II DYNAMIC DISCRETE PROBABILITY MODELS

3 Statistical Models for Discrete Panel Data

James J. Heckman

3.1 Introduction

This chapter formulates a general dynamic model for the analysis of discrete panel data that can be used to analyze the structure of discrete choices made over time. A rich group of discrete time-discrete outcome stochastic processes is generated by imposing restrictions on the general model developed here. Markov models, renewal processes, Pólya schemes, Bernoulli models, and other familiar stochastic processes emerge as special cases of this model. The model is sufficiently flexible to accommodate time-varying explanatory variables, quite general serial correlation patterns for unobservable variables, and complex structural economic interrelationships among decisions taken at different times.

The analysis in this chapter generalizes previous work by McFadden (1976) and others that considers consumer choice among a collection of discrete alternatives at a point in time. The models considered here focus on relationships among choices over time, or more generally, intertemporal relationships among discrete variables.

The procedures proposed here are used to investigate the following important problem: in a variety of contexts, such as in the study of the incidence of accidents (Bates and Neyman 1951), labor force participation (Heckman and Willis 1977) and unemployment (Layton 1978), it is often noted that individuals who have experienced the event under study in the past are more likely to experience the event in the future than are individuals who have not experienced the event. The conditional probability that an individual will experience the event in the future is a function of past experience. There are two distinct explanations for this empirical regularity.

One explanation is that as a consequence of experiencing an event, preferences, prices or constraints relevant to future choices are altered. In

This research was supported by NSF Grant SOC 77-27136, Grant 10-P-90748/9-01 from the Social Security Administration, and a Fellowship from the J. S. Guggenheim Memorial Foundation. Discussions with Tom MaCurdy have been valuable at all stages of the work. I have also benefited from comments by Gary Chamberlain, Chris Flinn, Zvi Griliches, Jan Hoem, Samuel Kotz, Charles Manski, Jerzy Neyman, Marc Nerlove, Guillherme Sedlacek, Donald Waldman and David Wise. I retain responsibility for any errors that remain. The first draft of this chapter circulated in July, 1977. I have benefited from assorted comments received at seminars at Harvard-MIT (the Joint Econometrics Workshop), The University of Iowa, Columbia University, the University of Wisconsin, and the University of Chicago.

this case past experience has a genuine behavioral effect in the sense that an otherwise identical individual who did not experience the event would behave differently in the future than an individual who experienced the event. This explanation applies even in an environment of perfect certainty so that all relevant information is available to the individual but not necessarily to the observing economist. Structural relationships of this sort give rise to true state dependence, as defined in this chapter.

A second explanation for the phenomenon is that individuals may differ in their propensity to experience the event. If individual differences are correlated over time, and if these differences are not properly controlled, previous experience may appear to be a determinant of future experience solely because it is a proxy for temporally persistent unobservables that determine choices. Improper treatment of unmeasured variables gives rise to a conditional relationship between future and past experience that is termed spurious state dependence.

The problem of distinguishing between spurious and true state dependence is somewhat analogous to the familiar problem of estimating a distributed lag model in the presence of serial correlation (Griliches 1966, Malinvaud 1970, Nerlove 1978). It is also closely related to previous work on the mover-stayer model that appears in the literature on discrete stochastic processes (Goodman 1961, Singer and Spilerman 1976).

This substantive problem is of considerable practical interest. Two examples are offered to illustrate this point. The first is drawn from recent work in the theory of unemployment. Phelps (1972) has argued that short-term economic policies that alleviate unemployment tend to lower aggregate unemployment rates in the long run by preventing the loss of work-enhancing market experience. His argument rests on the assumption that unemployment has a real and lasting effect on the future probability of unemployment of the currently unemployed. Cripps and Tarling (1974) maintain the opposite view in their analysis of the incidence and duration of unemployment. They assume that individuals differ in their propensity to experience unemployment and in their unemployment duration times and that differences cannot be fully accounted for by measured variables. They further assume that the actual experience of having been unemployed or the duration of past unemployment does not affect future incidence or duration. Hence in their model short-term economic policies have no effect on long-term unemployment. The model developed in this chapter is sufficiently flexible to accommodate both views of unemployment and can be used to test the two competing theories.

As another example, recent work on the dynamics of female labor supply assumes that entry and exit from the labor force can be described by a Bernoulli probability model (Heckman and Willis 1977). This view of female labor supply dynamics ignores considerable evidence that work experience raises wage rates and hence that such experience may raise the probability that a woman works in the future, even if initial entry into the work force is determined by a random process. The general model outlined in this chapter extends the econometric model of Heckman and Willis by permitting (1) unobservable variables that determine labor force choices to be freely correlated, in contrast with the rigid permanent-transitory error scheme for the unobservables assumed in their model, (2) observed explanatory variables to change over time and (3) previous work experience to determine current participation decisions. Empirical work based on the general model developed in this chapter (Heckman 1978b, 1981) reveals that these three extensions are important in correctly assessing the determinants of female labor supply and in developing models that can be used in policy simulation analysis.

Since this chapter is long, and a number of new ideas are developed in it, an outline of the topics covered is in order. The first sections discuss the general model proposed here. This model is an extension of previous work by the author (1978a) that incorporates dummy endogenous variables into a simultaneous equation system. This chapter extends that framework to develop a very general choice theoretic model for the analysis of discrete decisions made over time. Many different discrete time-discrete outcome stochastic processes are developed as special cases of a more general model.

The models considered here are based on the notion that discrete outcomes are generated by continuous variables that cross thresholds. In certain applications these continuous variables correspond to well-defined economic concepts. For example, in the work of Domencich and McFadden (1975) the continuous variables that generate discrete choices are differences in utilities of possible choices. In work on labor supply the continuous variable that generates labor force participation is the difference between market wages and reservation wages (Heckman and MaCurdy 1980).

The main novelty in this chapter comes in the treatment of consumer decision making over time. With the exception of the few papers mentioned here, previous work has only considered consumer decision making at a point in time. This chapter develops a flexible statistical model that

considers the relationship between current choices (or discrete outcomes) and choices (or outcomes) in other periods. Variation in the specification of the interrelationship among choices (or outcomes) in different periods gives rise to a variety of stochastic processes. For example, if choices made last period are the only prior choices relevant to current choice, a first-order Markov model is generated. If the entire history of the process is relevant to current decision making, as is assumed in certain human capital models in labor economics, a Pólya process (Feller 1957, Johnson and Kotz 1977) emerges. If the current continuous duration in one state is a determinant of the decision to remain in or exit the state, a renewal process is generated that captures the essence of many models of firm specific investment recently advanced in the literature on worker turnover (Jovanovic 1978).

In formulating any econometric model, the treatment of unobservables is an important ingredient of the specification. This chapter extends previous work on estimating discrete stochastic processes by permitting the unobservables that generate the stochastic process to be freely correlated over time. Within the context of the models considered here, previous work assumes the unobservables that generate the underlying continuous variables that cross thresholds (and thus generate discrete outcomes) follow a "components of variance" scheme. Virtually all of the available literature on discrete data stochastic processes (implicitly) defines heterogeneity in this way. This chapter broadens the definition of heterogeneity to allow for more general correlation patterns among the unobservables. The greater generality of the model developed here permits the analyst to relax the (implicit) assumption—maintained in previous work—that the unmeasured variables that determine discrete outcomes are a combination of an immutable person specific component and a temporally independently identically distributed component. Unobservables are permitted to be characterized by a more general scheme so that conventional specifications of heterogeneity can be tested against more general models.

A major advantage of the models for discrete stochastic processes that are developed in this chapter is that they are sufficiently flexible to accommodate the introduction of time-varying explanatory variables. This feature improves on previous models advanced in the literature in which explanatory variables cannot be introduced at all, or special assumptions on their structure must be invoked—such as their assumed constancy over time.

Another advantage of the models presented here is that they are computationally tractable and hence useful in practical work. This is especially true of the factor analytic schemes and fixed effect schemes discussed in sections 3.5 and 3.6. The random factor model is the discrete data analogue of the MIMIC model of Joreskog and Goldberger (1975). The fixed effect probit model is a conditional version of the random factor model. Both models are very simple to compute but neither is without its limitations. These limitations are discussed briefly in the text and are spelled out in greater detail in the appendix and in chapter 4. The appendix also develops more general factor analytic schemes than those presented in the text.

Special cases of the general model that are likely to be of practical interest are developed in sections 3.3 through 3.10. Markov models, renewal models, Bernoulli models, "latent Markov" models, Pólya processes, and other schemes emerge as restricted versions of the general model. Very general types of population heterogeneity for unobserved variables are considered. Comparisons are made among models in terms of data requirements, identification criteria, and implications for runs patterns.

One important topic is only briefly covered in this chapter: the problem of initial conditions. In formulating any stochastic process with structural dependence among time-ordered outcomes, it is necessary to initialize the process. In much applied work in social science this problem is treated somewhat casually. Typically the initial conditions or the relevant presample history of the process are assumed to be predetermined or exogenous. This assumption is valid only if the unobservables that generate the process are serially independent or if a genuinely new process is (fortuitously) observed at the beginning of the sample at the analyst's disposal, and the relevant presample history is unrelated to the unobservables that generate the process in the sample period. Neither assumption is very appealing in applied work.

If the process has been in operation prior to the time it is sampled (e.g., a labor force participation process for middle-age women), and the unobservables that generate the process are serially correlated, the standard treatment of initial conditions results in biased and inconsistent parameter estimates. The confluence of heterogeneity and true (structural) state dependence leads to an important and neglected problem. Because of the importance of the problem, it is given special treatment in chapter 4.

Sections 3.12 through 3.15 are devoted to a discussion of the concepts of heterogeneity and state dependence. These concepts are defined, and their applicability to models of perfect foresight and models of uncertainty is discussed. The limitations of the multivariate probit framework for measuring separate effects of heterogeneity and state dependence are considered. The main points raised in these sections are (1) the concepts of heterogeneity and state dependence do not require the multivariate probit framework for their definition, but the multivariate probit framework is sufficiently flexible to permit empirical discrimination between the two concepts; (2) analogies between the classical time-series problem of discriminating between a distributed lag model and a serial correlation model and the problem of discriminating between heterogeneity and state dependence in a discrete data model, while of some heuristic value, are not precise and, if pushed too far, are misleading; (3) the concept of structural state dependence defined here is applicable to an environment of perfect certainty, in which there is no revision of plans, as well as to an environment of imperfect certainty, or an environment of stimulus-response conditioning of the sort considered by mathematical psychologists.

3.2 A Framework for Analyzing Dynamic Choice

All of the statistical models considered in this chapter are based on the following ideas: the analyst has access to a random sample of I individuals. On each of these persons there is a record that registers the presence or absence of an event under study in each of T equispaced time periods. The event occurs in period t for individual i if and only if a continuous latent random variable $Y(i, t)$ crosses a threshold, assumed to be zero for convenience. The event occurs, and dummy variable $d(i, t) = 1$ if and only if $Y(i, t) \geq 0$. Otherwise the event does not occur and $d(i, t) = 0$. The model developed in this chapter is confined to only two states, although it can readily be extended to accommodate more states.

Introducing a latent continuous random variable into the analysis simplifies the analysis, links the current work to previous work in econometrics, and provides a natural framework for formulating choice theoretic econometric models. Several examples are offered.

In an analysis of the labor force participation of women, $Y(i, t)$ may be interpreted as the difference between the lifetime utility of woman i at time t if she is in the labor force at t and her lifetime utility if she does not

participate, the assumption being that she chooses the best sequence of lifetime labor force participation in the remainder of her lifetime given participation or nonparticipation in t. In an analysis of the purchase of consumer durables, $Y(i, t)$ may be interpreted as the difference between lifetime utility if consumer i purchases a durable at time t and lifetime utility if he does not. In both of these examples it is natural to assume that the difference in utilities is a continuous latent random variable (McFadden 1976). In analyses of labor market search decisions or female labor supply decisions (Heckman and MaCurdy 1980), it is sometimes natural to formulate a model in terms of the difference between reservation wages and offered market wages. If this difference is positive in period t, an individual chooses to continue searching (or remain out of the labor force) in the period. In certain cases it is possible to observe the continuous random variable that generates the discrete random variable $\varepsilon(i, t)$ so that $Y(i, t)$ is more than a theoretical construct. For example, if a person is classified to be in poverty ($d(i, t) = 1$) when income at time $t(E(i, t))$ is below some cutoff value C, the latent variable that generates the dynamics of poverty status is $Y(i, t) = C - E(i, t)$ (Fase 1971).

Random variable $Y(i, t)$ may be decomposed into two components: a purely stochastic disturbance component, $\varepsilon(i, t)$, and a function of exogenous, predetermined, and measured endogenous variables that affect current choices, $V(i, t)$. $V(i, t)$ may or may not be independent of $\varepsilon(i, t)$. We may write

$$Y(i,t) = V(i,t) + \varepsilon(i,t), \tag{3.1}$$

$$Y(i,t) \geq 0 \quad \text{iff} \quad d(i,t) = 1, \tag{3.2}$$
$$Y(i,t) < 0 \quad \text{iff} \quad d(i,t) = 0.$$

The distribution of the $d(i, t)$, $t = 1, \ldots, T$, $i = 1, \ldots, I$, is generated by the distributions of $\varepsilon(i, t)$ and $V(i, t)$. To simplify the argument in this chapter, it is assumed throughout much of the discussion that the $\varepsilon(i, t)$ are jointly normally distributed when a distributional assumption is required so that this model is similar to the multivariate probit model of Ashford and Sowden (1970) as extended by Amemiya (1975), Domencich and McFadden (1975) and the author (1978a). Alternative specifications of $V(i, t)$ and $\varepsilon(i, t)$ give rise to a variety of interesting and important models useful in the analysis of discrete panel data.

In sections 3.3 through 3.12 content is given to the terms $V(i, t)$ and $\varepsilon(i, t)$ in equation (3.1). The next section presents a very general model and some

intuitive motivation for its constituent terms. Sections 3.4 through 3.12 deal with specific versions of the model in much greater detail.

3.3 The General Model

In this section content is given to the model of equations (3.1) and (3.2). $Y(i, t)$ is assumed to be a linear function of exogenous variables, $\mathbf{Z}(i, t)$, lagged values of $Y(i, t)$, and past outcomes $d(i, t')$, $t' \leq t$. The general model considered in this chapter may be written as

$$\begin{aligned} Y(i,t) = \mathbf{Z}(i,t)\boldsymbol{\beta} &+ \sum_{j=1}^{\infty} \gamma(t-j,t)\, d(i,t-j) \\ &+ \sum_{j=1}^{\infty} \lambda(j,t-j) \prod_{l=1}^{j} d(i,t-l) + G(L)\, Y(i,t) \\ &+ \varepsilon(i,t), \end{aligned} \tag{3.3}$$

$i = 1, \ldots, I$, $t = 1, \ldots, T$, where $G(0) = 0$ and $G(L)$ is a general lag operator of order K, $[G(L) = g_1 L + g_2 L^2 + \cdots + g_K L^K$, $L^K Y(i, t) = Y(i, t - K)]$, $d(i, t) = 1$ iff $Y(i, t) \geq 0$, $d(i, t) = 0$ otherwise, and initial conditions $d(i, t')$, $t' = 0, -1, \ldots$, $Y(i, t')$, $t' = 0, -1, \ldots$, are assumed to be fixed outside of the model. The term $\varepsilon(i, t)$ is a normally distributed disturbance with mean zero. The distribution of vector $\boldsymbol{\varepsilon}(i) = (\varepsilon(i, 1), \ldots, \varepsilon(i, T))$ is fully characterized by the assumption

$$\boldsymbol{\varepsilon}(i) \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a $T \times T$ positive definite covariance matrix. No assumption about stationarity of the disturbances is imposed. Random sampling is assumed across people, so that $\boldsymbol{\varepsilon}(i)$ is independent of $\boldsymbol{\varepsilon}(i')$, $i \neq i'$, $i, i' = 1, \ldots, I$. The components of vector $\mathbf{Z}(i, t)$ are assumed to be distributed independent of $\boldsymbol{\varepsilon}(i)$, so that these variables are exogenous.

The first term on the right-hand side of equation (3.3) represents the effect of exogenous variables on current utility comparisons. Vector $\mathbf{Z}(i, t)$ may include past exogenous variables, current exogenous variables, and expectations of future exogenous variables that determine current choices. In principle the $\boldsymbol{\beta}$ parameters may depend on time, but this generality is foregone in this chapter.

The second term on the right-hand side of the equation represents the effect of the entire past history of the process on current choice. This term is

assumed to be finite. To capture the idea that the effect of a past event on current choice may depend on the time period in which the event occurred as well as on the current time period, the coefficients on past events are assumed to be functions of the current period, t, and the period in which the event occurred, $t-j$. This characterization of the effect of the past on current choices is consistent with depreciation and the notion that the values of exogenous variables at the time events occur as well as current values of exogenous variables modify the effect of previous choices on current choices. Various restrictions imposed on the coefficients $\gamma(t-j, t)$ generate a variety of interesting stochastic processes.[1]

The third term on the right-hand side represents the cumulative effect on current choices of the most recent continuous experience in a state. This term is introduced to capture the notion that, once an individual is in a state, an accumulation process begins. For example, in human capital theory specific capital may be accumulated and accumulation continues until the individual leaves the state, at which time the state specific capital is lost. This term generates a renewal process (see Karlin and Taylor 1975) of the sort considered by Jovanovic (1978). It is assumed to be finite. In principle one could generalize this term to allow for depreciation and other forms of time dependence. Moreover, one could introduce another term representing state specific capital that is accumulated when an individual is in the state corresponding to $d(i, t) = 0$. These generalizations are not pursued in this chapter.

The fourth term in the equation is introduced to capture the notion of habit persistence. This term represents the effect of *previous relative evaluations of the two states* on current choices. This term captures the essential idea in Coleman's "latent Markov" model (Coleman 1964) in which prior propensities to select a state rather than prior occupancy of a state determine the current probability that a state is occupied.

The information that $Y(i, t) \geq 0$ is equivalent to the information that $Y(i, t)/\sigma(t, t)^{1/2} \geq 0$. For notational convenience it is useful to work with the normalized latent variables.

1. This term could be augmented to include the effect of future outcomes of the process on current choice. Structural dependence of this sort, while unfamiliar in the literature on applied probability, naturally arises in economic models of life cycle decision making under perfect certainty of the sort considered by Polachek (1975). If the range of the j subscript on the second term is changed to range from -1 to $-\infty$, this sort of dependence can be captured by the model. A certain technical difficulty arises if both forward and past dependence are introduced in the model simultaneously. This difficulty is discussed in note 26.

