. In that case, the maximum likelihood estimator has the computational disadvantage of requiring numerical integration over M-1 dimensions.

Our strategy is to avoid numerical integration. We estimate $\underset{\sim}{\pi}_m$ by maximizing the marginal likelihood function that is based on the distribution of y_{im} conditional on $\underset{\sim}{x}_i$.

$$P(y_{im} = 1 | \underset{\sim}{x}_i) = F(\underset{\sim}{\pi}_m' \underset{\sim}{x}_i),$$

where F is the standard **normal** distribution function. Then-under standard assumptions we have $\hat{\underset{\sim}{\pi}}_m \xrightarrow{a.s.} \underset{\sim}{\pi}_m^o$, the true value. If $\sqrt{N}(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}^o) \xrightarrow{D} N(\underset{\sim}{0}, \underset{\sim}{\Omega})$, then we can impose the restriction that $\underset{\sim}{\pi} = \underset{\sim}{f}(\underset{\sim}{\delta})$ by choosing $\hat{\underset{\sim}{\delta}}$ to minimize

$$[\hat{\underset{\sim}{\pi}} - \underset{\sim}{f}(\underset{\sim}{\delta})]' \hat{\underset{\sim}{\Omega}}^{-1} [\hat{\underset{\sim}{\pi}} - \underset{\sim}{f}(\underset{\sim}{\delta})].$$

We only need to derive a **formula** for $\underset{\sim}{\Omega}$.[40]

Our estimator of $\underset{\sim}{\pi}$ is solving the following equation:

$$\underset{\sim}{s}(\hat{\underset{\sim}{\pi}}) = \frac{\partial Q(\hat{\underset{\sim}{\pi}})}{\partial \underset{\sim}{\pi}} = \underset{\sim}{0},$$

where

$$Q(\underset{\sim}{\pi}) = \sum_{i=1}^{N} \{ \sum_{m=1}^{M} y_{im} \ln F(\underset{\sim}{\pi}_m' \underset{\sim}{x}_i) + (1-y_{im}) \ln [1-F(\underset{\sim}{\pi}_m' \underset{\sim}{x}_i)] \}.$$

Hence the asymptotic distribution of $\hat{\underset{\sim}{\pi}}$ **can** be obtained from the theory of "M-estimators." **Huber** (1967) provides general results, which do not impose differentiability restrictions on $\underset{\sim}{s}(\underset{\sim}{\pi})$. His results cover. for example, regression estimators based on minimizing the residual sum of absolute deviations. We shall not need **this** generality here and shall sketch the derivation for the simpler, differentiable case. This case has been considered by Hansen **(1982), MaCurdy (1981a),** and White (1982).[41]

Let $\underset{\sim}{z}_i$ be **i.i.d.** according to a distribution with support $Z \subset R^q$. Let Θ be an open, convex subset of R^p and let $\underset{\sim}{\psi}(\underset{\sim}{z}, \underset{\sim}{\theta})$ **be** a function from $Z \times \Theta$ into R^p; its $k^{\underline{th}}$ component is $\psi_k(\underset{\sim}{z}, \underset{\sim}{\theta})$. For each $\underset{\sim}{\theta} \in \Theta$, $\underset{\sim}{\psi}$ is a measurable function of $\underset{\sim}{z}$, and there is a $\underset{\sim}{\theta}^o \in \Theta$ with

$$E[\underset{\sim}{\psi}(\underset{\sim}{z}_1, \underset{\sim}{\theta}^0)] = \underset{\sim}{0}, \quad E[\underset{\sim}{\psi}(\underset{\sim}{z}_1, \underset{\sim}{\theta}^0)\underset{\sim}{\psi}'(\underset{\sim}{z}_1, \underset{\sim}{\theta}^0)] = \underset{\sim}{\Delta} < \infty.$$

