Editors' Introduction

This volume deals with parametric statistical inference on structural conditional probability models in which some or all of the endogenous variables are discrete valued. Within this broad theme the models posed and inferential questions addressed arise out of each author's work in econometric analysis. Taken together, these chapters provide a methodological foundation for the analysis of economic problems involving discrete data and chart the current frontiers of this subject. Some chapters are also relevant to other literatures concerned with structural analysis of discrete data: biometrics, psychometrics, sociometrics, discrete multivariate analysis, and applied subjects such as finance, marketing, geography, and transportation. Workers in these areas will recognize that econometric methods for discrete data analysis have benefited from their own literatures. This volume is intended to be useful not only for econometricians but also for the wider community of researchers involved in the structural analysis of discrete data.

In econometrics, research on models with discrete endogenous variables has two primary sources: discrete choice analysis, the study of behavior in situations where decision makers must select from finite sets of alternatives, and discrete simultaneous system modeling, the study of economic processes which may be described by systems of equations in which some endogenous variables are structurally or observationally discrete.

Discrete Choice Analysis

The canonical discrete choice model has the form $P(i \mid \mathbf{z})$, where i is an alternative in a finite choice set $\mathbf{C}$, $\mathbf{z}$ is a real vector characterizing the choice set and decision maker, and P gives the conditional probability that in the choice context characterized by $\mathbf{z}$ alternative i will be selected. The econometric literature on discrete choice generally assumes that P has been specified up to a real parameter vector $\boldsymbol{\theta}$, in which case we write $P(i \mid \mathbf{z}, \boldsymbol{\theta})$. The concerns of the literature are (1) formulation of models $P(i \mid \mathbf{z}, \boldsymbol{\theta})$ consistent with rational choice behavior and tractable, (2) inference on the parameters $\boldsymbol{\theta}$ from observations of the choices made by samples of decision makers, and (3) application of estimated probabilistic choice models to predict the behavior of populations of decision makers in given choice contexts, such as occupation, travel mode, labor force participation, or migration to new locations.

Econometric discrete choice analysis has numerous connections with other literatures. In particular the notion of a probabilistic discrete choice model originates in psychometrics with the work of Thurstone (1927) on the probit random utility model. The modern psychometric literature on probabilistic choice, as exemplified by Luce (1959), Luce and Suppes (1965), and Tversky (1972), has greatly influenced econometric model specification. Conversely, the chapters in this volume by McFadden and by Fischer and Nagin should prove of interest to psychometricians.

McFadden generalizes the Luce (1959) strict utility model and demonstrates that the generalization is consistent with an underlying random utility model of specified form. He also offers a constructive approach to the problem, first addressed by Block and Marschak (1960), of determining when an arbitrary probabilistic choice model has a random utility interpretation.

Fischer and Nagin present intriguing empirical evidence on the usefulness of the random coefficients multinomial probit model as a probabilistic description of behavior. The multinomial probit model is an important generalization of the familiar binary probit model. Lerman and Manski address computational issues associated with the calculation of multinomial probit probabilities as well as of more general choice probability forms.

Stripped of its behavioral interpretation, a probabilistic discrete choice model is simply a multinomial or quantal response model. Quantal response models have long found use in biometrics, particularly in bioassay. Indeed the biometric literature on statistical inference in such models, as developed early on by Berkson (1944) and later by Finney (1971), Cox (1970) and others, provided the initial inferential tools for discrete choice analysis.

Recent developments in the statistical analysis of discrete choice should be very valuable to biometricians. The canonical discrete choice model presumes an extant population of decision makers $\mathbf{T}$, each member τ of whom must select an alternative $i \in \mathbf{C}$ and each of whom has his choice context characterized by an attribute vector $\mathbf{z} \in \mathbf{Z}$, $\mathbf{Z}$ being the attribute space. The joint distribution of choices and attributes in the population is described by the generalized density $f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z})$, where p is the marginal attribute distribution.

The primary inferential approach investigated in the literature is natural observation rather than experimentation: a sample of decision makers,

each with associated choice and attributes, is drawn from T by a specified sampling rule, and $\boldsymbol{\theta}$ is estimated from this sample of observations. Taken together, the chapters by Manski and McFadden and by Cosslett in this volume provide a quite general, rigorous treatment of sample design and estimation using natural observations. In particular they consider maximum likelihood and pseudomaximum likelihood approaches to the estimation of $\boldsymbol{\theta}$ under sampling processes in which the population is stratified into choice-attribute subsets and observations are drawn at random within the subpopulations defined by these subsets.

