Chapter Four

Conditional logit analysis of qualitative choice behavior

DANIEL McFADDEN[1]

UNIVERSITY OF CALIFORNIA AT BERKELEY
BERKELEY, CALIFORNIA

I. Preferences and Selection Probabilities
II. Conditional Logit Estimation
III. Statistical Properties
IV. An Empirical Application
Shopping Choice of Mode
Shopping Choice of Destination
Shopping Trip Frequency
Appendix: Proofs of Statistical Properties
References

A fundamental concern of economics is understanding human choice behavior. Models or hypotheses are formed on the nature of decision processes, and are evaluated in the light of observed behavior. This task is complicated

[1] The research for this paper was done with the support of National Science Foundation Grant GS-27226, the Ford Foundation, the University of Chicago, and the U.S. Department of Transportation. I am indebted to P. Cottingham, J. Craig, G. Debreu, T. Domencich, R. Hall, M. Richter, and B. Saffran for useful comments at various stages of this research, but retain sole responsibility for errors.

because the econometrician cannot observe or control all the factors influencing behavior, and because the process of observation itself influences acts of the decision-maker through the vehicle of experience. It becomes necessary to make statistical inferences on a model of *individual* choice behavior from data obtained by sampling from a *population* of individuals (or sampling from a population of "experience levels" for a single individual). When the model of choice behavior under examination depends on unobserved characteristics in the population, the testable implications of the individual choice model are obscured. However, it is possible to deduce from the individual choice model properties of population choice behavior which can be subjected to empirical test.

The link between models of individual behavior and data on population choices is most critical when the decision-maker's alternatives are qualitative, or "lumpy." In conventional consumer analysis with a continuum of alternatives, one can often plausibly assume that all individuals in a population have a *common* behavior rule, except for purely random "optimization" errors, and that systematic variations in aggregate choice reflect common variations in individual choice at the *intensive* margin. By contrast, systematic variations in aggregate choice among lumpy alternatives must reflect shifts in individual choice at the *extensive* margin, resulting from a *distribution* of decision rules in the population.

This paper outlines a general procedure for formulating econometric models of population choice behavior from distributions of individual decision rules. A concrete case with useful empirical properties, conditional logit analysis, is developed in detail. The relevance of these methods to economic analysis can be indicated by a list of the consumer choice problems to which conditional logit analysis has been applied: choice of college attended, choice of occupation, labor force participation, choice of geographical location and migration, choice of number of children, housing choice, choice of number and brand of automobiles owned, choice of shopping travel mode and destination.

Section I of this paper derives the relation between individual behavioral models and the distribution of population choices, and discusses the behavioral axiom which leads to the conditional logit model. Section II discusses estimation of the conditional logit model, and Section III discusses its statistical properties. Section IV summarizes an application of the method to the problem of shopping travel mode and destination choice.

I. Preferences and Selection Probabilities

A study of choice behavior is described by (1) the objects of choice and sets of alternatives available to decision-makers, (2) the observed attributes of decision-makers, and (3) the model of individual choice and behavior and

distribution of behavior patterns in the population. Observed data are assumed to be generated by the *trial* of drawing an individual randomly from the population and recording his attributes, the set of alternatives available to him, and his actual choice. A sample is obtained by a sequence of independent trials, with or without *replications* in which a sequence of choices are observed for individuals with the same measured attributes and alternative sets.

We let X denote the universe of objects of choice, S the universe of vectors of measured attributes of decision-makers. An individual drawn at random from the population will have some attribute vector $s \in S$ and will face some set of available alternatives, which we now assume to be finite and denote by $B \subseteq X$. Let $P(x \mid s, B)$ denote the conditional probability that an individual drawn at random from the population will choose alternative x, given that he has measured attributes s and faces the alternative set B. The observed choice in a trial with attributes s and alternatives B can then be viewed as a drawing from a multinomial distribution with selection probabilities $P(x \mid s, B)$ for $x \in B$.

An individual behavior rule is a function h which maps each vector of measured attributes s and possible alternative set B into a chosen member of B. A *model* of individual behavior is a set of behavior rules H. For example, h may be a demand function resulting from maximization of a specific utility function, and H may be the set of demand functions which result from maximization of *some* utility function. With unmeasured attributes varying across the population, a model H can contain many behavior rules.

If a model H truly describes a population, then there exists a probability π defined on the (measurable) subsets of H specifying the distribution of behavior rules in the population.[2] The selection probability that an individual drawn at random from the population will choose x, given measured attributes s and alternative set B, equals the probability of occurrence of a decision rule yielding this choice, or

$$P(x \mid s, B) = \pi[\{h \in H \mid h(s, B) = x\}]. \tag{1}$$

An econometric model of qualitative choice behavior can be constructed for a specified model of individual behavior by assuming π to be a member of a

[2] To be precise, each trial represents a drawing of a triple (s, B, ω) from an underlying universe, where s is a vector of measured attributes, B is an alternative set, and ω determines a unique decision rule h_ω, with $h_\omega(s', B') \in B'$ for all possible arguments (s', B'). A probability defined on the underlying universe induces a probability π on the set of h_ω, conditioned on values of (s, B). When the pair (s, B) and ω are statistically independent [e.g., the underlying probability is a product of a probability defined on the universe of (s, B) and a probability defined on the universe of ω], the probability π is independent of the conditioning values (s, B). We confine our attention to this case, noting that satisfaction of this condition is one of the criteria for a carefully designed laboratory experiment or sample survey.

parametric family of probability distributions and using the fact that the observed choices are multinomially distributed with the probabilities (1) to obtain estimators of the underlying parameters. The following paragraphs carry through this program for the classical model of the utility-maximizing economic consumer.

Suppose an individual in the population has a vector of measured attributes s, and faces J alternatives, indexed $j = 1, \ldots, J$ and described by vectors of attributes x_j. The individual has a utility function that can be written in the form

$$U = V(s, x) + \varepsilon(s, x),$$

where V is nonstochastic and reflects the "representative" tastes of the population, and ε is stochastic and reflects the idiosyncracies of this individual in tastes for the alternative with attributes x. The individual chooses the alternative which maximizes utility; let h_ε denote his behavior rule, and $B = \{x_1, \ldots, x_J\}$. The probability that an individual drawn randomly from the population, with attributes s and alternative set B, will choose x_i equals

$$\begin{aligned} P_i \equiv P(x_i \mid s, B) &= \pi[\{h_\varepsilon \in H \mid h_\varepsilon(s, B) = x_i\}] \\ &= P[\varepsilon(s, x_j) - \varepsilon(s, x_i) < V(s, x_i) - V(s, x_j) \quad \text{for all} \quad j \neq i]. \end{aligned} \tag{2}$$

The probability π induces a joint cumulative distribution function $F(\varepsilon_1, \ldots, \varepsilon_J)$ over the values $\varepsilon_j = \varepsilon(s, x_j)$ for $j = 1, \ldots, J$; i.e.,

$$F(\varepsilon_1, \ldots, \varepsilon_J) = \pi[\{h_\varepsilon \in H \mid \varepsilon(s, x_j) \leqslant \varepsilon_j \quad \text{for} \quad j = 1, \ldots, J\}].$$

Let F_i denote the partial derivative of F with respect to its ith argument, and let $V_i = V(s, x_i)$. Then Equation (2) can be written

$$P_i = \int_{\varepsilon=-\infty}^{+\infty} F_i(\varepsilon + V_i - V_1, \ldots, \varepsilon + V_i - V_J)\, d\varepsilon. \tag{3}$$

We may proceed by specifying a joint distribution, such as joint normal, which will yield a family of probabilities depending on the unknown parameters of the distribution. It will generally be necessary to impose rather stringent maintained hypotheses on the unknown parameters to make them identifiable in a choice experiment, particularly in the absence of repetitions.

In practice, it is difficult to define joint distributions F which allow the computation of econometrically useful formulas for the P_i in Equation (3). An alternative approach is to specify formulas for the selection probabilities and then examine the question of whether these formulas could be obtained via Equation (3) from *some* distribution of utility-maximizing consumers. This problem is the population analog of the conventional theory of revealed preference for individual consumers. The author and Professor Marcel K.

Richter have elsewhere (1971) characterized the necessary and sufficient condition on selection probabilities for satisfaction of Equation (3). We shall follow this method, using a particular specification of the selection probabilities that allows direct verification of condition (3).

We consider a powerful axiom on selection probabilities introduced by Luce (1959) which states that the relative odds of one alternative being chosen over a second should be independent of the presence or absence of unchosen third alternatives. Formally, we make the following assumption.

AXIOM 1. (Independence of Irrelevant Alternatives). For all possible alternative sets B, measured attributes s, and members x and y of B,

$$P(x \mid s, \{x, y\})\, P(y \mid s, B) = P(y \mid s, \{x, y\})\, P(x \mid s, B). \tag{4}$$

We show below that this axiom is consistent with condition (3) and leads to a simple econometric specification of the selection probabilities. Luce has presented evidence that the axiom is consistent with behavior in some choice experiments; we shall point out later some of its limitations.

When $P(x \mid s, B)$ is positive, Equation (4) implies $P(x \mid s, \{x, y\})$ positive, and

$$\frac{P(y \mid s, \{x, y\})}{P(x \mid s, \{x, y\})} = \frac{P(y \mid s, B)}{P(x \mid s, B)}. \tag{5}$$

This condition states that the odds of y being chosen over x in a multiple choice situation B, where both are available, equals the odds of a binary choice of y over x.

Since empirically a zero probability is indistinguishable from one that is extremely small, there is little loss of generality in assuming that the selection probabilities are all positive for the possible alternative sets in an experiment.

AXIOM 2 (Positivity). $P(x \mid s, B) > 0$ for all possible alternative sets B, vectors of measured attributes s, and $x \in B$.

Consider a choice set B containing alternatives x, y, z, and let $p_{xy} = P(x \mid s, \{x, y\})$. Define $p_{xx} = \frac{1}{2}$. From Equation (4),

$$P(y \mid s, B) = \frac{p_{yx}}{p_{xy}} P(x \mid s, B) \tag{6}$$

and

$$1 = \sum_{y \in B} P(y \mid s, B) = \left(\sum_{y \in B} \frac{p_{yx}}{p_{xy}} \right) P(x \mid s, B). \tag{7}$$

Hence, the multiple choice selection probabilities can be written in terms of

binary odds,

$$P(x \mid s, B) = \frac{1}{\sum_{y \in B} (p_{yx}/p_{xy})}. \tag{8}$$

Permuting the indices x, y, z in Equation (6) and multiplying yields the condition

$$\frac{p_{yx}}{p_{xy}} = \frac{p_{yz}/p_{zy}}{p_{xz}/p_{zx}}. \tag{9}$$

Taking z to be a "benchmark" member of the alternative set B and defining $V(s, x, z) = \log(p_{xz}/p_{zx})$, Equation (8) can be written

$$P(x \mid s, B) = \frac{e^{V(s,x,z)}}{\sum_{y \in B} e^{V(s,y,z)}}. \tag{10}$$

In the function $V(s, x, z)$, one may think of the argument s as giving a "measured taste effect," the argument x as giving a "choice alternative effect," and the argument z as giving an "alternative set effect." In an experiment with sufficient variation in measured attributes s and the alternative set B, and replications for each (s, B) pair, one can normally identify each of these effects. In the absence of replications, it is impossible to identify the "alternative set effect," and an identifying restriction is necessary to isolate the "choice alternative effect"; we shall assume the following.[3]

AXIOM 3 (Irrelevance of Alternative Set Effect). The function $V(s, x, z)$ determining the selection probabilities in Equation (10) has the additively separable form

$$V(s, x, z) = v(s, x) - v(s, z). \tag{11}$$

Then, Equation (10) becomes

$$P(x \mid s, B) = e^{v(s,x)} \Big/ \sum_{y \in B} e^{v(s,y)}. \tag{12}$$

and the function v can be interpreted as a "utility indicator" of "representative" tastes. The following result justifies this terminology in terms of the behavior of a population of consumers.

[3] Axiom 3 follows from Axioms 1 and 2 if there exists some "universal benchmark" alternative z such that if B is a possible alternative set, then $B \cup \{z\}$ is also. This follows by noting that Equation (9) holds for $z \notin B$, provided Axioms 1 and 2 holds for $B \cup \{z\}$. Then, taking z to be the universal benchmark in Equation (10) and defining $v(s, x) = V(s, x, z)$ for all alternative sets yields the result.