For each $\underset{\sim}{z} \in Z$, $\underset{\sim}{\psi}$ is a twice continuously differentiable function of θ. In addition,

$$\underset{\sim}{J} = E\left[\frac{\partial \underset{\sim}{\psi}(\underset{\sim}{z}_1, \underset{\sim}{\theta}^0)}{\partial \underset{\sim}{\theta}'}\right]$$

is nonsingular, and

$$\left|\frac{\partial \psi_k^2(\underset{\sim}{z}, \underset{\sim}{\theta})}{\partial \theta_\ell \partial \theta_m}\right| \leq h(\underset{\sim}{z}) \qquad (k, \ell, m = 1, \ldots, p)$$

for $\underset{\sim}{\theta} \in \Theta$, where $E[h(z_1)] < \infty$.

Suppose that we have a (measurable) estimator $\hat{\underset{\sim}{\theta}}_N \in \Theta$ such that $\hat{\underset{\sim}{\theta}}_N \xrightarrow{a.s.} \underset{\sim}{\theta}^0$ and

$$\sum_{i=1}^{N} \underset{\sim}{\psi}(\underset{\sim}{z}_i, \hat{\underset{\sim}{\theta}}_N) = \underset{\sim}{0}$$

for sufficiently large N a.s. By Taylor's theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_k(\underset{\sim}{z}_i, \underset{\sim}{\theta}^0) + [\underset{\sim}{j}'_{Nk} + \frac{1}{2}(\hat{\underset{\sim}{\theta}}_N - \underset{\sim}{\theta}^0)' \underset{\sim}{C}_{Nk}][\sqrt{N}(\hat{\underset{\sim}{\theta}}_N - \underset{\sim}{\theta}^0)] = \underset{\sim}{0},$$

where

$$\underset{\sim}{j}_{Nk} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \psi_k(\underset{\sim}{z}_i, \underset{\sim}{\theta}^0)}{\partial \underset{\sim}{\theta}}, \quad \underset{\sim}{C}_{Nk} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \psi_k(\underset{\sim}{z}_i, \underset{\sim}{\theta}^*_{Nk})}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'},$$

and $\underset{\sim}{\theta}^*_{Nk}$ is on the line segment joining $\hat{\underset{\sim}{\theta}}_N$ and $\underset{\sim}{\theta}^0$ (k=1, ..., P). (The measurability of $\underset{\sim}{\theta}^*_{Nk}$ follows from lemma 3 of Jennrich (1969).) By the strong law of large numbers, $\underset{\sim}{j}'_{Nk}$ converges a.s. to the k^{th} row of $\underset{\sim}{J}$, and

$$\left| \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \psi_k(z_i, \theta^*_{Nk})}{\partial\theta_\ell \partial\theta_m} \right| \leq \frac{1}{N} \sum_{i=1}^{N} h(z_i) \xrightarrow{a.s.} E[h(z_1)]$$

(k, ℓ, m=1, ... p). Hence $(\hat{\theta}_N - \theta^o)' C_{Nk} \to 0$ a.s. and

$$\sqrt{N}(\hat{\theta}_N - \theta^o) = -D_N^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \theta^o) \right]$$

for N sufficiently large a.s., where $D_N \xrightarrow{a.s.} J$. By the central limit theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \theta^o) \xrightarrow{D} N(0, \Delta).$$

Hence

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \xrightarrow{D} N(0, J^{-1}\Delta J'^{-1}).$$

Applying this result to our multivariate probit estimator gives

$$\sqrt{N}\,(\hat{\pi} - \pi^o) \xrightarrow{D} N(0, J^{-1}\Delta J^{-1}),$$

where $J = \text{diag}\{J_1, \ldots, J_M\}$ is a block-diagonal matrix with

$$J_m = E[\{(F')^2/[F(1-F)]\}\, x_1 x_1']$$

(F and its derivative F' are evaluated at $\pi_m^{o\prime} x_1$);