The general model may be summarized in a compact expression that is useful in computational work as well as in the theoretical analysis. Array the $d(i, t)$, $t = 1, \ldots, T$ into a $1 \times T$ vector $\mathbf{d}(i) = (d(i, 1), \ldots, d(i, T))$. Define $V(i, t)$ as the right-hand side of equation (3.3) exclusive of the disturbance term

$$V(i,t) = \mathbf{Z}(i,t)\boldsymbol{\beta} + \sum_{j=1}^{\infty} \gamma(t-j,t)\, d(i,t-j) + \sum_{j=1}^{\infty} \lambda(j,t-j) \prod_{l=1}^{j} d(i,t-l) + G(L)\, Y(i,t). \tag{3.4}$$

The $V(i, t)$ may be normalized by $\sigma(t, t)^{1/2}$. Thus $\tilde{V}(i, t) = V(i, t)/\sigma(t, t)^{1/2}$ Array the $\tilde{V}(i, t)$ into a $1 \times T$ vector, $\tilde{\mathbf{V}}(i)$,

$$\tilde{\mathbf{V}}(i) = [\tilde{V}(i,1), \ldots, \tilde{V}(i,T)]. \tag{3.5}$$

For convenience the vectors of exogenous variables may be collected into a super vector $\mathbf{Z}(i)$, where $\mathbf{Z}(i) = (\mathbf{Z}(i, 1), \mathbf{Z}(i, 2), \ldots, \mathbf{Z}(i, T))$.

The correlation matrix $\tilde{\boldsymbol{\Sigma}}$ is derived from the covariance matrix $\boldsymbol{\Sigma}$ by the equation

$$\tilde{\boldsymbol{\Sigma}} = (\operatorname{diag} \boldsymbol{\Sigma}^{-1})^{1/2}\, \boldsymbol{\Sigma}\, (\operatorname{diag} \boldsymbol{\Sigma}^{-1})^{1/2},$$

where diag $\boldsymbol{\Sigma}$ is the diagonal matrix formed from the diagonal of $\boldsymbol{\Sigma}$.

Letting $\boldsymbol{\iota}$ denote a $1 \times T$ vector of ones, the probability of $\mathbf{d}(i)$, given $\mathbf{Z}(i, t)$, $t = 1, \ldots, T$ and the nonstochastic initial conditions specified below, equation (3.3) may be written as

$$\begin{aligned} &\operatorname{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i), d(i,0), d(i,-1), \ldots, Y(i,0), Y(i,-1), \ldots] \\ &= F\{\tilde{V}(*)(2\mathbf{d}(i) - \boldsymbol{\iota});\ \tilde{\boldsymbol{\Sigma}}(*)[(2\mathbf{d}(i)) - \boldsymbol{\iota})'(2\mathbf{d}(i) - \boldsymbol{\iota})]\}, \end{aligned} \tag{3.6}$$

where $F(\mathbf{a};\ \tilde{\boldsymbol{\Sigma}})$ is the cumulative distribution function of a T-variate standardized multivariate normal random variable with correlation matrix $\tilde{\boldsymbol{\Sigma}}$ evaluated at an upper limit by vector $\mathbf{a}$, and where $(*)$ denotes the operation of a Hadamard product.[2] Expression (3.6) is a simple, shorthand summary of all of the possible probabilities associated with the 2^T possible values of $\mathbf{d}(i)$ that exploits the symmetry of the multivariate normal density.

2. A Hadamard product of two vectors $\mathbf{a}(*)\mathbf{b}$ is defined as a vector $\mathbf{C} = \mathbf{a}(*)\mathbf{b}$, where $(C_i) = (a_i b_i)$. A Hadamard product of two matrices $\mathbf{C} = \mathbf{A}(*)\mathbf{B}$ is defined by $(C_{ij}) = (a_{ij} b_{ij})$; e.g., see Rao (1973).

Given specific values for the exogenous variables, and given the initial conditions, the sample likelihood function may be written as

$$\mathscr{L} = \prod_{i=1}^{I} \text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i,1), \ldots, \mathbf{Z}(i,T), d(i,0), d(i,-1), \ldots, Y(i,0), Y(i,-1), \ldots]. \tag{3.7}$$

Maximizing the log likelihood produces estimators that are consistent, asymptotically normally distributed, and efficient.

To make the discussion more specific, and also to link the general model with previous work, it is helpful to consider the variety of special cases that arise from the general model by imposing restrictions on the coefficients and the admissible distribution of the error term in equation (3.7). In investigating these models, we consider the following issues of model identification: (1) What are the data requirements for the estimation of each model? In particular, when can cross section data be used to characterize fully a dynamic process? If cross section data cannot be so used, what information about the dynamic process can be retrieved from a cross section? (2) From observed sequences of discrete events (runs patterns) is it possible to infer the underlying stochastic process that generates the data?

We consider a sequence of models which are specializations of the general model starting with the simplest and most familiar: a Bernoulli model.

3.4 An Independent Trials Bernoulli Model

Let $V(i, t) = \bar{V}$, and assume that $\varepsilon(i, t)$ is independently identically distributed, iid. Each person has the same probability of experiencing the event $(d(i, t) = 1)$ in each period:

$$\text{Prob}[\varepsilon(i,t) \geq -\bar{V}] = \Phi\left(\frac{\bar{V}}{\sigma_\varepsilon}\right) = \bar{P}, \tag{3.8}$$

where

$$E(\varepsilon(i,t)^2) = \sigma_\varepsilon^2,$$

Φ is the cumulative distribution function of a standard normal random variable, and the symmetry of Φ is exploited ($\Phi(b) = 1 - \Phi(-b)$).[3]

Average continuous duration in the state ($d(i,t) = 1$) is $\bar{P}/(1 - \bar{P})$. From data on duration in the state, one can determine $\bar{P}$ uniquely. In a panel of T periods, the expected number of periods in the state for any person is $\bar{P}T$ with variance $\bar{P}(1 - \bar{P})T$. $\bar{P}$ can be consistently estimated from a single cross section or from a long time series on one person by the method of maximum likelihood. If the cross section and panel samples are the same size, estimators are equally efficient.

In T trials, the probability of J successes ($\Sigma d(i,t) = J$) and $T - J$ failures in a particular order is

$$\bar{P}^J(1 - \bar{P})^{T-J}.$$

The random variables $d(i,t)$, $t = 1, \ldots, T$, are exchangeable, in the sense that the probability of any sequence with J successes in T trials is the same as any other sequence with the same number of successes in the same number of trials.

This model can be modified to take account of measured differences in personal characteristics. If $V(i,t)$ is assumed to be a linear function of known exogenous variables ($\mathbf{Z}(i,t)$) distributed independent of $\varepsilon(i,t)$, one may write

$$V(i,t) = \mathbf{Z}(i,t)\boldsymbol{\beta}. \tag{3.9}$$

Depending on the content of the $\mathbf{Z}(i,t)$ regressor vector, one can generate a nonstationary time inhomogeneous stochastic process at the micro level (e.g., $\mathbf{Z}(i,t)$ may include "age" or "calendar time" variables). Provided that there is sufficient variation in the sample regressors, so that the cross product matrix for the data is nonsingular, and the expectation of the Hessian of the log likelihood is negative definite at true parameter values, one can estimate the parameters of the model from a cross section of individuals or a time series on a single person.[4]

3. If $\varepsilon(i,t)$ is distributed logit, Φ would be the cumulative distribution of the logit. Obviously $\varepsilon(i,t)$ may have *any* distribution, and Φ is the corresponding distribution of the standardized variate. However, for nonsymmetric variates the notation in the text would have to be altered in an obvious way.

4. For example, if education is included as a regressor in $\mathbf{Z}(i,t)$, and education does not change over the sample period, a time series for one person would not yield estimates of the effect of education on the probability of experiencing the event. If there are year effects, data from a single cross section would not permit estimation of the year effect.

Of course, if the $\mathbf{Z}(i, t)$ change over the sample period, the exchangeability property of the model disappears. Depending on the distribution of the exogenous variables, runs patterns with identical numbers of successes may have different probabilities.

The assumption that the $\varepsilon(i, t)$ are identically distributed can be relaxed. Suppose that the disturbances are independent (over time and people) but come from different distributions in different time periods. For example, suppose that

$$E(\varepsilon(i,t)^2) = \sigma(t,t),$$

so that the variance is different in each time period and the underlying disturbance is nonstationary. In this case the probability that $d(i, t) = 1$ given $\mathbf{Z}(i, t)$ is

$$\text{Prob}[\varepsilon(i,t) \geq -\mathbf{Z}(i,t)\boldsymbol{\beta}] = \Phi[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t)], \tag{3.10}$$

where

$$\tilde{\boldsymbol{\beta}}(t) = \frac{\boldsymbol{\beta}}{\sigma(t,t)^{1/2}}.$$

The probability that $d(i, t) = 0$ given $\mathbf{Z}(i, t)$ is the complement of this probability. Subject to the identification conditions previously stated, if the analyst has access to a series of successive cross sections, he can estimate $\tilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}/\sigma(t, t)^{1/2}$, $t = 1, \ldots, T$ by applying probit analysis to each cross section. In this case it is clearly possible to estimate the ratio of variances in successive cross sections. Of course this procedure requires that $\boldsymbol{\beta}$ be time invariant.

The likelihood function for this model is a special case of the general likelihood function given in equation (3.7):[5]

$$\mathscr{L} = \prod_{i=1}^{I} \text{Prob}[\mathbf{d}(i) \mid Z(i)] = \prod_{i=1}^{I} \prod_{t=1}^{T} \Phi\{[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t)][2d(i,t) - 1]\}. \tag{3.11}$$

5. Clearly, if $\varepsilon(i, t)$ is assumed to be logit distributed, or generated by any other symmetric (around zero) distribution, equation (3.11) applies with Φ as the relevant cumulative distribution function. The modification of (3.11) for asymmetric random variables is straightforward.

3.5 A Random Effect Bernoulli Model and One-Factor Schemes

Unobserved temporally correlated error components are now introduced into the analysis. Such components are often termed heterogeneity in the applied literature on stochastic processes. Initially it is assumed that individuals all have the same values for the time invariant exogenous variables so that $V(i, t) = \bar{V}$. Also it is assumed initially that $\varepsilon(i, t)$ has a components of variance structure:

$$\varepsilon(i,t) = \tau(i) + U(i,t), \tag{3.12}$$

where $U(i, t)$ is iid with mean zero and variance σ_U^2 and $\tau(i)$ is distributed independent of the $U(i, t)$.

Individual i has a fixed component $\tau(i)$. Given $\tau(i)$, the probability that person i experiences an event at time $t(d(i, t) = 1)$ is

$$\begin{aligned}\text{Prob}[\varepsilon(i,t) \geq -\bar{V} \mid \tau(i)] &= \text{Prob}[U(i,t) \geq -(\tau(i) + \bar{V})] \\ &= P(\tau(i)) = \Phi\left[\frac{\tau(i) + \bar{V}}{\sigma_U}\right]\end{aligned} \tag{3.13}$$

The mean probability in the population is

$$\begin{aligned}\bar{P} &= \text{Prob}[\varepsilon(i,t) \geq -\bar{V}] = \int \text{Prob}[U(i,t) \geq -(\tau(i) + \bar{V})] f(\tau)\, d\tau \\ &= \Phi\left[\frac{\bar{V}}{(\sigma_U^2 + \sigma_\tau^2)^{1/2}}\right],\end{aligned} \tag{3.14}$$

where $f(\tau)$ is the frequency distribution of τ, and where $\text{Prob}[U(i, t) \geq -(\tau(i) + \bar{V})]$ is shorthand for the probability that $U(i, t)$ exceeds minus $(\tilde{\tau}(i) + \bar{V})$ given $\tilde{\tau}(i)$ and $\bar{V}$—a shorthand notation that will be used in the rest of this chapter. The mean probability in the population $\bar{P}$, and hence $\Phi[\bar{V}/(\sigma_U^2 + \sigma_\tau^2)^{1/2}]$, can be estimated from a single cross section by ordinary probit analysis. At least two years of panel data must be obtained to estimate the correlation coefficient between $\varepsilon(i, t)$ and $\varepsilon(i, t')$, $t \neq t'$. This is known as the intraclass correlation coefficient, $\rho = \sigma_\tau^2/(\sigma_\tau^2 + \sigma_U^2)$. Using probit analysis, the expected number of periods in the state, $\bar{P}T$, can be estimated, but at least two periods of panel data are required to estimate the population variance, $(\int P(\tau)(1 - P(\tau)) f(\tau) d\tau) T$, unless $f(\tau)$ is degenerate $(\sigma_\tau^2 = 0)$.

Maximum likelihood estimators of $\bar{P}$ based on a single cross section are consistent estimators of $\bar{P}$ as the cross section sample size I becomes large.

Unlike the situation in the preceding model, maximum likelihood estimators based on a long time series on one person or a large cross section at a point in time estimate different parameters. If both samples become large, the first sample estimates $(\bar{V} + \tau(i))/\sigma_U$ while the second sample estimates $\bar{V}/(\sigma_U^2 + \sigma_\tau^2)^{1/2}$. The first sample is conditioned on a specific value of $\tau(i)$, so that τ is a fixed effect indistinguishable from $\bar{V}$. The second sample is not conditioned on a specific value of $\tau(i)$.

As a consequence of Jensen's inequality the average duration in the state cannot be estimated from cross section data, because expected continuous duration in the state satisfies the following inequality:

$$E_\tau((1 - P(\tau)) \sum_{j=0}^{\infty} jP(\tau)^j) = E_\tau\left(\sum_{j=1}^{\infty} P(\tau)^j \right) \geq \sum_{j=1}^{\infty} \bar{P}^j,$$

where E_τ denotes expectation with respect to the density of τ, $f(\tau)$. Estimates of the average duration based on an estimated cross section probability (an estimate of $\bar{P}$) understate the average length of duration in the state.

Panel data can be used to estimate a separate $P(\tau(i))$ for each person by the method of maximum likelihood. This estimate is consistent as T becomes large. The estimated probabilities can be used to generate consistent estimators of the average duration in a state for each person: insert the estimated $P(\tau(i))$ into the mathematical formula for average duration.[6]

The probability of J successes ($\Sigma d(i, t) = J$) and $T - J$ failures is the same for any sequence with J successes in any order. To see this, note that conditional on $\tau(i)$ the model in this section is the same as in the preceding section. Removing the conditioning (by integrating out $\tau(i)$), leads to the probability of J successes and $T - J$ failures in a particular sequence as

$$\int P(\tau)^J (1 - P(\tau))^{T-J} f(\tau)\, d\tau.$$

As in a case without heterogeneity, any of the $\binom{T}{J}$ sequences with J successes have the same probability.

It is possible to account for measured differences in personal characteristics in exactly the same way as is done in the model presented in the preceding section. If $V(i, t)$ is assumed to be a linear function of known exogenous variables, one may write

6. This example illustrates the point that panel data can be used to relax the ergodicity assumption maintained in much work in stationary time-series analysis.

$$V(i,t) = \mathbf{Z}(i,t)\boldsymbol{\beta}. \tag{3.15}$$

Under the identification conditions specified in section 3.3, $\boldsymbol{\beta}$ is estimable. This model has been estimated by Heckman and Willis (1975). Using maximum likelihood, they estimate $\boldsymbol{\beta}$ and ρ under the normalizing assumption that $\sigma_\tau^2 + \sigma_U^2 = 1$. This final assumption may be relaxed. Exactly as in the model of the preceding section it is possible to permit the disturbance variances to differ among time periods and estimate the ratio among disturbance variances in different periods. Thus a nonstationary version of the model can be estimated. If the $\mathbf{Z}(i, t)$ are permitted to vary arbitrarily, and disturbance variances are permitted to assume a free structure, the exchangeability property of the random effects model disappears.

Defining the probability of a given sequence of events given $\mathbf{Z}(i)$ for the random effect model is straightforward. For convenience it is useful to work with the standardized value of τ, $\tilde{\tau} = (\tau/\sigma_\tau)$, which has mean zero and variance one. Define $\tilde{\boldsymbol{\beta}}$ as $\boldsymbol{\beta}/\sigma_U$. In this notation the probability of sequence $\mathbf{d}(i)$ given $\mathbf{Z}(i)$ is

$$\text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i)]$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\left\{\left[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}} + \tilde{\tau}\left(\frac{\rho}{1-\rho}\right)^{1/2}\right][2d(i,t) - 1]\right\} f(\tilde{\tau})\, d\tilde{\tau}, \tag{3.16}$$

where $f(\tilde{\tau})$ is the density of the standard normal distribution and $\rho < 1$.[7] Subject to the given identification conditions maximum likelihood estimators of $\tilde{\boldsymbol{\beta}}$ and ρ are consistent and efficient. The likelihood formed from the product of the probabilities is relatively easy to compute since it involves only one numerical integration per observation of products of cumulative normal error functions which are available on most computers.

7. The probability that $d(i, t) = 1$ given $\mathbf{Z}(i, t)$ and $\tau(i)$ is

$$\text{Prob}[U(i,t) \geq -\mathbf{Z}(i,t)\boldsymbol{\beta} - \tau(i)] = \text{Prob}\left[\frac{U(i,t)}{\sigma_U} \geq -\mathbf{Z}(i,t)\frac{\boldsymbol{\beta}}{\sigma_U} - \frac{\tau(i)}{\sigma_U}\right].$$

Since $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma_U$, and since $\sigma_\tau/\sigma_U = (\rho/1-\rho)^{1/2}$, this probability is

$$\text{Prob}\left[\frac{U(i,t)}{\sigma_U} \geq -\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}} - \left(\frac{\rho}{1-\rho}\right)^{1/2}\tilde{\tau}(i)\right] = \Phi\left[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}} + \left(\frac{\rho}{1-\rho}\right)^{1/2}\tilde{\tau}(i)\right].$$

Removing the conditioning on $\tilde{\tau}(i) = \tau(i)/\sigma_\tau$, which in this context is equivalent to

The components of variance error specification can be generalized to a one-factor scheme. This generalization leads to a discrete data analogue of the MIMIC model of Joreskog and Goldberger (1975). One-factor representations of the cumulative normal integral have been considered by Gupta (1963) and others (see the references in Johnson and Kotz 1972, vol. 4, pp. 47–50). In this model the disturbance is written as

$$\varepsilon(i,t) = \alpha(t)\tau(i) + U(i,t), \tag{3.17}$$

$t = 1, \ldots, T$, $i = 1, \ldots, I$, where $\tau(i)$ is distributed independent of $U(i,t)$, $E(\tau(i)) = E(U(i,t)) = 0$, and $E(U(i,t)^2) = \sigma_U(t,t) > 0$, $E(\tau(i)^2) = \sigma_\tau^2$. The components of variance structure is a special case of this scheme with $\alpha(t) = 1$ and $\sigma_U(t,t) = \sigma_U$ for all t.