LEMMA 1. Suppose each member of a population of utility-maximizing consumers has a utility function $U(s,x) = v(s,x)+\varepsilon(s,x)$, where v is a nonstochastic function reflecting "representative" tastes and $\varepsilon(s,x)$ is a function that varies randomly in the population with the property that in each possible alternative set $B = \{x_1, \ldots, x_J\}$, the values $\varepsilon(s,x_j)$ are independently identically distributed with the Weibull (Gnedenko, extreme value) distribution[4]

$$P(\varepsilon(s,x_j) \leqslant \varepsilon) = e^{-e^{-\varepsilon}}. \tag{13}$$

Then the selection probabilities given by Equation (3) satisfy Equation (12).

Proof: From Equation (13), letting $V_i = v(s,x_i)$,

$$F_i(\varepsilon+V_i-V_1, \ldots, \varepsilon+V_i-V_J)$$

$$= \exp(-\varepsilon) \prod_{j=1}^{J} \exp(-\exp(-\varepsilon-V_i+V_j))$$

$$= \exp(-\varepsilon) \exp\left\langle -[\exp(-\varepsilon)]\left[\sum_{j=1}^{J} \exp(V_j-V_i)\right]\right\rangle .$$

Substituting this expression in Equation (3) yields the result.

A nonconstructive proof of this result was first given by Marschak (1959); the argument above appears in Luce and Suppes (1965), and is attributed to E. Holman and A. Marley. The next lemma establishes that under mild conditions the distribution (13) characterizes the population choice models whose selection probabilities satisfy Equation (12). A random variable ε is said to be *translation complete* if for a function h of bounded absolute variation with $h(\pm\infty) = 0$, the condition $Eh(\varepsilon+a) = 0$ for all real a implies $h \equiv 0$ (except possibly on a set of measure zero). Most common distributions have this property; in particular, the Weibull distribution above is translation complete.[5]

[4] Monotone increasing transformations of the utility function $U(s,x)$ do not affect utility maximization or the selection probabilities, but transform the distribution of the random component. In particular, $e^{U(s,x)} = e^{V(s,x)}\eta(s,x)$ has η distributed with the reciprocal exponential distribution $P(\eta \leqslant y) = e^{-1/y}$ $(y \geqslant 0)$; $-e^{-U(s,x)} = e^{-V(s,x)}\eta(s,x)$ has η distributed with the negative exponential distribution $P(\eta \leqslant y) = e^{y}$ $(y \leqslant 0)$; and $e^{-\beta\exp[-U(s,x)]} = \eta(s,x)^{\exp[-V(s,x)]}$ has η distributed with the power distribution $P(\eta \leqslant y) = y^{1/\beta}(0 \leqslant y \leqslant 1)$. These examples demonstrate that the moments of the distribution of utility in the population (or their existence) do not provide a useful guide to the degree of dispersion of tastes. We note for later reference that the Weibull distribution (13) has the characteristic function $\Gamma(1+it)$, which is nonzero for real t, and has all positive moments finite.

[5] A distribution whose characteristic function is nonzero for real arguments is translation complete [apply Feller (1966), p. 479]; the Weibull distribution satisfies this condition.

LEMMA 2. Suppose selection probabilities are given by Equation (12) for all finite alternative sets B in a universe X, and suppose that for each vector of measured attributes s, the values of $v(s, x)$ range over the real line; i.e., $v(s, X) = (-\infty, +\infty)$. Suppose the selection probabilities satisfy Equation (3) with independently identically distributed $\varepsilon(s, x_i)$ having a translation complete cumulative distribution function G. Then, $G(\varepsilon) = e^{-\alpha \exp(-\varepsilon)}$, where α is an arbitrary positive parameter. Fixing the parameter α by specifying $G(0) = e^{-1}$ yields the distribution (13).

Proof: Consider the choice between an alternative yielding "utility" $v_x = v(s, x)$ and K alternatives, each yielding v_y. Equations (3) and (12) imply that the probability P_x of choosing x is

$$P_x = \frac{e^{v_x}}{e^{v_x} + Ke^{v_y}} = \int_{\varepsilon=-\infty}^{+\infty} G(\varepsilon + v_x - v_y)^K \, dG(\varepsilon). \tag{14}$$

On the other hand, consider a binary choice between x and an alternative z yielding $v_z = v(s, z) = v_y + \log K$, implying

$$P_{xz} = \frac{e^{v_x}}{e^{v_x} + e^{v_z}} = \int_{\varepsilon=-\infty}^{+\infty} G(\varepsilon + v_x - v_z) \, dG(\varepsilon). \tag{15}$$

The construction of v_z makes Equations (14) and (15) equal, implying

$$\int_{\varepsilon=-\infty}^{+\infty} [G(\varepsilon + v_x - v_y - \log K) - G(\varepsilon + v_x - v_y)^K] \, dG(\varepsilon) = 0.$$

But this can be true for all values of $v_x \in (-\infty, +\infty)$ only if the term in brackets is zero, since G is translation complete, implying

$$G(v_x - \log K) = G(v_x)^K.$$

Taking $v_x = 0$ implies $G(-\log K) = e^{-\alpha K}$, where $\alpha = -\log G(0) > 0$, and taking $v_x = \log K - \log L$ implies $G(-\log L) = G(\log K/L)^K$. Hence, $G(\log K/L) = e^{-\alpha L/K}$ for all positive integers K, L. Since G is monotone, it follows in the limit that $G(\log k) = e^{-\alpha/k}$ for all positive real k. Then $G(\varepsilon) = e^{-\alpha \exp(-\varepsilon)}$.

We summarize the advantages, and then the limitations, of the axioms leading to the formula (12) for the selection probabilities. First, this formula allows a ready interpretation of the selection probabilities in terms of the relative representative utilities of alternatives, and is relatively amenable to computation. Second, the formula makes it simple to ascertain the effect of introducing a new alternative to an alternative set; the proportional decrease in the selection probability of each old alternative equals the selection probability of the new alternative. This also points out a weakness of the model in that one cannot postulate a pattern of differential substitutability and complementarity between alternatives. Third, the axioms provide the identifying

restrictions necessary to estimate choice alternative effects without replications, and to predict choice behavior resulting from extrapolation of observed alternative sets. Any set of identifying restrictions meeting these conditions will require powerful axioms on behavior, and care must be exercised in avoiding application of these models in situations where the axioms are implausible. The model above is subject to this general caveat.

The primary limitation of the model is that the independence of irrelevant alternatives axiom is implausible for alternative sets containing choices that are close substitutes. An example illustrates this point. Suppose a population faces the alternatives of travel by auto and by bus, and two-thirds choose to use auto. Suppose now a second "brand" of bus travel is introduced that is in all essential respects the same as the first. Intuitively, two-thirds of the population will still choose auto, and the remainder will split between the bus alternatives. However, if the selection probabilities satisfy Axiom 1, only half the population will use auto when the second bus is introduced. The reason this is counter-intuitive is that we expect individuals to lump the two bus alternatives together in making the auto-bus choice. This example suggests that application of the model should be limited to situations where the alternatives can plausibly be assumed to be distinct and weighed independently in the eyes of each decision-maker.

II. Conditional Logit Estimation

Formula (12) for the selection probabilities, obtained from Axioms 1–3, can be adapted for empirical analysis by specifying the functional form of "representative" utility $v(s, x)$. A particularly convenient assumption is that v is linear in unknown parameters.

AXIOM 4. The function $v(s, x)$ has the form

$$v(s, x) = \theta_1 v^1(s, x) + \cdots + \theta_K v^K(s, x),$$

where the $v^k(s, x)$ are specified numerical functions and the θ_k are unknown parameters.

A choice experiment yields observations on N distinct trials (s_n, B_n), where s_n is a vector of measured attributes of an individual and B_n is an alternative set. Let B_n contain J_n alternatives, indexed $j = 1, \ldots, J_n$, with vectors of attributes x_{jn}. Define $z_{jn}^k = v^k(s_n, x_{jn})$ and $z_{jn} = (z_{jn}^1, \ldots, z_{jn}^K)$. From Equation (12), the selection probabilities then satisfy

$$P_{in} = P(x_{in} \mid s_n, B_n) = \frac{e^{z_{in}\theta}}{\sum_{j=1}^{J_n} e^{z_{jn}\theta}}, \tag{16}$$

where $\theta' = (\theta_1, \ldots, \theta_K)$.[6] The experiment provides R_n repetitions of trial n, and the ith alternative is observed to be chosen S_{in} times. Of particular interest is the case without repetition,

$$R_n = \sum_{j=1}^{J_n} S_{jn} = 1.$$

We term this the *conditional logit model.* Note that it is an immediate generalization to the case of unequal, possibly unranked, alternatives of the multinomial logit model appearing in the literature.[7] The derivation of this model from a theory of population choice behavior appears to be new.

[6] The generality and limits of this form deserve emphasis. A variable $z^k = v^k(s, x)$ may be a component of x, a function specifying a nonlinear transformation or interaction between components of x, or a function specifying an interaction between x and s variables. It cannot be a component of s (or x) that is invariant over each alternative set, as this shifts the origin of the "representative" utility function leaving all the selection probabilities unchanged, and the associated θ_k is nonidentified. In general, the alternatives x_j have no natural ranking, and the indexing j is arbitrary. We would then say the attributes of the alternatives are generic, or "hedonic." However, in some application the alternatives are ranked, and the rank j is a component of the vector of attributes of alternative x_j summarizing the "unique" characteristics of this position. Then z^k may be a variable such that $z_j^k = 1$ and $z_l^k = 0$ for $l \neq j$, yielding a "specific jth alternative" effect θ_k. Further, the interaction of such a variable with other components of x can give variable alternative-specific interaction effects. An extreme case is that in which the specific alternative effect is the only attribute varying across the alternative set, and all variables are as an example of the form $z_j^i = s^i$ and $z_l^i = 0$ for $l \neq j$, implying $v(s, x_j) = \sum_{i=1}^{J} s^i \theta_j$, where only the parameters θ_j vary in j. Since translation of all θ_j leaves the selection probabilities unchanged, identification requires a normalization, say $\theta_1 = 0$.

[7] Binomial logit analysis was popularized by Berkson (1951, 1955) and has been analyzed extensively in the statistical literature: Antle (1970), Cox (1958, 1966, 1970), Gart (1967), Gilbert (1968), Grizzle (1962, 1971), Gupta (1967), Harter and Moore (1967), and Walker and Duncan (1967). Multinomial logit was developed for a special case by Gurland (1960), and more generally by Bloch (1967), Bock (1969), Rassam (1971), McFadden (1968), Stopher (1969), and Theil (1969, 1970). An analogous development has occurred for probit analysis, in which the cumulative normal rather than logistic distribution is used to determine the selection probabilities (Aitchison and Silvey, 1957; Aitchison and Bennett, 1970; Amemiya, 1972).

The notion of a distribution of tastes in a population of consumers as a source of stochastic components of demand has been implicit in much of the literature on consumer demand theory, particularly in random coefficients models of demand. The use of this concept in analyzing qualitative choice has been made explicit in the work of Quandt (1968, 1970; Quandt and Young, 1969; Quandt and Baumol, 1966), where selection probabilities are assumed to result from maximization of a log-linear utility function with random parameters. The relationship of logit models to distributions of utility functions was worked out in the context of models of stochastic choice behavior by Marschak (1960) and Block (1960), and explored further by Luce and Suppes (1965); the econometric implications of this work were apparently first noted by the author (1968). The foundations of the theory of testing hypotheses on individual behavior from population data were developed in a later paper by the author and Richter (1971).

The vector $(S_{1n}, \ldots, S_{J_n n})$ can be viewed as the result of R_n independent drawings from a multinomial distribution with probabilities given by Equation (16) for $i = 1, \ldots, J_n$. Hence, the likelihood of the given sample is a function $L = L(\theta) = L((S_{jn}, z_{jn}); \theta)$ satisfying

$$e^L = \prod_{n=1}^{N} \frac{R_n!}{S_{1n}! \cdots S_{J_n n}!} \prod_{i=1}^{J_n} P_{in}^{S_{in}}. \tag{17}$$

Substitution of Equation (16) yields the log-likelihood function

$$\begin{aligned} L &= C - \sum_{n=1}^{N} \sum_{i=1}^{J_n} S_{in} \log \sum_{j=1}^{J_n} e^{(z_{jn} - z_{in})\theta} \\ &= C + \sum_{n=1}^{N} \left[\left(\sum_{j=1}^{J_n} S_{jn} z_{jn} \right) \theta - R_n \log \sum_{j=1}^{J_n} e^{z_{jn}\theta} \right], \end{aligned} \tag{18}$$

where

$$C = \sum_{n=1}^{N} \left[\log R_n! - \sum_{j=1}^{J_n} \log S_{jn}! \right].$$

An estimator for θ with good large sample properties under very general conditions is obtained by a vector $\hat{\theta}$, depending on the observations, which maximizes the likelihood (18) of the given sample. We discuss the computation and statistical properties of the maximum likelihood estimator. Several alternative estimation methods are discussed at the end of this section.