and

$$\Delta = E[H \otimes x_1 x_1'],$$

where the m, n element of the M×M matrix H is $h_{mn} = e_m e_n$ with

$$e_m = \frac{y_{1m} - F}{F(1-F)} F' \qquad (m=1 \ldots, M)$$

(F and F' are evaluated at $\underset{\sim}{\pi}^{o}{}'\underset{\sim}{x}_1$). We obtain a 'consistent estimator $(\hat{\underset{\sim}{\Omega}})$ of $\underset{\sim}{J}^{-1}\underset{\sim}{\Delta}\underset{\sim}{J}^{-1}$ by replacing expectations by sample means and using $\hat{\underset{\sim}{\pi}}$ in place of π^{o}. Then we can apply the minimum distance theory of Section 4.2 to impose restrictions on π.

5. EMPIRICAL APPLICATIONS

5.1 Linear Models: Union Wage Effects

We shall present an empirical example that illustrates some of the preceding results.[42] The data come from the panel of Young Men in the National Longitudinal Survey (Parnes). The sample consists of 1454 young men who were not enrolled in school in 1969, 1970, or 1971, and who had complete data on the variables listed in Table 1. Table 2.1 presents an unrestricted least squares regression of the logarithm of wage in 1969 on the union, SMSA, and region variables for all three years. The regression also includes a constant, schooling, experience, experience squared, and race. This regression is repeated using the 1970 wage and the 1971 wage.

In Section 2 we discussed the implications of a random intercept (c). If the leads and lags are due just to c, then the submatrices of Π corresponding to the union, SMSA, or region coefficients should have the form $\beta I + \underset{\sim}{\ell}\underset{\sim}{\lambda}'$. Consider, for example, the 3×3 submatrix of union coefficients -- the off-diagonal elements in each column should be equal to each other. So we compare .048 to .046, .042 to .041, and'-.009 to .010; not bad.

In Table 2.2 we add a complete set of union interactions, so that, for' the union variables at least, we have a general regression function. Now *the* submatrix of union coefficients *is* 3×7. If it equals $(\beta\underset{\sim}{I}_3, 0) + \underset{\sim}{\ell}\underset{\sim}{\lambda}'$, then

Table 1

CHARACTERISTICS OF NATIONAL LONGITUDINAL SURVEY YOUNG MEN, NOT ENROLLED IN SCHOOL IN 1969, 1970, 1971:

Means and Standard Deviations

N = 1454

Variable	Mean	Standard Deviation
LW1	5.64	.423
LW2	5.74	.426
LW3	5.82	.437
U1	.336	
u2	.362	
u3	.364	
U1U2	.270	
U1U3	.262	
U2U3	.303	
U1U2U3	.243	
SMSA1	.697	
SMSA2	.627	
SMSA3	.622	
RNS1	.409	
RNS2	.404	
RNs3	.410	
S	11.7	2.64
EXP69	5.11	3.71
$EXP69^2$	39.8	46.6
RACE	.264	

Notes to *Table 1:*

LW1, LW2, LW3 -- logarithm of hourly earnings (in cents) on the current or last job in 1969, 1970, 1971; **U1, U2,** U3 -- I if wages on current or last job set by collective bargaining, 0 if not, in 1969, 1970, 1971; **SMSA1, SMSA2, SMSA3** -- 1 if respondent in SMSA. 0 if not, in 1969, 1970, 1971; **RNS1, RNS2, RNS3** -- 1 if respondent in South, 0 if not, in 1969, 1970, 1971; **S** -- years of schooling completed; **EXP69** -- (age in 1969 - S - 6); RACE -- 1 if respondent black, 0 if not.

