UC Berkeley

Title

Conditional Logit Analysis of Qualitative Choice Behavior

Permalink

https://escholarship.org/uc/item/61s3q2xr

Author

McFadden, Daniel

Publication Date

1972-11-01

CONDITIONAL LOGIT ANALYSIS OF

QUALITATIVE CHOICE BEHAVIOR

Daniel McFadden

November 1972

Working Paper No. 199/BART 10

CONDITIONAL LOGIT ANALYSIS OF QUALITATIVE CHOICE BEHAVIOR

by

Daniel McFadden
University of California, Berkeley

A fundamental concern of economics is understanding human choice be-
havior. Models or hypotheses are formed on the nature of decision proc-
esses, and are evaluated in the light of observed behavior. This task is
complicated 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 indivi-
dual 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 quali-
tative, 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 "optimi-
zation" 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 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, set of available

alternatives, and 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 $P(x|s,B)$ denote the conditional prob-
ability 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 $P(x|s,B)$ 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.[1] The selection probability

---

1. To be precise, each trial represents a drawing of a triple (s,B,ω)

that an individual drawn at random from the population will choose x ,

given measured attributes   s   and alternative set   B , equals the probab-

ility of occurence of a decision rule yielding this choice, or

(1) $P(x|s,B) = \pi(\{h \in H | h(s,B) = x\})$.

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 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,...,J and described

by vectors of attributes x_j . The individual has a utility function which

––––––––––––––––––

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 $h_\omega(x',B') \in B'$ 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 satis-

faction of this condition is one of the criteria for a carefully designed

laboratory experiment or sample survey.

can be written in the form

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

where V is non-stochastic 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

$$P_i \equiv P(x_i|s,B) = \pi(\{h_\varepsilon \in H|h_\varepsilon(s,B) = x_i\})$$

(2)
$$= P[\varepsilon(s,x_j) - \varepsilon(s,x_i) < V(s,x_i) - V(s,x_j) \text{ for all } j \neq i] \quad.$$

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.,

$$F(\varepsilon_1,\ldots,\varepsilon_J) = \pi(\{h_\varepsilon \in H|\varepsilon(s,x_j) \leq \varepsilon_j \text{ for } j=1,\ldots,J\}) \quad. \quad \text{Let } F_i \text{ denote}$$

the partial derivative of F with respect to its i-th argument, and let $V_i = V(s,x_i)$. Then equation (2) can be written

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

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 formulae for the P_i in equation (3). An alternative approach is to specify formulae for the

selection probabilities and then examine the question of whether these formulae could be obtained via equation (3) from some distribution of utility-maximizing consumers. This problem is the population analogue 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 which allows direct verification of condition (3).

We consider a powerful axiom on selection probabilities introduced by D. 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 assume

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

(4) $$P(x|s,\{x,y\})P(y|s,B) = P(y|s,\{x,y\})P(x|s,B) \quad .$$

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 $P(x|s,B)$ is positive, equation (4) implies $P(x|s,\{x,y\})$ positive, and

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

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). $P(x|s,B) > 0$ 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 $p_{xy} = P(x|s,\{x,y\})$. Define $p_{xx} = 1/2$. From (4),

$$(6) \qquad P(y|s,B) = \frac{p_{yx}}{p_{xy}} P(x|s,B) \quad ,$$

and

$$(7) \qquad 1 = \sum_{y \in B} P(y|s,B) = \left[\sum_{y \in B} \frac{p_{yx}}{p_{xy}} \right] P(x|s,B) \quad .$$

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

$$(8) \qquad P(x|s,B) = 1/ \sum_{y \in B} (p_{yx}/p_{xy}) \quad .$$

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

$$(9) \qquad p_{yx}/p_{xy} = (p_{yz}/p_{zy}) / (p_{xz}/p_{zx}) \quad .$$

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

$$(10) \qquad P(x|s,B) = e^{V(s,x,z)} / \sum_{y \in B} e^{V(s,y,z)} \quad .$$

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:[2]

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

(11) $$V(s,x,z) = v(s,x) - v(s,z) \quad .$$

Then, equation (10) becomes

(12) $$P(x|s,B) = e^{v(s,x)} \sum_{y \in B} e^{v(s,y)} \quad ,$$

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.