Differentiation of Equation (18) with respect to θ yields the formulas

$$\frac{\partial L}{\partial \theta} = \sum_{n=1}^{N} \left[\sum_{j=1}^{J_n} (S_{jn} - R_n P_{jn}) z_{jn} \right], \tag{19}$$

$$\frac{\partial^2 L}{\partial \theta \, \partial \theta'} = - \sum_{n=1}^{N} R_n \sum_{j=1}^{J_n} (z_{jn} - \bar{z}_n)' P_{jn} (z_{jn} - \bar{z}_n), \tag{20}$$

where

$$\bar{z}_n = \sum_{i=1}^{J_n} z_{in} P_{in}.$$

Since Equation (20) is the negative of a weighted moment matrix of the independent variables, it is negative semidefinite and the log-likelihood function is concave in θ. Then L is maximized at any critical point θ where $\partial L / \partial \theta = 0$.

Binomial logit and probit analyses have been used in a number of economic applications: Allouche (1972), Amemiya and Boskin (1972), Fisher (1962), Korbel (1966), Lave (1968), Lee (1963), Lisco (1967), McGillivray (1970), Moses *et al.* (1967), Reichman and Stopher (1971), Stopher (1969), Stopher and Lisco (1970), Talvitie (1972), Thomas and Thompson (1971), Uhler (1968), Walker (1968), Warner (1967), and Zellner and Lee (1965).

If, further, the matrix $\partial^2 L/\partial\theta\,\partial\theta'$ is nonsingular, L has a unique maximum in θ (provided one exists). A necessary and sufficient condition for $\partial^2 L/\partial\theta\,\partial\theta'$ to be negative definite is the following.

AXIOM 5 (Full Rank). The $\sum_{n=1}^{N} J_n \times K$ matrix whose rows are $(z_{in}-\bar{z}_n)$ for $i = 1,\ldots,J_n$ and $n = 1,\ldots,N$ is of rank K.

Since N linear dependency conditions are present in this matrix due to the subtraction of weighted means, a necessary order condition is $\sum_{n=1}^{N} J_n \geqslant K+N$. This will hold in particular if $N \geqslant K$ since $J_n \geqslant 2$, but may also hold for $N < K$ if the J_n are large. Analogously to the hypothesis of full rank in the linear statistical model, we can expect Axiom 5 to hold when the order condition is satisfied provided the data vary across alternative sets and are not collinear.

We next introduce an inequalities condition that guarantees the existence of a vector θ maximizing L.

AXIOM 6. There exists no nonzero K-vector γ satisfying $S_{in}(z_{jn}-z_{in})\gamma \leqslant 0$ for $i,j = 1,\ldots,J_n$ and $n = 1,\ldots,N$.

Note that there is a positive probability that Axiom 6 may fail in a finite sample since the S_{in} are random. We show later that this probability is negligible in samples of reasonable size and approaches zero asymptotically. The following result establishes the existence of a θ maximizing L.

LEMMA 3. Suppose Axioms 1–5 hold. Then Axiom 6 is necessary and sufficient for the existence of a vector θ maximizing L.

Proof: We first show Axiom 6 to be necessary. Suppose L has a maximum at $\hat{\theta}$, but Axiom 6 fails for some $\gamma \neq 0$. Recall that

$$\log P_{in}(\theta) = -\log\left(\sum_{j=1}^{J_n} e^{(z_{jn}-z_{in})\theta}\right).$$

If $S_{in} > 0$, then $(z_{jn}-z_{in})(\hat{\theta}+\gamma) \leqslant (z_{jn}-z_{in})\hat{\theta}$ and $\log P_{in}(\hat{\theta}+\gamma) \geqslant \log P_{in}(\hat{\theta}_n)$. Then $L(\hat{\theta}+\gamma) \geqslant L(\hat{\theta})$. Since L is strictly concave, $L(\hat{\theta}+\gamma/2) > L(\hat{\theta})$, contradicting the definition of $\hat{\theta}$. Hence, Axiom 6 is necessary.

Next suppose that Axiom 6 holds. Define $A = \{\gamma \mid \gamma'\gamma = 1\}$. For each $\gamma \in A$, there exists j, i, n such that $S_{in}(z_{jn}-z_{in})\gamma > 0$. Define

$$b(\gamma) = \max_{n=1,\ldots,N} \max_{i,j=1,\ldots,J_n} S_{in}(z_{jn}-z_{in})\gamma. \tag{21}$$

Then b is a positive continuous function on the compact set A, and has a positive lower bound b^* on this set. Let $|\theta| = (\theta'\theta)^{1/2}$ and define

$$D = \{\theta \mid |\theta| \leqslant [-L(0)+C]/b^*\}.$$

Consider any $\theta \neq 0$, and let $\gamma = \theta/|\theta|$. Then $\gamma \in A$, and there exist indices i, j, n such that $b(\gamma) = S_{in}(z_{jn} - z_{in})\gamma$. From Equation (18),

$$\begin{aligned} L(\theta) - C \leqslant S_{in} \log P_{in} &= -S_{in} \log \sum_{k=1}^{J_n} e^{(z_{kn} - z_{in})\gamma|\theta|} \\ &\leqslant -S_{in}(z_{jn} - z_{in})\gamma|\theta| = -b(\gamma)|\theta| \\ &\leqslant -b^*|\theta|. \end{aligned}$$

For $\theta \notin D$, $L(\theta) - C \leqslant -b^*|\theta| < L(0) - C$. Hence, L can be maximized on the compact set D, and an optimal θ exists.

The following lemma establishes that Axiom 6 can be tested by solving a quadratic programming problem. This can be done by using a finite computational algorithm such as Lemke's method. In practice it is unnecessary to carry out this computation for sample sizes N exceeding the number of parameters K, as the probability of nonexistence rapidly becomes negligible.

LEMMA 4. Suppose Axioms 1-5 hold. Then Axiom 6 holds if and only if the minimum in the following quadratic programming problem is zero:

$$\min_{y, \alpha} y'y$$

subject to

$$y' = \sum_{n=1}^{N} \sum_{i,j=1}^{J_n} \alpha_{ijn} S_{in}(z_{jn} - z_{in}) \quad \text{and} \quad \alpha_{ijn} \geqslant 1. \tag{22}$$

Proof: Suppose the program has a zero minimum, achieved at some $y' = 0$, but that Axiom 6 fails. Then there exists $\gamma \neq 0$ such that $S_{in}(z_{jn} - z_{in})\gamma \leqslant 0$, with at least one inequality strict by Axiom 5. Then

$$0 = y'\gamma = \sum_{n=1}^{N} \sum_{i,j=1}^{J_n} \alpha_{ijn} S_{in}(z_{jn} - z_{in})\gamma < 0,$$

a contradiction. Thus, if the program has a zero minimum, Axiom 6 holds.

Let K denote the convex cone generated by the vectors $S_{in}(z_{jn} - z_{in})$ for $n = 1, \ldots, N$ and $i, j = 1, \ldots, J_n$. If the origin is in the interior of K, then there exist positive scalars α_{ijn} such that $y' = \sum_{n=1}^{N} \alpha_{ijn} S_{in}(z_{jn} - z_{in}) = 0$, and the quadratic program achieves a minimum of zero. If the origin is not in the interior of K, then there exists a separating hyperplane with normal $\gamma \neq 0$ such that $S_{in}(z_{jn} - z_{in})\gamma \leqslant 0$ for all $n = 1, \ldots, N$ and $i, j = 1, \ldots, J_n$. Hence, Axiom 6 fails. This proves the lemma.

Computation of the maximum likelihood estimator can be carried out using a variety of standard programs for unconstrained nonlinear optimization.

Since the likelihood function is strictly concave, any algorithm which converges will attain the maximum. Experience has shown that a standard Newton–Raphson algorithm may converge slowly for this problem; we have found that one efficient procedure is to use Davidon's variable metric method to determine direction of search and a one-dimensional procedure employing a cubic approximation to determine the optimal step size.[8]

The maximum likelihood procedure has proved practical for problems of up to 20 variables and 2000 observations, but is relatively costly for large samples. A quick procedure which can be used for screening models is to make a linear expansion of the gradient of the likelihood function (19) in θ about some initial vector $\bar{\theta}$, and then solve for the value of θ equating this approximate gradient to zero, or

$$\tilde{\theta} = \bar{\theta} + \left[\sum_{n=1}^{N} R^n \sum_{j=1}^{J_n} (z_{jn} - \bar{z}_n)' \bar{P}_{jn} (z_{jn} - \bar{z}_n)\right]^{-1} \times \left[\sum_{n=1}^{N} \sum_{j=1}^{J_n} (z_{jn} - \bar{z}_n)'(S_{jn} - R_n \bar{P}_{jn})\right], \tag{23}$$

where $\bar{P}_{jn} = P_{jn}(\bar{\theta})$ and $\bar{z}_n = \sum_{j=1}^{J_n} z_{jn} \bar{P}_{jn}$. Note that $\tilde{\theta}$ is the result of one iteration of a Newton–Raphson procedure for maximizing the likelihood function, and can also be interpreted as the ordinary least squares estimator in the linear model (with R_n observations for each n)

$$(\bar{P}_{jn})^{-1/2}[(S_{jn}/R_n) - \bar{P}_{jn}] = (\bar{P}_{jn})^{1/2}(z_{jn} - \bar{z}_n)(\theta - \bar{\theta}) + \varepsilon_{jn}. \tag{24}$$

Equation (24) is termed the *linear probability model*, and is sometimes taken as a specification of selection probabilities $P_{jn} = E(S_{jn}/R_n)$. The estimator $\tilde{\theta}$ is not a consistent estimate of the true parameter vector θ when the specification of Axioms 1–6 is valid; however, as a practical matter it usually agrees in magnitude and sign with the maximum likelihood estimator provided the terms $|(z_{jn} - \bar{z}_n)(\theta - \bar{\theta})|$ are less than unity. Equation (24) is inappropriate for use in forecasting selection probabilities because the requirement that the forecasts lie in the unit interval is not met.

When the number of repetitions for each trial is large, a method of estimation developed by Berkson (1951) and generalized to the multinomial case by Theil (1969) can be employed. When S_{in}, S_{ln} are large,[9] $\log(S_{in}/S_{ln})$ is a close approximation to the left-hand side of

$$\log(P_{in}/P_{ln}) = (z_{in} - z_{ln})\theta, \tag{25}$$

and an estimate of θ can be obtained by applying ordinary or weighted least

[8] The author is indebted to H. Wills and H. Varian for work on the numerical methods and programming of this problem.

[9] A rule-of-thumb is $S_{in} \geq 5$.

squares to the model

$$\log(S_{in}/S_{ln}) = (z_{in}-z_{ln})\theta + \varepsilon_{in}, \tag{26}$$

taking into account linear restrictions across equations.[10] This procedure is asymptotically equivalent to maximum likelihood estimation as the S_{in} approach infinity (for appropriate weights in the regression), and is to be preferred to the maximum likelihood procedure on computational grounds. It should be noted, however, that grouping observations that are not exact replications in order to achieve the cell frequencies required for application of the Berkson–Theil method introduces an "errors in variables" component that makes the estimator inconsistent and may make it seriously biased. In such cases the maximum likelihood procedure should be more reliable.

III. Statistical Properties

Maximum likelihood estimation of the conditional logit model can be shown under very general conditions to provide estimators that are asymptotically efficient and normally distributed. Examples suggest that the approximation is reasonably good even in quite small samples. These results can be used to construct approximate large-sample confidence bounds and tests of hypotheses for the parameters.

We have noted that for finite sample sizes, there will be a positive probability that a maximum of the likelihood function cannot be attained. This corresponds to the case where the system of inequalities in Axiom 6 has a nontrivial solution and the sample is "explained" by maximization of this linear combination of the independent variables. We first show that when the sample is in fact generated by probabilities satisfying Axioms 1–5, then the probability that the likelihood function has a maximum approaches one as the sample size increases. We impose the following condition on the data.

AXIOM 7. The numbers of alternatives J_n are uniformly bounded by an integer J_*. The independent variables z_{in} are uniformly bounded by a scalar

[10] Some improvement in the statistical properties of the unweighted Berkson-Theil estimator can be obtained by replacing Equation (26) with the regression equation

$$\log\left(\frac{S_{in}+\frac{1}{2}}{S_{ln}+\frac{1}{2}}\right) = (z_{in}-z_{ln})\theta + \varepsilon_{in}. \tag{26a}$$

This modification, suggested by Haldane (1955) for the binomial logit model, makes $E\log[(S_{in}+\frac{1}{2})/(S_{ln}+\frac{1}{2})]$ equal to $\log(P_{in}/P_{ln})$ up to a term of order $1/R_n^2$, rather than of order $1/R_n$, as R_n approaches infinity. This improves the speed of convergence of the estimators to their large-sample values. Minor modifications of the Haldane argument establish its validity in the multinomial case.