TABLE 2

UNRESTRICTED LEAST SQUARES REGRESSIONS

2.1

Dependent Variable	Coefficients (and Standard Errors) of:								
	U1	U2	u3	SMSA1	SMSA2	SMSA3	RNS1	RNS2	RNS3
LW1	.171	.042	-.009	.135	-.001	.032	-.016	-.020	-.108
	(.025)	(.026)	(.025)	(.028)	(.055)	(.054)	(.081)	(.081)	(.070)
LW2	.048	.150	.010	.086	.053	.020	.065	-.039	-.155
	(.023)	(.028)	(.026)	(.027)	(.065)	(.061)	(.099)	(.109)	(.092)
LW3	.046	.041	.132	.083	.003	.088	.074	.056	-.232
	(.023)	(.030)	(.030)	(.031)	(.058)	(.056)	(.079)	(.093)	(.078)

Notes to Table 2.1:

All regressions include (1, S, EXP69, $EXP69^2$, RACE). The standard errors are calculated using $\hat{\Omega}$ in (4.2).

Dependent Variable	Coefficients (and Standard Errors) of:						
	U1	U2	u3	U1U2	Ulu3	U2U3	U1U2U3
LW1	.127	-.047	-.072	.128	.092	.156	-.182
	(.044)	(.042)	(.041)	(.072)	(.075)	(.070)	(.104)
LW2	-.019	.014	-.085	.181	.118	.227	-.229
	(.040)	(.045)	(.040)	(.074)	(.092)	(.066)	(.116)
LW3	-.050	-.072	-.022	.110	.264	.246	-.256
	(.037)	(.053)	(.052)	(.079)	(.081)	(.079)	(.113)

Notes to Table 2.2:

All regressions include (SMSA1, SMSA2, SMSA3, RNS1, RNS2, RNS3, 1, S, EXP69, $EXP69^2$, RACE). The standard errors are calculated using $\hat{\Omega}$ in (4.2).

in the first three columns, the off-diagonal elements within a column should be equal; in the last four columns, all elements within a column should be equal.

I first imposed the restrictions on the SMSA and region coefficients, using the minimum distance estimator. $\hat{\Omega}$ is estimated using the formula in (4.2), and $A_N = \hat{\Omega}^{-1}$. The minimum distance statistic (Proposition 3) is 6.82, which is not a surprising value from a $\chi^2(10)$ distribution. If we impose the restrictions on the union coefficients as well, then the 21 coefficients in Table 2.2 are replaced by 8: one β and seven λ's. This gives an increase in the minimum distance statistic (Proposition 3') of 19.36 - 6.82 = 12.54, which is not a surprising value from a $\chi^2(13)$ distribution. So there is no evidence here against the hypothesis that all the lags and leads are generated by c. In the terminology of Section 3.3, the (linear predictor) relationship of x to y appears to be static conditional on c.

Consider a transformation of the model in which the dependent variables are LW1, LW2-LW1, and LW3-LW2. Start with a multivariate regression on all of the lags and leads (and union interactions); then impose the restriction that U, SMSA, and RNS appear in the LW2-LW1 and LW3-LW2 equations only as contemporaneous changes ($E(y_t - y_{t-1} | x_1, x_2, x_3) = \beta(x_t - x_{t-1})$). This is equivalent to the restriction that c generates all of the lags and leads, and we have seen that it is supported by the data. I also considered imposing all of the restrictions with the single exception of allowing separate coefficients for entering and leaving union coverage in the wage change equations. The estimates (standard errors) are .097 (.019) and

-.119 (.022). The standard error on the sum of the coefficients is .024, so again there is no evidence against the simple model with $E(y_t|x_1,x_2,x_3,c) = \beta x_t + c$.[43]

Table 3.1 exhibits the estimates that result from imposing the restrictions using the optimal minimum distance estimator.[44] We also give the conventional generalized least squares estimates. They are minimum distance estimates in which the weighting matrix (A_N) is the inverse of

$$(5.1) \qquad \hat{\Omega}_s = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\Pi}x_i)(y_i - \hat{\Pi}x_i)' \otimes \left(\frac{1}{N}\sum_{i=1}^{N}x_i x_i'\right)^{-1}.$$

We give the conventional standard errors based on $(F'\hat{\Omega}_s^{-1}F)^{-1}$ and the standard errors calculated according to Proposition 2. which do not require an assumption of homoskedastic linear regression. These standard errors are larger than the conventional ones, by about 30%. The estimated gain in efficiency from using the appropriate metric is not very large; the standard errors calculated according to Proposition 2 are about 10% larger when we use conventional GLS instead of the optimum minimum distance estimator.