Before elaborating the one-factor model, it is useful to introduce some notation that simplifies the exposition. It is analytically convenient to work with the square root of the proportion of the variance of disturbance $\varepsilon(i,t)$, $t = 1, \ldots, T$, that is explained by the factor $\tau(i)$, defined as $\tilde{\alpha}(t)$, where

$$\tilde{\alpha}(t) \equiv \frac{\alpha(t)\sigma_\tau}{(\alpha^2(t)\sigma_\tau^2 + \sigma_U(t,t))^{1/2}}.$$

(Positive values of square roots are used.) In this notation the correlation between disturbances in periods t and t' for a randomly selected person is

$$\sigma(t,t') = 1, \quad t = t',$$

$$\sigma(t,t') = \tilde{\alpha}(t)\tilde{\alpha}(t'), \quad t \neq t'.$$

It is also convenient to define $\eta(t)$, the ratio of permanent to transitory variance, by

$$\eta(t) \equiv \left[\frac{\tilde{\alpha}(t)^2}{1 - \tilde{\alpha}(t)^2}\right]^{1/2} = \left[\frac{\alpha(t)^2\sigma_\tau^2}{\sigma_U(t,t)}\right]^{1/2}.$$

integrating out $\tau(i)$, leads to

$$\text{Prob}[d(i,t) = 1 \mid \mathbf{Z}(i)] = \int_{-\infty}^{\infty} \Phi\left(\mathbf{Z}(i,t)\boldsymbol{\beta} + \left(\frac{\rho}{1-\rho}\right)^{1/2}\tilde{\tau}\right) f(\tilde{\tau})\, d\tilde{\tau}.$$

The probability of any sequence of events conditional on $\tilde{\tau}(i)$ can be expressed as the product of cumulative distributions (see the kernel of the integral of equation 3.16). Removing the conditioning (integrating out $\tilde{\tau}$) leads to the expression in the text.

Clearly neither $\tilde{\tau}$ nor $U(i,t)/\sigma_U$ is restricted to be a normal random variable.

Thus $\eta(t)$ is the ratio of the standard deviation of the permanent component to the standard deviation of the transitory component in $\varepsilon(i, t)$.

Finally, it is notationally convenient to work with the normalized coefficient vector $\tilde{\boldsymbol{\beta}}(t)$, defined as

$$\tilde{\boldsymbol{\beta}}(t) \equiv \frac{\boldsymbol{\beta}}{\sigma_U(t,t)^{1/2}}.$$

In this notation the probability that $d(i, t) = 1$ given $\mathbf{Z}(i, t)$, and $\tau(i)$ is

$$\begin{aligned}\text{Prob}[d(i,t) = 1 \mid \mathbf{Z}(i,t), \tau(i)] &= \text{Prob}[U(i,t) \geq -\mathbf{Z}(i,t)\boldsymbol{\beta} - \alpha(t)\tau(i) \mid \tau(i), \mathbf{Z}(i,t)] \\ &= \Phi[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \eta(t)\tilde{\tau}(i)], \end{aligned} \tag{3.18}$$

where $\tilde{\tau}(i)$ is the standardized $\tau(i)$ variable and $|\eta(t)| < \infty$.[8] For proof of this proposition see appendix 3.18. This expression corresponds to the probability that $d(i, t) = 1$ in the components of variance model; $\eta(t)$ corresponds to $(\rho/(1-\rho))^{1/2}$.

The probability of $\mathbf{d}(i)$ given $\mathbf{Z}(i)$ for the one-factor model is

$$\text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i)] = \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\{[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \eta(t)\tilde{\tau}][2d(i,t) - 1]\} f(\tilde{\tau})\, d\tilde{\tau}. \tag{3.19}$$

Subject to the normalization restriction $\sigma_U(1, 1) = 1$, it is possible to maximize the sample likelihood to estimate $\boldsymbol{\beta}$, $\sigma_U(t, t)$, $t = 2, \ldots, T$, and the $\eta(t)$, $t = 1, \ldots, T$, for $T \geq 3$.[9,10] The $\eta(t)$, $t = 1, \ldots, T$, are uniquely identified up to a sign change for the entire set of values (e.g., see Lawley and Maxwell 1971). From these parameters it is possible to identify $\alpha(t)\sigma_\tau$, $t = 1, \ldots, T$, given the normalization $\sigma_U(1, 1) = 1$ and the estimates of $\sigma_U(t, t)$, $t = 2, \ldots, T$.

8. The final assumption is relaxed in appendix 3.18.

9. The choice of $\sigma_U(1, 1) = 1$ is arbitrary. One could normalize any of the $\sigma_U(j, j)$ to unity, or one could normalize $\alpha(1)\sigma_\tau = 1$ (or any $\alpha(j)\sigma_\tau$).

10. This restriction is familiar in factor analysis (e.g., see Joreskog and Goldberger 1975).

An alternative normalization sets $\alpha(1)\sigma_\tau = 1$. In this case it is possible to estimate $\boldsymbol{\beta}$, $\sigma_U(t, t)$, $t = 1, \ldots, T$, and the $\eta(t)$, $t = 1, \ldots, T$, for $T \geq 3$.[11,12]

Further results on the one-factor model and generalizations to higher factor schemes are given in appendix 3.18.

In the one-factor model the random variables $d(i, t)$, $t = 1, \ldots, T$, are not exchangeable even if $\mathbf{Z}(i, t)\boldsymbol{\beta} = \bar{V}$ unless the period specific factor-loading coefficients are identical ($\alpha(t) = \alpha$, $t = 1, \ldots, T$) and the variances of the unique components are equal ($\sigma_U(t, t) = \sigma_U$), conditions which generate the simple random effect model.

The one-factor model permits the generalization of the unobserved heterogeneity concept beyond the components of variance scheme initially suggested in this section. Other generalizations of the heterogeneity concept are considered in section 3.7. Both the one-factor and components of variance models are simply computed, since they require only one numerical integration of products of cumulative normal functions which are already available on most computers.

11. The statements about identification of parameters made in the text are readily verified. An intuitive argument is as follows: For $T \geq 3$ it is possible to estimate the correlation matrix of the unobservables $\tilde{\Sigma}$, by multivariate probit analysis. From the estimated correlation matrix it is possible to estimate $\tilde{\alpha}(t)$, $t = 1, \ldots, T$, up to a sign change for the entire set of values of these parameters. From cross section probit analysis applied to each of the T cross sections, one can estimate

$$\tilde{\tilde{\boldsymbol{\beta}}}(t) = \frac{\boldsymbol{\beta}}{(\alpha^2(t)\sigma_\tau^2 + \sigma_U(t, t))^{1/2}},$$

$t = 1, \ldots, T$. From the ratio of the coefficients in $\boldsymbol{\beta}(t)$ to the corresponding coefficients in $\tilde{\tilde{\boldsymbol{\beta}}}(t')$, it is possible to estimate

$$\frac{[\alpha^2(t)\sigma_\tau^2 + \sigma_U(t, t)]^{1/2}}{[\alpha^2(t')\sigma_\tau^2 + \sigma_U(t', t')]^{1/2}}$$

for all t and t'. Set $\sigma_U(1, 1) = 1$. From the estimated value of $\tilde{\alpha}(1)$ one can estimate $\alpha(1)\sigma_\tau$, and hence $\boldsymbol{\beta}$. From the ratio of the coefficients in $\tilde{\tilde{\boldsymbol{\beta}}}(t)$ to the corresponding coefficients in $\tilde{\tilde{\boldsymbol{\beta}}}(1)$ one can estimate $(\alpha^2(t)\sigma_\tau^2 + \sigma_U(t, t))^{1/2}$, $t = 2, \ldots, T$. This piece of information in conjunction with $\tilde{\alpha}(t)$ is sufficient to identify $\alpha(t)\sigma_\tau$, and hence $\sigma_U(t, t)$, $t = 2, \ldots, T$.

An alternative normalization sets $\alpha(1)\sigma_\tau = 1$. From the estimated value of $\tilde{\alpha}(1)$ one can estimate $\sigma_U(1, 1)$, and hence $\boldsymbol{\beta}$, and proceed, following the logic of the case in which $\sigma_U(1, 1) = 1$, to estimate $\alpha(t)\sigma_\tau$, and hence $\sigma_U(t, t)$, $t = 2, \ldots, T$.

12. For $T = 2$ it is necessary to normalize $\eta(1) = 1$ and $\sigma_U(1, 1) = 1$ (obviously 2 can be substituted for 1). This follows from well-known results in factor analysis; any two-period model can be one-factor analyzed. In this case $\eta(2) = (\rho/1 - \rho)$.

An alternative normalization is $\eta(1) = \eta(2) = (\rho/1 - \rho))^{1/2}$.

Note finally that in either the components of variance model or the one-factor model it is not necessary to assume that $\tau(i)$ or $U(i, t)$ are normal variates to write down the expression given in equations (3.16) and (3.19). The only assumption required is that the density of $U(i, t)$ be symmetric, and even this condition can easily be relaxed at the cost of minor notational inconvenience. An example of non-normal factor analysis for continuous data is found in the work of Mandelbrot (1962).

The one-factor model may be generalized in several ways. First, the period specific components may have zero variance ($\sigma_U(t, t) = 0$). Second, multiple factor schemes may be developed in a fairly straightforward way. These topics and examples of common error processes that can be one-factor analyzed are discussed in appendix 3.18, where certain restrictions inherent in a one-factor scheme are noted and a multiple factor model is introduced.

3.6 A Fixed Effect Bernoulli Model

Earlier $\tau(i)$, the person specific effect, was treated as a random variable. Following Mundlak's interpretation of the fixed effect regression model (1978), it is possible to derive conditional (on $\tau(i)$) fixed effect versions of the random effect and one-factor models. A fixed effect logit model has been considered by E. B. Andersen (1973). The advantages of such models are threefold: they are simple to compute; they provide one solution to the problem of initial conditions (discussed in chapter 4); and they permit the analyst to estimate rather than impose the population density of τ.

The essential ingredients of the fixed effect model are to be found in equation (3.16). The probability of sequence $\mathbf{d}(i)$ given $\mathbf{Z}(i)$ and $\tilde{\tau}(i)$ is

$$\text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i), \tilde{\tau}(i)]$$

$$= \prod_{t=1}^{T} \Phi\left\{\left[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}} + \tilde{\tau}(i)\left(\frac{\rho}{1-\rho}\right)^{1/2}\right][2d(i,t) - 1]\right\}.$$

The sample likelihood formed from the probabilities can be maximized with respect to $\tilde{\boldsymbol{\beta}}$ and $\tilde{\tau}(i)(\rho/1 - \rho)^{1/2} = l(i)$, $i = 1, \ldots, I$. Note, however, that the constant term in $\tilde{\boldsymbol{\beta}}$ and the correlation parameter $(\rho/1 - \rho)^{1/2}$ are absorbed into the estimated fixed effect $l(i)$. However, it is possible to estimate the correlation parameter from the square root of the sample

variance of the estimated $l(i)$. (Recall that $\tilde{\tau}(i)$ is restricted to have unit variance in the population.) From the mean of the estimated $l(i)$ one can retrieve the intercept or constant term in $\tilde{\boldsymbol{\beta}}$.[13] If $I \to \infty$ and $T \to \infty$, these estimators are consistent and asymptotically normally distributed. Estimates of $l(i)$ can be used to construct an empirical density that converges to the population density of person specific effects.

This model is very simple to compute. Holding $\tilde{\boldsymbol{\beta}}$ fixed, $l(i)$ can be estimated for each person. The log likelihood function is globally concave for $l(i)$ and hence tends to converge rapidly to an optimum in practice. Note, however, that if individual i does not change state in the course of the sample, so that $\Sigma_t d(i, t) = T$ or $\Sigma_t d(i, t) = 0$, the estimated value of $l(i)$ is $\pm\infty$, respectively. As $T \to \infty$, this becomes an improbable event (assuming that $\mathbf{Z}(i, t), t = 1, \ldots, T, i = 1, \ldots, I$, are bounded exogenous variables).

Given $l(i)$, the likelihood is an ordinary probit likelihood function and so is concave in the parameters in $\tilde{\boldsymbol{\beta}}$ (with constant term absorbed in $l(i)$). Sequential estimation of $l(i)$ and $\tilde{\boldsymbol{\beta}}$ results in rapid convergence to an optimum.[14]

The principal disadvantage of the fixed effects estimator is that if T does not become large, maximum likelihood estimators of $l(i)$ are inconsistent (Neyman and Scott 1948). Due to the nonlinearity of the model, the estimator of $\tilde{\boldsymbol{\beta}}$ is solved jointly with that of $l(i)$ to secure estimates. The inconsistency in $l(i)$ is transmitted to $\tilde{\boldsymbol{\beta}}$, unlike the situation in linear regression theory in which an estimator of $\tilde{\boldsymbol{\beta}}$ that does not depend on the estimated fixed effect can be found (Andersen 1973). Further discussion of this point is deferred to chapter 4.

3.7 Models with General Correlation in the Errors: The Concept of Heterogeneity Extended

A great advantage of the multivariate probit models considered in this chapter is that they admit a more general characterization of heterogeneity than is conventional in the literature (e.g. see Singer and Spilerman 1976).

13. Note that if there are exogenous variables that are constant for the person over the sample period (e.g., education), these variables and their coefficients are absorbed into the estimated fixed effect. One can regress the estimated $l(i)$ on an intercept and the means of all exogenous variables to estimate the coefficients of such variables. Under the conditions stated in the text such estimators are consistent and asymptotically normally distributed.

14. A copy of the fixed effects probit program is available from the author on request for a fee covering duplication and processing charges.

The standard treatment of heterogeneity assumes a components of variance scheme with $f(\tau)$ as a mixing distribution (see equation 3.16) or empirical Bayes density (e.g., see Maritz 1970). Although this treatment is generalized, somewhat, in the one-factor model (see equation 3.19) it is clearly possible, and in many economic models desirable, to entertain a more general correlation structure for the unobservables that generate discrete choices.[15] For example, a simple first-order Markov model for the unobservables is ruled out by a components of variance or a one-factor scheme (for $T > 3$, see appendix 3.18). Yet it is natural in many economic contexts to assume that the unobserved variables obey such a correlation scheme.

The errors $\varepsilon(i, t)$ can be given an unrestricted covariance structure, more general than that described by the one-factor model. Both stationary and nonstationary distributions of the error process may be entertained. Using the multivariate probit model of Ashford and Sowden (1970), Domencich and McFadden (1975), or Dutt (1976), it is possible to estimate the unrestricted $T \times T$ correlation matrix $\tilde{\Sigma}$, and, if regressors (or just a time invariant intercept) are present, $\sigma(t, t)$, $t = 2, \ldots, T$, where the first disturbance variance ($\sigma(1, 1)$) is normalized to unity. A general nonstationary error process can thus be estimated, and it is possible to test specific models of the error structure against the unrestricted general model.[16]

To illustrate these points, an example is given. Consider a stationary Markov process of order one with a permanent component for the disturbances of the model. This error process was first considered by Balestra and Nerlove (1966):

$$\varepsilon(i,t) = \rho\varepsilon(i,t-1) + \tau(i) + U(i,t),$$

$I = 1, \ldots, I, \; t = 1, \ldots, T,$ where $E(\tau(i)) = 0$, $E(U(i, t)) = 0$, $E(\tau(i)^2) = \sigma_\tau^2$, $E(U(i, t)^2) = \sigma_U^2$, and $E(U(i, t)\tau(i)) = 0$, $|\rho| < 1$ and stationarity is assumed.

15. The restrictions imposed by the one-factor model are investigated in appendix 3.18. For $T > 3$, a one-factor model implies a nonstationary error process unless a random effects model is assumed. Many interesting processes, such as first-order Markov, cannot be analyzed by the one-factor scheme for $T > 3$.

16. Consistent estimators of the ratio of disturbance variances are achieved if $I \to \infty$. One does not require $T \to \infty$.

$$E(\varepsilon(i,t)\varepsilon(i,t')) = \frac{\sigma_\tau^2}{1-\rho^2} + \frac{\sigma_U^2 \rho^{|t-t'|}}{1-\rho^2};$$

$$\tilde{\sigma}(t,t') = \left[\frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_U^2}\right] + \left[\frac{\sigma_U^2}{\sigma_\tau^2 + \sigma_U^2}\right]\rho^{|t-t'|}.$$

The correlation matrix $\tilde{\boldsymbol{\Sigma}}$ can be parameterized in terms of $\sigma_\tau^2/(\sigma_\tau^2 + \sigma_U^2)$ and ρ, and these two combinations of parameters can be estimated. Since the disturbance variance is assumed to be identical in all time periods, no further parameters can be estimated. This restriction on the correlation matrix can be tested against the unrestricted covariance matrix. For $T > 3$, this error scheme cannot be transformed into one-factor form (see appendix 3.18). Hence heterogeneity cannot be treated by classical mixing distribution methods. Nonetheless a model with this error structure can be estimated by multivariate probit analysis.[17]

The probability that randomly selected person i experiences an event at time period $t(d(i,t) = 1)$ in a population with identical and constant values of the exogenous variables ($V(i,t) = \bar{V} = \mathbf{Z}(i,t)\boldsymbol{\beta} \neq 0$) is

$$\bar{P}(t) = \text{Prob}[\varepsilon(i,t) \geq -\bar{V}] = \Phi\left[\frac{\bar{V}}{\sigma(t,t)^{1/2}}\right].$$

17. As a second example, a nonstationary first-order Markov process is considered. The process starts up with initial disturbance $W(i)$ assumed independent of $U(i,t)$. $E(U(i,t)) = E(W(i)) = 0$. $E(W(i)^2) = \sigma_W^2$. Thus

$$\varepsilon(i,t) = \sum_{j=0}^{t-1} U(i,t-j)\rho^j + \rho^{t-1}W(i).$$

For $t' < t$,

$$E(\varepsilon(i,t)\varepsilon(i,t')) = \rho^{|t-t'|}\sigma_U^2 \sum_{j=0}^{t'-1} \rho^{2j} + \sigma_W^2 \rho^{t'+t-2},$$

$$\tilde{\sigma}(t,t') = \frac{\rho^{|t-t'|}\sigma_U^2 \sum_{j=0}^{t'-1} \rho^{2j} + \sigma_W^2 \rho^{t'+t-2}}{\left[\sigma_U^2 \sum_{j=0}^{t-1} \rho^{2j} + \sigma_W^2 \rho^{2t-2}\right]^{1/2}\left[\sigma_U^2 \sum_{j=0}^{t'-1} \rho^{2j} + \sigma_W^2 \rho^{2t'-2}\right]^{1/2}}.$$

The covariance matrix can be parameterized in terms of ρ and σ_W^2/σ_U^2. These combinations of parameters can be consistently estimated by multivariate probit analysis as $I \to \infty$, irrespective of the value of T so long as $T > 2$. Note that $\rho = 1$ (a random walk process) is a special case of this model. It is possible to test this hypothesis using classical likelihood ratios or Wald statistics (Rao 1973) based on the estimated information matrix for the model.

This probability can be estimated from a single cross section (at time t). Panel data are required to estimate the temporal correlation pattern among the unobservables. A series of successive cross sections can be used to estimate the ratio of error variances. The expected number of periods in the state over panel period T for a randomly sampled individual is

$$\sum_{t=1}^{T} \bar{P}(t).$$

The average duration in the state cannot be estimated from cross section estimates of $\bar{P}(t)$. If the intertemporal correlations among all disturbances are positive,[18] the true average duration exceeds the duration estimated from cross section data under the assumption of no intertemporal correlation in the errors.