M.[11] The limit of the weighted moment matrix, as $\sum_{n=1}^{N} R_n \to +\infty$,

$$\lim\left(\sum_{n=1}^{N} R_n\right)^{-1} \sum_{n=1}^{N} R_n \sum_{i=1}^{J_n} (z_{in}-\bar{z}_n)' P_{in}(z_{in}-\bar{z}_{in}) = \Omega, \tag{27}$$

exists and is positive-definite.

The last part of this axiom strengthens the full-rank condition assumed earlier, implying that an infinite number of blocks of K trials can be found satisfying Axiom 5.[12] One can expect Axiom 7 to hold provided the data are not multicollinear and do not tend to become explosive or degenerate as the sample size increases. The following results are proved in the Appendix.

LEMMA 5. Suppose Axioms 1-4 and 7 hold. Then the probability that Axiom 6 holds and the maximum likelihood estimator exists approaches unity as $\sum_{n=1}^{N} R_n$ approaches infinity.

LEMMA 6. Suppose Axioms 1-4 and 7 hold, θ^0 is the true parameter vector, and $\hat{\theta}^M$ is the maximum likelihood estimator for a sample of size $M = \sum_{n=1}^{N} R_n$. Then $\hat{\theta}^M$ is consistent and asymptotically normal, with

$$\left(\sum_{n=1}^{N} R_n\right)^{1/2} \Omega^{1/2}(\hat{\theta}^M - \theta^0), \tag{28}$$

tending to a multivariate normal distribution with mean zero and a covariance matrix equal to the identity matrix.

This lemma implies that $\hat{\theta}^M$ tends to be distributed normally with mean θ^0 and covariance matrix $(\sum_{n=1}^{N} R_n)^{-1}\Omega^{-1}$, and that the quadratic form

$$Q(\hat{\theta}^M) \equiv \left(\sum_{n=1}^{N} R_n\right)(\hat{\theta}^M - \theta^0)'\Omega(\hat{\theta}^M - \theta^0)$$

tends to be chi-square distributed with K degrees of freedom. These statistics can be used to carry out large sample tests of hypotheses on θ^0. In particular, diagonal elements of the inverse of the information matrix $(\sum_{n=1}^{N} R_n)^{-1}\Omega^{-1}$ provide estimates of the variances of the estimators.[13] To test a hypothesis

[11] I.e., $|z_{in}| \leq M$, where the norm $|A|$ of any array A is the sum of the absolute values of its elements.

[12] Otherwise, all but a finite number of vectors $z_{in}-\bar{z}_n$ can be written as linear combinations of less than K linearly independent vectors. Then, Ω must also have this property, contradicting the hypothesis that it is nonsingular.

[13] Some improvement in the speed of convergence can be attained by multiplying these estimates by a correction factor for degrees of freedom,

$$\frac{\sum_{n=1}^{N} R_n(J_n-1)}{\sum_{n=1}^{N} R_n(J_n-1)-K}.$$

that the true parameter vector θ^0 lies in a $(K-K_1)$ dimensional manifold, calculate the maximum likelihood estimator $\hat{\theta}^H$ under the null hypothesis and the unconstrained maximum likelihood estimator $\hat{\theta}$. Then the statistic

$$(29) \quad -2[L(\hat{\theta}^H)-L(\hat{\theta})] = (\hat{\theta}^H-\hat{\theta})' \sum_{n=1}^{N} R_n \sum_{i=1}^{J_n} (z_{in}-\bar{z}_n)' P_{in}(z_{in}-\bar{z}_n)(\hat{\theta}^H-\hat{\theta}),$$

with P_{in} evaluated at $\hat{\theta}$, is distributed approximately chi-square with K_1 degrees of freedom. If the null hypothesis is that θ^0 is zero, or that it is zero except for pure alternative effects, then this statistic provides a test of the significance of an estimation equation, indicating respectively the "mean square error" explained or the "variance" explained. Noting that the extreme case is $L(\hat{\theta}) \approx 0$, we can define a coefficient of determination that is analogous to the multiple-correlation coefficient in the linear statistical model,

$$(30) \quad \rho^2 = 1 - \frac{L(\hat{\theta})}{L(\hat{\theta}^H)}.$$

If $\hat{\theta}^H$ is zero, or if $\hat{\theta}^H$ is zero except for pure alternative effects and the model contains such effects, then ρ^2 lies in the unit interval. If, in the latter case, the model has no pure alternative effects, it is possible for ρ^2 to be negative.

A second measure of goodness of fit is based on deviations of observed from fitted relative frequencies. Define the weighted residuals

$$(31) \quad D_{in} = \frac{S_{in}-R_n P_{in}}{(R_n P_{in})^{1/2}},$$

for $i = 1, \ldots, J_n$ and $n = 1, \ldots, N$, where P_{in} is evaluated at the maximum likelihood estimate. These residuals satisfy the first-order conditions for maximization of the likelihood function,

$$(32) \quad \sum_{n=1}^{N} (r_n)^{1/2} \sum_{i=1}^{J_n} D_{in}(P_{in})^{1/2}(z_{in}-\bar{z}_n) = 0,$$

where

$$(33) \quad r_n = \frac{R_n}{\sum_{m=1}^{N} R_m},$$

and the conditions

$$(34) \quad \sum_{i=1}^{J_n} (P_{in})^{1/2} D_{in} = 0,$$

for $n = 1, \ldots, N$, a total of $N+K$ restrictions. Now suppose $\sum_{n=1}^{N} R_n$ approaches infinity, with each r_n approaching a limit. We show in the Appendix that the

D_{in} are distributed asymptotically with mean zero and covariances

$$\Lambda_{in,jm} \equiv ED_{in}D_{jm} = \delta_{nm}[\delta_{ij}-(P_{in}^0 P_{jn}^0)^{1/2}] - (r_m r_n P_{in}^0 P_{jm}^0)^{1/2}(z_{in}-\bar{z}_n)\Omega^{-1}(z_{jm}-\bar{z}_m). \tag{35}$$

Consider the case in which N remains finite and the R_n approach infinity. Then Λ is an idempotent matrix of rank $N^* = \sum_{n=1}^{N} J_n - N - K$, and the asymptotic distribution of the D_{in} is multivariate normal. Hence,

$$G = \sum_{n=1}^{N}\sum_{i=1}^{J_n} D_{in}^2 \tag{36}$$

has an asymptotic chi-square distribution with N^* degrees of freedom.[14] The statistics D_{in} and G can be used to carry out large-sample tests of the model specification. For example, regression of the D_{in} on potential independent variables provides evidence on the validity of their exclusion from the model. A test of the significance of G provides evidence on the validity of the IIA specification of the selection probabilities and the absence of "alternative set" effects. Further, one can define an analog of the multiple-correlation coefficient,

$$R^2 = 1 - G/G^H, \tag{37}$$

where G is given by Equation (36) and G^H is given by the same equation when the numerators of the residuals are evaluated under the hypothesis that the parameter vector is zero, or is zero except for pure alternative choice effects.

In evaluating the results of regressions of the D_{in} on potential independent variables, one should adjust for the nonindependence and heteroskedasticity of the D_{in}. This can be achieved in part by using the linearly transformed residuals

$$Y_{in} = D_{in} - \frac{D_{1n}(P_{in})^{1/2}[1-(P_{1n})^{1/2}]}{1-P_{1n}}, \tag{38}$$

defined for $i = 2, \ldots, J_n$ and $n = 1, \ldots, N$. The Y_{in} are asymptotically multivariate normal with mean zero and covariances

$$\Gamma_{in,jm} \equiv EY_{in}Y_{jm} = \delta_{nm}\delta_{ij} - q_{in}' q_{jm}, \tag{39}$$

[14] Treating G as a function of θ and minimizing it at a value $\hat{\theta}$ provides a *minimum chi-square* estimator of the parameter vector. The first-order conditions for this minimization coincide with Equation (32) except for a term, reflecting the effect of changing θ on the weights in the denominator of Equation (31), which has probability limit zero when the $R_n \to +\infty$. Thus, the maximum likelihood and minimum chi-square estimators are asymptotically equivalent under these limiting conditions. On the other hand, the minimum chi-square procedure is not consistent under the limiting conditions that $N \to +\infty$ and R_n remain finite.

where

$$(40) \qquad q'_{in} = (r_n P^0_{in})^{1/2}\left((z_{in}-\bar{z}_n) + \frac{P^0_{1n}-(P^0_{1n})^{1/2}}{1-P^0_{1n}}(z_{1n}-\bar{z}_n)\right)\Omega^{-1/2}.$$

The matrix Γ is idempotent of rank N^*. When the r_n are small, the matrix Γ is nearly diagonal, and the regression of any subset of N^* of the Y_{in} on potential independent variables can, as a good approximation, be treated as independent and homoskedastic with unit variance.

We next consider the case in which N approaches infinity and the limiting values of the r_n are zero. Then, the residuals D_{in} have an asymptotic multinomial distribution with mean zero and covariances given by Equation (35), with the second term in this expression vanishing. The D_{in} and the transformed residuals Y_{in} defined in Equation (38) are independent across n, and the Y_{in} have zero mean, unit variance, and zero covariances. Suppose integers N_m satisfy $\sum_{m=1}^{M} N_m = N$ and $N_m \to +\infty$. Then the statistics

$$(41) \qquad Y^m = \frac{\sum_{n=N_{m'}+1}^{N_{m'}+1}\sum_{j=2}^{J_n} Y_{jn}}{\left(\sum_{n=N_{m'}+1}^{N_{m'}+1} J_n - N_m\right)^{1/2}},$$

where $N_m' = N_1 + \cdots + N_{m-1}$, are asymptotically independent standard normal, and can be used to test the specification of the absence of alternative set effects. When the R_n remain small, the distributions of the D_{in} and Y_{in} depart substantially from asymptotic normality. The statistics G and R^2 defined in Equations (36) and (37) remain useful summary measures, although the robustness of the asymptotic distributions obtained in the previous case has not been investigated. Since the Y_{in} satisfy the Gauss–Markov assumptions when the model is specified correctly, the usual asymptotic theory for the linear statistical model can be applied to test the validity of excluding potential independent variables.

The small sample properties of the maximum likelihood estimator of the conditional logit model are unknown except for a few special cases. Monte Carlo studies of related models suggest that maximum likelihood, minimum chi-square, and Berkson–Theil estimators are all reasonably well behaved in small samples, even when the number of repetitions is small.[15] We next consider several simple examples in which the maximum likelihood estimator can be calculated analytically. These examples suggest that the maximum likelihood estimator is well behaved in samples of sizes likely to be encountered in applications, 50 and greater, but may be inferior to the linear probability model estimator in very small samples provided the range of the data is not too large.

[15] Berkson (1955), Gart and Zweifel (1967), Gilbert (1968), and Talvitie (1972).

