COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 474

QUANTITATIVE METHODS FOR ANALYZING TRAVEL BEHAVIOUR OF INDIVIDUALS: SOME RECENT DEVELOPMENTS

Daniel McFadden

November 22, 1977

QUANTITATIVE METHODS FOR ANALYZING TRAVEL BEHAVIOUR OF INDIVIDUALS: SOME RECENT DEVELOPMENTS

Daniel McFadden, University of California, Berkeley

ABSTRACT

This paper is concerned with quantitative methods for the analysis of travel behaviour of individuals. It reviews some of the recent developments in model specification, estimation, model evaluation and testing, and aggregation and forecasting. Topics in model specification include the multinomial probit model and its computation, and generalized extreme value models and their relation to sequential models. Topics in estimation methods include the use of choice-based samples, sample designs, and incomplete choice sets. Model evaluation topics include prediction success tables and diagnostic tests of specification. Aggregation and forecasting topics include aggregation by the Clark method, synthesis of the distribution of explanatory variables, and the calculus of demand elasticities.

I INTRODUCTION

This paper is concerned with quantitative methods for the analysis of travel behaviour of individuals. It reviews some recent developments in model specification, estimation, model evaluation and testing, and aggregation and forecasting. The reader is assumed to be familiar with the general foundations of disaggregate choice theory,* the historical development and properties of the multinomial logit model,** and the use of behavioural models in travel demand analysis.†

Model specification is discussed in Section II, and statistical estimation methods in Section III. Section IV discusses methods for hypothesis testing and model evaluation. Section V concludes the paper with a discussion of aggregation and forecasting.

---

*See, for example, McFadden (1976b, 1976d).

**McFadden (1973).

†Meyburg and Stopher (1975, 1976); Domencich and McFadden (1975).

II MODEL SPECIFICATION

1 Choice Models

The choice models which have received serious consideration in travel demand applications are multinomial logit (MNL), multinomial probit (MNP), and a sequential — or tree — version of multinomial logit.

2 The Multinomial Logit Model

A typical MNL model for joint choice of mode, destination, and auto availability is

$$(1) \qquad P_{mda} = e^{V_{mda}} \Big/ \sum_{n,c,b} e^{V_{ncb}} \quad ,$$

where m = mode;
d = destination;
a = auto availability;

and $V_{mda} = \alpha x_{mda} + \beta y_{da} + \gamma z_a$ = utility, with α , β , γ parameter vectors and x_{mda} , y_{da} , z_a variable vectors describing the decision-maker and the alternative. Letting $P_{m|da}$ denote a conditional choice probability and P_m denote a marginal choice probability, one derives from (1) the formulae:

$$(2) \qquad P_{m|da} = e^{V_{mda}} \Big/ \sum_{n} e^{V_{nda}} = e^{\alpha x_{mda}} \Big/ \sum_{n} e^{\alpha x_{nda}}$$

$$(3) \qquad P_{d|a} = \sum_{n} e^{V_{nda}} \Big/ \sum_{n,c} e^{V_{nca}} = \sum_{n} e^{\alpha x_{nda} + \beta y_{da}} \Big/ \sum_{n,c} e^{\alpha x_{nca} + \beta y_{ca}}$$

$$(4) \qquad P_a = \sum_{n,c} e^{V_{nca}} \Big/ \sum_{n,c,b} e^{V_{ncb}} = \sum_{n,c} e^{\alpha x_{nca} + \beta y_{ca} + \gamma z_a} \Big/ \sum_{n,c,b} e^{\alpha x_{ncb} + \beta y_{cb} + \gamma z_b}$$

Define inclusive values

$$(5) \qquad I_{da} = \log \sum_{n} e^{\alpha x_{nda}} \quad ;$$

$$(6) \qquad J_a = \log \sum_{n,c} e^{\alpha x_{nca} + \beta y_{ca}} = \log \sum_{c} e^{I_{ca} + \beta y_{ca}} \quad .$$

Then the choice probabilities can be written

$$(7) \qquad P_{mda} = P_{m|da} P_{d|a} P_a \quad ;$$

$$(8) \qquad P_{m|da} = e^{\alpha x_{mda}} / e^{I_{da}} \quad ;$$

$$(9) \qquad P_{d|a} = e^{I_{da} + \beta y_{da}} / \sum_{c} e^{I_{ca} + \beta y_{ca}} = e^{I_{da} + \beta y_{da}} / e^{J_a} \quad ;$$

$$(10) \qquad P_a = e^{J_a + \gamma z_a} / \sum_{b} e^{J_b + \gamma z_b} \quad .$$

Discussion of the historical development of the MNL model can be found in McFadden (1976b); the properties of the model, including its derivation from the theory of individual utility maximization, are given in McFadden (1973).

3 Sequential MNL Models

Next consider the sequential or nested MNL model. A typical sequential model differs from the joint MNL model solely in that the coefficients of inclusive values are not constrained to equal one. Hence, the joint MNL model is a linear restriction on any of the sequential models. Specifically, a sequential model is defined by

$$(11) \qquad P_{mda} = P_{m|da} P_{d|a} P_a \quad ;$$

$$(12) \qquad P_{m|da} = e^{\alpha x_{mda}} / \sum_{n} e^{\alpha x_{nda}} \quad ;$$

$$(5') \qquad I_{da} = \log \sum_{n} e^{\alpha x_{nda}} \quad ;$$

$$(13) \qquad P_{d|a} = e^{\theta I_{da} + \beta y_{da}} / \sum_{c} e^{\theta I_{ca} + \beta y_{ca}} \quad ;$$

$$(14) \qquad J_a = \log \sum_c e^{\theta I_{ca} + \beta y_{ca}} \quad ;$$

$$(15) \qquad P_a = e^{\lambda J_a + \gamma z_a} \Big/ \sum_b e^{\lambda J_b + \gamma z_b} \quad .$$

When $\theta = \lambda = 1$, this model is identical to the joint MNL model. More generally, when $\theta \neq 1$, equations (13) and (14) differ in the two models, and when $\lambda \neq 1$, equation (15) differs in the two models.

The sequential model was introduced by Domencich and McFadden (1975), and studied by Ben-Akiva (1973), and is discussed in greater detail in part 7 of this section.

4 The Multinomial Probit Model

A typical model for the mode-destination-auto choice problem is obtained by assuming that each alternative mda has a utility $u_{mda} = V_{mda} + \lambda_{mda} + \eta_{da} + \nu_a$, where λ_{mda}, η_{da}, and ν_{da} summarize the influence of unobserved attributes and taste variations, and are assumed to be jointly normally distributed over the population. If each individual maximizes utility, the proportion of the population choosing mda is

$$(16) \qquad P_{mda} = \int_{\varepsilon_{mda}=-\infty}^{+\infty} \cdots \int_{\varepsilon_{ncb}=-\infty}^{V_{mda}-V_{ncb}+\varepsilon_{mda}} \cdots \int_{\varepsilon_{MDA}=-\infty}^{V_{mda}-V_{MDA}+\varepsilon_{mda}} n(\varepsilon;0,\Omega)d\varepsilon \quad ,$$

where the number of integrals equals the number of alternatives, $n(\varepsilon;0,\Omega)$ is the multivariate normal density with mean vector 0 and covariance matrix Ω, and $\varepsilon_{mda} = \lambda_{mda} + \eta_{da} + \nu_a$, with the joint normal distribution of λ, η, and ν determining Ω.

The MNP model generalizes a classical model of Thurstone (1927) for binary choice. Bock and Jones (1969) applied the model to the three-alternative case. The model was suggested for transportation analysis by Domencich and McFadden (1975), and first applied to transportation data by Hausman and Wise (1976). Further discussion is given in part 8 of this section.

5 Other Choice Models

Several other models deserve passing note. McFadden (1975a) has proposed a universal, or "mother," MNL model which can approximate an arbitrary choice model with a function of the form (1), except that V_{mda} functions will depend on the attributes of all

alternatives, and not solely on the attributes of mda . This model is useful for testing particular specifications, but is in general inconsistent with utility maximization.

McLynn (1973) has proposed the fully competitive model which is a one-parameter mapping of MNL choice probabilities into a second vector of probabilities. This model is in general inconsistent with individual utility maximization, yet it shares with the MNL model restrictive structural properties which render it implausible in some applications.*

6 The Generalized Extreme Value Model

McFadden (1977) has recently proposed a family of generalized extreme value (GEV) choice models which allow a general pattern of dependence among alternatives and yield a closed form for the choice probabilities. The following result characterizes the family:

THEOREM. Suppose $G(y_1,\ldots,y_J)$ is a nonnegative, homogeneous-of-degree-one function of $(y_1,\ldots,y_J) \geq 0$. Suppose $\lim_{y_i \to +\infty} G(y_1,\ldots,y_J) = +\infty$ for $i = 1,\ldots,J$. Suppose for any distinct $(i_1,\ldots,i_k)$ from $\{1,\ldots,J\}$, $\partial^k G/\partial y_{i_1},\ldots,\partial y_{i_k}$ is nonnegative if k is odd and non-positive if k is even. Then,

$$(17) \qquad P_i = e^{V_i} G_i(e^{V_1},\ldots,e^{V_J}) \,/\, G(e^{V_1},\ldots,e^{V_J})$$

defines a choice model which is consistent with utility maximization.

The special case $G(y_1,\ldots,y_J) = \sum_{j=1}^{J} y_j$ yields the MNL model. An example of a more general G function satisfying the hypotheses of the theorem is

---

*The model satisfies "simple scalability" = "order independence," which is closely related to the "independence from irrelevant alternatives" property of the MNL model. See McFadden (1975b).

$$(18) \qquad G(y) = \sum_{m=1}^{M} a_m \left(\sum_{i \in B_m} Y_i^{\frac{1}{1-\sigma_m}} \right)^{1-\sigma_m} ,$$

where $B_m \subseteq \{1,\ldots,J\}$, $\bigcup_{m=1}^{M} B_m = \{1,\ldots,J\}$, $a_m > 0$, and $0 \leq \sigma_m < 1$. The parameter σ_m is an index of the similarity of the unobserved attributes of alternatives in B_m . The choice probabilities for this function satisfy

$$(19) \qquad P_i = \sum_{m=1}^{M} P(i|B_m)P(B_m) ,$$

where $P(i|B_m)$ is the conditional probability that alternative i is chosen, given the event B_m , with

$$(20) \qquad P(i|B_m) = e^{\frac{V_i}{1-\sigma_m}} \Big/ \sum_{j \in B_m} e^{\frac{V_j}{1-\sigma_m}} \quad \text{if } i \in B_m ;$$

$$0 \qquad \text{if } i \notin B_m ;$$

and $P(B_m)$ is the probability of the event B_m , with

$$(21) \qquad P(B_m) = a_m \left\{ \sum_{j \in B_m} e^{\frac{V_j}{1-\sigma_m}} \right\}^{1-\sigma_m} \Big/ \sum_{n=1}^{M} a_n \left\{ \sum_{k \in B_n} e^{\frac{V_k}{1-\sigma_n}} \right\}^{1-\sigma_n} .$$

Functions of the form in (18) can also be nested to yield a wider class satisfying the theorem hypotheses. For example, the function

$$(22) \qquad G = \sum_{q=1}^{Q} a_q \left(\sum_{m \in D_q} \left(\sum_{j \in B_m} y_j^{\frac{1}{1-\sigma_m}} \right)^{\frac{1-\sigma_m}{1-\delta_q}} \right)^{1-\delta_q} ,$$

where $\cup B_m = \{1,\ldots,J\}$, satisfies the hypotheses provided $1 > \sigma_m \geq \delta_q \geq 0$ for $m \in D_q$. The choice probabilities for (22) and analogous functions can be written as sums of products of conditional and marginal probabilities, in a manner generalizing (19), with each probability element having a multinomial logit form, and the denominator in each element equalling a representative term in the succeeding element.