In T trials the probability of J successes ($\Sigma d(i, t) = J$) and $T - J$ failures is not the same for any sequence with J successes, even if $V(i, t) = \bar{V} = \mathbf{Z}(i, t)\boldsymbol{\beta}$. Hence the random variables $d(i, t)$, $t = 1, \ldots, T$, are not exchangeable. However, if the latent variables that generate the process are stationary (in the weak sense, e.g., see Koopmans 1974, p. 38), sequences of events that are reflections of each other have identical probabilities, assuming $V(i, t) = \bar{V}$. The reflection of a sequence of T outcomes ($d(i, t)$, $t = 1, \ldots, T$) is defined as another sequence $d(i, t')$, $t' = 1, \ldots, T$, with $d(i, t') = d(i, T - t + 1)$.[19] For example, a sequence of trials recorded as (1, 0, 1, 1) has as its reflection sequence (1, 1, 0, 1).

To establish this point on reflection sequences, array $\varepsilon(i, t)$, $t = 1, \ldots, T$, into a $1 \times T$ vector $\boldsymbol{\varepsilon}(i)$, assumed to be normally distributed with mean zero and variance $\boldsymbol{\Sigma}$. As a consequence of assumed stationarity $\sigma(t, t') = \sigma(|t - t'|)$ and $\sigma(t, t) = \sigma$. The reflection of $\boldsymbol{\varepsilon}$ is $\boldsymbol{\varepsilon}^R$ defined by

$$\boldsymbol{\varepsilon}^R = \mathbf{P}\boldsymbol{\varepsilon},$$

where $\mathbf{P}$ is a traverse diagonal permutation matrix ($P(i, j) = 1$ for $j = T - i + 1, P(i,j) = 0$ otherwise). The covariance matrix of $\boldsymbol{\varepsilon}$ is $\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}'$. From stationarity, $\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}' = \boldsymbol{\Sigma}$, since $\sigma(T - i, T - j) = \sigma(|i - j|) = \sigma(i, j)$. Thus any dichotomization of the elements of $\boldsymbol{\varepsilon}$ that generates an observed sequence of events $d(i, t)$, $t = 1, \ldots, T$, has the same probability as the

18. This condition is sufficient but not necessary.

19. The term "mirror image" is more suggestive. Imagine holding the first sequence up to a mirror and noting its reflection.

identical dichotomization applied to the elements of ε^R. Hence a sequence and its reflection have equal probability.[20]

Runs tests can thus be used to distinguish between the exchangeable models considered in sections 3.4 and 3.5 and the general stationary model considered here. In the former models all sequences with J successes in T trials have equal probability. In the general model for stationary disturbances only those subsequences that are reflections of each other have identical probability.[21] In a general nonstationary model reflection sequences do not have identical probability. Runs tests to distinguish between exchangeable, stationary, and nonstationary models are developed more completely elsewhere (Heckman 1978b).

Observable characteristics that determine choices can be incorporated into the model with general heterogeneity in precisely the same way as has been done in the models developed in the previous sections. $V(i, t)$ may be parameterized, so that $V(i, t) = \mathbf{Z}(i, t)\boldsymbol{\beta}$ (thus $V(i, t)$ in equation 3.6 is equal to $\mathbf{Z}(i, t)\boldsymbol{\beta}$). Given an intercept (or other exogenous variables), it is possible to estimate $\sigma(t, t)$, $t = 2, \ldots, T$, subject to the normalization that $\sigma(1, 1) = 1$.

3.8 Models with Structural State Dependence

The structural relationship between discrete outcomes in different periods is termed structural state dependence. All of the models considered in the previous sections assume no structural state dependence once heterogeneity is properly accounted for.

This is not to say that in the preceding models the conditional probability that $d(i, t) = 1$ given $d(i, t') = 1(t' \neq t)$ is the same as the marginal probability that $d(i, t) = 1$. If there are unmeasured, serially correlated components in the errors, or measured, serially correlated components not adequately controlled for, such a conditional relationship will arise. However, controlling for the serially correlated components in the error and in the measured variables, no conditional relationship will arise.

20. The assumption of weak stationarity and normality implies strong stationarity (Koopmans 1974, p. 38). The results in the text are a consequence of strong stationarity. Any strongly stationary series has a time reversibility property required to establish the results.

21. Thus in an exchangeable model the sequences (1, 0, 1), (1, 1, 0), and (0, 1, 1) have equal probability of occurrence, but in the stationary model in general only the last two sequences have equal probability of occurrence.

To illustrate this point, consider the random effect model developed in section 3.5. Assume that there is no variation in measured exogenous variables in the population. However, assume that the probability of experiencing the event is a function of an unobserved component $P = P(\tau)$.

The probability that $d(i, 2) = 1$, given $d(i, 1) = 1$, is

$$\text{Prob}[d(i,2) = 1 \mid d(i,1) = 1] = \frac{\int_{-\infty}^{\infty} P^2(\tau) f(\tau) d\tau}{\int_{-\infty}^{\infty} P(\tau) f(\tau) d\tau},$$

which is not the same as the marginal probability $\text{Prob}[d(i, 2) = 1] = \int_{-\infty}^{\infty} P(\tau) f(\tau) d\tau$. However, the probability that $d(i, 2) = 1$ given $d(i, 1)$ and $\tau(i)$ is the same as the probability that $d(i, 2) = 1$ given $\tau(i)$:

$$\text{Prob}[d(i,2) = 1 \mid d(i,1) = 1 \text{ and } \tau(i)] = \frac{P^2(\tau(i))}{P(\tau(i))}$$

$$= P(\tau(i)) = \text{Prob}[d(i,2) = 1 \mid \tau(i)].$$

Controlling for temporally correlated unobserved components (the τ), there is no conditional relationship between the probability that $d(i, 2) = 1$ and the value of $d(i, 1)$. It is in this sense that the models developed in the preceding sections do not generate structural relationships between outcomes in different periods. The models presented in this section do.[22]

To focus on essential ideas, assume initially that there is no heterogeneity in measured or unmeasured variables, so that $\mathbf{Z}(i, t)\boldsymbol{\beta} = \beta_0$ and the $\varepsilon(i, t)$ are independently identically distributed random variables and $E[\varepsilon(i, t)^2] = 1$. To commence the analysis, assume that only previous outcomes affect current choice. This leads to the following expression for $Y(i, t)$, the difference in remaining lifetime utilities at time t:

$$Y(i,t) = \beta_0 + \sum_{j=1}^{\infty} \gamma(t-j, j) d(i, t-j) + \varepsilon(i,t). \tag{3.20}$$

22. The example offered in this section is simple and thus has considerable pedagogical appeal. The validity of the point is not confined to a simple random effect or one-factor model. The general models developed in the preceding section also generate a conditional relationship between events in different periods, solely as a consequence of temporal correlation in the errors.

Presample values of $d(i, t')$, $t' = 0, -1, \ldots$, assume fixed, nonstochastic values. If $Y(i, t) \geq 0, d(i, t) = 1$. Otherwise $d(i, t) = 0$. The second term on the right-hand side is assumed to be finite.

The probability that $d(i, t) = 1$, given $d(i, t-1), \ldots$, is

$$\text{Prob}[d(i,t) = 1 \mid d(i,t-1), d(i,t-2), \ldots] = \Phi\left[\beta_0 + \sum_{j=1}^{\infty} \gamma(t-j, j) d(i, t-j)\right].$$

Thus the sample likelihood for a given sequence of outcomes arrayed in a $1 \times T$ vector $\mathbf{d}(i)$ is

$$\mathscr{L} = \prod_{i=1}^{I} \prod_{t=1}^{T} \Phi\left\{\left[\beta_0 + \sum_{j=1}^{\infty} \gamma(t-j, j) d(i, t-j)\right](2d(i,t) - 1)\right\}.$$

If $\gamma(t-j, j) = \gamma(1)$ for $j = 1$, and $\gamma(t-j, j) = 0$ for $j > 1$, the model generates a first-order Markov process.[23]

$$\text{Prob}[d(i,t) = 1 \mid d(i,t-1)] = \Phi[\beta_0 + \gamma d(i, t-1)].$$

If $\gamma(t-j, j) = \gamma(j)$ for $j \leq K$, $\gamma(t-j, j) = 0$ for $j > K$, a Kth-order Markov process is generated. If $\gamma(t-j, j) = \gamma$, a generalization of a Pólya process (e.g., see Feller 1957) is generated in which the entire history of the process is relevant to current choices.[24] Allowing for geometric decay of effects in the generalized Pólya model, one may parameterize $\gamma(t-j, j) = \gamma_0(\sigma)^j$, $0 < \sigma < 1$.

Permitting the γ coefficients to depend on calendar time t as well as age generates time inhomogeneous versions of the Markov and generalized Pólya models. Clearly the γ coefficients may be parameterized to depend on values of the exogenous variables at the time the event occurs and on current values of the exogenous variables (or for that matter values in other periods).

23. For logit Φ Boskin and Nold (1975) have presented a Markov model with exogenous variables. They ignore heterogeneity in unmeasured, serially correlated components. See also Amemiya (1978) who investigates the properties of maximum likelihood estimators for this model.

24. A related model for the Pólya type process has been developed by Chaddha (1963). I am indebted to Jerzy Neyman for this reference. The Pólya model is similar to the linear-learning probability model of Bush and Mosteller (1955). See also Massy, Montgomery, and Morrison (1970) and Wilson (1977). I am indebted to Frank O'Connnor and Abel Jeuland for these references.

Conditions for identification of parameters in Markov models (both time homogenous and time inhomogenous) are well known (c. f., Anderson and Goodman 1957). Without invoking special assumptions, such as stationarity of the process, panel data are required to estimate the model. Runs tests can be performed to discriminate between Bernoulli and Markov models (e.g., see David 1947, Goodman 1958, and Denny and Yakowitz 1978).

The parameters of the generalized Pólya model can be estimated from data available from a single cross section, provided that the number of past events ($\Sigma_{j=1}^{\infty} d(i, t-j)$) is known. One does not need to know when the past events occurred. For the generalized Pólya process with geometric decay one requires knowledge of the entire past history of the process to identify the parameters of the model.

In T trials the probability of J successes ($\Sigma_{t=1}^{T} d(i, t) = J$) and $T - J$ failures is not the same for any sequence of J successes in any order. Because of the time irreversability inherent in the nonstationary process induced by the random variable $\Sigma_{j=1}^{\infty} d(i, t-j)$, reflection sequences do not have identical probabilities. In the generalized Pólya model (without decay), if $\gamma > 0$, a sequence $\mathbf{d}(i) = (1, 0, 1, 1)$ is more probable than a sequence $(1, 1, 0, 1)$. (Recall that the sequences are ordered in time from left to right, starting with the earliest outcome.) Since occupancy of a state raises the probability of future occupancy, a later failure is less likely than an earlier one.

To see this, note that

$$\text{Prob}(1,0,1,1) = \Phi[\beta_0]\Phi[-(\beta_0 + \gamma)]\Phi[\beta_0 + \gamma]\Phi[\beta_0 + 2\gamma],$$

$$\text{Prob}(1,1,0,1) = \Phi[\beta_0]\Phi[\beta_0 + \gamma]\Phi[-(\beta_0 + 2\gamma)]\Phi[\beta_0 + 2\gamma]. \tag{3.21}$$

Since $\gamma > 0$, the first sequence is more probable. (Compare the second term in the first sequence with the third term in the second sequence.) Runs tests can be used to distinguish among exchangeable models, a model with stationary errors, and the generalized Pólya process (see Heckman 1978b).

Heterogeneity in unmeasured variables can be introduced into the models considered in this section in exactly the same way it has been introduced in the models considered in sections 3.5 through 3.7. No new idea is introduced by merging models that allow for heterogeneity with models that allow for structural state dependence. For example, in each of the models considered here, the components of variance error structure given in equation (3.22),

$$\varepsilon(i,t) = \tau(i) + U(i,t), \tag{3.22}$$

can be specified. This error structure generates to the mixing distribution representation of heterogeneity which leads to the probability for $\mathbf{d}(i)$ (given the fixed nonstochastic initial conditions of the process) of

$$\text{Prob}[\mathbf{d}(i) \mid d(i,0), d(i,-1), \ldots]$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\left\{\left[\beta_0 + \sum_{j=1}^{\infty} \gamma(t-j,j)d(i,t-j) + \left[\frac{\rho}{1-\rho}\right]^{1/2}\tilde{\tau}\right]\right.$$

$$\left.[2d(i,t)-1]\right\} f(\tilde{\tau})d\tilde{\tau}, \tag{3.23}$$

where ρ and $\tilde{\tau}$ are as defined in section 3.5.

The one-factor model given in equation (3.17) can be applied to the disturbances of the models considered in this section in a straightforward way, as can the fixed effect and fixed factor models considered in section 3.6. Nonstationary disturbances of the sort considered in sections 3.4 and 3.5 can also be introduced in these models, and ratios of disturbance variances in different periods can be estimated if $\beta_0 \neq 0$, even for general values of $\gamma(j, t-j)$.[25]

The results on runs patterns established for the generalized Pólya process (see equation 3.21 and the surrounding discussion) remain intact if heterogeneity of the components of variance type is introduced. To see this, note that in equation (3.23), if $\gamma(t-j,j) = \gamma, d(i,t') = 0, t' \leq 0$ (so that the generalized Pólya process is generated), the ordering among the probabilities of runs patterns previously established continues to be valid, since the relative ranking in probability of any two runs sequences is not affected by integration with respect to $f(\tilde{\tau})$.

The preceding analysis does not require that $\varepsilon(i, t)$ be normally distributed. Φ can be the cumulative distribution of any latent variable (symmetry can be relaxed at the cost of minor notational inconvenience). The principal advantage of the normality assumption is that it generates a model that can readily be generalized to accommodate a rich variety of error structures for serially correlated unobserved components.

25. With more structure on the $\gamma(j, t-j), j = 1, \ldots$, the ratio of disturbance variances can be identified even if $\beta_0 = 0$.

General heterogeneity, of the sort considered in section 3.7, can be introduced into the models considered in this section. The probability of $\mathbf{d}(i)$ given $\mathbf{Z}(i)$ and the nonstochastic initial conditions of the process is given by the general cumulative normal density; the expression for it is given in equation (3.6), with $V(i, t) = \beta_0 + \Sigma_{j=1}^{\infty} \gamma(t-j, t) d(i, t-j)$. Given $\beta_0 \neq 0$ (or specific structure on the $\gamma(t-j, t)$ coefficients), a nonstationary version of the model can be identified. In all of the models with nonindependent disturbances, panel data are required to estimate the serial correlation structure of the unobservables.

Introducing exogenous variables into the models considered in this section does not involve any new principle. In place of β_0 in the preceding expressions, one can substitute $\mathbf{Z}(i, t)\boldsymbol{\beta}$.

It is important to stress that the assumption made throughout this section that initial conditions are known and nonstochastic is neither innocuous nor especially plausible. In many contexts the analyst has access to data on a process that is sampled midstream, so that the initial conditions are determined by the same stochastic process that generates the panel data. In this case it is inappropriate to assume that the initial conditions are nonstochastic. Maximum likelihood estimators of the parameters of the models conditioned on presample realizations of the process are not consistent unless the disturbances are truly independent. This problem and various solutions to it are considered in chapter 4.

In all of the models considered in this section, it is possible to reverse the sense of the j subscript (in equation 3.20) and allow for future outcomes to determine current choices. This sort of structural dependence arises in certain life cycle models of decision making under perfect certainty. For example, in an analysis of labor supply behavior future work may determine the current probability of working if current labor supply raises future wage rates. The greater the volume of future labor supply, the more profitable is current work activity.[26]

26. For a discussion of certain technical problems that arise from simultaneous introduction of the effect of all past and future outcomes on current choices, see Heckman (1978a, pp. 936 and 957) and Schmidt (chapter 12). The basic problem is one of internal inconsistency in probit probability statements. The requirement for internal consistency of the model is that, through a suitable permutation of subscripts of the coefficients of the dummy variables denoting state occupancy, the equation system generating the model can be brought into lower triangular form for the coefficients of the dummy variables denoting state occupancy. The models given in the text satisfy this requirement.

3.9 A Renewal Model

The essential feature of the renewal model of structural state dependence is that the only effect of previous state occupancy on current choices is from the most recent current spell in the state. In an analysis of specific human capital of the sort considered by Jovanovic (1978), workers acquire wage-enhancing experience which makes them less likely to leave the work state. However, once the worker leaves the state, the experience is lost and hence is irrelevant to his future choices. The simplest way to capture this effect is with the term

$$\lambda \sum_{j=1}^{\infty} \prod_{l=1}^{j} d(i, t-l). \tag{3.24}$$

Closely related to the concept of specific capital is the concept of fixed costs. Such costs may be incurred once an individual enters a state (e.g., retraining costs for a woman who has entered the labor market). Having incurred the cost, the individual's choice set for subsequent decisions changes, in the sense that the fact she no longer has to incur the cost as long as she remains in the state is taken into account in her subsequent sequential decision making. This concept may be captured by the term

$$\lambda d(i, t-1).$$

This term also generates a renewal process.

Introduction of such effects into the preceding models raises no new conceptual issues, apart from those just discussed. A general expression for relative utility that captures both of these effects in a simple choice theoretic model is

$$Y(i,t) = \sum_{j=1}^{\infty} \lambda(t-j, j) \prod_{l=1}^{j} d(i, t-l) + \varepsilon(i,t).$$

The case of fixed costs corresponds to $\lambda(t-j, j) = \lambda(1)$, $j = 1$, $\lambda(t-j, j) = 0, j \geq 2$. The simplest model of specific human capital accumulation sets $\lambda(t-j, j) = \lambda$ for all j. Depreciation of these effects can be accommodated in the general model.

The fixed cost model is indistinguishable from a first-order Markov model.[27] The general renewal model is distinguishable from the general finite state Markov models and the generalized Pólya models considered in the preceding section. Heterogeneity and the effect of exogenous variables on choices may be introduced into the renewal models in exactly the same way as discussed in the preceding sections.

3.10 A Model with Habit Persistence

The key feature of the models with structural state dependence is that occupancy of a state in another period determines current choices, controlling for the effect of unmeasured heterogeneity. The model considered in this section ignores this form of dependence but permits relative utility evaluations in other periods ($Y(i, t')$, $t \neq t'$) to determine current choices. Models with habit formation have been considered by Pollak (1970) and are implicit in Coleman's latent Markov model (1964). The model considered here is the discrete data analogue of the classical distributed lag model in econometrics.

The basic idea of habit persistence can be captured by the following model for current relative utility, $Y(i, t)$,

$$Y(i,t) = G(L)\,Y(i,t) + \varepsilon(i,t), \tag{3.25}$$

where $G(0) = 0$, and $G(L)$ is a polynomial lag of order K. ($G(L) = g_1 L + g_2 L^2 + \cdots + g_K L^K$, $L^K Y(i,\ t) = Y(i,\ t-K)$.) One can introduce distributed leads as well, but this is not done here. Assuming that $(1 - G(L))$ is invertible (e.g., see Granger and Newbold 1977), the model may always be rewritten as

$$Y(i,t) = [1 - G(L)]^{-1}\,\varepsilon(i,t).$$

If the $\varepsilon(i, t)$ are iid, the coefficients of $G(L)$ may be estimated (up to an unknown factor of proportionality) by multivariate probit analysis, provided the available panel is of suitable length ($T \geq K$) and that the initial conditions for $Y(i, t')$, $t' < 0$, are specified. If the $\varepsilon(i, t)$ are not iid, and the process determining $\varepsilon(i, t)$ is unknown, the model is not identified. This identification problem is exactly the same problem that arises in

27. Indeed the fixed cost model provides a rationalization for a first-order Markov model.

estimating a distributed lag model in the presence of serial correlation (see Griliches 1967, p. 35).