TABLE 1

SAMPLING PROPERTIES OF THE MAXIMUM LIKELIHOOD ESTIMATOR: EXAMPLE 1

Parameter value (θ_1)	Sample size	Probability of existence of maximum likelihood estimator	Conditional expectation of maximum likelihood estimator	Percent bias	Conditonal variance of maximum likelihood estimator	Calculated variance of maximum likelihood estimator	Percent bias	Linear probability model estimator	Percent bias
0.01	5	0.93748	0.00828	−17.2	0.75018	1.21529	62.0	0.01000	a
	10	0.99814	0.01103	10.3	0.49988	0.50943	1.9	0.01000	a
	20	1.0	0.01057	5.7	0.22422	0.22314	−0.5	0.01000	a
	30	1.0	0.01036	3.6	0.14334	0.14308	a	0.01000	a
	50	1.0	0.01021	2.1	0.08343	0.08338	a	0.01000	a
	75	1.0	0.01014	1.4	0.05482	0.05481	a	0.01000	a
	100	1.0	0.01010	1.0	0.04083	0.04082	a	0.01000	a
0.1	5	0.93594	0.08272	−17.3	0.74726	1.21643	67.0	0.09992	−0.1
	10	0.99782	0.11023	10.2	0.50044	0.51128	2.2	0.09992	−0.1
	20	1.0	0.10575	5.8	0.22503	0.22387	−0.5	0.09992	−0.1
	30	1.0	0.10364	3.6	0.14379	0.14349	a	0.09992	−0.1
	50	1.0	0.10211	2.1	0.08366	0.08360	a	0.09992	−0.1
	75	1.0	0.10138	1.4	0.05497	0.05497	a	0.09992	−0.1
	100	1.0	0.10103	1.0	0.04093	0.04093	a	0.09992	−0.1
0.5	5	0.89888	0.40054	−20.0	0.68164	1.24250	83.0	0.48984	−2.0
	10	0.99121	0.54433	8.9	0.50916	0.55472	8.9	0.48984	−2.0
	20	0.99992	0.53020	6.0	0.24536	0.24215	−1.3	0.48984	−2.0
	30	1.0	0.51912	3.8	0.15515	0.15391	−0.8	0.48984	−2.0
	50	1.0	0.51101	2.2	0.08955	0.08922	a	0.48984	−2.0
	75	1.0	0.50720	1.4	0.05865	0.05852	a	0.48984	−2.0
	100	1.0	0.50535	1.1	0.04361	0.04354	a	0.48984	−2.0

1.0	5	0.78978	0.73234	−26.8	0.52531	1.30757	148.0	0.92423	−7.6
	10	0.95639	1.04040	4.0	0.49435	0.67376	36.5	0.92423	−7.6
	20	0.99810	1.06643	6.6	0.30950	0.30826	−0.4	0.92423	−7.6
	30	0.99992	1.04437	4.4	0.19662	0.19146	−2.6	0.92423	−7.6
	50	1.0	1.02525	2.5	0.11041	0.10881	−1.5	0.92423	−7.6
	75	1.0	1.01641	1.6	0.07145	0.07083	−0.8	0.92423	−7.6
	100	1.0	1.01216	1.2	0.05285	0.05252	−0.6	0.92423	−7.6
2.0	5	0.46885	1.12824	−44.5	0.23655	1.44030	510.0	1.5232	−23.8
	10	0.71897	1.71841	−14.1	0.30827	0.95333	209.0	1.5232	−23.8
	20	0.92102	2.04023	2.0	0.37726	0.60123	59.4	1.5232	−23.8
	30	0.97780	2.09760	4.9	0.35233	0.41411	17.6	1.5232	−23.8
	50	0.99825	2.08151	4.0	0.23948	0.23167	−3.3	1.5232	−23.8
	75	0.99993	2.05384	2.2	0.15052	0.14352	−4.7	1.5232	−23.8
	100	1.0	2.03925	2.0	0.10754	0.10383	−3.5	1.5232	−23.8
3.0	5	0.21568	1.28953	−57.1	0.09262	1.51321	1655.0	1.8103	−39.7
	10	0.38484	2.01670	−32.8	0.13505	1.11987	830.0	1.8103	−39.7
	20	0.62158	2.59216	−13.6	0.21698	0.88161	307.0	1.8103	−39.7
	30	0.76721	2.84486	−5.2	0.27726	0.74378	162.0	1.8103	−39.7
	50	0.91191	3.04643	1.5	0.33452	0.54907	64.4	1.8103	−39.7
	75	0.97385	3.10492	3.5	0.32443	0.38745	19.4	1.8103	−39.7
	100	0.99224	3.10337	3.4	0.27741	0.28459	2.5	1.8103	−39.7

[a] Negligible.

Example 1. Suppose N observations are taken of a binary choice with a selection probability $P_1 = 1/(1+e^{-\theta_1})$ for the first alternative, and suppose this alternative is chosen S times. The maximum likelihood estimator exists if $0 < S < N$, and equals $\hat{\theta}_1 = \log[S/(N-S)]$. Table 1 gives the actual expectation and variance of $\hat{\theta}_1$, conditioned on existence, and the large sample variance calculated from the information matrix. The last columns give the linear probability model approximation (23) to the logit model from starting value zero. For sample sizes exceeding 20, the maximum likelihood estimator and its calculated variance have expectations that are within 10% of true values except for extreme selection probabilities (e.g., $\theta_1 = 2.0$ yields $P_1 = 0.88$). The linear probability model approximation is quite accurate for small parameter values, even for small sample sizes. The bias is severe however for extreme selection probabilities, and is independent of sample size. The probability of existence of the maximum likelihood estimator rises rapidly with sample size, even for extreme selection probabilities.

Example 2. Suppose $N = 2R$ observations are taken of a binary choice with selection probabilities $P_{1n} = 1/(1+e^{-\theta_1-\theta_2 x_n})$, where $x_n = 0$ for $n = 1, \ldots, R$ and $x_n = 1$ for $n = R+1, \ldots, N$. Suppose the first alternative is chosen S_1 times in the first R observations and S_2 times in the second R observations. The maximum likelihood estimator exists if $0 < S_1 < R$ and $0 < S_2 < R$, and equals $\hat{\theta}_1 = \log[S_1/(R-S_1)]$ and $\hat{\theta}_2 = \log[S_2/(R-S_2)] - \log[S_1/(R-S_1)]$. Table 2 gives the conditional expectations of these estimators and the expectations of the variances calculated from the information matrix for selected parameter values. The pattern of the biases generally conforms to that of the previous example. For a sample size of 10, the estimator and its calculated variance are substantially biased, the former downward and the latter upward. As sample size increases, the bias in the estimator swings positive, but never more than 10% and then approaches zero. The calculated variances show similar behavior with reversed sign, their bias going from positive to negative as sample size increases and then approaching zero. As the parameter values and selection probabilities become more extreme, there is an increase in the sample size at which the maximum positive bias in the estimator occurs. The linear probability model approximation provides an accurate estimator for small sample sizes and parameter values, and indicates correctly signs and orders of magnitude of parameters even for extreme values, but with substantial biases. In samples of size 100 or 200 the biases and probabilities of nonexistence of the maximum likelihood estimator are acceptably small even for extreme selection probabilities.

One must be cautious in generalizing too far the conclusions drawn from these examples. In particular, we have not explored the behavior of the estimators in samples in which the observations are generated by mixtures of

TABLE 2

SAMPLING PROPERTIES OF THE MAXIMUM LIKELIHOOD ESTIMATOR: EXAMPLE 2

	Sample size					
	10	20	40	80	140	200
First parameter (all cases)						
True value	0.10000	0.10000	0.10000	0.10000	0.10000	0.10000
Expect. of M.L. estimate	0.08272	0.11023	0.10575	0.10267	0.10148	0.10103
Per cent bias of est.	−17.28364	10.22808	5.74896	2.66732	1.48170	1.02609
True variance of M.L. est.	0.74726	0.50044	0.22503	0.10576	0.05902	0.04094
Calc. variance of M.L. est.	0.80200	0.40100	0.20050	0.10025	0.05729	0.04010
Per cent bias of var. est.	7.32580	−19.87098	−10.90107	−5.21060	−2.93154	−2.04035
Linear prob. model est.	0.09992	0.09992	0.09992	0.09992	0.09992	0.09992
Per cent bias of linear est.	−0.08325	−0.08325	−0.08325	−0.08325	−0.08325	−0.08325
Second parameter						
True value	−0.50000	−0.50000	−0.50000	−0.50000	−0.50000	−0.50000
Expect. of M.L. estimate	−0.40695	−0.54780	−0.52948	−0.51365	−0.50757	−0.50524
Per cent bias of est.	−18.60931	9.56094	5.89654	2.72972	1.51402	1.04790
True variance of M.L. est.	1.45241	1.00740	0.46265	0.21631	0.12049	0.08353
Calc. variance of M.L. ext.	1.63443	0.81722	0.40861	0.20430	0.11675	0.08172
Per cent bias of var. est.	12.53225	−18.87888	−11.68067	−5.54934	−3.11136	−2.16302
Linear prob. model est.	−0.49467	−0.49467	−0.49467	−0.49467	−0.49467	−0.49467
Per cent bias of linear est.	−1.06652	−1.06652	−1.06652	−1.06652	−1.06652	−1.06652
Prob. of existence of M.L. est.	0.85421	0.99181	0.99996	1.00000	1.00000	1.00000
True value	0.10000	0.10000	0.10000	0.10000	0.10000	0.10000
Expect. of M.L. estimate	0.08204	0.10990	0.10582	0.10270	0.10150	0.10104
Per cent bias of est.	−17.96158	9.90456	5.82398	2.69820	1.49773	1.03691
True variance of M.L. est.	1.48579	1.00250	0.45254	0.21246	0.11852	0.08220
Calc. variance of M.L. ext.	1.61003	0.80501	0.40251	0.20125	0.11500	0.08050
Per cent bias of var. est.	8.36201	−19.69933	−11.05610	−5.27618	−2.96650	−2.06423
Linear prob. model est.	0.09942	0.09942	0.09942	0.09942	0.09942	0.09942
Per cent bias of linear est.	−0.58076	−0.58076	−0.58076	−0.58076	−0.58076	−0.58076
Prob. of existence of M.L. est.	0.87160	0.99496	0.99999	1.00000	1.00000	1.00000

TABLE 2—continued

	Sample size					
	10	20	40	80	140	200
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	0.70474	1.01873	1.06783	1.03421	1.01864	1.01282
Per cent bias of est.	−29.52618	1.87276	6.78330	3.42109	1.86412	1.28241
True variance of M.L. est.	1.23857	0.98412	0.55049	0.25613	0.14019	0.09663
Calc. variance of M.L. ext.	1.86941	0.93470	0.46735	0.23368	0.13353	0.09347
Per cent bias of var. est.	50.93327	−5.02084	−15.10210	−8.76667	−4.75197	−3.26642
Linear prob. model est.	0.90112	0.90112	0.90112	0.90112	0.90112	0.90112
Per cent bias of linear est.	−9.88763	−9.88763	−9.88763	−9.88763	−9.88763	−9.88763
Prob. of existence of M.L. est.	0.71254	0.94144	0.99680	0.99999	1.00000	1.00000
First parameter (all cases)						
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	0.73234	1.04040	1.06643	1.03222	1.01764	1.01216
Per cent bias of est.	−26.76589	4.04022	6.64324	3.22242	1.76425	1.21578
True variance of M.L. est.	0.52531	0.49435	0.30950	0.14142	0.07686	0.05285
Calc. variance of M.L. est.	1.01723	0.50862	0.25431	0.12715	0.07266	0.05086
Per cent bias of var. est.	93.64346	2.88570	−17.83370	−10.08782	−5.46810	−3.76027
Linear prob. model est.	0.92423	0.92423	0.92423	0.92423	0.92423	0.92423
Per cent bias of linear est.	−7.57657	−7.57657	−7.57657	−7.57657	−7.57657	−7.57657
Second parameter						
True value	−0.50000	−0.50000	−0.50000	−0.50000	−0.50000	−0.50000
Expect. of M.L. estimate	−0.33180	−0.49607	−0.53624	−0.51826	−0.50991	−0.50681
Per cent bias of est.	−33.64020	−0.78652	7.24721	3.65173	1.98177	1.36160
True variance of M.L. est.	1.20695	1.00351	0.55487	0.25493	0.13986	0.09646
Calc. variance of M.L. ext.	1.86828	0.93414	0.46707	0.23354	0.13345	0.09341
Per cent bias of var. est.	54.79349	−6.91277	−15.82285	−8.39383	−4.58190	−3.15818
Linear prob. model est.	−0.43440	−0.43440	−0.43440	−0.43440	−0.43440	−0.43440
Per cent bias of linear est.	−13.12060	−13.12060	−13.12060	−13.13060	−13.12060	−13.12060
Prob. of existence of M.L. est.	0.70992	0.94799	0.99802	1.00000	1.00000	1.00000

True value	0.10000	0.10000	0.10000	0.10000	0.10000	0.10000
Expect. of M.L. estimate	0.05511	0.08855	0.10715	0.10465	0.10248	0.10169
Per cent bias of est.	−44.88654	−11.44656	7.14974	4.65396	2.48035	1.69241
True variance of M.L. est.	1.01662	0.97802	0.63496	0.29179	0.15804	0.10854
Calc. variance of M.L. ext.	2.08464	1.04232	0.52116	0.26058	0.14890	0.10423
Per cent. bias of var. est.	105.05601	6.57426	−17.92244	−10.69591	−5.78006	−3.96928
Linear prob. model est.	0.07681	0.07681	0.07681	0.07681	0.07681	0.07681
Per cent bias of linear est.	−23.19389	−23.19389	−23.19389	−23.19389	−23.19389	−23.19389
Prob. of existence of M.L. est.	0.60127	0.90235	0.99491	0.99999	1.00000	1.00000
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	0.39590	0.67801	0.97380	1.06299	1.04043	1.02706
Per cent bias of est.	−60.40966	−12.19939	−2.62020	6.29931	4.04250	2.70572
True variance of M.L. est.	0.76186	0.80262	0.68676	0.43638	0.24022	0.16047
Calc. variance of M.L. ext.	2.92211	1.46106	0.73053	0.36526	0.20872	0.14611
Per cent bias of var. est.	283.54796	82.03495	6.37315	−16.29632	−13.11233	−8.94926
Linear prob. model est.	0.59895	0.59895	0.59895	0.59895	0.59895	0.59895
Per cent bias of linear est.	−40.10460	−40.10460	−40.10460	−40.10460	−40.10460	−40.10460
Prob. of existence of M.L. est.	0.37108	0.68762	0.91927	0.99376	0.99986	1.00000

extreme and nonextreme selection probabilities. On the other hand, we anticipate that the qualitative structure of biases will be unchanged by the addition of independent variables or of multiple-choice alternatives, as the estimator is analogous to a linear statistical estimator except for the nonlinear dependence of the selection probabilities on the parameters.