Choice probabilities of the form (19) were apparently first derived, for the case of three alternatives and $B_1 = \{1\}$, $B_2 = \{2,3\}$ by Scott Cardell (1975). For the case of disjoint B_m , the form (19) has been discovered, independently, by Daly and Zachary (1977), Williams (1977), and Ben-Akiva and Lerman (1977). The demonstration by Daly and Zachary that this choice model is consistent with random utility maximization is particularly noteworthy in that it permits generalization of the GEV model and provides a powerful tool for testing the consistency of choice models: Suppose alternative i has a utility $U_i = w_i + y(z_i) + \varepsilon_i$, where $V_i \equiv w_i + y(z_i)$ is the systematic component of utility and $(\varepsilon_1,\ldots,\varepsilon_J)$ is a jointly distributed random vector, with a distribution function not depending on $(w_1,\ldots,w_J)$, but in general depending on $(z_1,\ldots,z_J)$. Suppose the choice probabilities satisfy

$$(23) \qquad P_i \equiv P_i(V_1,\ldots,V_J;z_1,\ldots,z_J) = \text{Prob}\,[V_i + \varepsilon_i \geq V_j + \varepsilon_j \text{ for } j \neq 1]$$

$$\equiv \text{Prob}\,[V_i - V_j \geq \nu_j - \nu_i \text{ for } j \neq i] ,$$

where $\nu_i = \varepsilon_i - \varepsilon_1$. Define the expected value of the maximum of the utilities U_j ,

$$(24) \qquad \bar{U}(V_1,\ldots,V_J;z_1,\ldots,z_J) = E\,\text{Max}_j\, U_j .$$

Then, the choice probabilities satisfy*

$$(25) \qquad P_i(V_1,\ldots,V_J;z_1,\ldots,z_J) = \frac{\partial}{\partial V_i}\bar{U}(V_1,\ldots,V_J;z_1,\ldots,z_J) \quad ,$$

and the joint distribution of the differences of the random components of utility, $(\nu_2,\ldots,\nu_J)$, satisfies

$$(26) \qquad F(\nu_2,\ldots,\nu_J) = P_1(0,-\nu_2,\ldots,-\nu_J) \quad .$$

Conversely, any choice probability functions $P_i(V_1,\ldots,V_J;z_1,\ldots,z_J)$ which satisfy the necessary and sufficient conditions for $(P_1,\ldots,P_J)$ to be the gradient of a potential** $(\bar{U})$ and for $P_1(0,-\nu_2,\ldots,-\nu_J)$ to be a distribution function,† satisfy (23), and are consistent with stochastic utility maximization. The key assumption, and only significant restriction, underlying this result is that the random utilities U_i have linear components w_i with the property that the joint distribution of the stochastic components does not depend on $(w_1,\ldots,w_J)$.††

---

*This condition has been used by Domencich and McFadden (1975) and Harris and Tanner (1975) to establish a classical identity between social welfare, defined by the expected value (or average over the population) of the maximum utility for each individual, and consumer surplus, defined by the area under the market demand curves, or choice probabilities. The identity can be verified directly by writing out the definition of expected maximum utility and differentiating. The basic assumption required for the social welfare identity is a linear "transferable" numéraire commodity. A consequence of the additively separable structure of errors specified in (23) is that the choice probabilities are invariant with respect to location; i.e.,

$$P_i(V_1 + a,\ldots,V_J + a;z_1,\ldots,z_J) \equiv P_i(V_1,\ldots,V_J;z_1,\ldots,z_J) \quad .$$

**Suppose $P_i(V_1,\ldots,V_J)$ is continuous and continuously differentiable. Then, a necessary and sufficient condition for $(P_1,\ldots,P_J)$ to be the gradient of a potential $(\bar{U})$ is that $\partial P_i/\partial V_j = \partial P_j/\partial V_i$.

(continued on page)

The GEV model satisfies (25) with $\bar{U} = \log G(e^{V_1},\ldots,e^{V_J})$.

7 Relation of Sequential MNL and GEV Models

The choice probabilities corresponding to (22) can be specialized to the sequential MNL model described in (11) — (15), as we shall now show. This result establishes that sequential MNL models are consistent with individual utility maximization for appropriate parameter values, and that the coefficients of inclusive values can be used to obtain estimates of the similarity parameters σ and δ . It is hence possible to estimate some GEV models using sequential MNL models and inclusive values. Further, the GEV class provides a generalization containing alternative sequential MNL models, and could be estimated directly to test the presence of a sequential or tree structure.

To obtain the sequential model (11) — (15) from (22), index alternatives by mda for mode m , destination d , and auto availability a , and specialize (22) to the form

---

(continued from page)

If P_i is invariant with respect to location, then $\bar{U}$ is homogeneous with respect to location; i.e., $\bar{U}(V_1 + a,\ldots,V_J + a) = \bar{U}(V_1,\ldots,V_J) + a$.

†See Cramer (1946, Sect. 8.4). The key condition for $F(\nu_2,\ldots,\nu_J)$ to be a distribution is that the $(J - 1)$st difference, $\Delta^{J-1}F$, be nonnegative. If F is continuous and almost everywhere J-1 times differentiable, this condition reduces to the requirement that the density $\partial^{J-1}F/\partial\nu_2\ldots\partial\nu_J$ be nonnegative.

††Strictly, the condition is that the joint distribution of differences of the random components of utility, $F(\nu_2,\ldots,\nu_J)$, not depend on $(w_1,\ldots,w_J)$.

$$(27) \qquad G = \sum_a \left(\sum_d \left(\sum_m y_{mda}^{\frac{1}{1-\sigma}} \right)^{\frac{1-\sigma}{1-\delta}} \right)^{1-\delta} , \quad 0 \leq \sigma \leq \delta < 1 .$$

Assume $V_{mda} = (1-\sigma)\alpha' x_{mda} + (1-\delta)\beta' y_{da} + \gamma' z_a$. Then (17) yields

$$(28) \qquad P_{mda} = \frac{e^{\frac{V_{mda}}{1-\sigma}}}{\sum_n e^{\frac{V_{nda}}{1-\sigma}}} \cdot \frac{\left(\sum_n e^{\frac{V_{nda}}{1-\sigma}} \right)^{\frac{1-\sigma}{1-\delta}}}{\sum_c \left(\sum_n e^{\frac{V_{nca}}{1-\sigma}} \right)^{\frac{1-\sigma}{1-\delta}}} \cdot \frac{\left(\sum_c \left(\sum_n e^{\frac{V_{nca}}{1-\sigma}} \right)^{\frac{1-\sigma}{1-\delta}} \right)^{1-\delta}}{\sum_b \left(\sum_c \left(\sum_n e^{\frac{V_{ncb}}{1-\sigma}} \right)^{\frac{1-\sigma}{1-\delta}} \right)^{1-\delta}}$$

$$= \frac{e^{\alpha' x_{mda}}}{\sum_n e^{\alpha' x_{nda}}} \cdot \frac{e^{\beta' y_{da} + \frac{1-\sigma}{1-\delta} I_{da}}}{\sum_c e^{\beta' y_{ca} + \frac{1-\sigma}{1-\delta} I_{ca}}} \cdot \frac{e^{\gamma' z_a + (1-\delta) J_a}}{\sum_b e^{\gamma' z_b + (1-\delta) J_b}}$$

where

$$I_{da} = \log \sum_n e^{\alpha' x_{nda}}$$

$$J_a = \log \sum_c e^{\beta' y_{ca} + \frac{1-\sigma}{1-\delta} I_{ca}} .$$

This is precisely the sequential model (11) — (15), with $\theta = (1-\sigma)/(1-\delta)$ and $\lambda = 1 - \delta$. Hence, we have established that a sufficient condition for a nested logit model to be consistent with individual utility maximization is that the coefficient of each inclusive value lie between zero and one, $0 < \theta , \lambda \leq 1$.* Application of the Daly-Zachary test shows that this condition is also necessary for consistency with random utility maximization if the domain of $(V_2 - V_1, \ldots, V_J - V_1)$

---

*The preceding demonstration for three-level trees is readily generalized to trees of any depth. The simplest proof is by induction.

is unrestricted. When the necessary and sufficient condition $0 < \theta$, $\lambda \leq 1$ is satisfied, $1 - \theta$ is an index of the similarity of alternative modes, while $1 - \lambda$ is an index of the similarity of alternative destinations.

8 Computation of the MNP Model

The multinomial probit model, introduced in part 4 of this section, is an appealing conceptual model. It allows consideration of stochastic components for tastes and unobserved attributes within an alternative, and provides a way of specifying the structure of dependence between alternatives. However, MNP choice probabilities can be expressed exactly only as multivariate or iterated integrals of dimension $J - 1$, where J is the number of alternatives. Exact calculation by numerical integration is very fast for $J = 2$ or 3, moderately costly for $J = 4$, and impractical on a large scale for $J \geq 5$. One of the more effective direct numerical integration methods, adapted for transportation applications, is due to Hausman and Wise (1976).

Two recent contributions have provided techniques for approximating MNP choice probabilities at moderate cost. This has made MNP a practical alternative for many transportation applications. The first method, due to Manski (1976), applies a Monte Carlo procedure directly to the utilities of alternatives. Suppose $J + 1$ alternatives, with utilities $U_i = V_i + \varepsilon_i$, where $(\varepsilon_1, \ldots, \varepsilon_{J+1})$ is multivariate normal with zero means and covariance matrix $\Sigma = (\sigma_{ij})$. For given values of V_i, vectors $(\varepsilon_1, \ldots, \varepsilon_{J+1})$ from the multivariate distribution can be drawn, and the frequency with which utility is maximized at alternative i recorded. These frequencies approximate the exact MNP probabilities when the number of Monte Carlo repetitions is large. Because this method involves repetitive simple calculations, it can be programmed in computer assembly language to operate quite efficiently. The approach is appealing in its generality — any joint distribution of the unobserved effects can be assumed. In practice, the method is most effective when a relatively good initial approximation to the frequencies is available.

The second approximation method, due to Daganzo, Bouthelier, and Sheffi (1976), uses a procedure suggested by Clark (1961) to approximate the maximum of bivariate normal variables by a normal variable. When the correlation of the variables is nonnegative, this approximation is accurate within a few percent. Suppose $J + 1$ alternatives, with utilities $U_i = V_i + \varepsilon_i$ and $(\varepsilon_1, \ldots, \varepsilon_{J+1})$ distributed multivariate normal, zero means and covariance matrix Σ. The probability that the first alternative is chosen is then

$$(29) \qquad P_1 = \text{Prob}\,[V_1 + \varepsilon_1 > V_j + \varepsilon_j \quad \text{for} \quad j = 2,\ldots,J+1]$$

$$= \text{Prob}\,[V_1 - V_{J+1} + \varepsilon_1 - \varepsilon_{J+1} > V_j - V_{J+1} + \varepsilon_j - \varepsilon_{J+1}$$

$$\text{for} \quad j = 2,\ldots,J \quad \text{and} \quad V_1 - V_{J+1} + \varepsilon_1 - \varepsilon_{J+1} > 0] \quad .$$

Define $v_j = V_j - V_{J+1}$ and $y_i = \varepsilon_i - \varepsilon_{J+1}$. Then, $(y_1,\ldots,y_J)$ is multivariate normal with mean zero and covariance matrix $\Omega = (\omega_{ij})$, where $\omega_{ij} = \sigma_{ij} + \sigma_{J+1,J+1} - \sigma_{i,J+1} - \sigma_{j,J+1}$. Hence

$$(30) \qquad P_1 = \text{Prob}\,[v_1 + y_1 > 0 \quad \text{and} \quad v_1 + y_1 > v_j + y_j \quad \text{for} \quad j = 2,\ldots,J]$$

$$= \int_{y_1=-v_1}^{\infty} n_1(y_1) N_{(1)}((v_1 - v_j + y_1)|y_1)dy_1$$

where $n_{Y(X)}(y|x)$ denotes the normal density for the vector of variables indexed by Y , conditioned on the vector of variables indexed by X ; $N_{Y(X)}(y|x)$ denotes the corresponding cumulative distribution function, $N_{Y(X)}(y|x) = \int_{-\infty}^{y} n_{Y(X)}(y'|x)dy'$; and $n_Y(y)$ is the marginal density of the variables indexed by Y .* The form (30), involving J integrals, is the basis for exact calculations of P_1 . Alternately, write

$$(31) \qquad P_1 = \text{Prob}\,[v_1 + y_1 > 0 \quad \text{and} \quad v_1 + y_1 > \max_{j=2,\ldots,J} (v_j + y_j)] \quad .$$