Introduction of exogenous variables into the model aids in identification. If the model of equation (3.25) is augmented to include exogenous variables,

$$Y(i,t) = \mathbf{Z}(i,t)\boldsymbol{\beta} + G(L)\,Y(i,t) + \varepsilon(i,t), \tag{3.26}$$

it is possible to estimate (variance normalized) elements of $G(L)$ and $\boldsymbol{\beta}$ as well as the correlations among the disturbances. This is so because in reduced form

$$Y(i,t) = [1 - G(L)]^{-1}\mathbf{Z}(i,t)\boldsymbol{\beta} + [1 - G(L)]^{-1}\varepsilon(i,t), \tag{3.27}$$

so that from the estimated coefficients on the lagged values of the $\mathbf{Z}(i, t)$ variables it is possible to solve for the normalized coefficients of $G(L)$, provided that the $\mathbf{Z}(i, t), t = 1, \ldots, T$, are not linear combinations of each other for all i, and initial conditions $Y(i, t'), t \leq 0$, are specified.[28]

It is interesting to note that, if at least one variable in $\mathbf{Z}(i, t)$ changes over time, and exact linear dependency among the $\mathbf{Z}(i, t)$ does not exist, a probit model fit on one cross section can be used to test for habit persistence. The test consists of entering lagged values of $\mathbf{Z}(i, t)$ into the probit model based on equation (3.27). If the lagged values of $\mathbf{Z}(i, t)$ are statistically significantly different from zero, one can reject the hypothesis of no habit persistence. Cross section probit models can be used to estimate the normalized coefficients of $G(L)$, provided the analyst has access to lagged values of the $\mathbf{Z}(i, t)$.

The model for habit persistence may be grafted onto the models with structural state dependence developed earlier. General conditions for identification in this model are presented elsewhere (Heckman 1978a, p. 956). The important point to note is that subject to exclusion (or other identification) restrictions, even though $Y(i, t')$ is never observed, its effect on current choice can be estimated and distinguished from the effect of structural state dependence. Thus one can separate the effect of past propensities to occupy a state on current choices from the effect of past occupancy of a state on current choices.

28. A model with lagged latent variables appears in Heckman (1978a, pp. 932 and 956).

3.11 Computation in the General Model[29]

One factor and fixed effect schemes have already been proposed. In the appendix, multifactor schemes are discussed as well. All of these models are fairly cheap to compute and on these grounds are recommended.

In the random effect model it is only necessary to use two periods of data, not necessarily adjacent, to estimate ρ and $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma_U$ (see equation 3.16). All the parameters in this model may be estimated by standard bivariate probit programs. Estimates obtained from this procedure are presumably good starting values for optimization of the full likelihood function.

Even cheaper estimates are possible. From each cross section, $t = 1, \ldots, T$, it is possible to estimate $\tilde{\boldsymbol{\beta}}(1 - \rho^2)^{1/2}$ by probit analysis (recall that $(1 - \rho^2)^{1/2} = [\sigma_U^2/(\sigma_\tau^2 + \sigma_U^2)]^{1/2}$). Substituting for $\tilde{\boldsymbol{\beta}}$ in likelihood function (3.16), and optimizing the function with respect to ρ conditional on $\tilde{\boldsymbol{\beta}}(1 - \rho^2)^{1/2}$, reduces the computational task to a one-parameter problem. (In practice it is preferable to use an average of cross-sectional estimates of $\tilde{\boldsymbol{\beta}}(1 - \rho^2)^{1/2}$). Estimates of ρ obtained in this fashion are consistent but inefficient. Such estimates of $\tilde{\boldsymbol{\beta}}$ and ρ are consistent starting values for full system optimization. One can further simplify this procedure by optimizing only a two-period likelihood function (for any two periods of data) with respect to ρ conditional on $\tilde{\boldsymbol{\beta}}$.

These principles can also be applied to the other models considered earlier. For example, in the random factor model developed in section 3.5, one can utilize any two periods of data (say, for time t and t') to estimate $\tilde{\boldsymbol{\beta}}(t)$ and $\tilde{\boldsymbol{\beta}}(t')$, as well as $\tilde{\alpha}(t)\tilde{\alpha}(t')(= \tilde{\sigma}(t, t'))$, by bivariate probit analysis. These estimators are inefficient but can be used to compute all the parameters of the model by estimating all possible two-period models. (The periods need not be adjacent.)

It is possible to use cross section probit exactly as in the random factor model to estimate $\tilde{\boldsymbol{\beta}}(t)(1 - \eta^2(t))^{1/2}$, $t = 1, \ldots, T$, in the one-factor model. Bivariate probit (conditional on estimated values of $\tilde{\boldsymbol{\beta}}(t)(1 - \eta^2(t))^{1/2}$ can be used to estimate $\tilde{\alpha}(t)\tilde{\alpha}(t')(= \tilde{\sigma}(t, t'))$ for any two periods of data t and t'. This requires optimization of a bivariate probit model with respect to one parameter. By this bootstrap method, it is possible to estimate inexpensively all the parameters of the model. Of course, given estimated values of $\tilde{\boldsymbol{\beta}}(t)(1 - \eta^2(t))^{1/2}$, it is possible (but more

29. This section draws on Heckman (1976, pp. 245–246).

costly) to estimate the $\tilde{\alpha}(t)$ parameters from the likelihood function for the complete sample.

In a similar fashion bivariate probit may be used to compute the parameters of the model with general heterogeneity given in section 3.7. Again the full correlation matrix, $\tilde{\Sigma}$, $\sigma(t, t)$, $t = 2, \ldots, T$, and β can be estimated from all possible combinations of bivariate calculations, and cross section probit used to compute $\tilde{\beta}(t) = \beta/\sigma(t,t)^{1/2}, t = 1, \ldots, T$. The bivariate probit computations can be made conditional on the estimated values of $\tilde{\beta}(t)$, so that only optimization with respect to a single parameter (the correlation coefficient for the disturbances of the two periods selected) is required. Application of these methods to the model with habit persistence (section 3.10) is straightforward (see also Heckman 1976, pp. 245–246).

Fewer shortcut methods are available for the models with structural state dependence considered in sections 3.8 and 3.9. Given nonstochastic initial conditions, it is possible to use the first cross section in the panel data to estimate (variance normalized) structural coefficients. Given the (normalized) structural parameters, it is possible to use the remainder of the available panel data to estimate the correlation structure and the variances $(\sigma(t, t), t = 2, \ldots, T)$.

Recent advances in computing the multivariate normal integral (Albright, Lerman, and Manski 1977) make direct maximum likelihood estimation of the general model feasible for T as large as 10. The consistent estimators proposed in this section provide good starting values for this maximum likelihood algorithm. It is well known that, starting with consistent estimators, one Newton step toward the likelihood optimum yields asymptotically efficient estimators.

3.12 A Summary of Sections 3.2 through 3.11

A general model for the analysis of discrete choices made over time has been presented and special cases have been considered in detail. A variety of discrete time-discrete data stochastic processes emerges as special cases of the general model of section 3.3. The cases likely to be of interest in applied work are presented in table 3.1. The restrictions on the general model required to generate the special models are presented under the appropriate column headings.

Table 3.1
Restrictions on parameters required to generate some specific models from the general model

General model: $Y(i,t) = Z(i,t)\boldsymbol{\beta} + \sum_{j=1}^{\infty} \gamma(j, t-j)d(i,t-j) + \lambda \sum_{j=1}^{\infty} \prod_{l=1}^{j} d(i,t-l) + G(L)Y(i,t) + \varepsilon(i,t)$

$Y(i,t) \geq 0$ iff $d(i,t) = 1$
$Y(i,t) < 0$ iff $d(i,t) = 0$.

	Bernoulli model		Markov model (Kth order)		Renewal model		Polya-type model		Model with habit persistence	
	Homogeneous	Heterogeneous[a]	Homogeneous	Heterogeneous[a]	Homogeneous	Heterogeneous[a]	Homogeneous	Heterogeneous[a]	Homogeneous	Heterogeneous[a]
$\boldsymbol{\beta}$	0[e]	Free[b]	0[e]	Free[b]	0[e]	Free[b]	0[e]	Free[b]	0[e]	Free[b]
$\gamma(j, t-j)$	0	0	$\gamma(j)$[f]	$\gamma(j)$[f]	0	0	Free	Free	0	0
λ	0	0	0	0	λ	λ	0	0	0	0
$G(L)$	0	0	0	0	0	0	0	0	Free	Free[c]
$\varepsilon(i,t)$	iid[d]	Free	iid[d]	Free	iid[d]	Free	iid[d]	Free	iid[d]	Free

[a]Heterogeneous in this table refers to heterogeneity in measured variables and heterogeneity in unobserved serially correlated components.
[b]"Free" means that the parameter may assume unrestricted finite values. Thus one may replace $\boldsymbol{\beta}$ with $\mathbf{Z}(i, t)\,\tilde{\boldsymbol{\beta}}(t)$.
[c]As noted in the text, one requires regressors to distinguish between $G(L)$ and an arbitrary correlation pattern for the $\varepsilon(i, t)$, except in special cases in which $G(L)$ and/or the error process are restricted.
[d]As noted in the text, nonstationary, independently nonidentically distributed versions of these models can be estimated.
[e]Except for intercept which may be zero.
[f]$\gamma(j), j \leq K$, zero otherwise.

The concept of heterogeneity has been generalized in these models beyond the mixing distribution, or convolution, concept which appears in the literature to a broader definition of serial correlation among unobservable variables.[30] Each of the models considered here can accommodate heterogeneity of a very general sort, as well as time-varying explanatory variables. It is possible to test for nonstationarity of the errors as well as special hypotheses about the correlation structure of the unobservables.

Given current computing technology, the models are estimable. The one-factor and fixed effect models are particularly simple to implement.

3.13 Heterogeneity versus Structural State Dependence: An Application of the Preceding Models[31]

In the introduction to this chapter the following empirical regularity is noted: individuals who experience an event in the past are more likely to experience the event in the future than are individuals who have not experienced the event in the past. This observation is based on many studies of series of discrete events taken from individual histories, such as records of illness, unemployment, accidents, or labor force participation. There are two conceptually distinct explanations for this empirical regularity. One is that individuals who experience the event are altered by their experience in that the constraints, preferences, or prices (or any combination of the three) that govern future outcomes are altered by past outcomes. Such an effect of past outcomes on future outcomes is termed structural state dependence. A second explanation is that individuals differ in some unmeasured propensity to experience the event and this propensity is either stable over time or, if it changes, values of the propensity are autocorrelated. Broadly defined, the second explanation is a consequence of population heterogeneity.

The problem of distinguishing between these two explanations for the empirical regularity has a long history. The earliest systematic discussion of this problem appears in the analysis of accident proneness. The seminal work on this topic is due to Feller (1940) and Bates and Neyman (1951).[32] Bates and Neyman are especially clear in pointing out the need for panel

30. The two concepts of mixing distribution and convolution, while equivalent in the models considered in this chapter, are not always identical. See Blischke (1963).

31. The comments of Zvi Griliches and Tom MaCurdy have been very helpful in preparing the revision to the remaining sections of the chapter.

32. I am indebted to Jan Hoem for the Feller reference.

data on individuals to distinguish between the two explanations. Work that preceded the Feller and Bates-Neyman papers attempted to use cross section distributions of accident counts to distinguish between true and spurious state dependence. (See Feller for references to this work.)

In the balance of this chapter the apparatus developed in the preceding sections is applied to address this problem. Before becoming absorbed in the details of the solution, it is important to distinguish the solution, which relies on special techniques and assumptions, from the problem, which can be defined more generally.

To this end it is useful to consider four simple urn models which provide a useful framework within which to introduce intuitive notions about heterogeneity and state dependence. In the first scheme there are I individuals who possess urns with the same content of red and black balls. On T independent trials individual i draws a ball and then puts it back in the urn. If a red ball is drawn at trial t, person i experiences the event $(d(i,t) = 1)$. If a black ball is drawn, person i does not experience the event $(d(i,\ t) = 0)$. This model corresponds to the simple Bernoulli model presented in section 3.4 and captures the essential idea underlying the choice process in McFadden's (1976) work on discrete choice. From data generated by this urn scheme, one would not observe the empirical regularity previously described.

The second urn scheme generates data that would give rise to the empirical regularity solely due to heterogeneity. In this model individuals possess distinct urns which differ in their composition of red and black balls. As in the first model sampling is done with replacement. However, unlike the first model information concerning an individual's past experience of the event provides information on the composition of his urn.

The person's past record can be used to estimate the person specific urn composition. The conditional probability that individual i experiences the event at time t is a function of past experience of the event. The contents of each urn are unaffected by actual outcomes and in fact are constant. There is no true state dependence. This model corresponds to the random effect model presented in section 3.5.

The third urn scheme generates data characterized by true state dependence. In this model individuals start out with identical urns. On each trial the contents of the urn change *as a consequence of the outcome of the trial.* For example, if a person draws a red ball, and experiences the event, additional new red balls are added to his urn. If he draws a black ball, no

new black balls are added to his urn. Subsequent outcomes are affected by previous outcomes because the choice set for subsequent trials is altered as a consequence of experiencing the event. This model corresponds to the generalized Pólya model described in section 3.8.[33]

A variant of the third urn scheme can be constructed that corresponds to the renewal model presented in section 3.9. In this scheme new red balls are added to an individual's urn on successive drawings until a black ball is drawn, and then all of the red balls added in the most recent continuous run of drawings of red balls are removed from the urn. The composition of the urn is the same as it was before the first red ball in the run was drawn. The fixed cost model is a variant of the renewal scheme in which new red balls are added to an individual's urn only on the first draw of a red ball.

The crucial concept that distinguishes the third scheme from the second is that the contents of the urn (the choice set) are altered as a consequence of previous experience. The key point is not that the choice set changes across trials but that it changes in a way that depends on previous outcomes of the

33. For a complete description of the Pólya process and its generalizations see Johnson and Kotz (1977, chapter 4). They note (pp. 180–181) that, in the special case in which a person draws a ball and receives the *same number* of the balls of the color drawn whether a black or red ball is drawn, urn model three (in this case a strict Pólya model) generates sequences of outcomes *identical* in probability with the same sequences generated from urn model two provided that the population distribution of the proportion of red and black balls in the urn is *Beta*. In this case panel data cannot be used to distinguish between the two urn models. In a stationary environment, in which urn contents are not exogenously changed, as long as the number of red balls placed in the urn differs from the number of black balls placed in the urn when a black ball is drawn, it is possible to use panel data to discriminate between the two models. This observation is one of the key insights in the Bates-Neyman paper (1951).

A similar result appears in the multivariate probit model. For example, consider the following generalization of the model of equation (3.20) with β_0 replaced by $\beta_0 t$ where $t < \infty$ is the length of time the process has been in operation. Assume $\gamma(t-j, j) = \gamma$. Suppose that if individual i does not experience the event in time period $t'(< t)$, so that $d(i, t') = 0$, he receives a "dose" γ'. The relative utility evaluation for this model may be written as

$$Y(i, t) = \beta_0 t + \gamma \sum_{j=1}^{t} d(i, t-j) + \gamma' \sum_{j=1}^{t} (1 - d(i, t-j)) + \varepsilon(i, t)$$

$$= \beta_0 t + (\gamma - \gamma') \sum_{j=1}^{t} d(i, t-j) + \gamma' t + \varepsilon((i, t),$$

$d(i, t) = 1$ if $Y(i, t) \geq 0$, $d(i, t) = 0$ otherwise. If $\gamma = \gamma'$, there is no structural state dependence as defined in the text, although there is a trend effect (so long as $\beta_0 + \gamma' \neq 0$). Thus even though the individual receives a "dose" of γ when he experiences the event and a dose of γ' when he does not, if the doses are of equal strength there is no way to measure the dose. In the special case of a stationary environment ($\beta_0 = 0$), it is clearly possible to estimate $\gamma(= \gamma')$ from the coefficient on t.

choice process. To clarify this point, it is useful to consider a fourth urn scheme that corresponds to the models with more general types of heterogeneity considered in sections 3.5 and 3.7.

In this model individuals start out with identical urns, exactly as in the first urn scheme. After each trial, but independent of the outcome of the trial, the contents of each person's urn are changed by discarding a randomly selected portion of balls and replacing the discarded balls with a randomly selected group of balls from a larger urn (say, with a very large number of balls of both colors). Assuming that the individual urns are not completely replenished on each trial, information about the outcomes of previous trials is useful in forecasting the outcome of future trials, although the information from a previous trial declines with its remoteness in time. Like the situation in the second and third urn models, previous outcomes give information about the contents of each urn. Unlike the situation in the second model, the information depreciates since the contents of the urn are changed in a random fashion. Unlike the third model the contents of the urn do not change as a consequence of any outcome of the choice process.

The general model presented in section 3.3 is sufficiently flexible that it can be specialized to generate data on the time series of individual choices consistent with samplings from each of the four urn schemes just mentioned as well as more general schemes (including combinations of the four). The principal advantage of this model over models considered in previous work is that it accommodates very general sorts of heterogeneity and state dependence as special cases of the general model and permits the introduction of explanatory exogenous variables in a natural way. The generality of the framework proposed here permits the analyst to combine models and test among competing specifications within a unified framework.

In section 3.14 a simple example is offered to illustrate how the models presented in sections 3.2 through 3.12 can be used to distinguish betweeen heterogeneity and state dependence. Section 3.15 examines the superficially appealing analogy between the problem of distinguishing heterogeneity from state dependence and the classical time-series problem of distinguishing a distributed lag model from a model with serial correlation. The analogy is found to be somewhat misleading. A more appropriate analogy is proposed. The final section 3.16 offers three examples of how structural state dependence may arise. The most interesting example is one with state dependence generated in an environment of perfect certainty.

3.14 Testing for Heterogeneity versus State Dependence

Suppose that there is access to a sample of I randomly selected individuals who are observationally identical at time $t = 1$. There are two observations per person, so that $T = 2$. The process is assumed to start up with no history at $t = 1$. Equivalently $d(i, t') = 0$, $t' \leq 0$, and these values are fixed and independent of the process.