IV. An Empirical Application

The theory of qualitative choice behavior outlined above has been applied to several areas of consumer choice. The author (1968) has investigated the criteria employed by a state highway department in selecting urban freeway routes. The determinants of college choice have been studied by Professor Miller and Professor Radner (1970), and the results have been used to forecast the effects of changing educational policy on college enrollment. Professor Boskin (1972) has applied the model to the problem of occupational choice. Studies in progress are investigating urban trip generation, distribution, and modal choice; labor force participation and job search decisions; housing location and type; recidivism; child-bearing decisions and the implications of population control policy; choice of consumer durables; and rural-urban migration decisions. To illustrate the method, we reproduce here selected results on shopping trip mode and destination decisions obtained in a study of travel demand models by T. Domencich and D. McFadden (1972).[16]

The objective of this study is to develop disaggregated, policy-oriented, behavioral models of urban trip generation, distribution, and mode. The behavioral unit studied is the individual trip-maker, faced with decisions on whether to take a trip, mode, and destination. The empirical analysis is based on a household survey in Pittsburgh conducted by the Southwestern Pennsylvania Regional Planning Commission in 1967, supplemented with time and cost data collected by the study authors. A detailed description of the sample frame and variables collected is given in the study.

The analysis of shopping travel behavior is separated into three decisions: (1) choice of mode for trips actually made at the observed time and to the observed destination; (2) choice of destination for trips made at an observed time by preferred mode; and (3) choice of whether or not to make a trip, given a preferred time, mode, and destination.[17] The results of each analysis are summarized in turn.

[16] The results below are reproduced with the permission of Charles River Associates, Inc.

[17] The separation of these decisions is justified in the study by postulating a "tree" utility structure; we shall not repeat the argument here.

Shopping choice of mode

Choice between public transit and auto mode is examined for a sample of 140 individual shopping trips. For each observation, walk access to transit is available. The sample is drawn from a southern suburban corridor of Pittsburgh and a central city corridor running from downtown to the east. A number of alternative models were fitted; those giving the most satisfactory results in terms of fitting the base data are described in Table 3. All coefficients in these models are of the expected sign. The coefficients of transit walk time, auto in-vehicle less transit station-to-station time, and auto operating costs less transit fares imply a value of walk time of \$5.46 per hour and a value of in-vehicle time of \$0.95 per hour in Model 1. These values are in close accord

TABLE 3

CONDITIONAL LOGIT MODEL OF SHOPPING MODE CHOICE; DEPENDENT VARIABLE EQUALS THE LOG ODDS OF CHOICE OF AUTO MODE; BINARY LOGIT MAXIMUM LIKELIHOOD ESTIMATES; STANDARD ERRORS IN PARENTHESES

Independent variable	Model 1	Model 2	Model 3	Model 4
Pure auto mode preference effect (constant)[a]	−6.77 (1.66)	−6.20 (2.10)	−6.65 (1.54)	−6.37 (1.82)
Transit walk time (minutes)	0.374 (0.328)	0.398 (0.410)	0.30 (0.351)	0.274 (0.612)
Transit wait plus transfer time (minutes)	—	—	0.0647 (0.0403)	—
Transit station-to-station time (minutes)	—	—	—	0.0532 (0.0455)
Auto in-vehicle time (minutes)	—	—	—	−0.0486 (0.0956)
Auto in-vehicle time less transit station-to-station time (minutes)	−0.0654 (0.0320)	−0.0636 (0.0398)	—	—
Auto in-vehicle time less transit line-haul time (minutes)	—	—	−0.0287 (0.0715)	—
Auto operating cost less transit fares (dollars)	−4.11 (1.67)	−4.66 (2.06)	−4.10 (2.13)	−4.06 (1.74)
Ratio of number of autos to number of workers in the household	2.24 (1.11)	2.26 (1.14)	2.01 (1.04)	1.89 (0.76)
Race of respondent (0 if white, 1 if non-white)	—	−2.18 (1.26)	—	—
Occupation of head of household (0 if blue-collar, 1 if white collar)	—	−1.53 (1.10)	—	—

[a] Because of the sample selection procedure and the presence of the last three variables giving socioeconomic-auto mode interaction effects, this constant cannot be interpreted as a "transit" bias that would be replicated in a random sample of the population.

with shopping trip value of time studies. Thus, the estimates seem quite stable despite the relatively large standard errors. The models provide excellent fits of the base-line data; in Model 1, the probability of selecting the actual mode is greater than one-half for 133 of the 140 observations, and greater than 0.9 for 116 of the 140 observations.

Shopping choice of destination

The choice of shopping destination is analyzed for 63 auto-mode trips starting from the southern suburban corridor. The possible alternative destinations for each observation are selected by dividing the city into zones and choosing all destination zones to which there is a trip in the sample from the origin zone. The number of alternatives varies from three to five in this sample.

This model is estimated using only three explanatory variables, an inclusive index of the "price" of a trip in terms of time and cost, an index of the "attractiveness" of each shopping destination, and an interaction of the inclusive price index and a socioeconomic variable, the number of preschool children. The inclusive price is defined from the shopping mode choice Model 1 to be

$$[\text{Inclusive price}] = 0.0654[\text{Auto in-vehicle time}] + 4.11[\text{Auto operating cost}].$$

The measure of destination attractiveness is taken to be the retail employment in the zone as a percentage of total retail employment in the region. Because the alternative destinations are unranked and vary from one observation to the next, the explanatory variables enter generically. In particular, it is assumed that there are no "specific destination" effects. The results of the estimation are given in Table 4. The two independent variables above are both found to be

TABLE 4

CONDITIONAL LOGIT MODEL OF SHOPPING DESTINATION CHOICE; DEPENDENT VARIABLE EQUALS THE LOG ODDS THAT ONE DESTINATION ZONE IS CHOSEN OVER A SECOND; MULTINOMIAL LOGIT MAXIMUM LIKELIHOOD ESTIMATES; STANDARD ERRORS IN PARENTHESES

Independent variable	Model 5	Model 6
Inclusive price of trip (weighted time and cost using Model 1 weights)	−1.06 (0.28)	−0.602 (0.159)
Index of attractiveness of destination	0.844 (0.227)	0.832 (0.224)
Interaction effect equals the inclusive price of trip times the number of preschool children	—	−0.521 (0.343)

highly significant. Model 6 yields calculated selection probabilities which are maximum for the actual destination in 29 cases, as opposed to a match for 16 cases which would be expected by chance.

Shopping trip frequency

The decision of whether to take a shopping trip on a given day is analyzed for a sample of 80 households in the southern suburban corridor, of whom 59 recorded a shopping trip on the survey day. An inclusive price of a trip for nontrip takers is calculated by assuming that the distribution of destination preferences is that determined in Model 6, and that utility has a separable form implying this distribution is independent of the distribution of tastes for taking auto trips. The independent variables in the model are the preference-distribution-weighted inclusive price, the measure of attractiveness of shopping zone used above, similarly weighted, and a household-income shopping trip

TABLE 5

CONDITIONAL LOGIT MODEL OF SHOPPING TRIP FREQUENCY; DEPENDENT VARIABLE EQUALS THE LOG ODDS OF MAKING A SHOPPING TRIP ON SAMPLED DAY; BINOMIAL LOGIT MAXIMUM LIKELIHOOD ESTIMATES; STANDARD ERRORS IN PARENTHESES

Independent variable	Model 7	Model 8
Inclusive price of trip (weighted time and cost using Model 6 weights)	−1.72 (0.54)	−2.25 (0.68)
Index of attractiveness of destination	3.90 (1.08)	2.85 (1.19)
Family income	—	−0.199 (0.195)

interaction variable. The estimates are given in Table 5. Model 7 predicts the actual decision with probability 0.5 or better for 60 of the 80 observations.

The models above of shopping mode, destination, and frequency decisions can be combined with distributions of the independent variables in an urban area to produce trip generation and distribution tables by mode.[18] These

[18] Such tables could also be generated by aggregating over individuals for a random sample of the population, a procedure that requires a smaller sample than that necessary to obtain accurate cell frequencies for a detailed classification of multiple-independent variables. In particular, the sample used to calibrate the models may be utilized to produce trip tables. On the other hand, when samples of sufficient size are available to obtain cell frequencies, it may be possible to calibrate the model using the Berkson-Theil estimation procedure.

tables are functions of policy variables such as transit fares and wait times, and can be recalculated to forecast the effects of policy changes on the transportation system. Because the parameters are estimated from data at the level of the individual decision, they do not suffer from the "fallacy of composition" that could occur in attempting to infer response elasticities from data on behavior of heterogeneous groups. Thus, this modeling approach has the potential of providing accurate forecasts of the response of shopping travel demand to policy variables, in a framework that exploits the common thread of utility maximization and taste distribution in a variety of choice situations.

The empirical study summarized above represents a typical application of the theory of qualitative choice behavior of populations of consumers, with the conditional logit specification of the distribution of tastes. For applications in which the independence of irrelevant alternatives is plausible, this statistical procedure provides an analog for qualitative dependent variables of the conventional linear statistical model.

Appendix: Proofs of Statistical Properties

This section outlines proofs of Lemmas 5 and 6, and the properties of statistics based on weighted residuals.

LEMMA 5. Suppose Axioms 1-4 and 7 hold. Then the probability that the maximum likelihood estimator exists approaches unity as $\sum_{n=1}^{N} R_n$ approaches infinity.

Proof: As noted in the text, Axiom 7 implies that Axiom 5 holds when $\sum_{n=1}^{N} R_n$ is large. We next show that Axiom 7 implies a second linear independence condition. Let m be a serial index of trials and repetition; e.g., m identifies the r_mth repetition of trial n_m. We shall show that there exists an infinite subset M of the indices m, and integers i_m, j_m satisfying $1 \leqslant i_m, j_m \leqslant J_{n_m}$ such that each sequence of K successive vectors $z_{i_m n_m} - z_{j_m n_m}$ for $m \in M$ are linearly independent. We proceed by induction. Axiom 7 implies there is some m, i, j such that $z_{in_m} - z_{jn_m} \neq 0$; set $M_1 = \{m\}$, $i_1 = i$, and $j_1 = j$. Suppose that we have constructed a set M_{l-1} containing $l-1$ indices that satisfy the required property. Suppose there does not exist an index m_l such that $M_l = M_{l-1} \cup \{m_l\}$ has the desired property. Then, for all $m > m_{l-1}$ and $1 \leqslant i, j \leqslant J_{n_m}$, the vector $z_{in_m} - z_{jn_m}$ can be written as a linear combination of vectors $z_{i_p n_p} - z_{j_p n_p}$ for the last $K-1$ or fewer elements p of M_{l-1}. But then $z_{in_m} - \bar{z}_{n_m} = \sum_{j=1}^{J_{n_m}} P_{jn_m}(z_{in_m} - z_{jn_m})$ also has this property, implying that the limiting matrix Ω in Axiom 7 is singular, for a contradiction. Hence, by induction, the set $M = \bigcup_{l=1}^{\infty} M_l$ has the desired property.

Partition the set M into successive subsets $M^1, M^2, \ldots$, each containing K

elements. Let W^q denote the matrix with rows $z_{i_m n_m} - z_{j_m n_m}$ for $m \in M^q$. Then, W^q is nonsingular and the linear transformation $(W^{q-1})(W^q)^{-1}$ is continuous. Hence, for even q one can define a strictly positive vector a^{q-1} and a vector a^q with no zero elements such that $a^q W^q = a^{q-1} W^{q-1}$.

For q even, consider the event in which alternative i_m is chosen for $m \in M^{q-1}$; alternative i_m is chosen for $m \in M^q$ if $a_m{}^q$ is negative; and alternative j_m is chosen for $m \in {}_iM^q$ if $a_m{}^q$ is positive. Suppose this event occurs, but Axiom 6 fails, and let $\gamma \neq 0$ be such that $S_{in}(z_{jn} - z_{in})\gamma \leqslant 0$ for all i, j, n. Then, $W^{q-1}\gamma$ is a nonnegative vector. Since W^{q-1} is nonsingular, it has at least one component positive. Hence, $a^q W^q \gamma = a^{q-1} W^{q-1} \gamma > 0$. But z^q and $W^q \gamma$ have opposite signs in each component, contradicting the last inequality. Therefore, when this event occurs, Axiom 6 holds and the maximum likelihood estimator exists.