The Clark method considers trivariate normal random variables (X_1,X_2,X_3) ,

---

*As a shorthand, the set of all indices, or the set of all indices excluding those on which a distribution is conditioned, are omitted. Thus, $N_{(1)}$ means $N_{2,\ldots,J(1)}$

and approximates the bivariate distribution of $(X_1, \max(X_2, X_3))$ by a bivariate normal distribution with the same first and second moments. The approximation rests on the fact that these moments for $(X_1, \max(X_2, X_3))$ can be calculated exactly in a straightforward manner. Applied recursively to the expression

$$(32) \qquad Y_0 = \max(v_2 + y_2, \max(v_3 + y_3, \ldots, \max(v_{J-1} + y_{J-1}, v_J + y_J) \ldots)) ,$$

the method allows the distribution of (y_1, y_0) to be approximated by a bivariate normal distribution $n_1(y_1) n_{0(1)}(y_0|y_1)$, so that (30) is approximated by the univariate integral

$$(33) \qquad P_1 = \int_{y_1=-v_1}^{\infty} n_1(y_1) N_{0(1)}(v_1 + y_1|y_1) dy_1 ,$$

where $N_{0(1)}(y_0|y_1) = \int_{-\infty}^{y_0} n_{0(1)}(y_0'|y_1) dy_0'$. Thus, an MNP choice probability for $J + 1$ alternatives is approximated by a univariate integral involving a univariate normal density and univariate normal cumulative distribution function (which can be accurately approximated computationally by a series expansion). The approximation requires $J - 2$ applications of the Clark formula.

Manski (1976) has reported good results in maximum likelihood search methods using the approximation above, with search directions determined by numerical evaluation of derivatives. This suggests that the bias caused by the approximation is relatively stationary for evaluation of probabilities at neighboring points. This fortuitous conclusion suggests that it is probably unnecessary to obtain analytic derivatives of P_1 with respect to parameters in statistical routines. On the other hand, it is possible that the use of analytic derivatives could decrease computation time. The following argument shows that the Clark procedure can be applied to yield quick approximations to analytic derivatives.

From (30),

$$(34)\qquad \frac{\partial P_1}{\partial \theta} = n_1(-v_1)N_{(1)}((-v_j)|-v_1)\,\frac{\partial v_1}{\partial \theta}$$

$$+ \sum_{j=2}^{J} \frac{\partial (v_1 - v_j)}{\partial \theta} \int_{y_1=-v_1}^{\infty} n_{1j}(y_1, v_1 - v_j + y_1) N_{(1j)}((v_1 - v_k + y_1)|y_1, v_1 - v_j + y_1)\,dy_1$$

The term $N_{(1)}((-v_j)|-v_1)$ can be approximated by applying the Clark procedure to the conditional distribution. The integrals in the last right-hand term of (34) each have the essential structure of (30), since $n_{1j}(y_1, v_1 - v_j + y_1)$ is proportional to a normal density for y_1 whose mean and variance are computed by a straightforward completion of the square. Then, each integral in this term can be approximated by the corresponding analogue of (33). We conclude that the analytic derivative $\partial P_1/\partial \theta$ can be computed by the evaluation of J univariate integrals, each with the generic form of (33), and each involving $J - 3$ applications of the Clark procedure. For most problems, where the number of parameters exceeds J, this computation should be considerably faster than numerical computation of the derivatives.

The probability P_1 also depends parametrically on the covariance matrix Ω. The requirement that Ω be positive definite can be imposed by writing $\Omega^{-1} = TT'$, where $T = (\tau_{ij})$ is a lower triangular matrix with positive diagonal elements. Then, $|\Omega|^{-1/2} = \tau_{11} \cdots \tau_{JJ}$. Alternately, Ω^{-1} may be represented as an unknown nonnegative linear combination, with full rank, of known positive semi-definite matrices. Analogues for the parameters of Ω^{-1} of the analytic derivatives above are computationally complex, and their use appears unlikely to improve significantly on numerical differentiation.

The key to the accuracy of the Daganzo-Bouthelier-Sheffi approximation is the accuracy of the Clark procedure. Because the true distribution of the maximum of two normal variates is skewed to the right, one would expect the procedure to tend to underestimate small probabilities. The approximation will be best when the variates are positively correlated, with widely differing means, and worse when they are negatively correlated with similar means. It may be possible to adjust the Clark formulae empirically to improve their accuracy for computation of small probabilities. Alternately, it would be interesting to explore the possibility of adapting the Clark methodology to other trivariate distributions. In particular, if the generalized extreme value distribution were utilized, then the only point of approximation would be the initial fit to the multivariate normal density, since maxima of GEV distributed variates are again GEV distributed. This

would limit approximation error as J increases, in contrast to the Clark procedure which becomes less accurate with large numbers of alternatives.

III STATISTICAL ESTIMATION METHODS AND SAMPLING STRATEGIES

1 Maximum Likelihood Estimation

The statistical estimation of disaggregate choice models by the maximum likelihood method is now well-established. For random samples of individuals, this procedure can be shown in general to produce estimates with good statistical properties, at least in large samples. The problems remaining in application of maximum likelihood estimation in this context are primarily computational — the issues of rapid computation of choice probabilities, the concavity or unimodality of the log likelihood function, and the relative convergence speed of alternative algorithms.

Estimation of sequential models, with inclusive values obtained using estimates from earlier stages of the model, has been carried out by many investigators (e.g., Domencich and McFadden (1975), Ben-Akiva (1973)) treating each stage as an independent estimation problem. This procedure neglects the fact that the use of inclusive value measures which are themselves statistics change the asymptotic distribution of the estimators, and leads to biased estimates of the standard errors of the estimators. This problem has been pointed out by Amemiya (1976), who provides the corrected asymptotic estimators for the standard errors of the estimates.

2 Estimation in Choice-Based Samples

Several recent papers have considered the problem of statistical estimation of choice models using data collected by sampling procedures other than random sampling. Of particular interest are choice-based samples, utilizing data collected from "on-board" or "destination" surveys. Such data sources are often available to transportation analysts from marketing and operations departments of operating agencies, or can be commissioned at low cost relative to random household surveys. Ierman and Manski (1976) have shown that treating choice-based samples as if they were random and calculating estimators appropriate to random samples will generally yield inconsistent estimates.* They introduce a weighted likelihood function whose maximization is shown to yield consistent estimates.

Manski and McFadden (1977) have considered more generally the problem of estimation of discrete choice models under alternative sample designs. The discrete choice problem can be defined by a finite set

---

*In an MNL model with alternative-specific dummy variables, the inconsistency is confined to the dummy variable coefficients.

C of mutually exclusive alternative responses, a space of attributes Z , assumed to be a subset of a finite-dimensional vector space, a generalized probability density, $p(z)$ $[z \in Z]$, g ving the distribution of attributes in the population, and a response probability, or choice probability, $P(i|z,\theta^*)$, specifying the conditional probability of selection of alternative $i \in C$, given attributes $z \in Z$. Prior knowledge of causal structure is assumed to allow the analyst to specify the response model $P(i|z,\cdot)$ up to a parameter vector θ^* contained in a subset Θ of a finite-dimensional vector space. The analyst's problem is to estimate θ^* from a suitable sample of subjects and their associated responses.

The probability density of (i,z) pairs in the population is given by

$$(35) \qquad f(i,z) = P(i|z,\theta^*)p(z) \quad . \qquad [(i,z) \in C \times Z]$$

The analyst can draw observations of (i,z) pairs from $C \times Z$ according to one of various sampling rules. The problem of interest is first, given any sampling rule, to determine how θ^* may be estimated, and second, to assess the relative advantages of alternative sampling rules and estimation methods.

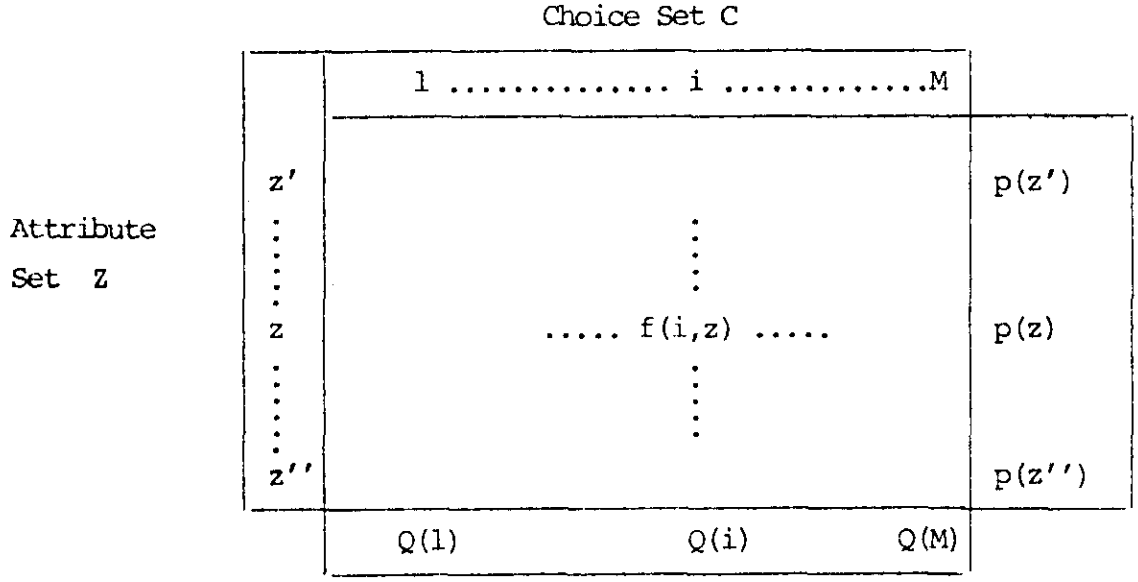
The data layout can be visualized using a contingency table, as illustrated in Figure 1. An observation (i,z) occurs in the population with frequency $f(i,z)$. The row sums give the marginal distributions of attributes $p(z)$, while the column sums give the population shares of responses $Q(i)$. The joint frequency $f(i,z)$ can be written either in terms of the conditional probability of i given z , or choice probability, or in terms of the conditional probability of z given i , as the formulae in the figure illustrate.

The feature of the quantal response problem which distinguishes it from the general analysis of discrete data is the postulate that the response probability $P(i|z,\theta^*)$ belong to a known parametric family, and reflects an underlying link from z to i which will continue to hold even if the distribution $p(z)$ of the explanatory variables changes.* Alternately, given a population $C \times Z$ with probability distribution specified by $f(i,z)$, one might, in the absence of any

---

*This postulate is fundamental to the concept of "scientific explanation." If the response probability function is invariant over populations with different distributions of attributes, then it defines a "law" which transcends the character of specific sets of data. Otherwise, the model provides only a device for summarizing data, and fails to provide a key ingredient of "explanation" — predictive power.

FIGURE 1. Contingency Table Layout of Observations



$$p(z) = \sum_{i \in C} f(i,z)$$

$$Q(i) = \int f(i,z)dz$$

$$f(i,z) = P(i|z,\theta^*)p(z) = q(z|i,\theta^*)Q(i)$$

knowledge of the process relating i's to z's , obtain a random sample from $C \times Z$ and directly examine the joint distribution $f(i,z)$. This exploratory data analysis approach is exemplified by the literature on associations in contingency tables, where it is assumed only that Z is finite. See, for example, Goodman and Kruskal (1954), Haberman (1974), and Bishop, Fienberg, and Holland (1975).

If one believes that the elements of C index conceptually distinct populations of z values, then the natural analytical approach is to decompose $f(i,z)$ into the product $f(i,z) = q(z|i)Q(i)$, where $q(z|i)$ gives the distribution of 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 $q(z|i)$ up to a parametric family, and a sample suitable for estimating the unknown parameters is obtained from the subpopulation i . See, for example, Anderson (1958) and Kendall and Stuart (1976).