Utilizing the notation established in section 3.3, individual i experiences an event ($d(i, 1) = 1$) if and only if $Y(i, 1) \geq 0$, where

$$Y(i,1) = \bar{V} + \varepsilon(i,1),$$
$$E(\varepsilon(i,1)) = 0,$$
$$E(\varepsilon(i,1)^2) = \sigma(1,1).$$

Thus $Y(i, 1) \geq 0$ iff $d(i, 1) = 1$. $Y(i, 1) < 0$ iff $d(i, 1) = 0$. The utility function consists of a deterministic component $\bar{V}$ and a stochastic component $\varepsilon(i, 1)$. The probability that $d(i, 1) = 1$ is

$$\text{Prob}[\varepsilon(i,1) \geq -\bar{V}] = \Phi\left[\frac{\bar{V}}{\sigma(1,1)^{1/2}}\right].$$

The hypothesis that there is a real effect of occupancy of a state on future behavior requires that individuals who experience the event in time period one have their relevant second-period choice set changed in a way that directly depends on choice in the preceding period so that second-period choice probabilities are altered.

One way to capture this idea which is a natural extension of the choice theoretic models of McFadden (1973, 1975, 1976); is to define random variable $Y(i, 2)$ in the following way:

$$Y(i,2) = \bar{V} + \gamma d(i,1) + \varepsilon(i,2).$$

If $Y(i, 2) \geq 0$, $d(i, 2) = 1$. If $Y(i, 2) < 0$, $d(i, 2) = 0$. $E(\varepsilon(i, 2)) = 0$, $E(\varepsilon(i, 2)^2) = \sigma(2, 2)$. $E(\varepsilon(i, 1)\varepsilon(i, 2)) = \sigma(1, 2)$, and $\rho = \sigma(1, 2)/[\sigma(1, 1)\sigma(2, 2)]^{1/2}$. In this specification the act of choosing $d(i, 1) = 1$ shifts up the mean utility function of the next period by an amount γ.

If $\gamma > 0$, or $\rho > 0$, or both, individuals who experience the event in the first period are more likely to experience the event in the second period. The ρ generates this effect because on average individuals with a high value of $\varepsilon(1)$ in the first period have a high value of $\varepsilon(2)$ in the second period. The γ in

the expression has this effect because of the shift in the choice set that arises from occupancy of the state in the past.

To see how $\rho > 0$ generates a conditional relationship between events, set $\gamma = 0$, and note that the conditional probability that a person experiences the event in the second period, given that he experiences the event in the first period, is

$$\text{Prob}[d(i,2) = 1 \mid d(i,1) = 1] = \frac{\int_{-\bar{V}}^{\infty} \int_{-\bar{V}}^{\infty} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1) d\varepsilon(2)}{\int_{-\infty}^{\infty} \int_{-\bar{V}}^{\infty} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1) d\varepsilon(2)},$$

where $f(\varepsilon(1), \varepsilon(2))$ is a bivariate density. If $\varepsilon(i, 1)$ and $\varepsilon(i, 2)$ are independent, so that $\rho = 0$,

$$\text{Prob}[d(i,2) = 1 \mid d(i,1) = 1] = \text{Prob}[d(i,2) = 1] = \Phi\left[\frac{\bar{V}}{\sigma(2,2)^{1/2}}\right].$$

Assuming normality, the conditional probability is a monotonically increasing function of ρ, so that the dependence grows with the value of ρ. If $\rho = 1$, individuals who experience the event in period one are certain to experience the event in period two. Even if the correlation is not perfect, the information that an individual has experienced the event at $t = 1$ conveys information about his likelihood of experiencing the event at $t = 2$.

If $\rho > 0$, $d(i, 1)$ and $\varepsilon(i, 2)$ are positively correlated, so that a simple probit model applied to the second period data would lead to upward biased estimates of γ. To estimate γ consistently, and to test for true state dependence, one must control for the effect of correlated disturbances.

The data at the analyst's disposal can be summarized in the following contingency table. Sample proportions are entered in each cell. I is assumed to be sufficiently large that sample proportions closely approximate population probabilities.

	$d(2) = 1$	$d(2) = 0$
$d(1) = 1$	P_{11}	P_{10}
$d(1) = 0$	P_{01}	P_{00}

The probability of the four events in the general case is

$$P_{11} = \text{Prob}[d(i,1) = 1 \wedge d(i,2) = 1]$$
$$= \int_{-\bar{V}-\gamma}^{\infty} \int_{-\bar{V}}^{\infty} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1) d\varepsilon(2),$$

$$P_{10} = \text{Prob}[d(i,1) = 1 \wedge d(i,2) = 0]$$
$$= \int_{-\infty}^{-\bar{V}-\gamma} \int_{-\bar{V}}^{\infty} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1) d\varepsilon(2),$$

$$P_{01} = \text{Prob}[d(i,1) = 0 \wedge d(i,2) = 1]$$
$$= \int_{-\bar{V}}^{\infty} \int_{-\infty}^{-\bar{V}} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1) d\varepsilon(2),$$

$$P_{00} = \text{Prob}[d(i,1) = 0 \wedge d(i,2) = 0]$$
$$= \int_{-\infty}^{-\bar{V}} \int_{-\infty}^{-\bar{V}} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1) d\varepsilon(2).$$

Assuming $\sigma(1, 1) = \sigma(2, 2) = 1$, and $|\rho| < 1$, one can utilize the three independent cells of data from the contingency table to estimate the parameters $\bar{V}$, γ, and ρ, by either the method of maximum likelihood or minimum chi-square. The restriction on ρ is necessary in order to get observations in both off diagonal cells. (Recall that $|\rho| = 1$ induces either a perfect positive or perfect negative correlation in status over time, and so in general results in empty cells and lack of identification for parameters of the model.)

To see how the method works, note that the probability density of $\varepsilon(2)$ given $d(i, 1) = 1$ is

$$g(\varepsilon(2) \mid d(i,1) = 1) = \frac{\int_{-\bar{V}}^{\infty} f(\varepsilon(1), \varepsilon(2)) d\varepsilon(1)}{\int_{-\bar{V}}^{\infty} f_1(\varepsilon(1)) d\varepsilon(1)},$$

where $f_1(\varepsilon(1))$ is the marginal density of $\varepsilon(1)$. The probability of the event $d(i, 2) = 1$ given $d(i, 1) = 1$ is generated by

$$\text{Prob}[d(i,2) = 1 \mid d(i,1) = 1] = \int_{-\bar{V}-\gamma}^{\infty} g(\varepsilon(2) \mid d(i,1) = 1)d\varepsilon(2).$$

Evaluating the probability of the event that $d(i, 2) = 1$ with respect to the conditional distribution of $\varepsilon(2)$ given $d(i, 1) = 1$ avoids the spurious correlation between $d(i, 1)$ and $\varepsilon(1)$ that arises from correlation between $\varepsilon(1)$ and $\varepsilon(2)$: this procedure "controls" for the sample selection bias that causes the mean disturbance (and general distribution) of $\varepsilon(2)$ to be different for people who have experienced different period one events.

In estimating the parameters $\bar{V}$, ρ, and γ, it is desirable to utilize all available information to secure efficient estimators. The contribution to sample likelihood of an observation with $d(i, 1) = 1$ and $d(i, 2) = 1$ is

$$\text{Prob}[d(i,2) = 1 \mid d(i,1) = 1]\text{Prob}[d(i,1) = 1] = P_{11}.$$

A similar argument for other sequences of events justifies the other cell probabilities. By correctly conditioning the period two distribution, the sample likelihood "controls" for spurious correlation running from $\varepsilon(i, 1)$ to $d(i, 2)$ via $\varepsilon(i, 2)$.

The source of the identification of the parameter γ comes from the following insight: from the outcomes of the choice process in the first period it is possible to estimate $\bar{V}$. Given $\bar{V}$, and hence $P_{1.}(= 1 - P_{0.})$ the probability that the event is experienced in time period one, it is possible to use the conditional probabilities that individuals in state zero in time period one transit to states one and zero in time period two ($P_{01}/P_{0.}$ and $P_{00}/P_{0.}$, respectively) to estimate ρ. Given ρ and $\bar{V}$, it is possible to estimate γ from the transit proportions from state one in period one from $P_{11}/P_{1.}$ and $P_{10}/P_{1.}$. If there is no true state dependence, $P_{01} = P_{10}$, and the proportion of the population in state one is the same in period one and period two, since the same proportion of the population leaves state zero as enters it in period two. Starting from arbitrary initial conditions, the process is always in equilibrium if $\gamma = 0$.[34]

34. Of course, if the process were started in equilibrium, and $\gamma \neq 0$, $P_{01} = P_{10}$. This case requires a different example and has been ruled out here by the assumption that presample values of $d(t')$, $t' \leq 0$, are fixed nonstochastic constants. As noted in chapter 4, first-period equilibrium probabilities are not probit probabilities. One does not require disequilibrium to identify γ.

Another way to show how an estimate of γ is secured in this example is to consider the regions of integration for the density $f(\varepsilon(1), \varepsilon(2))$ used to define the probabilities P_{01} and P_{10}. Figure 3.1 corresponds to the case of $\gamma = 0$. The area under the density in region DBC yields P_{01}. The area under the density in region $D'BC'$ yields P_{10}. Under the assumption that the variance of $\varepsilon(1)$ is the same as that of $\varepsilon(2)$, an assumption consistent with the assumption of underlying stationarity in the distribution of the latent variables, B lies on a 45° line from the origin, and $P_{01} = P_{10}$.

Next consider the case in which $\gamma > 0$, figure 3.2. The appropriate regions of integration are DBC (for P_{01}), which is the same as in the previous diagram, and $C'B'D''$ (for P_{10}), which has a smaller area than $D'BC'$ in figure 3.1. The reduction in area is given by the strip $DBB'D''$. Accordingly $P_{01} > P_{10}$. (Clearly if $\gamma < 0$, $P_{01} < P_{10}$.)

At the heart of the definition of true state dependence in this chapter is the nonlinear shift term $\gamma d(i, 1)$ which captures the notion that occupancy of a state affects the subsequent choice set. The distinction between true and spurious state dependence rests on the distinction between the association that arises from correlation between $\varepsilon(1)$ and $\varepsilon(2)$, giving rise to spurious state dependence and the association ($\gamma d(i, 1)$) between the event in the preceding period and utility levels in the current period. Note that in the example just given, if $\rho = 1$, it is not possible to estimate γ. There is no innovation in $\varepsilon(2)$ that permits one to identify γ. The outcome in the first period perfectly predicts the outcome in the second period whether or not $\gamma = 0$.

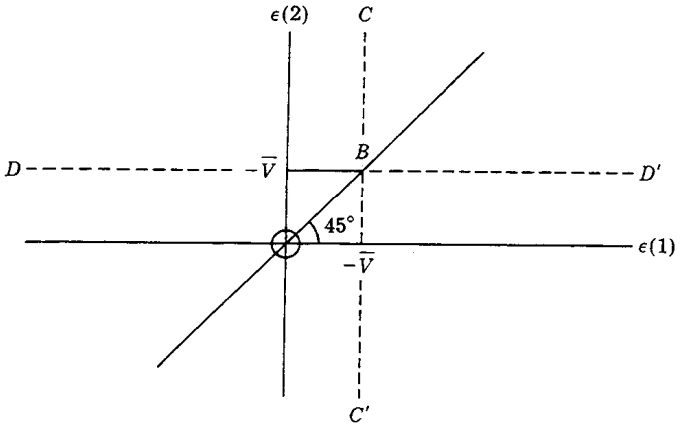

Figure 3.1
$\gamma = 0$

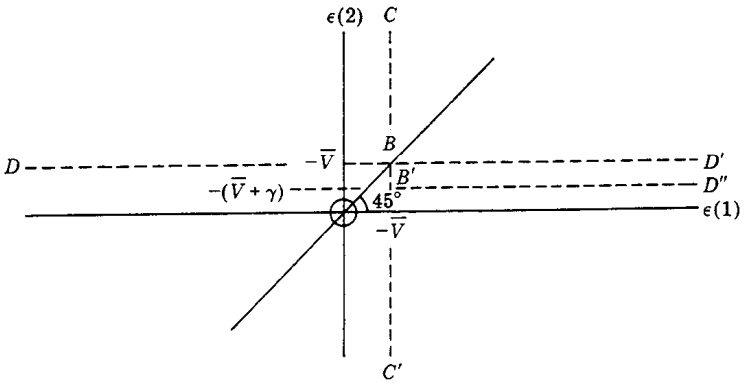

Figure 3.2
$\gamma > 0$

This example illustrates how the techniques developed in sections 3.1 through 3.12 can be used to address an important substantive problem. However, since the example is somewhat special, it is useful to separate the essential from the inessential assumptions that underlie it.

The assumptions in the example are (1) ρ is less than one in absolute value, (2) only two periods of data are available for each person, (3) the variance in $\varepsilon(2)$ is the same as that of $\varepsilon(1)$, an assumption of stationarity of the distribution of the disturbances, (4) everyone is observationally identical in terms of exogenous variables, and (5) the initial conditions of the process are fixed, nonstochastic constants, and the same for everyone.

The assumption that ρ is less than one in absolute value is essential. Without it the contingency table has empty cells, and state dependence parameters cannot be estimated. The restriction to two periods of data is made solely for convenience. If three periods of data are available, one has access to seven independent pieces of information, and a less restrictive model can be fit. It is straightforward to show that, if there are no empty cells, one can estimate the variances $\sigma(2, 2)$, $\sigma(3, 3)$ ($\sigma(1, 1) = 1$ is a required normalization), the correlation coefficients, $\tilde{\sigma}(1, 2)$, $\tilde{\sigma}(1, 3)$, $\tilde{\sigma}(2, 3)$, and γ from the seven cells of data.[35] With four periods of data one has fifteen independent cells that can be used to estimate six correlation coefficients, three variances (setting $\sigma(1, 1) = 1$), $\bar{V}$, and γ. In the four-period case more general forms of state dependence may be entertained (e.g., a fourth-order Markov process).

35. This statement and the following assume that $\beta_0 \neq 0$.

Thus, if $T > 3$, and there are no empty cells, one can separate out the effect of nonstationarity in the error process from state dependence so that the stationarity assumption invoked in the example is not essential.

The assumption that everyone is observationally identical with respect to the exogenous variables can be relaxed, and with profit. A linear combination of exogenous regressor variables, $\mathbf{Z}(i, t)\boldsymbol{\beta}$, may be substituted in place of $\bar{V}$. Assuming that the regressor matrix is of full rank, the addition of these variables permits identification of $\sigma(2, 2)$ even if $T = 2$.[36]

The assumption that initial conditions of the process are fixed and nonstochastic is essential and difficult to relax. Discussion of the important problem of initial conditions is deferred to chapter 4.

The entire discussion in this section has been conducted within the convenient framework of the normal distribution for the disturbances of the model. A parallel analysis could be performed within the general multivariate t family or for more general distributions. For example, first-order Markov state dependence could be generated in a logit model with a components of variance structure.

A complete analysis of the general, non-normal case is beyond the scope of this chapter. The normal framework is sufficiently flexible to accommodate behavior consistent with the urn schemes discussed and so is useful for testing among competing specifications. However, the concept of structural state dependence does not require the normal framework for its definition, although such a framework is convenient for measuring its effect.

The normality assumption is convenient primarily because the normal distribution can be parameterized to accommodate nonstationarity in the distribution of the disturbances in such a way that the nonstationarity can be removed or accounted for (e.g., one can estimate $\sigma(t, t)$ or introduce time trends as exogenous variables).

In the general case of a non-normal arbitrarily nonstationary distribution with unknown parameters, the measurement of state dependence effects will be a hopeless task.[37] In general for any contingency table with

36. Thus identification conditions in this model are analogous to identification conditions in a time-series model with first-order serial correlation, and a lagged value of the dependent variable. Identification of correlation and lag coefficients in that model is secured through sample variation in the exogenous variables. The proposition in the text follows from Heckman (1978a, part III).

37. This problem closely resembles the equally hopeless task of estimating parameters of distributed lag models in the presence of arbitrary serial correlation in the errors without the benefit of any a priori information (Hatanaka 1975).

choice process, any distribution with $2^T - 1$ or more parameters will in general fit the table. The methods proposed here secure identification of the state dependence effect by restricting the nonstationarity effect to operate through shifts in covariances and means of the distribution of the errors generating the model.

3.15 Analogies with Time-Series Models[38]

The analogy between the problem of distinguishing heterogeneity from state dependence and the classical time-series problem of distinguishing a serial correlation model from a distributed lag model, although superficially appealing, is not precise. As noted in section 3.10, an exact analogy can be made between the problem of distinguishing heterogeneity from habit persistence and the classical time-series problem.

A model with habit persistence is

$$Y(i,t) = G(L)\,Y(i,t) + \varepsilon(i,t),$$

$G(0) = 0$, $Y(i, t) \geq 0$ iff $d(i, t) = 1$. $Y(i, t) < 0$ otherwise. This model is exactly in the form of the classical time-series problem, except in that problem $Y(i, t)$ is observed. As noted in section 3.10, if regressors are present, it is possible to distinguish between the effects of habit persistence and serial correlation. Thus let

$$Y(i,t) = \mathbf{Z}(i,t)\boldsymbol{\beta} + G(L)\,Y(i,t) + \varepsilon(i,t).$$

Provided that the regressors in different periods for individual i ($i = 1, \ldots, I$) are not linear combinations of each other, one can compute the marginal probability that $d(i, t) = 1$ and determine if past values of $\mathbf{Z}(i, t)$ are determinants of current period choices. If they are, one can reject the hypothesis of no habit persistence. This is so because

$$Y(i,t) = [1 - G(L)]^{-1}\mathbf{Z}(i,t)\boldsymbol{\beta} + [1 - G(L)]^{-1}\varepsilon(i,t).$$

Only if $G(L) \equiv 0$ for all L will lagged $\mathbf{Z}$ not determine the current marginal probability that $d(i, t) = 1$. In principle one can approximate the marginal probability by a linear probability model (e.g., see Heckman 1978b) so that this test does not require a normality assumption for $\varepsilon(i, t)$.

38. This section has benefited from discussions with Zvi Griliches and Tom MaCurdy, and the incisive remarks of Marc Nerlove (1978).

A model with structural state dependence may be written as a nonlinear time series. For example, consider

$$Y(i,t) = \sum_{j=1}^{\infty} \gamma(j) d(i, t-j) + \varepsilon(i,t),$$

with $Y(i, t) \geq 0$ iff $d(i, t) = 1$, $Y(i, t) < 0$ otherwise. Unlike the autoregressive habit persistence model, an effect of past Y on current Y arises only if a threshold is crossed. In a model of discrete choice in which $Y(i, t) \geq 0$ corresponds to occupancy of a different state than that occupied when $Y(i, t) < 0$, this sort of nonlinearity is natural, although in an ordinary time-series model it may appear to be artificial.

Assuming no habit persistence, a test of state dependence against heterogeneity can be based on the marginal probability that $d(i, t) = 1$, provided that regressors are available that satisfy the conditions given in the test for habit persistence against serial correlation. If lagged $\mathbf{Z}(i, t)$ determine current marginal choice probabilities, state dependence is present.[39]

In the general case with habit persistence and state dependence, the finding that lagged values of $\mathbf{Z}(i, t)$ determine current marginal choice probabilities suggests that habit persistence or state dependence, or both, are present, except in the unusual case where the two effects cancel. The finding that lagged $\mathbf{Z}(i, t)$ does not determine current marginal choice probabilities is, except for the unusual case just stated, evidence against both habit persistence and state dependence.