The selection probabilities are bounded below by

$$P_{in} \geqslant 1/J_* e^{2M|\theta|} \equiv P_* > 0, \tag{42}$$

where J_* and M are the bounds given by Axiom 7. Hence, the probability that the event above occurs for an even q is at least P_*^{2K}, and the probability that this event occurs for some even $q \leqslant 2q'$ is at least $1-(1-P_*^{2K})^{q'}$. This last probability approaches unity as $q' \to +\infty$, proving the lemma.

LEMMA 6. Suppose Axioms 1–4 and 7 hold, θ^0 is the true parameter vector and $\hat{\theta}^m$ is the maximum likelihood estimator for a sample of size $m = \sum_{n=1}^{N} R_n$. Then $\hat{\theta}^m$ is consistent and asymptotically normal as $m \to +\infty$, with

$$\sqrt{m}\,\Omega^{1/2}(\hat{\theta}^m - \theta^0),$$

tending to a multivariate normal distribution with mean zero and identity covariance matrix.

Proof: We shall first establish that $\hat{\theta}^m$ is a consistent estimator of θ^0. From Axiom 7, $|z_{in}| \leqslant M$ for a positive scalar M. Differentiation of Equation (16) yields the bounds

$$\begin{aligned} \left|\frac{\partial \log P_{in}}{\partial \theta}\right| &\leqslant 2M, \\ \left|\frac{\partial^2 \log P_{in}}{\partial \theta\, \partial \theta'}\right| &\leqslant 4M^2, \\ \left|\frac{\partial^3 \log P_{in}}{\partial \theta\, \partial \theta'\, \partial \theta_k}\right| &\leqslant 8M^3, \end{aligned} \tag{43}$$

uniform in θ.

Let m be a serial index of trials and repetitions, and let S_{im} equal unity if alternative i is chosen at this observation and zero otherwise, for $1 \leqslant i \leqslant J_{n_m}$. Define a sequence of independent random variables

$$\lambda^m(\theta) = \sum_{i=1}^{J_{nm}} S_{im} \log P_{in}(\theta). \tag{44}$$

Then

$$L^q(\theta) = C_q + \sum_{m=1}^{q} \lambda^m(\theta), \tag{45}$$

with C_q a constant independent of θ, is the log likelihood function. From Equation (43), the derivatives of λ^m, denoted $\partial\lambda^m/\partial\theta = \lambda_\theta{}^m$, etc., satisfy the uniform bounds $|\lambda_\theta{}^m| \leqslant 2M$, $|\lambda^m_{\theta\theta'}| \leqslant 4M^2$, and $|\lambda^m_{\theta\theta'\theta_k}| \leqslant 8M^3$. Further,

$$E\lambda_\theta{}^m(\theta^0) = \left[\sum_{i=1}^{J_{nm}} \frac{\partial P_{in_m}}{\partial\theta}\right]_{\theta^0} = 0. \tag{46}$$

Define

$$\begin{aligned} \Omega_m &= -E\lambda^m_{\theta\theta'}(\theta^0) \\ &= \sum_{i=1}^{J_{nm}} (z_{in_m} - \bar{z}_{n_m})' P_{in_m} (z_{in_m} - \bar{z}_{n_m}). \end{aligned} \tag{47}$$

Then by Axiom 7,

$$\lim_{q\to\infty} \frac{1}{q} \sum_{m=1}^{q} \Omega_m = \Omega. \tag{48}$$

Let β denote the smallest characteristic value of Ω. Then there exists q_0 such that for $q \geqslant q_0$, the smallest characteristic value of $(1/q)\sum_{m=1}^{q} \Omega_m$ is at least $\beta/2$.

Given a small positive scalar ε, choose $\delta = \min\{\varepsilon, \beta/4(1+4KM^3)\}$. The strong law of large numbers (Feller, 1966, Vol. II, p. 233) implies that there exists $q_1 \geqslant q_0$ such that for $q \geqslant q_1$, $L_\theta{}^q(\theta^0) = \sum_{m=1}^{q} \lambda_\theta{}^m(\theta^0)$ satisfies

$$\left|\frac{1}{q} L_\theta{}^q(\theta^0)\right| < \delta^2, \tag{49}$$

with probability at least $1-\varepsilon$.

A second-order Taylor's expansion of $L^q_{\theta_k}$ about θ^0 yields

$$\frac{1}{q} L^q_{\theta_k}(\theta) - \frac{1}{q} L^q_{\theta_k\theta}(\theta^0)(\theta-\theta^0) = \frac{1}{q} L^q_{\theta_k}(\theta^0) + \frac{1}{2}\frac{1}{q}(\theta-\theta^0)' L^q_{\theta_k\theta\theta'}(\tilde{\theta})(\theta-\theta^0), \tag{50}$$

where $\tilde{\theta}$ lies between θ and θ^0. Consider $q \geqslant q_1$ and θ satisfying

$$(\theta-\theta^0)'(\theta-\theta^0) = \delta^2,$$

and suppose Equation (49) holds. Then

$$\left|\frac{1}{q}L^q_{\theta_k}(\theta) - \frac{1}{q}L^q_{\theta_k\theta}(\theta^0)(\theta-\theta^0)\right| \leqslant \delta^2(1+4K^2M^3) \leqslant \delta\beta/4. \tag{51}$$

Hence,

$$\frac{1}{q}|(\theta-\theta^0)'L_\theta{}^q(\theta) - (\theta-\theta^0)'L_{\theta\theta'}(\theta^0)(\theta-\theta^0)| \leqslant \delta^2\beta/4. \tag{52}$$

But $(1/q)(\theta-\theta^0)'L_{\theta\theta'}(\theta^0)(\theta-\theta^0) \leqslant -\delta^2\beta/2$, implying

$$\frac{1}{q}(\theta-\theta^0)'L_\theta{}^q(\theta) \leqslant -\delta^2\beta/4. \tag{53}$$

Hence, at each point on the sphere $(\theta-\theta^0)'(\theta-\theta^0) = \delta^2$, the gradient $L_\theta{}^q(\theta)$ is directed inward. Since L^q is concave in θ, this implies that a maximum $\hat{\theta}^q$ of L^q is achieved inside this sphere. Since this event occurs with probability at least $1-\varepsilon$, we have proved the estimator to be consistent.

We next show that $\sqrt{q}\,\Omega^{1/2}(\hat{\theta}^q-\theta^0)$ is asymptotically standard multivariate normal. Evaluating the Taylor's expansion (50) at the maximum likelihood estimator yields

$$0 = \frac{1}{q}L^q_{\theta_k}(\theta^0) + \frac{1}{q}L^q_{\theta_k\theta}(\theta^0)(\theta-\theta^0) + 4M^3|\hat{\theta}^q-\theta^0|\,a_k(\theta-\theta^0), \tag{54}$$

where a_k is a $1\times K$ vector depending on $\hat{\theta}^q$, which satisfies $|a_k|\leqslant 1$. Letting A denote the matrix with rows a_k and defining

$$D_q = \Omega^{-1/2}\left(\frac{1}{q}\sum_{m=1}^{q}\Omega_m - 4M^3|\hat{\theta}^q-\theta^0|\,A\right)\Omega^{-1/2}, \tag{55}$$

Equation (54) can be written

$$\frac{1}{\sqrt{q}}\Omega^{1/2}D_q[\sqrt{q}\,\Omega^{1/2}(\hat{\theta}^q-\theta^0)] = \frac{1}{q}L_\theta{}^q(\theta^0). \tag{56}$$

Then,

$$\operatorname*{plim}_{q\to\infty} D_q = I - 4M^3\Omega^{-1/2}A\Omega^{-1/2}\operatorname*{plim}_{q\to\infty}|\hat{\theta}^q-\theta^0| = I. \tag{57}$$

Hence, $\sqrt{q}\,\Omega^{1/2}(\hat{\theta}^q-\theta^0)$ has the same asymptotic distribution as

$$\frac{1}{q}\Omega^{-1/2}L_\theta{}^q(\theta^0) = \frac{1}{q}\sum_{m=1}^{q}\Omega^{-1/2}\lambda_\theta{}^m(\theta^0). \tag{58}$$

But the independent random variables $\Omega^{-1/2}\lambda_\theta{}^m(\theta^0)$ satisfy the Lindeberg–Levy theorem (Feller, 1966, Vol. II, pp. 256–258), implying that Equation (58) is asymptotic standard normal. This proves the lemma.

Since $\sqrt{q}\,\Omega^{1/2}(\hat{\theta}^q-\theta^0)$ is asymptotically standard multivariate normal, it follows that $q(\hat{\theta}^q-\theta^0)'\Omega(\hat{\theta}^q-\theta^0)$ is asymptotically chi-square with K degrees of freedom. Further, a second-order Taylor's expansion of the log-likelihood function about $\hat{\theta}^q$ can be used to establish that $q(\hat{\theta}^q-\theta^0)'\Omega(\hat{\theta}^q-\theta^0)$ and $2[L(\theta^0)-L(\hat{\theta})]$ converge in probability, and hence have the same asymptotic distribution. This argument justifies use of the statistic (29). Details of the proof can be found in Theil (1971, p. 396), Rao (1968, pp. 347–351), and Kendall and Stuart (1967, Vol. II, pp. 230–236). Rao gives several asymptotically equivalent forms for the test.

Consider the weighted residuals D_{in} in Equation (31). Define

$$D_{in}^0 = \frac{S_{in}-R_n P_{in}^0}{(R_n P_{in}^0)^{1/2}}, \tag{59}$$

$$D_{in}^1 = \sum_{m=1}^{N}\sum_{j=1}^{J_m}(r_n r_m P_{in}^0 P_{jm}^0)^{1/2}(z_{in}-\bar{z}_n)\Omega^{-1}(z_{jm}-\bar{z}_m)' D_{jm}^0, \tag{60}$$

where $P_{in}^0 = P_{in}(\theta^0)$. Then D_{in}^0 has a multivariate distribution with $ED_{in}^0 = 0$ and $ED_{in}^0 D_{jm}^0 = \delta_{nm}[\delta_{ij}-(P_{in}^0 P_{jn}^0)^{1/2}]$, implying $ED_{in}^1 = 0$,

$$ED_{in}^0 D_{jm}^1 = (r_n r_m P_{in}^0 P_{jm}^0)^{1/2}(z_{in}-\bar{z}_n)\Omega^{-1}(z_{jm}-\bar{z}_m), \tag{61}$$

and asymptotically

$$ED_{in}^1 D_{jm}^1 = (r_n r_m P_{in}^0 P_{jm}^0)^{1/2}(z_{in}-\bar{z}_n)\Omega^{-1}(z_{jm}-\bar{z}_m). \tag{62}$$

Making Taylor's expansions, one can show that the random variables

$$\frac{(R_n)^{1/2}(P_{in}^0-P_{in})}{(P_{in}^0)^{1/2}} \quad \text{and} \quad (r_n P_{in}^0)^{1/2}(z_{in}-\bar{z}_n)\left(\sum_{n=1}^{N} R_n\right)^{1/2}(\theta^0-\hat{\theta})$$

differ from $-D_{in}^1$ by terms with probability limits zero. It then follows, since $(\sum_{n=1}^{N} R_n)^{1/2}(\theta^0-\hat{\theta})$ is asymptotic normal with mean zero and covariance matrix Ω, that these three random variables have a common asymptotic normal distribution. Hence, $D_{in}^0-D_{in}^1$ has an asymptotic distribution with mean zero and covariance

$$\begin{aligned}\Lambda_{in,jm} &= \delta_{nm}[\delta_{ij}-(P_{in}^0 P_{jm}^0)^{1/2}] \\ &\quad -(r_n r_m P_{in}^0 P_{jm}^0)^{1/2}(z_{in}-\bar{z}_n)\Omega^{-1}(z_{jm}-\bar{z}_m).\end{aligned} \tag{63}$$

Write

$$D_{in} = \frac{S_{in}-R_n P_{in}^0}{(R_n P_{in})^{1/2}} + \frac{(R_n)^{1/2}(P_{in}^0-P_{in}^0)}{(P_{in})^{1/2}}. \tag{64}$$

The first term differs from D_{in}^0, and the second from $-D_{in}^1$, by factors with probability limits zero. Hence, D_{in} has the same asymptotic distribution as $D_{in}^0 - D_{in}^1$. When the R_n approach infinity, D_{in}^0 is asymptotically normal, implying D_{in} asymptotically normal. The covariance matrix can be written

$$\Lambda = I - \sum_{m=0}^{N} q_m q_m', \tag{65}$$

where

$$(q_0)_{in} = (r_n P_{in}^0)^{1/2}(z_{in} - \bar{z}_n)\Omega^{-1/2},$$
$$(q_m)_{in} = \delta_{mn}(P_{in}^0)^{1/2}, \qquad m = 1, \ldots, N.$$

The vectors q_m are orthonormal, implying Λ idempotent of rank $N^* = \sum_{n=1}^{N} J_n - N - K$. Then $G = \sum_{n=1}^{N}\sum_{j=1}^{J_n} D_{in}^2$ has an asymptotic chi-square distribution with N^* degrees of freedom (Rao, 1968, p. 149).