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 non-stochastic function reflecting "representative" tastes and $\varepsilon(s,x)$

---

2. 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 (9) holds for $z \notin B$, provided Axioms 1 and 2 hold for $B \cup \{z\}$. Then, taking z to be the universal benchmark in (10) and defining $v(s,x) = V(s,x,z)$ for all alternative sets yields the result.

is a function which 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[3]

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

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

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

$$F_i(\varepsilon + V_i - V_1, \ldots, \varepsilon + V_i - V_J) = e^{-\varepsilon} \prod_{j=1}^{J} e^{-e^{-\varepsilon + V_i - V_j}}$$

$$= e^{-\varepsilon} e^{-e^{-\varepsilon(\sum_{j=1}^{J} e^{V_i - V_j})}} .$$

Substituting this expression in (3) yields the result. $\qquad$ Q.E.D.

----

3. Monotone increasing transformations of the utility function $U(s, x)$ do

not affect utility maximization or the selection probabilities, but trans-

form 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 $P(\eta \leqq y) = e^{-1/y}$ $(y \geqq 0)$; $-e^{-U(s,x)} = e^{-V(s,x)} \eta(s,x)$ has

η distributed with the negative exponential distribution $P(\eta \leqq y) = e^{y}$

$(y \leqq 0)$; and $e^{-\beta e^{-U(s,x)}} = \eta(s,x)^{V(s,x)}$ has η distributed with the power

distribution $P(\eta \leqq y) = y^{1/\beta}$ $(0 \leqq y \leqq 1)$. These examples demonstrate that

the moments of the distribution of utility in the population (or their exist-

ence) 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 non-zero for real t , and has

all positive moments finite.

A non-constructive proof of this result was first given by J. Marschak (1968); the argument above appears in D. Luce and P. Suppes (19..), 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 (12). A random variable ε is said to be translation complete if for a function h of bounded absolute variation with $h(\underline{+}\infty) = 0$, 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.[4]

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, $G(\varepsilon) = e^{-\alpha e^{-\varepsilon}}$, 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

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

---

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

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

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

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

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

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

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

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 $G(\varepsilon) = e^{-\alpha e^{-\varepsilon}}$. $\hfill$ Q.E.D.

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 which 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 which 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

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

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 $z^k_{jn} = v^k(s_n, x_{jn})$ and $z_{jn} = (z^1_{jn}, \ldots, z^k_{jn})$. From equation (12), the selection probabilities then satisfy

$$(16) \qquad P_{in} = P(x_{in}|s_n, B_n) = e^{z_{in}\theta} \Big/ \sum_{j=1}^{J_n} e^{z_{jn}\theta} \quad ,$$

where $\theta' = (\theta_1, \ldots, \theta_K)$.[5] The experiment provides R_n repetitions of

---

5. 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 non-linear 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) which 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 non-identified. In general, the alternatives x_{jn} have no natural ranking, and the indexing j is arbitrary. We would then say the attributes of the alternative are generic, or "hedonic". However, in some applications the alternatives are ranked, and the rank j is a component of the vector of attributes of

trial n , and the i-th alternative is observed to be chosen S_{in} times. Of particular interest is the case without repetition,

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

We term this the <u>conditional logit model</u>. Note that it is an immediate generalization to the case of unequal, possibly unranked, alternatives of the multinomial logit model appearing in the literature.[6] The derivation of this model from a theory of population choice behavior appears to be new.

---

alternative x_j summarizing the "unique" characteristics of this position. Then z^k may be a variable such that $z_{jn}^k = 1$ and $z_{\ell n}^k = 0$ for $\ell \neq j$, yielding a "specific j-th 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 $z_{jn}^j = s_n^1$ and $z_{\ell n}^j = 0$ for $\ell \neq j$, implying $v(s_n, x_{jn}) = \sum_{j=1}^{J_1} s_n^1 \theta_j$ 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 $\theta_1 = 0$.

6. 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 (   ), Harter (1967), Walker (1967). Multinomial logit was developed for a special case by Gurland (1960), and more generally by Bloch (1967), Boch (1969), Rassom (1971), McFadden (1968), Stopher (1969), and Theil (1969,1970). An analogous

---

Footnote 6. (continued)

development has occurred for probit analysis, in which the cumulative normal rather than logistic distribution is used to determine the selection probabilities (Aitchison (1957, 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 (1966, 1968, 1969, 1970), 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 (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 (1971).