When a well-defined process generates a value from C given any $z \in Z$, then the decomposition $f(i,z) = P(i|z,\theta^*)p(z)$ is appropriate. This decomposition, and the attending focus on the structural relation embodied in $P(i|z,\theta^*)$, is clearly the natural one for the analysis of choice data. A separate and interesting question is whether specific parametric models permit estimation of the parameter vector θ^* of $P(i|z,\theta^*)$ from convenient parameterizations of $f(i,z)$ or $q(z|i)$.

Manski and McFadden attempt to provide a general theory of estimation for quantal response models. The scope of the investigation is as follows: Consider the problem of estimating θ^* from stratified samples of (i,z) observations. A stratified sampling process is one in which the analyst establishes an index set B , partitions $C \times Z$ into mutually exclusive and exhaustive measurable subsets $(C \times Z)_b$, $b \in B$, and specifies a suitable probability distribution over B . To obtain an (i,z) observation, he draws a subset of $C \times Z$ according to the specified distribution and then samples at random from within the drawn subset.

Within the class of all stratification rules, two symmetric types of stratification are of particular statistical and empirical interest. In "exogenous" sampling, the analyst partitions Z into subsets Z_b , $b \in B$, and lets $(C \times Z)_b = C \times Z_b$. In "endogenous" or "choice-based" sampling, he partitions C into subsets C_b , $b \in B$, and lets $(C \times Z)_b = C_b \times Z$. Less formally, in exogenous sampling the analyst selects decision-makers and observes their choices while in choice-based sampling the analyst selects alternatives and

observes decision-makers choosing them. In Figure 1, exogenous sampling corresponds to stratifying on rows, and then sampling randomly from each row, while choice-based sampling corresponds to stratifying on columns, and then sampling randomly from each column.

Manski and McFadden make a detailed statistical examination of maximum likelihood estimation of θ^* in both exogenous and choice-based samples. They find that application of maximum likelihood is wholly classical in exogenous samples. In choice-based samples, however, the form of the maximum likelihood estimate (MLE) depends crucially on whether the analyst has available certain prior information, namely, the marginal distributions $p(z)$, $z \in Z$, or $Q(i)$, $i \in C$, where $Q(i) = \int_Z P(i|z,\theta^*)p(z)dz$.

The maximum likelihood estimator of θ in a choice-based sample when p is known and Q is unknown satisfies

$$(36) \qquad \underset{\theta\in\Theta}{\text{Max}} \sum_{n=1}^{N} \log P(i_n|z_n,\theta) - \sum_{n=1}^{N} \int_Z P(i_n|z,\theta)p(z)dz \quad .$$

When Q and p are both known, (36) is maximized subject to the constraints

$$(37) \qquad Q(i) = \int_Z P(i_n|z,\theta)p(z)dz \quad .$$

When p is unknown, the classical conditions for maximum likelihood estimation are not met. However, several alternative non-classical maximum likelihood and pseudo-maximum likelihood methods are available which yield consistent estimators.

When Q is known and p is unknown, S. Cosslett (1977) has shown that the non-classical full information maximum likelihood estimator satisfies

$$(38) \qquad \underset{\theta\in\Theta}{\text{Max}} \underset{\lambda>0}{\text{Min}} \sum_{n=1}^{N} \log \left\{ P(i_n|z_n,\theta) \left(\sum_{j\in C} \lambda_j Q(j) \right) \Big/ \left(\sum_{j\in C} \lambda_j P(j|z_n,\theta) \right) \right\} \quad .$$

A second estimator, introduced by Lerman and Manski (1976) and termed WESML, satisfies

$$(39) \qquad \underset{\theta\in\Theta}{\text{Max}} \sum_{n=1}^{N} w(i_n) \log P(i_n|z_n,\theta) \quad ,$$

where $w(i) = Q(i)/H(i)$ and $H(i)$ is the sampling frequency for alternative i . Two other consistent estimators, introduced by Manski and McFadden, satisfy

$$(40)\qquad \underset{\theta\in\Theta}{\text{Max}} \left\{ \sum_{n=1}^{N} \log P(i_n|z_n,\theta) - \sum_{n=1}^{N} \log \sum_{j\in C} \frac{Q(j)}{N_j} \sum_{m\in N(j)} P(i_n|z_m,\theta) \right\} ,$$

where $N(j)$ is the set of observations where alternative j is chosen and N_j is the number of elements in $N(j)$; and

$$(41)\qquad \underset{\theta\in\Theta}{\text{Max}} \sum_{n=1}^{N} \log \left\{ P(i_n|z_n,\theta) \frac{H(i_n)}{Q(i_n)} \Big/ \sum_{j\in C} P(j|z_n,\theta) \frac{H(j)}{Q(j)} \right\} .$$

If both p and Q are unknown in a choice-based sample, then provided an identification condition is satisfied,* Manski and McFadden show that the non-classical full-information maximum likelihood estimator satisfies

$$(42)\qquad \underset{\theta\in\Theta}{\text{Max}}\ \underset{\lambda\geq 0}{\text{Max}} \sum_{n=1}^{N} \log \left\{ P(i_n|z_n,\theta)\lambda_{i_n} \Big/ \sum_{j\in C} P(j|z_n,\theta)\lambda_j \right\} .$$

A second consistent estimator for this case is obtained by maximizing (40), with Q determined as a solution to the equations

$$(43)\qquad Q(i) = \sum_{j\in C} Q(j) \frac{1}{N_j} \sum_{m\in N(j)} P(i|z_m,\theta) .$$

Note that one can, with some loss of efficiency, obtain consistent estimates for an information case by using a consistent estimator which ignores some available information. For example, the estimators (38) or (39) could be used in the case both p and Q known, and the estimator (42) could be used in any of the information cases.

---

*An important case in which the identification condition fails is the MNL model, where in the absence of a knowledge of Q there is a confounding of the effects of Q and alternative-specific dummies.

3 Selection of a Sample Design and Estimation Method

Sample designs and estimation methods differ in terms of sampling and computation costs, and precision in parameter estimates and forecasts. Cost comparisons are situation-specific, and only a few general observations can be made. Comparison of the precision of alternative estimators can be made for large samples using the asymptotic covariance matrices of the estimators. In a few cases, the difference of two covariance matrices is positive semidefinite for all possible parameter vectors, and a uniform ranking can be made. More generally, rankings will depend on the true parameter vector and on the true distribution of explanatory variables. Then, rankings of designs and estimators will usually require a Bayesian approach utilizing a priori beliefs on the distributions of parameters, perhaps based on pilot samples and previous studies.

Consider sampling costs. In general, substantial economies can be achieved by stratifications designed to make it easier to locate and observe subjects. For example, exogenous cluster sampling, in which respondents are clustered geographically, reduces interviewer access time. Stratification on other exogenous variables, such as employer, may also reduce location cost. In many applications, choice-based sampling greatly simplifies locating subjects. For example, subjects choosing alternative travel modes can be sampled economically at the site of choice. Choice-based sampling has the greatest potential economy in applications where some responses are rare (e.g., choice of a seldom used travel mode) or are difficult to observe accurately in an exogenously drawn sample.

Computation costs are comparable in most of the estimation methods considered by Manski and McFadden. The primary component of computation costs is usually the evaluation of response probabilities at each sample point. For some models (e.g., linear), this cost is minimal, for others (e.g., multinomial logit), moderate, and for some (e.g., multinomial probit), substantial.

4 Sample Designs

Consider the precision of estimates obtained by alternative methods from alternative sample designs. We note first that the level of precision, and possibly the ranking of alternatives, will depend on the prior information available on the marginal distributions p and Q. We shall assume the state of this information is fixed. However, it should be noted that in practice the question of drawing observations on p or Q at some cost in order to utilize more efficient estimators of the response probability function may be an important part of the overall design decision.

S. Cosslett (1977) has investigated the efficiency of alternative choice-based sample designs and estimators for binary probit, logit, and arctan models with a single explanatory variable. All three models have form $P(1|z,\theta) = \psi(\theta z)$, where

$$(44)\qquad \psi(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{y} e^{-x^2/2}\, dx & \text{for probit,} \\ 1/(1 + e^{-y}) & \text{for logit,} \\ \dfrac{1}{2} + \dfrac{1}{\pi} \tan^{-1} y & \text{for arctan,} \end{cases}$$

and z is assumed to be normally distributed with mean 2 and variance 1/2 . Choice-based sample designs vary in the proportion of the sample $H(1)$ drawn from the subpopulation choosing alternative 1 . The optimal sample design for any estimator is determined by the value of $H(1)$ which minimizes the asymptotic variance of the estimator.

We concentrate on the case with p unknown and Q known. The maximum likelihood estimator (38) with an optimal sample design provides a standard against which other estimators and sample designs can be measured. Define the asymptotic efficiency of an alternative estimator and sample design to be the asymptotic variance of the maximum likelihood estimator with optimal design, divided by the asymptotic variance of the alternative estimator.

Consider as alternative estimators the WESML estimator (39), the Manski-McFadden estimator (41), and the "conditional" maximum likelihood estimator (42), which does not use information on Q . Table 2 gives the asymptotic efficiencies of these estimators for each model for selected values of θ . Three sample designs are considered: "pseudo-random" sampling in proportion to population shares, $H(1) = Q(1)$; sampling equally from each alternative, $H(1) = 1/2$; and sampling optimally for the estimator. The optimizing values of $H(1)$ for these estimators are also given in the table.

The results in Table 2 suggest the following conclusions:

(1) Knowledge of the aggregate share $Q(1)$ is of great value when the maximum likelihood estimator (38) is used, as indicated by the low efficiency of the conditional maximum likelihood estimator (42) which does not utilize this knowledge. Note however, that the information contained in $Q(1)$ will be greatest for a one-variable model without an alternative-specific dummy, and in general the efficiency differential will be smaller.

(2) The Manski-McFadden estimator (41) is uniformly more efficient than the WESML estimator (39), but the differential is small when the true parameter value is small. Both (39) and (41) have low efficiency relative to maximum likelihood for small parameter values, but (41) is relatively efficient for large parameter values.

TABLE 2. Asymptotic Efficiency of Choice-Based Sample Designs and Estimators*

	Pseudo-Random Design, H(1)=Q(1)	Equal Shares, H(1) = 1/2	Optimal Design	Optimal Value of H(1)
Probit Model				
Q(1) = .75**				
MLE (38)	87.1%	95.0%	100.0%	0.13
MM (41)	3.1	4.5	4.5	0.46
WESML (39)	3.1	4.4	4.4	0.47
Cond ML (42)	0.4	0.6	0.6	0.49
Q(1) = .9				
MLE (38)	62.1%	95.2%	100.0%	0.30
MM (41)	6.3	20.5	21.0	0.42
WESML (39)	6.3	19.1	19.2	0.47
Cond ML (42)	1.3	3.6	3.7	0.45
Q(1) = .95				
MLE (38)	40.7%	95.5%	100.0%	0.34
MM (41)	6.1	37.9	39.2	0.39
WESML (39)	6.1	32.0	32.0	0.50
Cond ML (42)	1.6	7.5	7.6	0.43
Q(1) = .99				
MLE (38)	9.5%	96.9%	100.0%	0.38
MM (41)	3.4	78.2	81.2	0.38
WESML (39)	3.4	36.8	40.6	0.66
Cond ML (42)	1.4	17.8	17.8	0.46
Q(1) = .995				
MLE (38)	4.5%	98.4%	100.0%	0.42
MM (41)	2.6	91.2	92.9	0.41
WESML (39)	2.6	23.4	30.1	0.77
Cond ML (42)	1.4	23.0	23.0	0.49
Logit Model				
Q(1) = .75**				
MLE (38)	86.7%	94.5%	100.0%	0.09
MM (41)	2.9	4.0	4.0	0.47
WESML (39)	2.9	4.0	4.0	0.48
Cond ML (42)	0.3	0.4	0.4	0.50

