Annals of Economic and Social Measurement, 5/4, 1976

INTRODUCTION

BY JAMES J. HECKMAN

In a journal containing several excellent surveys of the literature on the analysis of discrete data, a survey of the surveys is an unnecessary addition. Accordingly, the reader is spared the usual magisterial overview and instead is offered a brief introduction to the contents of this volume. The papers published here are a partial collection of papers presented at two conferences on quantal choice organized by Dan McFadden that were held at Berkeley in March, 1974 and at the University of Chicago in May, 1975. These conferences provided a meeting ground for a group of econometricians and applied economists to share ideas and problems that arise in discrete data analysis. Both conferences were sponsored by the NBER/NSF Conference on Econometrics and Mathematical Economics.

Interest in discrete data has been stimulated by a growing interest in microeconomic problems and a growing availability of good microeconomic data. As economists attempt to make greater use of their theory to solve such practical problems as estimating the demand for new modes of travel and ascertaining the determinants of the labor supply of women, the analytical fiction of the representative consumer and its econometric analogue—the classical regression model—have become less useful. Increasingly, economists have begun to recognize that the analysis of choices at the extensive margin (i.e., discrete choices) are just as interesting and often of greater empirical importance than the analysis of choices at the intensive margin that is treated in traditional analysis. Because the source of sample variation critically affects the formulation and estimation of many models of discrete choice, the traditional schizophrenia of "Marshallian econometrics" that separates the formulation of an economic model for a "typical individual" from its stochastic specification is absent from many of the best papers in this literature.

There are several distinct styles of model formulation. Work by McFadden and Quandt discussed in McFadden's survey paper carefully specifies an economic model of discrete choice in which a consumer makes utility comparisons to select a most preferred alternative in a choice set. Both authors provide integrated econometric models with parameters that possess a well defined economic interpretation. In Quandt's work, the source of sample variation is individual differences in preference functions. In McFadden's work, the source of sample variation arises from randomness in the underlying choice process.

Other work by Nerlove and Press (1973) and Amemiya (1975) that is not adequately represented in this volume offers parametric schemes that are useful for investigating empirical relationships but which are less amenable to direct structural economic interpretation. Nonetheless, precisely because there is less economic structure imposed at the outset, their estimators may serve as better tools for exploratory data analysis.

An important issue in the practical use of most models for discrete data is their computational tractability. To date, the Nerlove-Press multivariate logistic model and McFadden's conditional logit model have proved to be more tractable than other models. Models based on the multivariate normal tend to be more unwieldy although the paper by Dutt offered in this volume reports on promising developments in evaluating the multivariate normal integral.

Even if one only seeks to analyze empirical relationships in discrete data, there is still the question of the best way to do so. McFadden's second paper addresses this question. The outcomes of any discrete choice experiment can be characterized by a joint distribution of the dummy indicator variables that represent choices made among alternatives and the explanatory variables that determine the choices. The logit model and related probability models represent the distribution of the dummy indicator variables conditional on the explanatory variables. Such conditional probability models are natural representations of causal models. Given the marginal distribution of the explanatory variables, these conditional probabilities fully characterize the data.

The discriminant function approach is based on the distribution of explanatory variables conditional on values of dummy indicator variables, and is a less natural tool for the analysis of causal models. However, the marginal distribution of the indicator variables and the conditional distribution of the explanatory variables uniquely determine the joint distribution of the data. Thus either conditional model supplemented by its appropriate marginal distribution uniquely determines the joint distribution of the data. McFadden argues that the choice between these models depends in part on the purpose of the investigation, and that conditional probability models have a natural causal interpretation. However, as Efron (1975) has recently demonstrated, when the classical normal discriminant model is appropriate (McFadden's paper demonstrates the severe assumptions required to justify its use), direct estimation of the discriminant model produces more efficient estimates of the parameters of the logit model than direct estimation of the logit model. Thus even if the conditional probability model is the object of the investigation, discriminant function estimation may yield more efficient estimates.

The Kohn–Manski–Mundel paper that follows McFadden's survey paper illustrates both the strengths and limitations of the conditional logit model. As the authors note, in estimating a discrete choice model specifying the choice set available to consumers may be exceedingly difficult and surely is a more complicated empirical procedure than specifying the budget set in traditional demand analysis. Moreover, the assumed source of sample variation—independent "disturbances" in preferences for each alternative—is restrictive, especially if data are missing on explanatory variables relevant to the selection of all alternatives. Nonetheless, the Kohn-Manski-Mundel paper illustrates the power of McFadden's methodology and its potential value in forecasting the demand for new alternatives.

Virtually all of the available models of quantal choice are cross sectional models although most may be adapted to handle panel data. To date, however, there has been little systematic work in econometrics on specifying and estimating discrete dynamic choice models with a clearly formulated stochastic structure.

The analysis of stochastic models for discrete panel data has been an active area of research in sociology dating back to pioneering work by James Coleman (1964). The paper by Singer and Spilerman reviews and extends this literature and presents some new ideas on model formulation and identification. Their work focuses solely on the stochastic structure of discrete models. As economists absorb this literature, they will recognize that there is room for improvement on purely stochastic models that do not possess a clear economic structural interpretation. Much work remains to be done on formulating dynamic models and investigating alternative methods of estimation.

The papers by Hausman and Wise, Heckman, Nelson and Maddala and Lee represent another strand of the econometric literature on qualitative data that takes its starting point from the seminal paper of Tobin (1958) on limited dependent variables. Heckman's paper displays the common structure of many recent models of sample selection and truncation and their intellectual debt to Tobin's pioneering paper. In his paper, a computationally simple estimator is proposed that does not rely on cumbersome full information maximum likelihood methods, and hence is useful in exploratory data analysis. The paper by Nelson presents a computationally efficient algorithm for such models if full information maximum likelihood estimates are desired. The Hausman and Wise paper applies a model of truncation to data from the New Jersey negative income tax experiment to demonstrate the relevance of recent concern about sample truncation. If samples are selected on the basis of the dependent variable in an analysis (earnings in the case of the New Jersey data) important biases may result and the empirical results of Hausman and Wise illustrate this bias (see also Crawford, 1975). The paper by Maddala and Lee is a "second generation" Tobin model that specializes previous work—discussed in the final section of McFadden's survey—on simultaneous equation systems with both continuous and discrete endogenous variables.

JAMES J. HECKMAN
University of Chicago and
National Bureau of Economic Research

REFERENCES

Amemiya, T., "Qualitative Response Models," *Annals of Economic and Social Measurement* Vol. 4, #3 (Summer 1975): 363–372.

Crawford, David, "Estimating Earnings Functions from Truncated Samples," *Institute for Research on Poverty*, discussion paper 287–75 (July 1975).

Coleman, James, *Introduction to Mathematical Sociology*. The Free Press of Glencoe, New York, 1964.

Efron, Bradley, "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association* Vol. 70, #352 (December 1975): 892–898.

Nerlove, Marc and James Press, "Univariate and Multivariate Log-Linear and Logistic Models," Report #R-1306-EDA/NIH, Rand Corporation, 1973.

Tobin, James, "Estimation of Relationships for Limited Dependent Variables," *Econometrica* Vol. 26, pp. 24–36.

ACKNOWLEDGEMENTS

The following referees provided valuable critical reviews of the papers in this volume. Their help is gratefully acknowledged, and has improved the quality of the final product.

Jerry Hausman, Massachusetts Institute of Technology
Charles Manski, Carnegie-Mellon University
Forest Nelson, California Institute of Technology
Randy Olsen, Yale University
David Wise, Harvard University