The identification of habit persistence effects requires exogenous variables.[40] The identification of state dependence effects does not, as the example of the preceding section has shown. For this reason the analogy between the problem of distinguishing between heterogeneity and state dependence and the problem of distinguishing between a distributed lag and serial correlation model is inexact.

39. This test for state dependence against heterogeneity presented in this paragraph was suggested to me by Gary Chamberlain and Tom MaCurdy.

3.16 Examples of Models that Generate Structural State Dependence

This section briefly considers how models with structural state dependence can be generated from well-defined economic models. Three examples are discussed: a model of stimulus-response conditioning of the sort developed by mathematical psychologists, a model of decision making under uncertainty, and a model of decision making under perfect foresight.

In the stimulus-response model developed by behavioral psychologists (e.g., see Bush and Mosteller 1955, Restle and Greeno 1970, or Johnson and Kotz 1977) the individual who makes a given "correct" response is rewarded so that he is more likely to make the response in the future. Decision making is myopic. This model closely resembles the generalized Pólya process discussed in sections 3.8 and 3.13. General heterogeneity can be introduced into the model along the lines discussed in sections 3.3 through 3.12. Models that resemble the stimulus-response model have been proposed by dual labor market economists who assume that individuals who are randomly allocated to one market are rewarded for staying in the market and are conditioned by institutions in that market so that their preferences are altered. The more time one has spent in a particular type of market, the more likely one is to stay in it.

The model of myopic sequential decision making just presented is unlikely to prove attractive to many economists. Nonmyopic sequential models of decision making under imperfect information also generate structural state dependence. Such models have been extensively developed in the literature on dynamic programming (e.g., see Dreyfus 1965, pp. 213–215, or Astrom 1970). An example is a model in which an agent at time t maximizes expected utility over the remaining horizon, given all the information at his disposal and his constraints as of time t. Transition to a state may be uncertain. As a consequence of being in a state, costs may be incurred or information may be acquired that alters the information set or opportunity set, or both, relevant for future decisions. In such cases the outcome of the process affects subsequent decision making, and structural state dependence is generated.

The disturbance in this model consists of unmeasured variables known to the agent but unknown to the observing economist as well as

40. See Hatanaka (1975). Restrictions on error covariances and/or admissible habit persistence effects can also secure identification of these effects.

unanticipated random components unknown to both the agent and the observing economist.

Structural state dependence can also be generated as *one representation* of a model of decision making under perfect certainty. In such a model there are no surprises. Given the inital conditions of the process, the full outcome of the process is perfectly predictable from information available to the agent (but not necessarily available to the observing economist).

Consider the following three-period model of consumer decision making under perfect certainty with indivisibility in purchase quantities: a consumer's strictly concave utility function is specified as

$$U(a(1)d(1), a(2)d(2), a(3)d(3)),$$

where the $a(i)$ are the fixed amounts that can be consumed in each period. The consumer purchases amount $a(i)$ if $d(i) = 1$, otherwise $d(i) = 0$. Resources are fixed so that

$$\sum a(i)d(i) \leq M.$$

The agent has full information and selects the $d(i)$ optimally. Optimal solutions are denoted by $d^*(i)$.

An alternative characterization of the problem is the following sequential interpretation. Given $d^*(1)$, maximize utility with respect to remaining choices.

Thus

$$\max_{d(2),d(3)} U(a(1)d^*(1), a(2)d(2), a(3)d(3))$$

subject to

$$\sum_{i=2}^{3} a(i)d(i) \leq M - a(1)d^*(1).$$

The demand functions (really the demand inequalities) for $d(2)$ and $d(3)$ may be written in terms of $d^*(1)$ and available resources $(M - a(1)d^*(1))$. This characterization is a discrete choice analogue of the Hotelling (1935), Samuelson (1960), Pollak (1969) treatment of ordinary consumer choice and demonstrates that the demand function for a good can be expressed as a function of quantities consumed of some goods, the "prices" of the remaining goods and income. (Pollak's term "conditional demand function" is felicitous.)

Either past choices $d^*(1)$ or past $a(1)$ determine current choices in conjunction with future prices and current resources.[41] The choice of which characterization of the decision problem to use is a matter of convenience. When the analyst knows current disposable resources $(M - a(1)d^*(1))$ and past choices $(d^*(1))$ but not $a(1)$ or M, the second form of the problem is econometrically more convenient. The conditional demand function gives rise to structural state dependence, in the sense that past choices influence current decisions. The essential point in this example is that past choices serve as a legitimate proxy for missing M and $a(1)$ variables known to the consumer but unknown to the observing econometrician. The conditional demand function is a legitimate structural equation.[42]

Both a model of decision making under uncertainty and a model of decision making under perfect foresight may be brought into sequential form so that past outcomes of the choice process may determine future outcomes. In principle one can distinguish between a certainty model and an uncertainty model if one has access to all the relevant information at the agent's disposal. In a model of decision making under perfect certainty, if all past prices are known and entered as explanatory variables for current choice, past outcomes of the choice process contribute no new information relevant to determining current choices. In a model of decision making under uncertainty, past outcomes would contribute information on current choices not available from past prices, since uncertainty necessarily makes the prediction of past outcomes from past prices inexact, and the unanticipated components of past outcomes alter the budget set and cause a revision of initial plans.[43] In practice it is difficult to distinguish between the two models given limitations of data. The observing economist usually has less information at his disposal than the agent being analyzed has at his disposal when he makes his decisions.

41. In this example, if the utility function is additive, $d^*(1)$ would have no effect on future choices except through its effect on current resources $(M - a(1)d^*(1))$. Thus a test of structural state dependence in this model is a test of intertemporal independence in preferences.

42. Another model that generates structural state dependence in an environment of perfect certainty is a model with fixed costs. In some dynamic models of labor supply, training costs are assumed to be incurred by labor force entrants. Once these costs are incurred, they are not incurred again until re-entry occurs. Labor force participation decisions taken by labor force participants take account of such costs. In this way structural state dependence is generated.

43. If the uncertainty comes in the form of price uncertainty, *ex ante* prices are required to perform the test.

The key point to extract from these examples is that structural state dependence as defined in this chapter may be generated from a variety of models. It is not necessary to assume myopic decision making to generate structural dependence. Nor does empirical evidence in support of structural state dependence prove that agents make their decisions myopically.

3.17 Summary and Conclusion

This chapter presents a general model for the analysis of discrete panel data. The model is sufficiently flexible that it can be used to generate a variety of models useful in applied work as special cases of the general model. Bernoulli, Pólya, Markov, and renewal models are produced by imposing restrictions on the general model. Time-varying exogenous variables and unobserved variables with a general serial correlation structure can be introduced into the model. The definition of heterogeneity used in previous work is generalized in this model.

Special cases of the model likely to be used in applied work are considered in detail. Issues of identification and data requirements needed to estimate these models are addressed. Simply computed versions of the models receive considerable attention. A fixed effect probit model and one-factor probit model are presented, and their strengths and limitations are evaluated. Inexpensive methods for estimating the general model are discussed.

A great advantage of the models developed in this chapter is that they can be used to generate choice theoretic discrete data models. The apparatus developed here extends the atemporal choice models of McFadden (1974, 1976) to an intertemporal setting. Structural dependence among time-ordered discrete events can be investigated by the models. Certain models widely used in the analysis of discrete data, such as Goodman's log linear probability model, defy structural interpretation and so are not useful for the analysis of structural discrete data models (Heckman 1978a, part IV, pp. 950–954).

The methodology developed here can be put to use to address a longstanding statistical problem: distinguishing between true and false contagion (Bates and Neyman 1951) or, in the language of this chapter, distinguishing between spurious and true state dependence. The problem can be stated simply. The existence of a conditional probability relationship between the occurrence of an event in one period and its

occurrence in previous periods may be due to serial correlation in the unobservables that generate the event or because past experience of the event affects the choice set and preferences relevant to choices taken in subsequent periods. The first reason for the existence of the conditional relationship is termed spurious state dependence. The second reason for the conditional relationship is termed structural state dependence. Methods for estimating structural relationships among time-ordered outcomes can be used to test for the presence of true state dependence and to measure the quantitative significance of the two sources of dependence.

Intuitively appealing analogies between this problem and the classical time-series problem of distinguishing between a distributed lag model and a serial correlation model are examined and are found to be somewhat misleading. Examples of choice theoretic models that generate structural state dependence are presented. It is demonstrated that it is possible to produce structural state dependence as one representation of a model of consumer decision making under perfect certainty.

Empirical work based on these models has been performed in other work. In Heckman (1981) data on the labor force participation of women are analyzed. A one-factor model is fit to three periods of panel data drawn from the Michigan Panel Survey of Income Dynamics. Tests for the existence of heterogeneity and state dependence are conducted. The major findings of the empirical analysis of female labor force activity are (1) Heterogeneity is not characterized by a components of variance scheme; a first-order Markov process for the unobserved variables fits the data better. (2) There is evidence of true state dependence. For an application of these models to the analysis of unemployment data, see Cave (1981). Heckman and Willis (1975) apply a simple version of these models to analyze dynamic fertility behavior.

3.18 Appendix: Factor Analytic Probit Models

Four topics are addressed: first, a more general treatment of the one-factor model is given. Equation 3.19 is derived, and the requirement that $\sigma_U(t, t) > 0$, $t = 1, \ldots, T$, is relaxed. Second, implicit restrictions inherent in the one-factor scheme are presented. It is demonstrated that in the general case the one-factor model implies that the disturbances generating the stochastic process are nonstationary. Third, specific examples of one-factor representations are given. Fourth, a multifactor model is considered.

A General One-Factor Model

The reader is referred to the text, especially the discussion following equation (3.17). The error structure is written as

$$\varepsilon(i,t) = \alpha(t)\tau(i) + U(i,t), \tag{3.28}$$

$t = 1, \ldots, T, i = 1, \ldots, I$. A key concept is the term $\tilde{\alpha}(t)$, the normalized factor loading,

$$\tilde{\alpha}(t) \equiv \frac{\alpha(t)\sigma_\tau}{[\alpha^2(t)\sigma_\tau^2 + \sigma_U(t,t)]^{1/2}}.$$

In this notation the following proposition can be verified:

Proposition 3.1: Given the one-factor structure, and the assumption $\sigma_U(t, t) > 0$, the probability of $\mathbf{d}(i)$ given $\mathbf{Z}(i)$ may be written as

$$\text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i)]$$

$$= \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\{[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \eta(t)\tilde{\tau}][2d(i,t) - 1]\} f(\tilde{\tau})\, d\tilde{\tau}, \tag{3.29}$$

where $\eta(t) = [\tilde{\alpha}(t)^2/1 - \tilde{\alpha}(t)^2]^{1/2}$, $\tilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}/\sigma_U(t, t)^{1/2}$, and $\tilde{\tau}$ is a standardized variate with variance one. (Positive square roots are to be used.)

PROOF: The probability that $d(i, t) = 1$ given $\tau(i)$ and $\mathbf{Z}(i, t)$ is

$$\text{Prob}[U(i,t) \geq -\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) - \alpha(t)\tau(i) \mid \tau(i), \mathbf{Z}(i,t)]$$
$$= \Phi[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \frac{\alpha(t)}{\sigma_U(t,t)^{1/2}}\tau(i)],$$

where $\tilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}/\sigma_U(t, t)^{1/2}$.

Define $\tilde{\tau}(i) = \tau(i)/\sigma_\tau$, and note that

$$\eta(t) = \left[\frac{(\tilde{\alpha}(t))^2}{(1 - (\tilde{\alpha}(t)^2)}\right]^{1/2} = \frac{\alpha(t)\sigma_\tau}{\sigma_U(t,t)^{1/2}}.$$

Thus the probability that $d(i, t) = 1$ given $\tilde{\tau}(i)$ and $\mathbf{Z}(i, t)$ is

$$\Phi[\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \eta(t)\tilde{\tau}(i)].$$

Expressing the model in general form, and integrating with respect to the density of $\tilde{\tau}$, the result follows immediately. Q.E.D.

Note that, if $\sigma_U(t', t') = 0$, so that $\tilde{\alpha}(t') = 1$, the probability that $d(i, t') = 1$ given $\mathbf{Z}(i, t')$ and $\tau(i)$ is either zero or one. Thus $d(i, t') = 1$ imposes the condition that

$$\alpha(t')\tau(i) \geq -\mathbf{Z}(i, t')\boldsymbol{\beta},$$

which is a restriction on $\tilde{\tau}(i)$, with

$$\tilde{\tau}(i) \geq -\frac{\mathbf{Z}(i, t')\boldsymbol{\beta}}{\alpha(t')\sigma_\tau}, \quad \alpha(t') > 0.$$

If $\sigma_U(t', t') = 0, \sigma_U(t, t) > 0, t \neq t'$, the probability of $\mathbf{d}(i)$, the vector of the $d(i, t)$, given $\mathbf{Z}(i)$, the vector of $\mathbf{Z}(i, t)$, *is*

$$\text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i)]$$

$$= \left[\int_{-\mathbf{Z}(i,t')\boldsymbol{\beta}/\alpha(t')\sigma_\tau}^{\infty} \prod_{\substack{t=1 \\ t\neq t'}}^{T} \Phi\{(\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \eta(t)\tilde{\tau})(2d(i,t) - 1)\} f(\tilde{\tau})\, d\tilde{\tau} \right]^{d(i,t')}$$

$$\left[\int_{-\infty}^{-\mathbf{Z}(i,t')\boldsymbol{\beta}/\alpha(t')\sigma_\tau} \prod_{\substack{t=1 \\ t\neq t'}}^{T} \Phi\{(\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) + \eta(t)\tilde{\tau})(2d(i,t) - 1)\} f(\tilde{\tau})\, d\tilde{\tau} \right]^{(1-d(i,t'))}$$

Under general conditions for $T \geq 3$, the $\tilde{\boldsymbol{\beta}}(t)$, $\eta(t)$, $t = 1, \ldots, T$, $t \neq t'$, and $\boldsymbol{\beta}/\alpha(t')\sigma_\tau$ can be estimated by the method of maximum likelihood. Normalizing $\sigma_U(1, 1) = 1$ (so that $t' \neq 1$), it is thus possible to estimate $\boldsymbol{\beta}$, $\alpha(t)\sigma_\tau$, $t = 1, \ldots, T$, and $\sigma_U(t, t)$, $t = 2, \ldots, T$. Other normalizations are possible (e.g., $\alpha(t)\sigma_\tau = 1$).[44]

If there are two zero variances ($\sigma_U(t', t') = 0$ and $\sigma_U(t'', t'') = 0$), two restrictions on the single factor are generated, and a special dependency between the $d(i, t')$ and $d(i, t'')$ is implied. For example, suppose

44. It is interesting to note that the extra information $\sigma_U(t', t') = 0$ does not permit any more parameters to be identified than if $\sigma_U(t', t') > 0$. This is because, if $\sigma_U(t', t') = 0$, $\tilde{\alpha}(t')$ is equal to one and is no longer a source of information on the parameters of the model.

$$\tilde{\tau}(i) \geq -\frac{\mathbf{Z}(i,t')\boldsymbol{\beta}}{\alpha(t')\sigma_\tau} = k(t'),$$

$$\tilde{\tau}(i) \geq -\frac{\mathbf{Z}(i,t'')\boldsymbol{\beta}}{\alpha(t'')\sigma_\tau} = k(t''), \quad \alpha(t'), \quad \alpha(t'') > 0,$$

where $k(t') < k(t'')$. Then, if $d(i,\ t'') = 1$, $d(i,\ t') = 1$. If $d(i,\ t') = 0$, $d(i, t'') = 0$. The only possibilities are $(d(i, t'') = 1, d(i, t') = 1; d(i, t') = 1$, $d(i, t'') = 0; d(i, t') = 0, d(i, t'') = 0)$. The outcome $d(i, t') = 0, d(i, t'') = 1$ is ruled out. The probability statement (3.19) must be modified in a nontrivial way to incorporate such possibilities. Such a modification, and also the modification required for more than two zero variances, are topics left for another occasion.

The General One-Factor Model Imposes Nonstationarity

The one-factor model imposes restrictions on the admissible error process. For $T > 3$, it implies nonstationarity. Only the permanent-transitory error process and a peculiar relative are stationary and one-factor analyzable.

Proposition 3.2: In the general case of weak stationarity of the process, if $T > 3$, no one-factor model is stationary except the process

$$\varepsilon(i,k) = b\tau(i) + U(i,k), \tag{3.30}$$

$k = 2t - 1$, $t = 1, \ldots,$ $[T/2,\ T$ even; $(T + 1)/2,\ T$ odd$]$;

$$\varepsilon(i,k) = -b\tau(i) + U(i,k),$$

$k = 2t$, $t = 1, \ldots,$ $[T/2,\ T$ even; $(T - 1)/2,\ T$ odd$]$; or the process

$$\varepsilon(i,t) = \tau(i) + U(i,t), \tag{3.31}$$

$t = 1, \ldots, T$, where $E(\tau) = 0 = E(U(i,t)), E(\tau^2) = \sigma_\tau^2, E(U(i,t)^2) = \sigma_U^2$, and $E(U(i,\ t)\tau(i)) = 0$.

PROOF: For a weakly stationary sequence of random variables, $\varepsilon(i,\ t)$, $t = 1, \ldots,\ T$, the autocovariances $(\tilde{\sigma}(t,\ t'))$ must satisfy $\tilde{\sigma}(t,\ t') = \tilde{\sigma}(|\ t - t'\ |)$ and $E(\varepsilon(i,\ t)^2) = \sigma_\varepsilon^2$ for all t, t'.

For a stationary process to be one-factor analyzable, it must be the case that

$$\tilde{\sigma}(j,j+1) = \tilde{\alpha}(j)\tilde{\alpha}(j+1) = \tilde{\alpha}(j+1)\tilde{\alpha}(j+2) = \tilde{\sigma}(j+1,j+2)$$

for $j = 1, \ldots, T-2$. Thus $\tilde{\alpha}(j) = \tilde{\alpha}(j+2), j = 1, \ldots, T-2$, so that all even- and odd-normalized factor loadings are equal, but the odd-numbered loadings need not equal the even-numbered loadings.

Stationarity also implies

$$\tilde{\sigma}(j,j+2) = \tilde{\alpha}(j)\tilde{\alpha}(j+2) = \tilde{\alpha}(j+1)\tilde{\alpha}(j+3) = \tilde{\sigma}(j+1,j+3)$$

for $T > 3$. Since $\tilde{\alpha}(j) = \tilde{\alpha}(j+2)$, it must be the case that $\tilde{\alpha}(j) = \pm\tilde{\alpha}(j+1)$. Thus either all normalized factor loadings are equal, or the odd-numbered normalized factor loadings are minus the even-numbered normalized factor loadings.