TABLE 2, continued

	Pseudo-Random Design, H(1)=Q(1)	Equal Shares, H(1) = 1/2	Optimal Design	Optimal Value of H(1)
Q(1) = .9				
MLE (38)	62.2%	94.3%	100.0%	0.26
MM (41)	5.2	16.1	16.2	0.44
WESML (39)	5.2	15.1	15.1	0.48
Cond ML (42)	0.8	1.8	1.8	0.50
Q(1) = .95				
MLE (38)	41.5%	94.7%	100.0%	0.30
MM (41)	4.9	28.9	29.5	0.42
WESML (39)	4.9	24.8	24.8	0.51
Cond ML (42)	0.9	3.4	3.4	0.50
Q(1) = .99				
MLE (38)	9.0%	95.0%	100.0%	0.35
MM (41)	2.7	66.5	69.3	0.38
WESML (39)	2.7	31.7	34.4	0.65
Cond ML (42)	0.9	8.9	8.9	0.50
Q(1) = .995				
MLE (38)	3.9%	95.7%	100.0%	0.37
MM (41)	2.1	83.4	86.9	0.37
WESML (39)	2.1	20.3	25.6	0.76
Cond ML (42)	1.0	12.9	12.9	0.51
Arctan Model				
Q(1) = .75**				
MLE (38)	83.5%	91.3%	100.0%	0.00
MM (41)	1.8	2.4	2.4	0.49
WESML (39)	1.8	2.4	2.4	0.49
Cond ML (42)	0.08	0.09	0.09	0.55
Q(1) = .9				
MLE (38)	52.9%	84.0%	100.0%	0.00
MM (41)	1.7	4.7	4.7	0.49
WESML (39)	1.7	4.6	4.6	0.51
Cond ML (42)	0.04	0.04	0.05	0.73
Q(1) = .95				
MLE (38)	27.0%	78.1%	100.0%	0.00
MM (41)	1.1	5.9	5.9	0.49
WESML (39)	1.1	5.4	5.4	0.53
Cond ML (42)	0.04	0.03	0.04	0.83

TABEL 2, continued

	Pseudo-Random Design, H(1)=Q(1)	Equal Shares H(1) = 1/2	Optimal Design	Optimal Value of H(1)
Q(1) = .99				
MLE (38)	2.2%	64.4%	100.0%	0.00
MM (41)	0.5	11.8	11.9	0.46
WESML (39)	0.5	5.4	5.9	0.65
Cond ML (42)	0.04	0.03	0.05	0.94
Q(1) = .995				
MLE (38)	0.8%	60.0%	100.0%	0.00
MM (41)	0.4	21.2	21.2	0.42
WESML (39)	0.4	4.1	5.1	0.74
Cond ML (42)	0.05	0.03	0.07	0.96

---

*Adapted from S. Cosslett (1977). Asymptotic efficiency is defined by the ratio of asymptotic variances, with the optimal choice-based sample design maximum likelihood estimator as the standard. Note that when Q(1) is not observed, the estimators (38), (39), and (41) are not available, and (42) is asymptotically efficient.

**For the one-parameter model, knowledge of p(z) and Q(1) determines θ ; for comparability of models, Q(1) rather than θ has been given.

(3) The equal shares sample design is generally quite efficient for maximum likelihood estimation, and for all the estimators yields efficiencies comparable to those for the optimal sample designs. The behaviour of optimal $H(1)$ is sensitive to the model and to the parameter value. Hence, in the absence of strong prior knowledge on parameter values, the equal shares sample design is recommended.

Table 3 compares the relative efficiencies of a choice-based sample design with equal shares and an exogenous random sample design. For $Q(1)$ known, the choice-based design is always more efficient. For $Q(1)$ unknown, the choice-based design is less efficient for the arctan model, and for small parameter values in the remaining models. Given prior beliefs on the correct model and on the value of $Q(1)$, and given a relative cost r of collecting an observation from an equal shares choice-based sample compared with an exogenous random sample, maximum efficiency subject to a fixed sampling budget will be achieved with the choice-based design if and only if the relative efficiency given in Table 3 exceeds r.

5 Estimation When Alternatives are Sampled Randomly from the Full Choice Set

A particularly advantageous use of choice-based sampling, either in primary data collection, or in synthesizing and reducing existing data sets, is in estimation of the MNL model from data on a strict subset of the full choice set. This method can greatly reduce the magnitude of data to be collected and analyzed, with attendant savings in time and cost. The property that a choice model can be calibrated consistently using data on a strict subset of the choice set is unique to the MNL model, and is a characterization of the independence from irrelevant alternatives (IIA) property of this model.

The following summary is drawn from McFadden (1977). Let C denote the full choice set. We shall assume it does not vary over the sample; however, this is inessential and can easily be generalized. Let $P(i|C,z,\theta^*)$ denote the true selection probabilities. We assume the choice probabilities satisfy the independence from irrelevant alternatives (IIA) assumption,

$$(45) \qquad i \in D \subseteq C \Rightarrow P(i|C,z,\theta) = P(i|D,z,\theta) \sum_{j \in D} P(j|C,z,\theta)$$

which characterizes the MNL model.

Now suppose for each case, a subset D is drawn from the set C according to a probability distribution $\pi(D|i,z)$ which may, but need not, be conditioned on the observed choice i. The observed choice may be either in or out of the set D. Examples of π distributions are (1) choose a fixed subset D of C, independent

TABLE 3. Relative Efficiency of Choice-Based Sample Design (with Equal Shares) and Exogenous Random Sample Design, with Maximum Likelihood Estimators*

	Q(1)	Q(1) known	Q(1) unknown
Probit	.75	1.09	0.19
	.9	1.53	0.57
	.95	2.35	1.23
	.99	10.23	5.30
	.995	21.73	8.83
Logit	.75	1.09	0.16
	.9	1.52	0.35
	.95	2.28	0.69
	.99	10.54	3.28
	.995	24.68	6.26
Arctan	.75	1.09	0.05
	.9	1.59	0.02
	.95	2.90	0.03
	.99	29.04	0.06
	.995	75.45	0.10

*Relative efficiency equals the asymptotic variance of the exogenous random sample maximum likelihood estimator divided by that of the choice-based equal share design maximum likelihood estimator. A ratio exceeding one indicates that the choice-based design is more efficient.

of the observed choice, (2) choose a random subset D of C, independent of the observed choice, and (3) choose a subset D of C, consisting of the observed choice i and one or more other alternatives, selected randomly.

We give several examples of distributions of type (3):

(3-1) Suppose D is comprised of i plus a sample of alternatives from the set $C \setminus \{i\}$, obtained by considering each element of this set independently, and including it with probability p. Then, the probability of D will depend solely on the number of elements $K = \#(D)$ it contains, and is given by the binomial formula

$$\pi(D|i,z) = p^{K-1}(1-p)^{J-K} \quad \text{if } i \in D \text{ and } K = \#(D), \tag{46}$$
$$= 0 \quad \text{if } i \notin D,$$

where J is the number of alternatives in C. For example, the probability that D will be any two-alternative set containing i as one alternative is $(J-1)p(1-p)^{J-2}$.

(3-2) Suppose D is always selected to be a two-element set containing i and one other alternative selected at random. If J is the number of alternatives in C, then

$$\pi(D|i,z) = \frac{1}{J-1} \quad \text{if } D = \{i,j\} \text{ and } j \neq i, \tag{47}$$
$$= 0 \quad \text{otherwise.}$$

(3-3) Suppose C has four elements, and

$$\pi(\{1,4\}|4) = \pi(\{1,4\}|1) = \pi(\{2,3\}|2) = \pi(\{2,3\}|3) = 1, \tag{48}$$
$$\text{and } \pi(D|i) = 0 \text{ otherwise.}$$

(3-4) Suppose C is partitioned into sets $\{C_1,\ldots,C_M\}$, with J_m elements in C_m, and suppose D is formed by choosing i (from the partition set C_n) and one randomly selected alternative from each remaining partition set. Then,

$$(49)\qquad \pi(D|i,z) = J_n \Big/ \prod_{m=1}^{M} J_m \quad \text{if } i \in D,\ M = \#(D),\ \text{and } D \cap C_m \neq \emptyset \text{ for } m = 1,\ldots,M,$$
$$= 0 \quad \text{otherwise.}$$

The π distributions of the type (1), (2), and (3-1) to (3-4) all satisfy the following basic property, which guarantees that if an alternative j appears in an assigned set D, then it has the logical possibility of being an observed choice from the set D, in the sense that the assignment mechanism could assign the set D if a choice of j is observed:

Positive conditioning property: If $j \in D \subseteq C$ and $\pi(D|i,z) > 0$, then $\pi(D|j,z) > 0$.

The π distributions (1), (2), and (3-1) to (3-3), but not (3-4), satisfy a stronger condition:

Uniform conditioning property: If $i,j \in D \subseteq C$, then $\pi(D|i,z) = \pi(D|j,z)$.

A distribution with the uniform conditioning property can be written $\pi(D|i,z) = \phi(D,z)X_D(i)$, where $X_D(i)$ equals one for $i \in D$, and zero otherwise.

Consider a sample $n = 1,\ldots,N$, with the alternative chosen on case n denoted by i_n, and D_n denoting the choice set assigned to this case from the distribution $\pi(D|i_n,z_n)$. Observations with an observed choice not in the assigned set of alternatives are assumed to be excluded from the sample. Write the multinomial logit model in the form

$$(50)\qquad P(i|C,z,\theta) = e^{V_i(z,\theta)} \Big/ \sum_{j \in C} e^{V_j(z,\theta)},$$

where $V_i(z,\theta)$ is the strict utility of alternative i.

THEOREM 2. If $\pi(D|i,z)$ satisfies the positive conditioning property and the choice model is multinomial logit, then maximization of the modified likelihood function

$$(51) \qquad L_N = \frac{1}{N} \sum_{n=1}^{N} \log \left\{ e^{V_{i_n}(z_n,\theta)+\log \pi(D_n|i_n,z_n)} \Big/ \sum_{j\in D_n^K} e^{V_j(z_n,\theta)+\log \pi(D_n|j,z_n} \right.$$

yields, under normal regularity conditions, consistent estimates of θ^*. When $\pi(D|i,z)$ satisfies the uniform conditioning property, then (51) reduces to the standard likelihood function,

$$(52) \qquad L_N = \frac{1}{N} \sum_{n=1}^{N} \log \left\{ e^{V_{i_n}(z_n,\theta)} \Big/ \sum_{j\in D_n} e^{V_j(z_n,\theta)} \right\} .$$

The theorem above assumes the assigned choice set for an observation may depend on the observed choice set and environment for the observation, but is independent of other observations. More generally, a set of observed choices may be used to define the assigned choice set for each observation. For example, a common procedure is to assign to all observations in a traffic analysis zone the set consisting of all the chosen alternatives observed for this zone. Assume there are N zones, with K_n observations in zone n. If $K_n \to \infty$ for each n, then every alternative in C will eventually be chosen by some subject in a zone, and estimators maximizing (52), with assigned sets equal to the set of observed choices for the zone, will have the same asymptotic properties as a maximum likelihood estimator for choice from the full set C. Thus, standard maximum likelihood estimation with assigned choice sets given by the set of chosen alternatives in a zone yields consistent estimates under normal regularity conditions and the usual sampling method where the number of observations in each zone becomes large when the overall sample size becomes large.