Table 3.1 also presents the estimated λ's. Consider, for example, an individual who was covered by collective bargaining in 1969. The linear predictor of c increases by .089 if he is also covered in 1970, and it increases by an additional .036 if he is covered in all three years. The predicted c for someone who is always covered is higher by .102 than for someone who is never covered.

Table 3.2 presents estimates under the constraint that $\lambda = 0$. The increment in the distance statistic is 89.08 - 19.36 = 69.72, which is a

TABLE 3

RESTRICTED ESTIMATES

3.1

	Coefficients (and Standard Errors) of:		
	U	SMSA	RNS
$\hat{\beta}$:	.107 (.016)	.056 (.020)	-.082 (.045)
$\hat{\beta}_{GLS}$:	.121 (.013) (.018)	.050 (.017) (.021)	-.085 (.040) (.052)

	U1	u2	U3	Ulu2	U1U3	U2U3	U1U2U3
λ:	-.02 (.03	-.067 (.040)	-.082 (.037)	.156 (.057)	.152 (.062)	.195 (.059)	-.229 (.085)
	SMSA1	SMSA2	SMSA3	RNS1	RNs2	RNS3	
	.086 (.025)	-.008 (.046)	.032 (.046)	.100 (.072)	-.021 (.077)	-.128 (.068)	

$\chi^2(23) = 19.36$

3.2 *Restrict* $\lambda = 0$.

	Coefficients (and Standard Errors) of:		
	U	SMSA	RNS
$\hat{\beta}$:	.157 (.012)	.120 (.013)	-.150 (.016)

$\chi^2(36) = 89.08$

Notes to Table 3:

$E^*(y \mid x) = \Pi x = \Pi_1 x_1 + \Pi_2 x_2$; x_1' = (U1, U2, U3, U1U2, U1U3, U2U3, U1U2U3, SMSA1, SMSA2, SMSA3, RNS1, RNS2, RNS3); x_2' = (1, S, EXP69, $EXP69^2$, FACE) . $\Pi_1 = (\beta_u I_3, 0, \beta_{SMSA} I_3, \beta_{RNS} I_3) + \ell \lambda'$; Π_2 is unrestricted. The restrictions are expressed as $\pi = F \delta$, where δ is unrestricted. $\hat{\beta}$ and $\hat{\lambda}$ are minimum distance estimates with $A_N^{-1} = \hat{\Omega}$ in (4.2); $\hat{\beta}_{GLS}$ and $\hat{\lambda}_{GLS}$ are minimum distance estimates with $A_N^{-1} = \hat{\Omega}_s$ in (5.1) ($\hat{\lambda}_{GLS}$ is not shown in the table). The first standard error for $\hat{\beta}_{GLS}$ is the conventional one based on $(F'\hat{\Omega}_s^{-1}F)^{-1}$; the second standard error for $\hat{\beta}_{GLS}$ is based on $(F'\hat{\Omega}_s^{-1}F)^{-1} F'\hat{\Omega}_s^{-1} \hat{\Omega} \hat{\Omega}_s^{-1} F(F'\hat{\Omega}_s^{-1}F)^{-1}$ (Proposition 2). The χ^2 statistics are computed from $N[\hat{\pi} - F \hat{\delta}]' \hat{\Omega}^{-1}[\hat{\pi} - F\hat{\delta}]$ (Proposition 3).

surprisingly large value to come from a χ^2 (13) distribution. If we constrain only the union λ's to be zero, then the increment is 57.06 - 19.36 = 37.7, which is surprisingly large coning from a $\chi^2(7)$ distribution. So there is strong evidence for heterogeneity bias.