Binomial logit and probit analysis 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 (1967), Reichman (1971), Stopher (1969, 1970), Talvitie (1972), Thomas (1971), Uhler (1968), Walker (1968), Warner (1967), Zellner (1965).

END FOOTNOTE 6

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 (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

$$(17) \qquad e^L = \prod_{n=1}^{N} \frac{R_n!}{S_{1n}! \cdot \ldots \cdot S_{J_n n}!} \prod_{i=1}^{J_n} P_{in}^{S_{in}} \quad .$$

Substitution of (16) yields the log likelihood function

$$(18) \qquad L = C - \sum_{n=1}^{N} \sum_{i=1}^{J_n} S_{in} \log \sum_{j=1}^{J_n} e^{(z_{jn} - z_{in})\theta}$$

$$= C + \sum_{n=1}^{N} \left[\left(\sum_{j=1}^{J_n} S_{jn} z_{jn} \right) \theta - R_n \log \sum_{j=1}^{J_n} e^{z_{jn}\theta} \right] \quad ,$$

where

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

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 (18) with respect to θ yields the formulae

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

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

where

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

Since (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 $\partial L/\partial \theta = 0$. If, further, the matrix $\partial^2 L/\partial\theta\partial\theta'$ is non-singular, 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:

<u>Axiom 5</u> (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 $\sum_{n=1}^{N} J_n \geq K + N$. This will hold in particular if $N \geq K$ since $J_n \geq 2$, 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 colinear.

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

<u>Axiom 6</u>. There exists no non-zero K-vector γ satisfying $S_{in}(z_{jn} - z_{in}) \gamma \leq 0$ 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 $(z_{jn} - z_{in})(\hat{\theta} + \gamma) \leq (z_{jn} - z_{in})\hat{\theta}$ and $\log P_{in}(\hat{\theta} + \gamma) \geq \log P_{in}(\hat{\theta})$. Then $L(\hat{\theta} + \gamma) \geq L(\hat{\theta})$. 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

$$(21) \qquad b(\gamma) = \underset{n=1,\ldots,N}{\text{Max}} \quad \underset{i,j=1,\ldots,J_n}{\text{Max}} \quad S_{in}(a_{jn} - z_{in})\gamma .$$

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

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

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 (18),

$$L(\theta) - C \leq S_{in}\log P_{in} = -S_{in} \log \sum_{k=1}^{J_n} e^{(z_{kn} - z_{in})\gamma|\theta|}$$

$$\leq -S_{in}(z_{jn} - z_{in})\gamma|\theta| = -b(\gamma)|\theta|$$

$$\leq -b^*|\theta| \quad .$$

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

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 non-existence 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:

$$\text{Min}_{y,\alpha} \quad y'y \quad \text{subject to}$$

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

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
$S_{in}(z_{jn} - z_{in}) \gamma \leqq 0$, 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 $S_{in}(z_{jn} - z_{in}) \gamma \leqq 0$ for all $n = 1,\ldots,N$ and $i,j = 1,\ldots,J_n$.
Hence, Axiom 6 fails. This proves the lemma. Q.E.D.

Computation of the maximum likelihood estimator can be carried out
using a variety of standard programs for unconstrained non-linear optimization.
Since the likelihood function is strictly concave, any algorithm which con-
verges 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.[7]

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

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

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)

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

Equation (24) is termed the <u>linear probability model</u>, 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 one. Equation (24) is inappropriate/because the requirement that the forecasts lie in the unit for use in forecasting selection probabilities

---

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

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 (1968) can be employed. When S_{in}, S_{ln} are large[8], $\log(S_{in}/S_{ln})$ is a close approximation to the left-hand side of

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

and an estimate of θ can be obtained by applying ordinary or weighted least squares to the model

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

taking into account linear restrictions across equations.[9] 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 which 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 which makes the estimator inconsistent and may make it seriously biased. In such cases the maximum likelihood procedure should be more reliable.