In the less common case where the number of zones N becomes large, but the number K of observations in each zone is fixed, the procedure above fails in general to yield consistent estimators.* Let $\lambda(\underset{\sim}{i}_n, \underset{\sim}{z}_n, \theta)$ denote the kernel of the "likelihood" function for zone n, where $\underset{\sim}{i}_n = (i_{1n},\ldots,i_{Kn})$ and $\underset{\sim}{z}_n = (z_{1n},\ldots,z_{Kn})$ are observed in the zone. Define $D_n = D(\underset{\sim}{i}_n) = \{j | j = i_k \text{ for some } k = 1,\ldots,K\}$

---

*I have benefitted from discussions with Joel Horowitz on this problem.

and $J(D) = \{\underset{\sim}{j} | D = \cup_k \{j_k\}\}$. In the standard case,

$$\lambda(\underset{\sim}{i}_n, \underset{\sim}{z}_n, \theta) = \prod_{k=1}^{K} P(i_{kn} | D_n, z_{kn}, \theta) = e^{\sum_k V_{i_{kn}}(z_{kn}, \theta)} \Big/ \sum_{\underset{\sim}{j} \in D_n} e^{\sum_k V_{j_k}(z_{kn}, \theta)} .$$

Asymptotically, the likelihood function is the expectation in $\underset{\sim}{z}$ of terms of the form

$$(53) \qquad \sum_{D \subseteq C} a(D) \sum_{\underset{\sim}{j} \in J(D)} \frac{e^{\sum_k V_{i_k}(z_k, \theta^*)}}{\sum_{\underset{\sim}{j} \in J(D)} e^{\sum_k V_{j_k}(z_k, \theta^*)}} \log \lambda(\underset{\sim}{i}, \underset{\sim}{z}, \theta) ,$$

where $a(D) = \prod_{k=1}^{K} P(D|C, z_k, \theta^*) \sum_{\underset{\sim}{j} \in J(D)} e^{\sum_k V_{j_k}(z_{kn}, \theta^*)}$. Consistency requires that (53) be maximized at $0 = \theta^*$. When $J(D) \neq D^K$, the standard case fails to give this result. However, consistency can be attained by using a modified likelihood function with the kernel

$$\lambda(\underset{\sim}{i}_n, \underset{\sim}{z}_n, \theta) = e^{\sum_k V_{i_{kn}}(z_{kn}, \theta)} \Big/ \sum_{\underset{\sim}{j} \in J(D_n)} e^{\sum_k V_{j_k}(z_{kn}, \theta)} .$$

To illustrate the impact of these results, consider a destination choice problem in which individuals face a CBD destination and a large number of suburban destinations. One is interested primarily in whether the CBD destination will be chosen. If an individual chooses the CBD destination, then he is assigned the choice set consisting of the CBD destination and one suburban destination chosen at random. (From the previous analysis, we may choose the suburban destination at random from the subset of suburban destinations chosen by some individual in the home zone of the case in question.) If an individual chooses a suburban destination, he is assigned a choice set consisting of this destination and the CBD destination. Assume J suburban destinations, with probability of selection $1/J$ for each in the case of a CBD choice. The π distribution is then

$$\begin{aligned}
&\pi(\{j, \text{CBD}\} | \text{CBD}) = 1/J && \text{for } j = 1, \ldots, J , \\
(54) \qquad &\pi(\{j, \text{CBD}\} | j) = 1 && \text{for } j = 1, \ldots, J , \\
&\pi\{D\} | j) = 0 && \text{otherwise .}
\end{aligned}$$

This distribution satisfies the positive conditioning property (but not the uniform conditioning property), and hence consistent estimates can be obtained by maximizing (51), which reduces to

$$(55) \qquad \text{Max} \frac{1}{N} \sum_{n=1}^{N} \left\{ V_{i_n}(z_n,\theta) - \log \left[e^{V_{CBD}(z_n,\theta)+\log J} + e^{V_{j_n}(z_n,\theta)} \right] \right\} ,$$

where j_n is the suburban alternative chosen or assigned on observation n, and a term involving $\log J$ but independent of θ has been dropped. Alternately, if the model contains a CBD-specific dummy, then unweighted maximum likelihood gives consistent estimates of all parameters except the CBD-specific dummy, and gives a consistent estimate of the true CBD-specific dummy plus $\log J$.

6 Weighting and Estimation in Composite Samples

Transportation samples may be the result of a complex mixture of exogenous and choice-based sampling, or of the amalgamation of surveys conducted using various sampling procedures. The techniques of Lerman and Manski (1976) and Manski and McFadden (1977) can be adapted to construct consistent estimators from these samples.

Consider first the problem of working with a composite survey, made up of subsamples collected by various procedures. Provided the subsamples are identified and the sampling procedure used for each is known, maximum likelihood estimation of parameters using the combined sample is straightforward: the sample likelihood is the sum of the likelihoods of each of the subsamples, taking into account the sampling process used in each subsample. For example, the likelihood function for an exogenous stratified sample which is "enriched" by a choice-based sample for minority modes is the sum of an exogenous likelihood function for the first subsample and a choice-based likelihood function for the second subsample.* Maximization of this composite likelihood function would require modification of most standard computer routines. An alternative consistent estimator which can be calculated using a maximum likelihood program which allows weighting of the choice variable is the Lerman-Manski estimator (39), with $W(i) = 1$ for the exogenous subsample and $W(i) = Q(i)/H(i)$ for the choice-based subsample. Interestingly, the result that applying unweighted exogenous maximum likelihood estimation to an MNL model and pure choice-based sample produces inconsistency only in the alternative dummy coefficients does not carry over to the case of a composite sample when the exogenous subsample is stratified.

Next consider the problem of complex stratifications, such as would result from choice-based subsampling from a large exogenous stratified

---

*S. Cosslett (1977) has pointed out that the kernel of the composite sample likelihood will include the marginal distribution $p(z)$.

transportation survey. The general theory of estimation from stratified samples of Manski and McFadden (1977) can be applied. In the example above, a consistent estimator would be (38), with Q(i) defined to equal the marginal share of alternative i in the exogenous stratified sample rather than in the population.

7 Non-Maximum Likelihood Estimation Methods

While maximum likelihood estimators have good asymptotic statistical properties under the conditions normally imposed in transportation applications, their finite sample properties are largely unknown. There is some evidence from very limited Monte Carlo studies that maximum likelihood estimators will be unduly sensitive to observations with low calculated probabilities, and hence relatively non-robust with respect to errors in model specification or data measurement which could yield low calculated probabilities for some observed choices. These limited studies suggest that when data grouping is possible, Berkson-Theil estimators may be preferable to maximum likelihood estimators (see Domencich and McFadden (1975), p. 112). However, plausible grouping is rarely possible with transportation data. An alternative approach is to develop more "robust" estimators for individual observations. Manski and McFadden (1977) have investigated a class of such estimators, including non-linear least squares (NLLS), which satisfies (for exogenous samples)

$$(56) \qquad \min_{\theta \in \Theta} \sum_{n=1}^{N} [S_{i_n} - P(i_n | z_n, \theta)]^2 \quad ,$$

where S_i is one if i is chosen and zero otherwise. This estimator is consistent, although not as efficient as maximum likelihood estimation, and appears in Monte Carlo studies to be less sensitive than maximum likelihood estimation to outliers caused by data measurement errors. Applications to transportation data sets have not, however, resulted in significant differences between maximum likelihood and NLLS estimators.

IV MODEL EVALUATION AND VALIDATION

1 Model Evaluation

The transportation analyst usually has a number of alternative model specifications he considers a priori plausible, and wishes to determine empirically which alternative best fits the data. This calls for statistics which measure goodness-of-fit, and procedures which allow tests of hypothesized specifications.

General goodness-of-fit measures for discrete choice models which are now widely used are the log likelihood function, the likelihood ratio index, a multiple correlation coefficient, and a prediction success index.

The likelihood ratio index ρ^2 is defined by the formula

$$\rho^2 = 1 - L/L_0 \quad , \tag{57}$$

where

$$L = \sum_{n=1}^{N} \sum_{i=1}^{J} S_{in} \log P(i|z_n,\theta) \tag{58}$$

is the log likelihood function, with the S_{in} equal to one if i is chosen, zero otherwise,

$$L_0 = \sum_{n=1}^{N} \sum_{i=1}^{J} S_{in} \log Q_i \quad , \tag{59}$$

and Q_i equals the sample aggregate share of alternative i .*

When the disaggregate model parameters are estimated by non-linear least squares, an appropriate goodness-of-fit measure is the sum of squared residuals,

$$SS = \sum_{n=1}^{N} \sum_{i=1}^{J} (S_{in} - R_n P_{in}(\hat{\theta}))^2 / R_n \quad , \tag{60}$$

where R_n is the sum of S_{in} . A transformation of this statistic yields a multiple correlation coefficient of the form familiar from regression analysis,

$$R^2 = 1 - \frac{SS}{SS_0} \quad , \tag{61}$$

where

$$SS_0 = \sum_{n=1}^{N} \sum_{i=1}^{J} (S_{in} - R_n Q_i)^2 / R_n \tag{62}$$

with Q_i the sample aggregate share of mode i as before.**

---

*This likelihood ratio index is defined "about aggregate shares," and measures the explanatory power of the model beyond that of a simple constant shares model. This index is preferable to a likelihood index "about zero" reported by some computer programs, which measures the power of the model beyond that of an equal shares model. A similar comment applies to the multiple correlation coefficient.

**While the R^2 index is a more familiar concept to planners who are experienced in ordinary regression analysis, it is not as well-behaved

(continued on page)

A third method of assessing the fit of a calibrated model is to examine the proportion of successful predictions, by alternative and overall. A success table can be defined as illustrated in Table 4, with the entry N_{ij} in row i and column j giving the number of individuals who are observed to choose i and predicted to choose j.* Column sums give predicted shares for the sample; row sums give observed shares. The proportion of alternatives successfully predicted, $N_{ii}/N_{\cdot i}$, indicates that fraction of individuals expected to choose an alternative who do in fact choose that alternative. An overall proportion successfully predicted, $(N_{11} + \ldots + N_{JJ})/N_{\cdot\cdot}$, can also be calculated.

Because the proportion successfully predicted for an alternative varies with the aggregate share of that alternative, a better measure of goodness-of-fit is the prediction success index,

$$\sigma_i = \frac{N_{ii}}{N_{\cdot i}} - \frac{N_{\cdot i}}{N_{\cdot\cdot}} \quad , \tag{63}$$

where $N_{\cdot i}/N_{\cdot\cdot}$ is the proportion which would be successfully predicted

---

(continued from page)

a statistic as the ρ^2 measure, for maximum likelihood estimation. Those unfamiliar with the ρ^2 index should be forewarned that its values tend to be considerably lower than those of the R^2 index and should not be judged by the standards for a "good fit" in ordinary regression analysis. For example, values of .2 to .4 for ρ^2 represent an excellent fit.

*The formula for N_{ij} is

$$N_{ij} = \sum_{n=1}^{N} S_{in} P_{jn} \quad .$$

An alternative prediction method is to forecast that the alternative with the highest probability will be chosen.

A dot subscript indicates summation over the corresponding index, e.g.,

$$N_{i\cdot} = \sum_{j} N_{ij} \quad .$$

TABLE 4. A Prediction Success Table

		Predicted Choice 1	2	...	J	Observed Count	Observed Share
Observed Choice	1	N_{11}	N_{12}		N_{1J}	$N_{1\cdot}$	$N_{1\cdot}/N_{\cdot\cdot}$
	2	N_{21}	N_{22}		N_{2J}	$N_{2\cdot}$	$N_{2\cdot}/N_{\cdot\cdot}$
	⋮						
	J	N_{J1}	N_{J2}		N_{JJ}	$N_{J\cdot}$	$N_{J\cdot}/N_{\cdot\cdot}$
Predicted Count		$N_{\cdot 1}$	$N_{\cdot 2}$		$N_{\cdot J}$	$N_{\cdot\cdot}$	1
Predicted Share		$\frac{N_{\cdot 1}}{N_{\cdot\cdot}}$	$\frac{N_{\cdot 2}}{N_{\cdot\cdot}}$		$\frac{N_{\cdot J}}{N_{\cdot\cdot}}$	1	
Proportion Successfully Predicted		$\frac{N_{11}}{N_{\cdot 1}}$	$\frac{N_{22}}{N_{\cdot 2}}$		$\frac{N_{JJ}}{N_{\cdot J}}$	$\frac{N_{11}+\ldots+N_{JJ}}{N_{\cdot\cdot}}$	
Success Index		$\frac{N_{11}}{N_{\cdot 1}} - \frac{N_{\cdot 1}}{N_{\cdot\cdot}}$	$\frac{N_{22}}{N_{\cdot 2}} - \frac{N_{\cdot 2}}{N_{\cdot\cdot}}$		$\frac{N_{JJ}}{N_{\cdot J}} - \frac{N_{\cdot J}}{N_{\cdot\cdot}}$	$\sum_{i=1}^{J} [\frac{N_{ii}}{N_{\cdot\cdot}} - (\frac{N_{\cdot i}}{N_{\cdot\cdot}})^2]$	
Proportional Error in Predicted Share		$\frac{N_{\cdot 1}-N_{1\cdot}}{N_{\cdot\cdot}}$	$\frac{N_{\cdot 2}-N_{2\cdot}}{N_{\cdot\cdot}}$		$\frac{N_{\cdot J}-N_{J\cdot}}{N_{\cdot\cdot}}$		

if the choice probabilities for each sampled individual were assumed to equal the observed aggregate shares.* This index will usually be nonnegative, with a maximum value of $1 - N_{\cdot i}/N_{\cdot\cdot}$. If an index normally lying between zero and one is desired, (63) can be normalized by $1 - N_{\cdot i}/N_{\cdot\cdot}$.