The union coefficient declines from .157 to .107 when we relax the $\underset{\sim}{A} = \underset{\sim}{0}$ restriction. The least squares estimates for the separate cross sections, with no leads or lags, give union coefficients of .195, .189, and .191 in 1969, 1970, and 1971.[45] So the decline in the union coefficient, when we allow for heterogeneity bias, is 32% or 44% depending on which biased estimate (.16 or .19) one uses. The SMSA and region coefficients also decline in absolute value. The least squares estimates for the separate cross sections give an average SMSA coefficient of .147 and an average region coefficient of -.131. So the decline in the SMSA coefficient is either 53% or 62%, and the decline in absolute value of the region coefficient is either 45% or 37%.

5.2. Nonlinear Models: *Labor Force Participation*

We shall illustrate some of the results in Section 3. The sample consists of 924 married women in the Michigan Panel Study of Income Dynamics. The sample selection criteria and the means and standard deviations of the variables are in Table 4. Participation status is measured by the question "Did _____ do any *work* for money last yeas?" We shall model participation in 1968, 1970, 1972, and 1974.

In terms of the model described in Section 3.1, the wage predictors are schooling, experience, and experience squared, where experience is measured

as age minus schooling minus six; the tastes for nonmarket time are predicted by these variables and by children. The specification for children is a conventional one that uses the number of children of age less than six (YS) and the total number of children in the family unit (K).[46] Variables that affect only the lifetime budget constraint in this certainty model are captured by c. In particular, nonlabor income and the husband's wage are assumed to affect the wife's participation only through the lifetime budget constraint. The individual effect (c) will also capture unobserved permanent components in wages or in tastes for nonmarket time.

Table 5 presents maximum likelihood (ML) estimates of cross-section probit specifications for each of the four years. Table 6 presents unrestricted ML estimates for all lags and leads in YK and K. If the residuals (u_{it}) in the latent variable model (3.1) have constant variance, then $\alpha_1 = \ldots = \alpha_4$ *in* (3.8), and the submatrices of Π corresponding to YK and K should have the form $\beta \underset{\sim}{I} + \underset{\sim}{\ell}\,\underset{\sim}{\lambda}'$. There may be some indication of this pattern in Table 6, but it is much weaker than in the wage regressions in Table 2.

we allow for unequal variances and provide formal tests by using the minimum distance estimator developed in Section 4.5. In Table 7.1 we impose the restrictions that

$$\underset{\sim}{\Pi} = \text{diag}\{\alpha_1, \ldots, \alpha_4\}\,[\beta_{YK}\,\underset{\sim}{I}_4 + \underset{\sim}{\ell}\,\underset{\sim}{\lambda}'_{YK},\ \beta_K\,\underset{\sim}{I}_4 + \underset{\sim}{\ell}\,\underset{\sim}{\lambda}'_K]$$

The minimum distance statistic is 53.8, which is a very surprising value coming from a $\chi^2(19)$ distribution. So the latent variable c does not appear to provide an adequate interpretation of the unrestricted leads and lags.

It may be that the distributed lag relationship between current participation and previous births is more general than the one implied by summing over the previous six years (YK) and over the previous eighteen years (K). It may be fruitful to explore this in more detail in future work. Perhaps strict exogeneity conditional on c will hold when we use a more general specification for lagged births. But we must keep in mind that this question is intrinsically tied to the functional form restrictions -- we saw in Section 3.3 that there always exist specifications in which y_t is independent of $x_1, \ldots, x_T$ conditional on c.

If we do impose the restrictions in Table 7.1, then there is strong evidence that $\underset{\sim}{\lambda} \neq \underset{\sim}{0}$. Constraining $\underset{\sim}{\lambda} = \underset{\sim}{0}$ in Table 7.2 gives an increase in the distance statistic of 78.4 - 53.8 = 24.6, which is surprisingly large to come from a $\chi^2(8)$ distribution.