---

8. A rule-of-thumb is $S_{in} \geq 5$.

9. Some improvement in the statistical properties of the unweighted Berkson-Theil estimator can be obtained by replacing (26) with the regression equation

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

This modification, suggested by Haldane (   ) for the binomial logit model,

III. Statistical Properties

Maximum likelihood estimation of the conditional logit model can be shown under very general conditions to provide estimators that are efficient and normally distributed in asymptotically large samples. 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 non-trivial 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 M .[10] The limit of the weighted moment matrix as $\sum_{n=1}^{N} R_n \to +\infty$.

---

makes $E \log \left(\dfrac{S_{in} + 1/2}{S_{ln} + 1/2} \right)$ equal to $\log \left(\dfrac{P_{in}}{P_{ln}} \right)$ 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.

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

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

exists and is positive definite.

The last part of this axiom strenthens the full-rank condition assumed earlier, implying that an infinite number of blocks of K trials can be found satisfying Axiom 5.[11] One can expect Axiom 7 to hold provided the data are not multicolinear 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 one 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

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

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

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

---

11. 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 non-singular.

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 $\left(\sum_{n=1}^{N} R_n \right)^{-1} \Omega^{-1}$ provide estimates of the variances of the estimators. To test a hypothesis 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,

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

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.

---

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

$$\left(\sum_{n=1}^{N} R_n(J_n - 1) \right) \Big/ \left(\sum_{n=1}^{N} R_n(J_n - 1) - K \right) \quad .$$

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

$$(31) \qquad D_{in} = (S_{in} - R_n P_{in}) / \sqrt{R_n P_{in}} \quad ,$$

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,

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

where

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

and the conditions

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

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

$$(35) \qquad \begin{aligned} \Lambda_{in,jm} \equiv E\, D_{in} D_{jm} &= \delta_{nm}[\delta_{ij} - \sqrt{P_{in}^0 \, P_{jn}^0}\,] \\ &\quad - \sqrt{r_m r_n P_{in}^0 P_{jm}^0}\,(z_{in} - \bar{z}_n)\Omega^{-1}(z_{jm} - \bar{z}_m). \end{aligned}$$

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 distritribution of the D_{in} is multivariate normal. Hence,

$$(36) \qquad G = \sum_{n=1}^{N} \sum_{i=1}^{J_n} D_{in}^2 \quad ,$$

has an asymptotic chi-square distribution with N^* degrees of freedom. [13]

---

13. Treating G as a function of θ and minimizing it at a value $\tilde{\theta}$ provides

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 logit specification of the selection probabilities and the absence of "alternative set" effects. Further, one can define an analogue of the multiple correlation coefficient,

$$(37) \qquad R^2 = 1 - G/G^H \quad ,$$

where G is given by (36) and G^H is given by the same equation when 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, on should adjust for the non-independence and heteroscedasticity of the D_{in}. This can be achieved in part by using the linearly transformed residuals

$$(38) \qquad Y_{in} = D_{in} - D_{1n} \sqrt{P_{in}} \, (1 - \sqrt{P_{1n}}) \, / \, (1 - P_{1n}) \quad ,$$

defined for $i = 2,\ldots,J_n$ and $n = 1,\ldots,N$. The Y_{in} are asymptotically

---

a $\underline{\text{minimum chi-square}}$ estimator of the parameter vector. The first order conditions for this minimization coincide with (32) except for a term, reflecting the effect of changing θ on the weights in the denominator of (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.

multivariate normal with mean zero and covariances

$$(39) \qquad \Gamma_{in,jm} \equiv E\, Y_{in} Y_{jm} = \delta_{nm} \delta_{ij} - q'_{in} q_{jm} \quad ,$$

where

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

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 homoscedastic 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 (35), with the second term in this expression vanishing. The D_{in} and the transformed residuals Y_{in} defined in (38) are independent across n, and the Y_{in} have zero mean, unit variance, and zero covariances. Suppose M integers N_m satisfy $\sum_{m=1}^{M} N_m = N$ and $N_m \to +\infty$. Then the statistics

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

where $N'_m = N_1 + \ldots + N_{m-1}$, 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 (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.[14] 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, fifty 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.