The focus of the discrete choice literature on inference from natural observations follows in part from the difficulties associated with experimentation in human populations. While much of the biometric literature is concerned with animal populations where experimentation is possible, many biometric investigations concern human populations where natural observation may often be the only feasible inferential approach. Given this, the Manski-McFadden and Cosslett chapters seem to us quite relevant to biometric practice. (For example, case-control sampling of the biometric literature is closely related to what is termed choice-based sampling here.)

The relationship between discrete choice analysis and the statistical literature on discrete multivariate analysis is also close. Consider again the population model $f(i, \mathbf{z})$, which is the starting point for discrete multivariate analysis as well as for formal discrete choice analysis. The feature of the discrete choice problem that distinguishes it from the general analysis of discrete data is the postulate that the conditional probability $P(i \mid \mathbf{z})$ belongs to a known parametric family and reflects an underlying link from $\mathbf{z}$ to i that will continue to hold even if the marginal distribution $p(\mathbf{z})$ changes. This postulate motivates our decomposition of $f(i, \mathbf{z})$ into the form $f(i, \mathbf{z}) = P(i \mid \mathbf{z}, \boldsymbol{\theta})p(\mathbf{z})$.

In general, given a population with a probability distribution $f(i, \mathbf{z})$, one might in the absence of any knowledge of the process relating i's to $\mathbf{z}$'s obtain a random sample and directly examine the joint distribution $f(i, \mathbf{z})$. This exploratory data analysis approach is exemplified by the literature on associations in contigency tables, where it is assumed that $\mathbf{Z}$ is finite. See, for example, Goodman and Kruskal (1954), Haberman (1974), and Bishop, Fienberg, and Holland (1975).

Alternatively, if one believes that the elements of $\mathbf{C}$ index conceptually distinct populations of $\mathbf{z}$ values, then the natural analytic approach is to decompose $f(i, \mathbf{z})$ into the product $f(i, \mathbf{z}) = q(\mathbf{z} \mid i)Q(i)$, where $q(\mathbf{z} \mid i)$

gives the distribution of $\mathbf{z}$ within the population indexed by i and $Q(i)$ is the proportion of the population with this index. This is the approach taken in discriminant analysis. There prior knowledge allows the analyst to specify $q(\mathbf{z} \mid i)$ up to a parametric family, and a sample suitable for estimating the unknown parameters is obtained from the subpopulation i. See, for example, Anderson (1959), Warner (1963), and Kendall and Stuart (1976).

Clearly discrete choice analysis, or more generally quantal response modeling, falls within and not outside the general statistical analysis of discrete data. This fact has sometimes been obscured because statisticians have analyzed via contingency tables or discriminant functions populations where the relation between i's and $\mathbf{z}$'s is more appropriately modeled using the quantal response decomposition of $f(i, \mathbf{z})$. Some examples are given in the chapter by Manski and McFadden.

Where the quantal response decomposition is in fact appropriate, the discrete choice literature makes practical contributions that should interest statisticians. First, it offers a variety of useful forms for the response probability $P(i \mid \mathbf{z}, \boldsymbol{\theta})$. The statistical literature appears to us excessively preoccupied with log-linear forms. (Note, however, that the multinomial logit model used in many discrete choice analyses is log-linear.) Second, it offers a range of sample designs and estimation methods for $\boldsymbol{\theta}$ and highlights the value of auxiliary information in the estimation process. Many of the technical results on sample design and estimation achieved in the discrete choice literature have not been explored by statisticians. Perhaps calling attention to the relations and distinctions among the contigency table, quantal response and discriminant analysis approaches to discrete data analysis will lead statisticians to examine more carefully which approach is the most appropriate in applications.

A further symbiosis has existed between choice analysis and mobility studies in sociometrics, geography, and regional science. Sociometricians have for some time applied descriptive Markov-modeling approaches to study the way individuals move within organizations and across space. See, for example, Blumen, Kogan, and McCarthy (1955), Ginsberg (1972), and Stewman (1976). In contrast, the discrete choice literature has generally confined its attention to static modeling. To workers in both econometrics and sociometrics, it has become increasingly apparent that the development of dynamic discrete choice models would constitute a significant advance over both the descriptive dynamic models of mobility studies and the structural static models of present discrete choice analysis.

An important step in this direction is taken by Heckman. In his work mobility arises as the outcome of sequences of choices made by individuals over time. The choice process may be behaviorally dynamic (exhibit true state dependence, in Heckman's terms), observationally dynamic (exhibit spurious state dependence), or both. A focus of Heckman's analysis is the development of inferential procedures for distinguishing true from spurious state dependence. A second focus is on the statistical problems that arise when one's observation of a dynamic choice process does not provide a complete history of the process.