In Table 7.3 we constrain all of the residual variances to be equal ($\alpha_t = 1$). An alternative interpretation of the time varying coefficients is provided in Table 7.4, where β_{YK} and β_K vary freely over time and $\alpha_t = 1$. In principle. we could also allow the α_t to vary freely, since they can be identified from changes over time in the coefficients of c. In fact that model gives very imprecise results and it is difficult to ensure numerical accuracy.

We shall interpret the coefficients on YK and K by following the procedure in (3.4). Table 8 presents estimates of the expected change in the participation probability when we assign an additional young child to a randomly chosen family, so that YK and K increase by one. We compute this measure for the models in Tables 7.1, 7.3, and 7.4. The average change in

the participation probability is -.096. We can get an indication of omitted variable bias by comparing these estimates with the ones based on Table 1.2, where $\underset{\sim}{\lambda}$ is constrained to be zero. Now the average change in the participation probability is -.122, so that the decline in absolute value when we control for c is 21%. An alternative comparison can be based on the cross-section estimates, with no leads or lags, in Table 5. Now the average change in the participation probability is -.144, giving an omitted variable bias of 33%.

Next we shall consider estimates from the logit framework of Section 3.2. Table 9 presents (standard) maximum likelihood estimates of cross-section logit specifications for each of the four years. We can use the cross-section probit results in Table 5 to construct estimates of the expected change in the log odds of participation when we add a young child to a randomly chosen family. Doing this in each of the four years gives -.502, -.598, -.683, and -.703. With the logit estimates, we simply add together the coefficients on YK and K in Table 9; this gives -.507, -.612, -.691, and -.729. The average over the four years is -.621 for probit and -.635 for logit. so at this point there is little difference between the two functional forms.

Now allow for the latent variable (c). Table 10 presents the conditional maximum likelihood estimates for the fixed effects logit model. The striking result here is that, unlike the probit case, allowing for c leads to an increase in the absolute value of the children coefficients. If we constrain β_{YK} and β_K to be constant over time (Table 10.1), the estimated change in the log odds of participation when we add an additional young child is -.898. If we allow β_{YK} and β_K to vary freely over time (Table 10.2), the average of

the estimated changes is -.883. So the absolute value of the estimates increases by about 40% when we control for c using the logit framework. The estimation method is having a first order effect on the results.

It is commonly found that probit and logit specifications, when properly interpreted, give very similar results; our cross-section estimates are an example of this. But our attempt to incorporate latent variables has turned up marked differences between the probit and logit specifications. There are a number of possible explanations for this. The probit specification restricts c to have a normal distribution conditional on x with a linear regression function and constant variance. The conditional likelihood approach in the logit model does not impose this possibly false restriction. On the other hand, the probit model has a more general specification for the residual covariance matrix.

We have seen that the restrictions on the probit $\underset{\sim}{\Pi}$ matrix, which underlie our estimate of β, appear to be false. An analogous test in the logit framework is based on (3.10). We use conditional ML to estimate a model that includes $YK_s \cdot D_t$, $K_s \cdot D_t$ ($s = 1,\ldots,4$; $t = 2,3,4$), where D_t is a dummy variable that is one in period t and zero otherwise. It is not restrictive to exclude $YK_s \cdot D_1$ and $K_s \cdot D_1$, since they can be absorbed in c. We include also D_t, $S \cdot D_t$, $EXP68 \cdot D_t$, and $EXP68^2 \cdot D_t$ ($t=2,3,4$). Then comparing the maximized conditional likelihoods for this specification and the specification in Table 10.2 gives a conditional likelihood ratio statistic of 47.5, which is a very surprising value to come from a $\chi^2(16)$ distribution. So the restrictions underlying our logit estimates of β also appear to be false.

It may be that the false restrictions simply imply different biases in the probit and logit specifications.