Next consider the linear transformations

$$Y_{in} = D_{in} - D_{1n}(P_{in})^{1/2}\alpha_n, \qquad i = 2, \ldots, J_n, \tag{66}$$

where $\alpha_n = [1 - (P_{1n})^{1/2}](1 - P_{1n})$. Then, Y_{in} has the same asymptotic distribution as the random variable Y_{in}^0 formed by replacing P_{in} with P_{in}^0 in Equation (66), and the latter random variable has asymptotic moments $EY_{in}^0 = 0$ and

$$\begin{aligned}\Gamma_{in,jm} = EY_{in}^0 Y_{jm}^0 &= ED_{in}D_{jm} - (P_{jm}^0)^{1/2}\alpha_m ED_{in}D_{1m} \\ &\quad - (P_{in}^0)^{1/2}\alpha_n ED_{1n}D_{jm} + (P_{in}^0 P_{jm}^0)^{1/2}\alpha_n\alpha_m ED_{1n}D_{1m} \\ &= \delta_{nm}\delta_{ij} - q_{in}' q_{jm},\end{aligned}$$

with

$$q_{in}' = (r_n P_{in}^0)^{1/2}\left((z_{in} - \bar{z}_n) + \frac{P_{1n}^0 - (P_{1n}^0)^{1/2}}{1 - P_{1n}^0}(z_{1n} - \bar{z}_n)\right)\Omega^{-1/2}.$$

Then, $\sum_{n=1}^{N}\sum_{i=2}^{J_n} q_{in}q_{in}' = I_K$ and Γ is idempotent of rank N^*. Hence,

$$\sum_{n=1}^{N}\sum_{i=2}^{J_n} Y_{in}^2 = \sum_{n=1}^{N}\sum_{i=1}^{J_n} D_{in}^2$$

has an asymptotic chi-square distribution with N^* degrees of freedom.

REFERENCES

Aitchison, J. and J. Bennett (1970). "Polychotomous quantal response by maximum indicant," *Biometrika* **57**, 253-262.

Aitchison, J. and S. Silvey (1957). "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika* **44**, 131-140.

Allouche, J. (1972). "Approach to Probability Distribution of Value of Walking Time and Pedestrian Circulation Models," *Highway Res. Rec.* **392**, 121-133.

Amemiya, T. (1972). "Bivariate Probit Analysis: Minimum Chi-Square Methods," Dept. of Econ., Stanford Univ., Stanford, California, unpublished.

Amemiya, T. and M. Boskin (1972). "Regression Analysis when the Dependent Variable is Truncated Lugnormal, with An Application to the Determinants of the Duration of Welfare Dependency," Inst. for Math. Stud. in the Soc. Sci., Stanford Univ., Stanford, California, Tech. Rep. No. 75.

Antle, C., L. Klimko and W. Harkness (1970). "Confidence intervals for the parameters of the logistic distribution," *Biometrika* **57**, 397–402.

Atkinson, A. (1972). "A Test of the Linear Logistic and Bradley-Terry Models," *Biometrika* **59**, 37–42.

Berkson, J. (1951). "Why I prefer Logits to Probits," *Biometrics* **7**, 327–339.

Berkson, J. (1955). "Maximum Likelihood and Minimum Chi-Square Estimates of the Logistic Function," *J. Amer. Statist. Ass.* **50**, 130–162.

Bloch, D. and G. Watson (1967). "A Bayesian Study of the Multinomial Distribution," *Ann. Math. Statist.* **38**, 1423–1435.

Block, H. and J. Marschak (1960). "Random Orderings and Stochastic Theories of Response," *In Contributions to Probability and Statistics* (I. Olkin, ed.). Stanford Univ. Press, Stanford, California.

Bock, R. (1969). "Estimating Multinomial Response Relations," *In Contributions to Statistics and Probability: Essays in Memory of S. N. Roy* (R. Bose, ed.). Univ. of North Carolina Press.

Boskin, M. (1972). "A Conditional Logit Model of Occupational Choice," Dept. of Econ., Stanford Univ., Stanford, California, unpublished.

Cox, D. (1958). "The Regression Analysis of Binary Sequences," *J. Roy. Statist. Soc. Ser. B*, **20**, 215–242.

Cox, D. (1966). "Some Procedures Connected with the Logistic Qualitative Response Curve," *Research Papers in Statistics* (F. David, ed.), pp. 55–71. Wiley, New York.

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

Cox, D. and E. Snell (1968). "A General Definition of Residuals," *J. Roy. Statist. Soc. Ser B* **30**, 248–265.

Cox, D. and E. Snell (1971). "On Test Statistics Calculated from Residuals," *Biometrika* **58**, 589–594.

Domencich, T. and D. McFadden (1972). *A Disaggregated Behavioral Model of Urban Travel Demand*, Report No. CRA-156-2, Charles River Associates, Inc., Cambridge, Massachusetts.

Ergun, G. (1971). "Development of a Downtown Parking Mode," *Highway Res. Rec.* **369**, 118–134.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York.

Finney, D. (1952). *Probit Analysis*. Cambridge Univ. Press, London and New York.

Fisher, J. (1962). "An Analysis of Consumer Goods Expenditures in 1957," *Rev. Econ. Statist.* **44**, 64–71.

Gart, J. and J. Zweifel (1967). "On the Bias of Various Estimators of the Logit and its Variance," *Biometrika* **54**, 181–187.

Gilbert, E. (1968). "On Discrimination Using Qualitative Variables," *J. Amer. Statist. Ass.* **63**, 1399–1412.

Goodman, L. (1970). "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications," *J. Amer. Statist. Ass.* **65**, 226–256.

Goodman, L. (1972). "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables," *Amer. Sociolog. Rev.* **37**, 28–46.

Grizzle, J. (1962). "Asymptotic Power of Tests of Linear Hypotheses Using the Probit and Logit Transformation," *J. Amer. Statist. Ass.* **57**, 877–894.

Grizzle, J. (1971). "Multivariate Logit Analysis," *Biometrics* **27**, 1057–1062.

Gupta, S., A. Qureishi and B. Shah (1967). "Best Linear Unbiased Estimators of the Parameters of the Logistic Distribution Using Order Statistics," *Technometrics* **9**, 43–56.

Gurland, J., I. Lee and P. Dolan (1960). "Polychotomous quantal response in biological assay," *Biometrics* **16**, 382–398.

Harter, J. and A. Moore (1967). "Maximum likelihood estimation, from censored samples, of the parameters of a logistic distribution," *J. Amer. Statist. Ass.* **62**, 675–683.

Kendall, M. and A. Stuart (1967). *The Advanced Theory of Statistics*. Vol. 2. Hafner, New York.

Korbel, J. (1966). "Labor Force Entry and Attachment of Young People," *J. Amer. Statist. Ass.* **61**, 117–127.

Lave, C. (1968). Modal Choice in Urban Transportation: A Behavioral Approach. Dept. of Econ., Stanford Univ., Stanford, California, Ph.D. dissertation.

Lee, T. (1963). "Demand for Housing: A Cross-Section Analysis," *Rev. Econ. Statist.* **45**, 190–196.

Leonard, T. (1972). "Bayesian Methods for Multinomial Data," *American College Testing Program*, Tech. Bull. No. 4.

Lisco, T. (1967). The Value of Commuters' Travel Time: A Study in Urban Transportation. Dept. of Econ., Univ. of Chicago, Chicago, Illinois, Ph.D. dissertation.

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

Luce, R. and P. Suppes (1965). "Preference, Utility, and Subjective Probability," *In Handbook of Mathematical Psychology* (R. Luce, R. Bush, and E. Galanter, eds.), Vol. 3. Wiley, New York.

Marschak, J. (1960). "Binary Choice Constraints on Random Utility Indicators," *In Stanford Symp. Math. Methods Soc. Sci.* (K. Arrow, ed.). Stanford Univ. Press, Stanford, California.

McFadden, D. (1968). "The Revealed Preferences of a Government Bureaucracy," Dept. of Econ., Univ. of California, Berkeley, California, unpublished.

McFadden, D. and M. Richter (1971). "On the extension of a set function to a probability on the Boolean algebra generated by a family of events, with applications," Working Paper 14, MSSB Workshop on the Theory of Markets under Uncertainty, Dept. of Econ., Univ. of California, Berkeley, California.

McGillivray, R. (1969). Binary Choice of Transport Mode in the San Francisco Bay Area, Dept. of Econ., Univ. of California, Berkeley, California, Ph.D. dissertation.

McGillivray, R. (1970). "Demand and Choice Models of Modal Split," *J. Transport Econ. Policy* **4**, 192–207.

Miller, L. and R. Radner (1970). "Demand and Supply in U.S. Higher Education," *Amer. Econ. Rev. Papers & Proceedings*, **60**, 326–340.

Miller, L. and R. Radner (forthcoming). *Demand and Supply in U.S. Higher Education*, Carnegie Commission on the Future of U.S. Higher Education.

Moses, L., R. Beals and M. Levy (1967). "Rationality and Migration in Ghana," *Rev. Econ. Statist.* **49**, 480–486.

Quandt, R. (1968). "Estimation of Modal Splits," *Transportat. Res.* **2**, 41–50.

Quandt, R. (1970). *The Demand for Travel*. Heath.

Quandt, R. and W. Baumol (1966). "The Demand for Abstract Transport Modes: Theory and Measurement," *J. Regional Sci.* **6**, 13–26.

Quandt, R. and K. Young (1969). "Cross Sectional Travel Demand Models: Estimation and Tests," *J. Regional Sci.* **9**, 201–214.

Quarmby, D. (1967). "Choice of Travel Mode for the Journey to Work: Some Findings," *J. Transport Econ. Policy.*

Rao, C. (1965). *Linear Statistical Inference and its Applications.* Wiley, New York.

Rassam, P., R. Ellis and J. Bennett (1971). "The n-Dimensional Logit Model: Development and Application," *Highway Res. Rec.* **369**, 135–147.

Reichman, S. and P. Stopher (1971). "Disaggregate Stochastic Models of Travel-Mode Choice," *Highway Res. Rec.* **369**, 91–103.

Stopher, P. (1969). "A Multinomial Extension of the Binary Logit Model for Choice of Mode of Travel," Northwestern Univ., unpublished.

Stopher, P. and T. Lisco (1970). "Modelling Travel Demand: A Disaggregate Behavioral Approach—Issues and Applications," *Transportat. Res. Forum Proc.* pp. 195–214.

Stopher, P. (1969). "A Probability Model of Travel Mode Choice for the Work Journey," *Highway Res. Rec.* **283**, 57–65.

Talvitie, A. (1972). "Comparison of Probabilistic Modal-Choice Models: Estimation Methods and System Inputs," *Highway Res. Rec.* **392**, 111–120.

Theil, H. (1969). "A Multinomial Extension of the Linear Logit Model," *Int. Econ. Rev.* **10**, 251–259.

Theil, H. (1970). "On the Estimation of Relationships involving Qualitative Variables," *Amer. J. Sociol.* **76**, 103–154.

Theil, H. (1971). *Principles of Econometrics*, Wiley, New York.

Thomas, T. and G. Thompson (1971). "Value of Time Saved by Trip Purpose," *Highway Res. Rec.* **369**, 104–113.

Thurstone, L. (1927). "A Law of Comparative Judgment," *Psycholog. Rev.* **34**, 273–286.

Tobin, J. (1958). "Estimation of Relationships for Limited Dependent Variables," *Econometrica* **26**.

Tversky, A. and J. Russo (1969). "Substitutability and Similarity in Binary Choice ," *J. Math. Psychol.* **6**, 1–12.

Uhler, R. (1968). "The Demand for Housing: An Inverse Probability Approach," *Rev. Econ. Statist.* 50, 129–134.

Walker, F. (1968). "Determinants of auto scrappage," *Rev. Econ. Statist.* **50**, 503–506.

Walker, S. and D. Duncan (1967). "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika* **54**, 167–179.

Warner, S. (1962). *Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice.* Northwestern Univ. Press, Evanston, Illinois.

Zellner, A. and T. Lee (1965). "Joint Estimation of Relationships Involving Discrete Random Variables," *Econometrica* **33**, 382–394.

Zinnes, J. (1969). "Scaling," *Ann. Rev. Psychol.* **20**, 447–478.