An overall prediction success index is

$$(64) \qquad \sigma = \sum_{i=1}^{J} \frac{N_{\cdot i}}{N_{\cdot\cdot}} \sigma_i = \sum_{i=1}^{J} \left(\frac{N_{ii}}{N_{\cdot\cdot}} - \left(\frac{N_{\cdot i}}{N_{\cdot\cdot}}\right)^2 \right) \quad .$$

Again, this index will usually be nonnegative, with a maximum value of $1 - \sum_{i=1}^{J} \left(\frac{N_{\cdot i}}{N_{\cdot\cdot}}\right)^2$, and can be normalized to have a maximum value of one if desired.

In tests of model specification, one is often concerned with questions such as whether certain variables enter the determination of choice, and whether certain coefficients are equal. For example, the question of whether on-vehicle travel time is generic, or homogeneous-effect, can be formulated as the hypothesis that the coefficients of alternative-specific travel times are all equal. Such problems, where the null hypothesis is a subset of a specified universe of alternatives, can be tested conveniently using likelihood ratios, as described in Theil (1971, p. 396), and McFadden (1973).**

2 Diagnostic Tests for the MNL Model

The MNL model has significant advantages over most alternative choice models in terms of simplicity and computational efficiency, and its independence from irrelevant alternatives (IIA) property greatly facilitates estimation and forecasting. On the other hand, the IIA restriction may be invalid in some applications, resulting in

---

*In a model with alternative-specific dummies and the calibration data set, estimation of parameters imposes the condition $N_{\cdot i} = N_{i\cdot}$. If one predicted the choice probabilities for each individual to equal aggregate shares, then $N_{\cdot i}/N_{\cdot\cdot}$ would be the proportion successfully predicted to choose i . This represents a "chance" prediction rate for a model in which no variables other than alternative-specific dummies enter. Thus, σ_i measures the net contribution to prediction success of variables other than the alternative-specific dummies.

**Specification tests which are less easily performed using classical statistical methods are those in which the model corresponding to

(continued on page)

erroneous forecasts. Hence, the validity of the IIA property should be tested in each application. McFadden, Tye, and Train (1976) have developed a series of diagnostic tests for this property. One is a test of the MNL model against a "universal" alternative, approximated by an MNL-like form in which attributes of all alternatives can enter the "utility" of each alternative — this is the "universal" logit model mentioned in Section II.5. A second test is based on the implications of the IIA property that the model can be estimated consistently from a random sample of the set of all available alternatives, as discussed in Section III.5. A third class of tests examines residuals from the fitted MNL model, i.e., the differences of indicators of observed choices and the estimated probabilities of these choices. Under the hypothesis that the MNL specification is correct, these residuals will have specific mean, variance, and correlation properties which can be utilized in statistical tests.

V AGGREGATION AND FORECASTING

1 Aggregate Forecasts

An important use of disaggregate models is in policy analysis of the impacts of alternative transportation plans on operating strategies. Evaluation of these impacts usually requires forecasts of the behaviour of the aggregate population, or of specific market segments. Given an estimated choice model $P(i|z,\theta)$, the aggregate share of alternative i satisfies

$$(65) \qquad Q(i) = \int_Z P(i|z,\theta)p(z)dz \quad ,$$

where $p(z)$ is the probability distribution of the explanatory variables in the population. For a market segment, this formula applies, with $p(z)$ interpreted as the distribution of explanatory variables in the segment.

A variety of methods have been proposed for the evaluation of (65) in applications; the most practical and flexible appears to be a "Monte Carlo" procedure in which $Q(i)$ is approximated by

$$(66) \qquad Q(i) = \frac{1}{N} \sum_{n=1}^{N} P(i|z_n,\theta) \quad ,$$

where $\{z_n\}$ is a sample drawn randomly from $p(z)$. The points z_n

---

(continued from page)

the universe of alternatives cannot be specified or estimated. Examples are the question of which of two alternative measures of travel time better explain mode choice, and tests of a particular model specification such as MNL against mutually exclusive alternatives such as MNP. Methods of statistical decision theory can be applied to some of these problems; an exposition is beyond the scope of this paper.

may be from a representative sample of the population, or may themselves be synthesized from incomplete data sources, as described in part 3 below. The formula (66) can be modified to accommodate non-uniform sampling weights. For computational purposes, it is often useful to group sample points into strata with homogeneous choice probabilities. Discussions of this and alternative aggregation procedures and their properties can be found in Koppelman (1975), McFadden (1976f), and Reid (1977).

2 Aggregation by the Clark Method

In general, direct evaluation of (65) requires numerical integration over the set Z, which may be of relatively high dimension. This may be impractical even if the choice probabilities are relatively easy to compute, and the problem is compounded if evaluation of the choice probabilities is expensive.

An approach which eliminates the intermediate calculation of choice probabilities has been suggested in a specific context by McFadden and Reid (1975), and generalized by Manski and Daganzo. Suppose individuals maximize utility, with utility functions $u_i = \beta' z_i + \varepsilon_i$ for alternative i. Given a probability distribution $p(z)$ for $z = (z_1,\ldots,z_J)$ and a distribution of $(\varepsilon_1,\ldots,\varepsilon_J)$, one can construct the probability distribution of $(u_1,\ldots,u_J)$ resulting from joint variation of z and the ε_i.* Let $H(u_1,\ldots,u_J)$ denote the cumulative distribution of $(u_1,\ldots,u_J)$ and H_i its derivative with respect to u_i,

---

*The distribution of $(u_1,\ldots,u_J)$ is obtained as a multivariate convolution of the probability densities of z and of $(\varepsilon_1,\ldots,\varepsilon_J)$. For some probability distributions, such as the multivariate normal case considered below, the distribution of the convolution is known. More generally, if $\phi(t_1,\ldots,t_{JK})$ is the characteristic function of the distribution of $z = (z_{11},\ldots,z_{K1},\ldots,z_{1J},\ldots,z_{KJ})$, and $\psi(t_1,\ldots,t_J)$ is the characteristic function of the distribution of $(\varepsilon_1,\ldots,\varepsilon_J)$, with z and $(\varepsilon_1,\ldots,\varepsilon_J)$ assumed independent, then $(u_1,\ldots,u_J)$ with $u_i = \beta' z_i + \varepsilon_i$ has the characteristic function $\gamma(t_1,\ldots,t_J) = \psi(t_1,\ldots,t_J)\phi(t_1\beta_1,\ldots,t_1\beta_K,t_2\beta_1,\ldots,t_J\beta_1,\ldots,t_J\beta_K)$, or more compactly, $\gamma(t) = \psi(t)\phi(t \otimes \beta')$. The density of $(u_1,\ldots,u_J)$ can then be obtained from the inversion formula $h(u_1,\ldots,u_J) = (2\pi)^{-J}\int e^{-itu}\gamma(t)dt$. Using this expression in (67), one could carry out the computation of $Q(i)$ with a numerical

(continued on page)

$$(67) \qquad Q(i) = \text{Prob}\{u_i \geq u_j \quad \text{for} \quad j = 1,\ldots,J\}$$
$$= \int_{u=-\infty}^{+\infty} H_i(u,u,\ldots,u)du \quad .$$

Evaluation of this integral requires only a single numerical integration when H_i can be obtained analytically, and at most a J-dimensional numerical integration is required to compute $Q(i)$ when the density of H is analytic.

The procedure outlined above can be applied with particular convenience to the case where z and ε are assumed multivariate normal. This assumption, which yields the MNP model of individual choice, implies $(u_1,\ldots,u_J)$ is multivariate normal, with mean of u_i equal to $\beta'\bar{z}_i$ where $\bar{z}$ is the mean of z, and covariances $\omega_{ij} = \beta'\Sigma_{ij}\beta + \sigma_{ij}$, where $\sigma_{ij} = \text{cov}(\varepsilon_i,\varepsilon_j)$ and $\Sigma_{ij} = \text{cov}(z_i,z_j)$. Then, $Q(1)$ can be obtained from a formula analogous to (30) for this multivariate normal distribution. The Clark approximation method discussed in Section II.8 then permits rapid computation of approximate aggregate shares. Further, application of the Clark formulae to the computation of analytic derivatives of $Q(i)$ with respect to $\bar{z}$, in a manner analogous to that described in (34), would allow rapid approximation of aggregate elasticities.

3 The Distribution of Explanatory Variables

The computation of aggregate shares or elasticities requires knowledge of the distribution of the explanatory variables in the population, or in a market segment of the population. A random sample from the population of sufficient size, say from a major population survey, can meet this data requirement. However, it is often difficult to obtain current data of this type. Forecasts at future dates present a further problem, since the distribution of explanatory variables used in the forecasts should take into account shifts in explanatory variables over time.

Cosslett, Duguay, Jung, and McFadden (1977) have proposed a method of synthesizing the distribution of explanatory variables at any

---

(continued from page)

integration of dimension at most $2J$, for an extremely broad class of distributions of z and ε. Application of approximation methods to the combined integral may then allow rapid computation of aggregate probabilities, even for complex choice models.

forecast date, integrating available data sources plus information on trends. The method is particularly useful when current random survey data is unavailable, and can be applied in most urban areas using only U.S. Census data. The method utilizes a classical statistical procedure for completing contingency tables, called iterative proportional fitting, due to Deming and Stephan (1940). This procedure allows the integration of marginal information from U.S. Census tract statistics, Public Use Samples, and the Urban Transportation Planning Package. Parametric models of some variable interactions, simple trend models for shifts in the distribution over time, and exogenous forecasts for some explanatory variables allow projection of the synthesized distribution to future dates. Sampling from the constructed distribution yields a synthesized random sample for the urban area at the forecast date.

4 Calculus for Demand Elasticities

Demand elasticities encapsulate considerable information on transport demand response, and are valuable tools for policy analysis. For the multinomial logit (MNL) model, the elasticities can be expressed in relatively simple formulae. However, great care must be taken to avoid mechanical use of these formulae, and to see that the computation performed corresponds to the policy question asked. The first rules set out below hold for any choice model. We use the notation P_i for the choice probability for alternative i, and z_k^i for the k-th component of the vector of attributes of alternative i.

Rule 1 — aggregation over market segments. Aggregate elasticity equals the sum of segment elasticities, weighted by segment shares of the market. If P_i^ℓ is the choice probability for segment ℓ, $\bar{P}_i$ is the aggregate choice probability, and q_ℓ is the proportion of the population in segment ℓ, then

$$(68) \qquad \left(\frac{z_k^j}{\bar{P}_i}\frac{\partial \bar{P}_i}{\partial z_k^j}\right) = \sum_\ell q_\ell \left(\frac{z_k^j}{P_i^\ell}\frac{\partial P_i^\ell}{\partial z_k^j}\right) .$$

(Note: This is a relevant elasticity only if a policy will result in equal percentage changes for each marget segment.)