A chapter that will be of interest to urban geographers, regional economists, and socioeconometricians pursuing mobility studies is by Ben-Akiva and Watanatada. These authors address the problem of characterizing the spatial distribution of destination alternatives faced by trip makers and the way trip makers choose among these destinations. Their MIT-TRANS modeling approach, which incorporates a continuous endogenous variable logit choice model, should be applicable to the analysis of intraurban residential and business location.

Discrete Simultaneous Systems Modeling

The literature on discrete simultaneous systems modeling is a natural outgrowth of the long-standing concern in econometrics with the estimation of linear model systems. Consider the two-equation linear system,

$$y_1 = \beta_1 y_2 + x_1 \gamma_1 + \varepsilon_1,$$
$$y_2 = \beta_2 y_1 + x_2 \gamma_2 + \varepsilon_2,$$

where the distribution of $(\varepsilon_1, \varepsilon_2)$, conditioned on (x_1, x_2), is multivariate normal with mean zero and covariance matrix Σ. A major theme of the literature on discrete systems is to investigate ways to estimate the parameters $(\beta_1, \beta_2, \gamma_1, \gamma_2, \Sigma)$ when an economic process is described by the two-equation (or a similar multi-equation) system but observations are influenced by discrete events involving y_1 and y_2.

To start with some relatively simple cases, Tobin (1958) and Amemiya (1973) examine estimators for γ_1 in the situation where $\beta_1 = 0$, x_1 is always observed but y_1 is observed only when $y_1 > \alpha_1$, a constant. Gronau (1974) and Heckman (1976) analyze the version of this situation in which $y_1 \leqq \alpha_1$ implies that neither y_1 nor x_1 is observed. The latter problem is one of truncated sampling; the former has been termed the tobit case.

In another type of problem $\beta_1 = \beta_2 = 0$, x_1 and x_2 are always observed, y_1 and y_2 are not observed, but the event $y_1 > y_2$ is observed. The reader familiar with discrete choice analysis will recognize that this is the observational situation faced when one attempts to infer preferences from choices. That is, if y_1 and y_2 are random utilities for alternatives 1 and 2, a decision-maker's choice of alternative 1 over 2 implies only that $y_1 > y_2$. See, for example, McFadden (1973) or Manski (1975).

A third class of problems that has received much attention is switching regression. Here (y_1, x_1) is observed if and only if $y_1 < y_2$; otherwise $(y_2, x_2)$is observed. Switching regressions, which have been studied by Fair and Jaffee (1972), Maddala and Nelson (1974), and by others, arise naturally in the analysis of markets in disequilibrium.

A great many variants and generalizations of observational problems have been identified and studied in recent years. Lee offers a unified framework for posing and resolving such problems. In particular Lee demonstrates that an estimation approach proposed by Amemiya (1978, 1979) in specific contexts can be usefully applied to a broad range of discrete observational conditions.

A second chapter by Hausman and Wise examines the sampling process used in data collection for a recent social experiment and presents alternative estimation methods appropriate under that process. Ostensibly the sampling process followed is endogenous censored sampling, in which a random sample is first drawn and then some observations are deleted, based on a discrete condition related to the value of endogenous variables. Hausman and Wise clarify some subtle distinctions among various stratified and censored sampling processes, which superficially appear quite similar, and develop tractable estimators.

Poirier in an interesting applied chapter analyzes various aspects of physician behavior. His behavioral model involves both discrete choice and linear model aspects. In the sampling process generating his data, the physicians' discrete choice determines what variables from the linear system are observed. Also the procedure by which physicians were drawn into the sample itself is choice based. Poirier's handling of this myriad of complexities demonstrates the power of discrete choice analysis and discrete simultaneous modeling as applied tools.

Recently the literature on discrete simultaneous modeling has developed a second major theme. Consider the two-equation mixed discrete-linear system,

$$y_1 = \beta_1 y_2 + \beta_1^* y_2^* + x_1 y_1 + \varepsilon_1,$$
$$y_2 = \beta_2 y_1 + \beta_2^* y_1^* + x_2 y_2 + \varepsilon_2,$$

where $y_1^* = 1$ if $y_1 > \alpha_1$, $y_1^* = 0$ otherwise, $y_2^* = 1$ if $y_2 > \alpha_2$, $y_2^* = 0$ otherwise, and ε is the same as before. This system is qualitatively different from the one posed earlier because discrete transformations of the endogenous variables are part of the system structure. Consequently the system now does not have a linear reduced form.

Even when observational problems do not exist, parameter estimation in models such as the mixed discrete-linear system poses difficulties. See Amemiya (1974) and Heckman (1978) for relevant analyses. The paper by Schmidt sets out several classes of mixed discrete-linear model systems. Schmidt finds that in each model internal consistency requires that a set of more or less restrictive parameter constraints be satisfied. Since these constraints often have no apparent economic interpretation, his results call into question the appropriateness of some of the model structures that have been posed in the literature.