6. CONCLUSION

Our discussion has focused on models that are static conditional on a latent variable. The panel aspect of the data has primarily been used to control for the latent variable. Much work needs to be done on models that incorporate uncertainty and Interesting dynamics. Exploiting the martingale implications of time-additive utility seems fruitful here, as' in Hall (1978) and Hansen and Singleton (1981). There is, however, a potentially important distinction between time averages and cross-section averages. A time average of forecast errors over T periods should converge to zero as $T \to \infty$. But an average of forecast errors across N individuals surely need not converge to zero as $N \to \infty$; there my be common components in those errors, due to economy-wide innovations. The same point applies when we consider covariances of forecast errors with variables that are in the agents' information sets. If those conditioning variables are discrete, we can think of averaging over subsets of the forecast errors; as $T \to \infty$, these averages should converge to zero, but not necessarily as $N \to \infty$.

As for controlling for latent variables, I think that future work will have to address the lack of identification that we have uncovered. It is not restrictive to assert that $(y_1, \ldots, y_T)$ and $(x_1, \ldots, x_T)$ are independent conditional on some latent variable c.

TABLE 4

CHARACTERISTICS OF MICHIGAN PANEL STUDY OF INCOME DYNAMICS MARRIED WOMEN

Means and Standard Deviations

N = 924

Variable	Mean	Standard Deviation
LFP1	.499	
LFP2	.530	
LFP3	.529	
LFP4	.566	
YK1	.969	1.200
YK2	.764	1.069
YK3	.551	.895
YK4	.363	.685
K1	2.38	1.69
K2	2.30	1.64
K3	2.11	1.61
K4	1.84	1.52
S	12.1	2.1
EXP68	17.2	8.5
$EXP68^2$	368.	301.

Notes to Table 4:

LFP1, ... LFP4 -- 1 if answered "yes" to "Did work for money last year, 0 otherwise, referring to 1968, 1970, 1972, 1974; YK1, ... YK4 -- number of children of age less than six in 1968, 1970, 1972, 1974; K1 , ..., K4 -- number of children of age less than eighteen living in the family unit in 1968, 1970, 1972, 1974; S -- years of schooling completed; EXP68 -- (age in 1968 - S-6). The sample selection criteria required that the women be married to the same spouse from 1968 to 1976; not part of the low income subsample; between 20 and 50

years old in 1968; white; out of school from 1968 to 1076; not disabled. We required complete data on the variables in the Table, and that there be no inconsistency between reported earnings and the answer to the participation question.

TABLE 5

ML PROBIT CROSS-SECTION ESTIMATES

Dependent Variable	Coefficients (and Standard Errors) of:							
	YK1	YK2	YK3	YK4	K1	K2	K3	K4
LFP1	-.246 (.046)	-	-	-	-.063 (.031)	-	-	
LFP2		-.293 (.055)	-			-.075 (.031)	-	-
LFP3			-.342 (.067)	-			-.077 (.032)	-
LFP4				-.366 (.081)	-			-.069 (.034)

NOTES TO TABLE 5:

separate ML estimates each year. All specifications include (1, S, EXP68, $EXP68^2$).

TABLE 6

UNRESTRICTED ML PROBIT ESTIMATES

Dependent Variable	Coefficients (and Standard Errors) of:							
	YK1	YK2	YK3	YK4	K1	K2	K3	K4
LFP1	-.205 (.081)	-.017 (.119)	-.160 (.141)	.420 (.144)	.176 (.076)	-.142 (.100)	-.196 (.110)	.063 (.090)
LFP2	-.047 (.079)	-.238 (.117)	-.047 (.140)	.093 (.142)	.320 (.077)	-.278 (.102)	-.250 (.110)	.177 (.090)
LFP3	-.254 (.080)	.214 (.116)	-.190 (.139)	-.209 (.141)	.204 (.077)	-.210 (.102)	-.045 (.112)	.030 (.090)
LFP4	-.195 (.079)	.252 (.118)	-.211 (.139)	-.282 (.138)	.120 (.075)	.083 (.100)	-.181 (.110)	.058 (.090)

NOTES TO TABLE 6:

Separate ML estimates each year. All specifications include (1, S, EXP68, $EXP68^2$).