Stationarity also requires that the variances in each period be the same or

$$\alpha(t)^2\sigma_\tau^2 + \sigma_U(t,t) = \alpha(t')^2\sigma_\tau^2 + \sigma_U(t',t').$$

Since by the previous argument $\tilde{\alpha}(t)^2 = (\tilde{\alpha})^2, \alpha(t)\sigma_\tau = k$ from the definition of $\tilde{\alpha}(t)$. Hence $\sigma_U(t, t) = \sigma_U$ for all t.

Therefore the error process must be either the ordinary permanent-transitory process, given by equation (3.31) or the alternating permanent-transitory process, given in equation (3.30). Q.E.D.

This result is discouraging. Many interesting error structures cannot be one-factor analyzed. Note, however, if $T = 3$, the proposition is not true.

Some Examples

This subsection considers some examples of error structures that can be one-factor analyzed. The first example is a stationary first-order Markov process for $T = 3$. This process can be one-factor analyzed and provides an interesting case in which $\sigma_U(t, t) = 0$.

Thus $E(\varepsilon(i, t)\varepsilon(i, t')) = \rho^{|t-t'|}\sigma_\varepsilon^2$, $E(\varepsilon(i, t)^2) = \sigma_\varepsilon^2$, $\tilde{\alpha}(1) = \tilde{\alpha}(3) = \rho = [E(\varepsilon(i, 1)\varepsilon(i, 2))]/E(\varepsilon(i, 1))^2$, $\tilde{\alpha}(2) = 1$. The joint probability of $d(i, 1) = 1$, $d(i, 2) = 0$, and $d(i, 3) = 1$ is

$$\text{Prob}[d(i,1) = 1, d(i,2) = 0, d(i,3) = 1 \mid \mathbf{Z}(i)]$$

$$= \int_{-\infty}^{-\mathbf{Z}(i,2)\boldsymbol{\beta}} \Phi[\mathbf{Z}(i,1)\boldsymbol{\beta} + \tilde{\tau}\eta]\,\Phi[\mathbf{Z}(i,3)\boldsymbol{\beta} + \tilde{\tau}\eta]\, f(\tilde{\tau})\, d\tilde{\tau}$$

where $\eta = (\rho/1-\rho)^{1/2}$.

As another example, consider the scheme of Balestra and Nerlove (1966):

$$\varepsilon(i,t) = \tau(i) + \frac{U(i,t)}{1-\rho L},$$

$t = 1, \ldots, 3$. For this model

$$\tilde{\alpha}(1) = \tilde{\alpha}(3) = \left(\frac{1+\rho^2 k}{1+k}\right)^{1/2},$$

$$\tilde{\alpha}(2) = \frac{1}{\tilde{\alpha}(1)}\left(\frac{1+\rho k}{1+k}\right),$$

where $k = \sigma_U^2/\sigma_\tau^2$.

It is easily verified that no first-order moving average scheme can be one-factor analyzed, but a first-order moving average scheme with a permanent component can be one-factor analyzed for the case $T = 3$.

A Model with Multiple Factors

The principal advantage of the one-factor model is that it is simple to compute. However, for $T > 3$, it imposes restrictions on the error process that may be inappropriate in certain applications. Higher-factor schemes are less restrictive and yet reduce the scale of the computing problem in comparison with the scale in the general case of an unrestricted correlation matrix. Assuming $Q < T$ independent factors, the probability integral can be written as Q univariate integrations of products of functions available on most computers. The general form of the Q-factor model is the topic of this subsection.

The disturbance $\varepsilon(i, t)$ can be Q factor analyzed if it can be written

$$\varepsilon(i,t) = \sum_{q=1}^{Q} \alpha(t,q)\tau(i,q) + U(i,t),$$

$t = 1, \ldots, T$, $i = 1, \ldots, I$, $Q < T$, where
$E(\tau(i, q)) = 0$,
$E(U(i, t)) = 0, \quad i = 1, \ldots, I, t = 1, \ldots, T, q = 1, \ldots, Q,$

$$E(\tau(i,j)\tau(i'',j')) = \sigma_{jj'}, \quad i = i'', j = j',$$
$$= 0, \quad i \neq i'' \text{ or } j \neq j',$$

$$E(U(i, t)\tau(i', j)) = 0 \quad \text{for all } i, t, i', \text{ and } j,$$
$$E(U(i, t)^2) = \sigma_U(t, t), \quad t = 1, \ldots, T.$$

By analogy with the one-factor case, define the normalized factor loading for factor q at time t as

$$\tilde{\alpha}(t, q) = \frac{\alpha(t, q)\sigma_{qq}}{\left[\sum_{q=1}^{Q} \alpha^2(t, q)\sigma_{qq} + \sigma_U(t, t)\right]^{1/2}}.$$

The square of $\tilde{\alpha}(t, q)$ is the proportion of the variance in $\varepsilon(i, t)$ that is due to factor q. Array the $\tilde{\alpha}(t, q)$ into a $1 \times Q$ vector $\tilde{\boldsymbol{\alpha}}(t)$. Then $\tilde{\boldsymbol{\alpha}}(t)\tilde{\boldsymbol{\alpha}}(t)'$ is the proportion of the variance in $\varepsilon(i, t)$ due to all Q factors.

In this notation it is straightforward to establish the following proposition:

Proposition 3.3: If the disturbances can be Q-factor analyzed, and if $\sigma_U(t, t) > 0$, $t = 1, \ldots, T$, the probability of $\mathbf{d}(i)$ given $\mathbf{Z}(i)$ may be written as

$$\text{Prob}[\mathbf{d}(i) \mid \mathbf{Z}(i)]$$
$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\left\{\left[\mathbf{Z}(i, t)\tilde{\boldsymbol{\beta}}(t) + \frac{\tilde{\boldsymbol{\alpha}}(t)\mathbf{l}}{(1 - \tilde{\boldsymbol{\alpha}}(t)\tilde{\boldsymbol{\alpha}}(t)')^{1/2}}\right](2d(i,t) - 1)\right\} f(\mathbf{l})d\mathbf{l},$$

where $\mathbf{l}$ is a $Q \times 1$ vector of independent standard normal variates, $f(\mathbf{l})$ is a product of Q standard normal densities, and $\tilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}/\sigma_U(t, t)^{1/2}$.

PROOF: The probability that $d(i, t) = 1$ given $\mathbf{Z}(i, t)$ and $\tau(i, q)$, $q = 1, \ldots, Q$, may be written as

$$\text{Prob}[d(i, t) = 1 \mid \mathbf{Z}(i), \tau(i, q), q = 1, \ldots, Q]$$
$$= \text{Prob}\left[U(i, t) \geq -\mathbf{Z}(i, t)\boldsymbol{\beta} - \sum_{q=1}^{Q} \alpha(t, q)\tau(i, q)\right]$$
$$= \text{Prob}\left[\frac{U(i,t)}{\sigma_U(t,t)^{1/2}} \geq -\mathbf{Z}(i,t)\tilde{\boldsymbol{\beta}}(t) - \sum_{q=1}^{Q} \frac{\tilde{\alpha}(t,q)l(i,q)}{(1 - \tilde{\boldsymbol{\alpha}}(t)\tilde{\boldsymbol{\alpha}}(t)')^{1/2}}\right]$$
$$= \Phi\left[\mathbf{Z}(i, t)\tilde{\boldsymbol{\beta}}(t) + \frac{\tilde{\boldsymbol{\alpha}}(t)\mathbf{l}(i)}{(1 - \tilde{\boldsymbol{\alpha}}(t)\tilde{\boldsymbol{\alpha}}(t)')^{1/2}}\right],$$

where $l(i, q) = \tau(i, q)/\sigma_{qq}^{1/2}$ and $\mathbf{l}(i)$ is a $Q \times 1$ vector of the $l(i, q)$.

Removing the conditioning on $\mathbf{l}(i)$, which in this problem is equivalent to integrating out the $\mathbf{l}(i)$ with respect to $f(\mathbf{l})$(the product of the Q independent standardized variates) and considering the probability of a given sequence of outcomes (a given value of $\mathbf{d}(i)$) leads to the expression given in proposition 3.3. The crucial point to note is that, given values of $\mathbf{l}(i)$, the random variables $d(i, t)$, $t = 1, \ldots, T$, are independent.

Note that it is not required that the components of $\mathbf{l}$ be normally distributed, nor is it necesssary for $U(i, t)$ to be normally distributed. In principle each component of $\mathbf{l}$ and $U(i, t)$ may have functionally different (independent) distributions. The expression in the proposition assumes that the $U(i, t)$ are distributed symmetrically around zero. Symmetry can be relaxed at the cost of only minor notational inconvenience.

For a fixed T any correlation matrix $\tilde{\boldsymbol{\Sigma}}$ can be factor analyzed, provided a sufficiently large number of factors are used. Utilizing the results of Anderson and Rubin (1956), it is straightforward to develop a likelihood ratio test for the appropriate number of factors in order to specify a parsimonious approximation to the true correlation matrix.

As in the one-factor case the restriction that $\sigma_U(t, t) > 0$ can be relaxed. The analysis of this case resembles that in the one-factor case, except that in the general case up to Q of the period specific variances may be zero before problems arise with regard to special dependence among outcomes of the sort discussed at the beginning of this appendix, in which the occurrence of one event may imply (with probability one) the occurrence of another event. Intuitively, if Q or fewer of the T disturbances have no period specific variance ($\sigma_U(t, t) = 0$), the events associated with those periods (the $d(i, t)$) are generated by (linear combinations of) the Q independent components. Hence in this case no special dependence among outcomes is created.[45]

A complete discussion of identification in the general Q-factor model is beyond the scope of this chapter. Identification conditions for the $\tilde{\boldsymbol{\alpha}}(t)$, $t = 1, \ldots, T$, follow from standard theorems on factor representations of correlation matrices (Anderson and Rubin 1956). Note further, assuming $\sigma_U(1, 1) = 1$, it is possible to estimate $\sigma_U(t, t)$, $t = 2, \ldots, T$, as long as $\boldsymbol{\beta} \neq 0$. This follows from the discussion in section 3.4.

45. This conclusion holds provided that no linear dependencies exist among the column vectors of normalized factor loadings for the periods with zero-period specific variances, $\{t \mid \sigma_U(t, t) = 0\}$, and provided that the system of Q equations generated by the column vectors is indecomposable. If the system of equations is decomposable, and there are no linear dependencies among the equations, the statement in the text must be altered in an obvious way.

References

Albright, R. L., S. R. Lerman, and C. F. Manski. 1977. Report on the Development of an Estimation Program for the Multinomial Probit Model. Carnegie-Mellon University, Pittsburgh, Pa. Prepared for Federal Housing Administration.

Amemiya, T. 1975. Qualitative Response Models. *Annals of Economic and Social Measurement*. 4: 363–372.

Amemiya, T. 1978. A Note on the Estimation of a Time Dependent Markov Chain Model. Stanford University, Stanford, Calif.

Andersen, E. B. 1973. *Conditional Inference and Models for Measuring*. Mentalhygiejnisk Forsknings Institut, Copenhagen.

Anderson, T. W., and L. Goodman. 1957. Statistical Inference about Markov Chains. *Annals of Mathematical Statistics*. 28: 89–110.

Anderson, T. W., and H. Rubin. 1956. Statistical Inference in Factor Analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. 5: 111–150.

Ashford, J. R., and R. R. Sowden. 1970. Multivariate Probit Analysis. *Biometrics*. 26: 535–546.

Astrom, K. J. 1970. *Introduction to Stochastic Control Theory*. New York: Academic Press.

Balestra, P., and M. Nerlove. 1966. Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model. *Econometrica*. 34: 585–612.

Bates, G., and J. Neyman. 1951. Contributions to the Theory of Accident Proneness II: True or False Contagion. *University of California Publications in Statistics*. 1: 215–253.

Blischke, W. R. 1965. Mixtures of Discrete Distributions. In *Classical and Contagious Discrete Distributions*, ed. G. Patil. Calcutta: Statistical Publishing Society.

Boskin, M., and F. Nold. 1975. A Markov Model of Turnover in Aid to Families with Dependent Children. *Journal of Human Resources*. 10: 467–481.

Bush, R., and F. Mosteller. 1955. *Stochastic Models for Learning*. New York: Wiley.

Cave, G. 1981. The Incidence and Duration of Youth Unemployment: Human Capital Theory and Longitudinal Analysis. Ph.D. dissertation. University of Chicago.

Chaddha, R. 1965. A Case of Contagion in Binomial Distribution. In *Classical and Contagious Discrete Distribution*, ed. G. Patil. Calcutta: Statistical Publishing Society.

Coleman, J. 1964. *Models of Change and Response Uncertainty*. Englewood Cliffs, N. J.: Prentice Hall.

Cripps, T. F., and R. J. Tarling. 1974. An Analysis of the Duration of Male Unemployment in Great Britain 1932–1973. *The Economic Journal* 84: 289–316.

David, F. N. 1947. A Power Function for Tests for Randomness in a Sequence of Alternatives. *Biometrika*. 34: 335–339.

Denny, J., and S. Yakowitz. 1978. Admissible Run-Contingency Type Tests of Independence and Markov Dependence. *Journal of the American Statistical Association*. 73: 171–181.

Domencich, T., and D. McFadden. 1975. *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North Holland.

Dreyfus, S. E. 1965. *Dynamic Programming and the Calculus of Variations*. New York: Academic Press.

Dutt, J. 1976. Numerical Aspects of Multivariate Normal Probabilities in Econometric Models. *Annals of Economic and Social Measurement*. 5: 547–562.

Fase, M. G. 1971. Estimation of Lifetime Income. *Journal of the American Statical Association*. 66: 686–693.

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications, vol 1*. New York: Wiley.

Feller, W. 1943. On a General Class of Contagious Distributions. *Annals of Mathematical Statistics*. 14: 389–400.

Goodman, L. 1961. Statistical Methods for the Mover-Stayer Model. *Journal of the American Statistical Association*. 56: 841–868.

Goodman, L. 1958. Simplified Runs Tests and Likelihood Ratio Tests for Markov Chains. *Biometrika*. 45: 181–197.

Granger, C. W. J., and P. Newbold. 1977. *Forecasting Economic Time Series*. New York: Academic Press.

Griliches, Z. 1967. Distributed Lags: A Survey. *Econometrica*. 35: 16–49.

Gupta, S. 1963. Probability Integrals of Multivariate Normal and Multivariate *t*. *Annals of Mathematical Statistics*. 34: 792–828.

Hatanaka, M. 1975. On the Global Identification of the Dynamic Simultaneous Equations Model with Stationary Disturbances. *International Economic Review*. 16: 545–554.

Heckman, J. 1976. Simultaneous Equations Models with Continuous and Discrete Endogenous Variables and Structural Shifts. *Studies in Nonlinear Estimation*, ed. S. Goldfeld and R. Quandt. Cambridge, Mass.: Ballinger.

Heckman, J. 1978a. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*. 46: 931–959.

Heckman, J. 1978b. Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence. *Anals de Insee*, Paris. 30-31: 227–270.

Heckman, J. 1981. Heterogeneity and State Dependence. In *Studies in Labor Markets*, ed. S. Rosen. Chicago: University of Chicago Press.

Heckman, J., and T. MaCurdy. 1980. A Dynamic Model of Female Labor Supply. *Review of Economic Studies*, in press.

Heckman, J., and R. Willis. 1975. Estimation of a Stochastic Model of Reproduction: An Econometric Approach. In *Household Production and Consumption*, ed. N. Terleckyj. National Bureau of Economic Research, Stanford, Calif.

Heckman, J., and R. Willis. 1977. A Beta Logistic Model for the Analysis of Sequential Labor Force Participation of Married Women. *Journal of Political Economy*. 85: 27–58.

Heckman, J., and B. Singer, 1980, eds. *Longitudinal Labor Market Studies; Theory, Methods and Empirical Results*. Social Science Research Council Monograph. New York: Academic Press, in press.

Hotelling, H. 1935. Demand Functions with Limited Budgets. *Econometrica*. 3: 66–78.

Johnson, N., and S. Kotz. 1972. *Distributions in Statistics; Continuous Multivariate Distributions*. New York: Wiley.

Johnson, N., and S. Kotz. 1977. *Urn Models and Their Application; An Approach to Modern Discrete Probability Theory*. New York: Wiley.

Joreskog, K., and A. Goldberger. 1975. Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable Model. *Journal of the American Statistical Association*. 70: 631–639.

Jovanovic, B. 1978. State Dependence in a Continuous Time Stochastic Model of Worker Behavior. Mimeographed. Columbia University, New York.

Karlin, S., and H. Taylor. 1975. *A First Course in Stochastic Processes*. 2nd ed. New York: Academic Press.

Koopmans, L. H. 1974. *The Spectral Analysis of Time Series*. New York: Academic Press.

Lawley, D., and A. Maxwell. 1971. *Factor Analysis as a Statistical Method*. 2nd ed. London: Buttersworths.

Layton, L. 1978. *Unemployment over the Work History*. Ph.D. dissertation. Department of Economics, Columbia University, New York.

Malinvaud, E. 1970. *Statistical Methods of Econometrics*. 2nd ed. Amsterdam: North Holland.

Mandelbrot, B. 1962. Paretian Distributions and Income Maximization. *Quarterly Journal of Economics*. 76: 57–83.

Maritz, J. 1970. *Empirical Bayes' Methods*. London: Metheun.

Massy, W. F., D. B. Montgomery, and D. G. Morrison. 1970. *Stochastic Models of Buying Behavior*. Cambridge, Mass.: The MIT Press.

Mundlak, Y. 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica*. 46: 69–86.

McFadden, D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1976. Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement*. 5: 363–390.

Nerlove, M. 1978. Econometric Analysis of Longitudinal Data: Approaches, Problems and Prospects. *The Econometrics of Panel Data. Annals de Insee*, Paris. 30-31: 7–22.

Neyman, J., and E. Scott. 1948. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*. 16: 1–32.

Phelps, E. 1972. *Inflation Policy and Unemployment Theory*. New York: Norton.

Polachek, S. 1975. Differences in Expected Post School Investment as a Determinant of Market Wage Differentials. *International Economic Review*. 16: 451–470.

Pollak, R. 1968. Conditional Demand Functions and Consumption Theory. *Quarterly Journal of Economics*. 83: 209–227.

Pollak, R. 1970. Habit Formation and Dynamic Demand Functions. *Journal of Political Economy*. 78: 745–763.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*. 2nd ed. New York: Wiley.

Restle, F., and J. G. Greeno. 1970. *Introduction to Mathematical Psychology*. Reading, Mass.: Addison-Wesley.

Samuelson, P. 1960. Structure of Minimum Equilibrium Systems. In *Essays in Economics and Econometrics; A Volume in Honor of Harold Hotelling*, ed. R. Pfouts. Chapel Hill, N. C.: University of North Carolina Press.

Singer, B., and S. Spilerman. 1976. Some Methodological Issues in the Analysis of Longitudinal Surveys. *Annals of Economic and Social Measurement*. 5: 447–474.

Wilson, R. D. 1977. Generalized and Embedded Versions of Heterogeneous Stochastic Models of Consumer Choice Behavior: An Empirical Test and Statistical Evaluation in a Dynamic Store Selection Context. Ph.D. dissertation. University of Iowa, Iowa City.