Avery deals with estimation of a mixed discrete-linear system in the presence of discrete observational problems. Avery's concern is with the measurement of racial differences in consumer credit demand and supply. His model assumes that a household's durable demand is a function of, among other things, its observed credit, which is itself the minimum of its unobserved demand for and supply of credit. His work illustrates the concerns of the discrete simultaneous modeling literature, and his empirical results are of substantive interest.

It will be noticed that, in discussing the papers in this volume that contribute to the discrete simultaneous modeling literature, we have not developed connections with other literatures as we did in our treatment of discrete choice analysis. This asymmetry arises because the simultaneous equations field has as a whole developed largely within econometrics. Certainly connections with other disciplines exist. In particular the reduced form of a linear simultaneous system is the multivariate regression model widely studied in statistics. Recursive simultaneous equations models are the path analysis models of sociometrics. Simultaneous systems models with unobserved (latent) exogenous variables are the factor analytic models of psychometrics. However, we are unaware of systematic efforts to go beyond the obvious similarities of the models used in econometrics and other disciplines and search for approaches that can be productively

transferred between subjects. We hope the readers of this volume will be motivated to further research that integrates the methods in various disciplines for structural analysis of discrete data.

References

Amemiya, T. 1973. Regression Analysis When the Dependent Variable is Truncated Normal. *Econometrica*. 41: 997–1016.

Amemiya, T. 1974. Multivariate Regression and Simultaneous Equation Models When the Dependent Variables are Truncated Normal. *Econometrica*. 42: 999–1012.

Amemiya, T. 1978. The Estimation of a Simultaneous Equation Generalized Probit Model. *Econometrica*. 46: 1193–1205.

Amemiya, T. 1979. The Estimation of a Simultaneous Equation Tobit Model, *International Economic Review*.

Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Berkson, J. 1944. Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*. 39: 357–365.

Bishop, Y., S. Fienberg, and P. Holland. 1975. *Discrete Multivariate Analysis*, Cambridge, Mass.: MIT Press.

Block, H., and J. Marschak. 1960. Random Orderings and Stochastic Theories of Response. In *Contributions to Probability and Statistics*, ed. I. Olkin. Stanford, Calif.: Stanford University Press.

Blumen, A., M. Kogan, and J. McCarthy. 1955. *The Industrial Mobility of Labor as a Stochastic Process*. Ithaca, N. Y.: Cornell University Press.

Cox, D. 1970. *Analysis of Binary Data*. London: Methuen.

Fair, R., and D. Jaffee. 1972. Methods of Estimation for Markets in Disequilibrium. *Econometrica*. 40: 497–514.

Finney, D. 1971. *Probit Analysis*. New York: Cambridge University Press.

Ginsberg, R. 1972. Incorporating Causal Structure and Exogenous Information with Probabilistic Models with Special Reference to Choice, Gravity, Migration and Markov-Chains. *Journal of Mathematical Sociology*. 2: 83–101.

Goodman, L., and W. Kruskal. 1954. Measures of Association for Cross-Classifications. *Journal of the American Statistical Association*. 49: 732–764.

Gronau, R. 1974. Wage Comparisons—A Selectivity Bias. *Journal of Political Economy*. 82: 1119–1143.

Haberman, S. 1974. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Heckman, J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*. 5: 475–492.

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

Kendall, M., and J. Stuart. 1976. *Advanced Theory of Statistics*, vol. 3. New York: Hafner.

Luce, R. 1959. *Individual Choice Behavior*. New York: Wiley.

Luce, R., and P. Suppes. 1965. Preference, Utility and Subjective Probability. In *Handbook of Mathematical Psychology*, ed. R. Luce, R. Bush, and E. Galanter, vol. 3, pp. 249–410. New York: Wiley.

Maddala, G. S., and F. Nelson. 1974. Maximum Likelihood Methods for Markets in Disequilibrium. *Econometrica*. 42: 1013–1030.

Manski, C. 1975. Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics*. 3: 205–228.

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

Stewman, S. 1976. Markov Models of Occupational Mobility: Theoretical Development and Empirical Tests. *Journal of Mathematical Sociology*. 6: 201–278.

Thurstone, L. 1927. A Law of Comparative Judgment. *Psychological Review*. 34: 273–286.

Tobin, J. 1958. Estimation of Relationships for Limited Dependent Variables. *Econometrica*. 26: 24–36.

Tversky, A. 1972. Elimination by Aspects: A Theory of Choice. *Psychological Review*. 79: 281–299.

Warner, S. 1963. Multivariate Regression of Dummy Variates Under Normality Assumptions. *Journal of the American Statistical Association*. 58: 1054–1063.