Rule 2 — aggregation over alternatives. Elasticity for a compound alternative equals the sum of component alternative elasticities, weighted by the component shares of the compound alternative. Let $\bar{P} = \sum_i P_i$ be the choice probability for a compound alternative (e.g., "all transit"). Then,

$$(69) \qquad \left(\frac{z_k^j}{\bar{P}} \frac{\partial \bar{P}}{\partial z_k^j}\right) = \sum_i \left(\frac{P_i}{\bar{P}}\right)\left(\frac{z_k^j}{P_i} \frac{\partial P_i}{\partial z_k^j}\right) .$$

(Note: This is a relevant elasticity only if a policy will result in equal percentage changes for each component alternative.)

Rule 3 — component effect. The elasticity with respect to a component of a variable equals the elasticity with respect to the variable times the component's share in the variable. Suppose $z_k^j = w_k^j + y_k^j$. Then

$$(70) \qquad \frac{y_k^j}{P_i} \frac{\partial P_i}{\partial y_k^j} = \left(\frac{y_k^j}{z_k^j}\right)\left(\frac{z_k^j}{P_i} \frac{\partial P_i}{\partial z_k^j}\right) .$$

(Note: It is particularly important in policy analysis to look only at components influenced by a policy, such as the transit fare component of total trip cost or the bus on-vehicle time component of a multi-mode trip total on-vehicle time.)

Rule 4 — multiple effect. The elasticity with respect to a policy that causes an equal percentage change in several variables equals the sum of the elasticities with respect to each variable. Suppose a policy changes z_k^j to $z_k^j(1 + t)$ for several j . Then,

$$(71) \qquad \frac{1}{P_i} \frac{\partial P_i}{\partial t} = \sum_j \left(\frac{z_k^j}{P_i} \frac{\partial P_i}{\partial z_k^j}\right) .$$

Since t is a proportional change, the left-hand-side of this equation is in elasticity form. Alternately, suppose $z_k^j = w_k^j + y_k$ for several j , and a policy changes y_k . Then,

$$(72) \qquad \frac{y_k}{P_i} \frac{\partial P_i}{\partial y_k} = \sum_j \left(\frac{y_k}{z_k^j}\right)\left(\frac{z_k^j}{P_i} \frac{\partial P_i}{\partial z_k^j}\right) .$$

This formula is obtainable from a combination of Rules 3 and 4 .

Often, combinations of these rules will be required to obtain the most relevant elasticity for a policy calculation. For example, suppose transit alternatives are disaggregated by access mode,

and the impact of a fare increase is to be assessed. The answer is given by combining Rules 2, 3, and 4 to obtain the formula

$$(73)\qquad \begin{bmatrix}\text{Elasticity of transit patronage with respect to fare}\end{bmatrix} = \sum_i \sum_j \begin{bmatrix}\text{Patronage on mode } i \text{ as a proportion of total transit patronage}\end{bmatrix} \cdot \begin{bmatrix}\text{Fare on mode } j \text{ as a proportion of total cost on } j\end{bmatrix} \cdot \begin{bmatrix}\text{Elasticity of mode } i \text{ patronage with respect to total cost on mode } j\end{bmatrix}$$

where i and j are summed over the transit access modes.

Consider elasticities for the sequential model defined in Section II.3. Taking $\theta = \lambda = 1$, these elasticities will also hold for the joint MNL model.

$$(74)\qquad \frac{x_{ncbk}}{P_{m|da}} \frac{\partial P_{m|da}}{\partial x_{ncbk}} = \alpha_k z_{ndak} (\delta_{mn} - P_{n|da}) \delta_{cd} \delta_{ab}$$

where x_{ndak} is component k of x_{nda} and $\delta_{mn} = 1$ if $m = n$, 0 otherwise, etc.

$$(75)\qquad \frac{x_{ncbk}}{P_{d|a}} \frac{\partial P_{d|a}}{\partial x_{ncbk}} = \theta \alpha_k x_{ncak} P_{n|ca} (\delta_{cd} - P_{c|a}) \delta_{ab} \ ,$$

$$(76)\qquad \frac{y_{cbk}}{P_{d|a}} \frac{\partial P_{d|a}}{\partial y_{cbk}} = \beta_k y_{cak} (\delta_{cd} - P_{c|a}) \delta_{ab} \ ,$$

where Y_{cak} is component k of Y_{ca} .

$$(77)\qquad \frac{x_{mdbk}}{P_a} \frac{\partial P_a}{\partial x_{mdbk}} = \lambda\theta\alpha_k x_{mdbk} P_{m|db} P_{d|b} (\delta_{ab} - P_b) \ ;$$

$$(78)\qquad \frac{y_{dbk}}{P_a} \frac{\partial P_a}{\partial y_{dbk}} = \lambda\beta_k y_{dbk} P_{d|b} (\delta_{ab} - P_b) \ ;$$

$$(79)\qquad \frac{z_{bk}}{P_a} \frac{\partial P_a}{\partial z_{bk}} = \gamma z_{bk} (\delta_{ab} - P_b) \ .$$

Other elasticities are readily derived from these formulae using Rules 1 -- 4 and the definitions of conditional and marginal probabilities. For example, $P_{da} = P_{d|a}P_a$ implies

$$(80) \qquad \frac{x_{mca}}{P_{da}} \frac{\partial P_{da}}{\partial x_{mca}} = \frac{x_{mca}}{P_{d|a}} \frac{\partial P_{d|a}}{\partial x_{mca}} + \frac{x_{mca}}{P_a} \frac{\partial P_a}{\partial x_{mca}} ,$$

and the preceding formulae can be substituted in the right-hand-side of this expression. Another example is the joint probability $P_{mda} = P_{m|da}P_{d|a}P_a$, which satisfies

$$(81) \qquad \frac{x_{ncbk}}{P_{mda}} \frac{\partial P_{mda}}{\partial x_{ncbk}} = \alpha_k x_{ncbk} \left\{ \delta_{mn}\delta_{cd}\delta_{ab} + \delta_{cd}\delta_{ab}(\theta - 1)P_{n|cb} + \theta\delta_{ab}(\lambda - 1)P_{n|cb}P_{c|b} - \theta\lambda P_{ncb} \right\} .$$

As a check, one sees that for $\theta = \lambda = 1$, this reduces to the conventional MNL elasticity formula.

VI CONCLUSION

This paper has surveyed selected recent developments in quantitative methods for travel demand analysis. This subject has developed rapidly in the past few years on a wide spectrum of topics. As a result, the transportation analyst now has available a greatly expanded "bag of tools" with which to address policy problems. Experience makes it clear that quantitative methods are not a panacea for solving the problems of transportation policy analysis. On the other hand, the use of techniques of modern mathematics and statistics relax one of the constraints which has limited the analyst in attacking a full range of transportation issues. A review of the state of the art of quantitative methods in transportation demand forecasting suggests that the task of establishing a firm analytic and statistical foundation for the subject has just begun.

REFERENCES

Amemiya, T. (1976), "Specification and Estimation of a Multinomial Logit Model," Technical Report 211, Institute of Mathematical Studies in the Social Sciences, Stanford University, Stanford, CA

Anderson, T.W. (1958), An Introduction to Multivariate Statistical Analysis, Wiley, NY

Ben-Akiva, M. (1973), "Structure of Passenger Travel Demand Models," Transportation Research Board Record No. 526, Washington, DC

Ben-Akiva, M. and S. Lerman (1977), "Disaggregate Travel and Mobility Choice Models and Measures of Accessibility," Third International Conference on Behavioural Travel Modelling, Australia, 1977

Bishop, Y., S. Fienberg, and P. Holland (1975), Discrete Multivariate Analysis, MIT Press, Cambridge, MA

Bock, R.D. and L.V. Jones (1968), The Measurement and Prediction of Judgement and Choice, Holden-Day, San Francisco, CA

Cardell, S. (1975), personal communication

Clark, C. (1961), "The Greatest of a Finite Set of Random Variables," Operations Research, Vol. 9, p. 145-162

Cosslett, S. (1977), "Efficient Estimation of Choice Probabilities from Choice-Based Samples," Alfred P. Sloan Foundation Workshop in Transportation Economics, Department of Economics, University of California, Berkeley, CA

Cosslett, S., G.E. Duguay, W.S. Jung, and D. McFadden (1977), "Synthesis of Household Transportation Survey Data : the SYNSAM Methodology," Working Paper No. 7705, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, CA

Cramér, H. (1946), Mathematical Methods of Statistics, Princeton University Press, Princeton, NJ

Daly, A. and S. Zachary (1976), "Improved Multiple Choice Models," mimeographed

Daganzo, C., F. Bouthelier, and Y. Sheffi (1976), "An Efficient Approach to Estimate and Predict with Multinomial Probit Models," MIT, Department of Civil Engineering, unpublished

Deming, W. and F. Stephan (1940), "On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known," Annals of Math. Stat., Vol. 11, p. 427-444

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

Goodman, L. and W. Kruskal (1954), "Measures of Association for Cross-Classification," J.A.S.A., Vol. 49, p. 732-764

Haberman, S. (1974a), The Analysis of Frequency Data, University of Chicago Press, Chicago, IL

Haberman, S. (1974b), "Log-Linear Fit for Contingency Tables," Appl. Stat., Vol. 21, p. 218-225

Harris, A. and J. Tanner (1974), "Transport Demand Models Based on Personal Characteristics," Transport and Road Research Laboratory Supplementary Report 65 UC

Hausman, J.A. and D.A. Wise (1976), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," Working Paper No. 173, Department of Economics, MIT, Cambridge, MA

Kendall, . and . Stuart (1976), Advanced Theory of Statistics, Vol. 3, Hafner

Koppelman, F. (1975), "The Structure of Aggregated Prediction Models," Ph.D. Dissertation, Department of Civil Engineering, Northwestern University, Evanston, IL

Lerman, S. and C. Manski (1976), "The Estimation of Choice Probabilities from Choice-Based Samples," forthcoming in Econometrica

Manski, C. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," J. of Econometrics, Vol. 3, p. 205-228

Manski, C. (1976), "Multinomial Probit Model" (title approx.), Cambridge Systematics, Inc., Cambridge, MA, internal memorandum

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.), Frontiers in Econometrics, Academic Press, NY

McFadden, D. (1975a), "On Independence, Structure, and Simultaneity in Transportation Demand Analysis," Working Paper No. 7511A, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, CA

McFadden, D. (1975b), "Economic Applications of Psychological Choice Models," Working Paper No. 7519, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, CA

McFadden, D. (1976a), "The Revealed Preferences of a Government Bureaucracy: Evidence," Bell Journal of Economics, Vol. 7, p. 55-72

McFadden, D. (1976b), "Quantal Choice Analysis: A Survey," Annals of Economic and Social Measurement, Vol. 5

McFadden, D. (1976c), "A Comment on Discriminant 'versus' Logit Analysis," Annals of Economic and Social Measurement, Vol. 5

McFadden, D. (1976d), "The Theory and Practice of Disaggregate Demand Forecasting for Various Modes of Urban Transportation," Working Paper 7623, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, CA

McFadden, D. (1976e), "Properties of the Multinomial Logit (MNL) Model," Working Paper No. 7617, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, CA

McFadden, D. (1976f), "The Mathematical Theory of Demand Models," Chapter 17 in A. Meyburg and P. Stopher (eds.), Behavioral Travel Demand Models, Heath-Lexington

McFadden, D., (1977), "Modeling the Choice of Residential Location," Department of Economics, University of California, Berkeley, CA

McFadden, D. and F. Reid (1975), "Aggregate Travel Demand Forecasting from Disaggregated Behavioral Models," Transportation Research Board Record No. 534, Washington, DC

McLynn, J. (1973), "A Technical Note on a Class of Fully Competitive Modal Choice Models," DTM Corp., Bethesda, MD, unpublished

Meyburg, A. and P. Stopher (1975), Urban Transportation Planning and Modelling, Heath-Lexington

Meyburg, A. and P. Stopher (1976), Behavioral Travel Demand Models, Heath-Lexington

Theil, H. (1971), Principles of Econometrics, Wiley, NY

Thurstone, L. (1927), "A Law of Comparative Judgment," Psychological Review, Vol. 34, p. 273-286

Williams, H.C.L. (1977), "On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit," Environment and Planning, Vol. A.9, p. 285-344