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 $\hat{\theta}_1 = \log(\frac{S}{N-S})$. Figure 1 is a graph of the logistic curve specifying P_1 as a function of θ_1. Figure 2 gives the actual expectation and variance of $\hat{\theta}_1$, conditioned on existence, and the large sample calculated variance 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 which are within ten percent of true values except for extreme selection probabilities (e.g., $\theta_1 = 2.0$ yields $P_1 = .88$). The linear probability model approximation is quite accurate for small parameter values, even for small sample

---

14. Berkson (1955), Gart (1967), Gilbert (1968), Talvitie (1972).
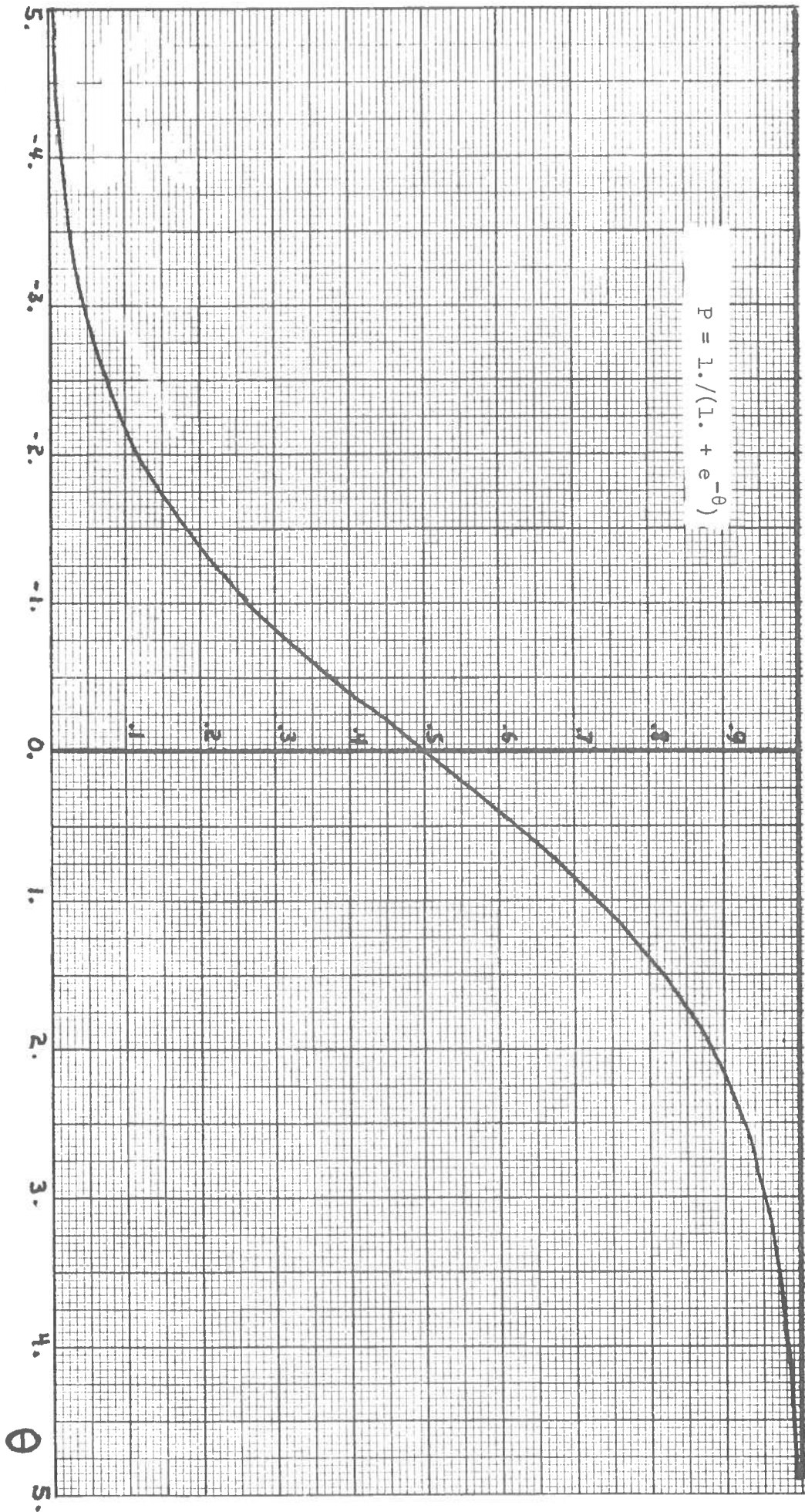
FIGURE 1

Logistic Function

$$P = 1./(1. + e^{-\theta})$$

PARAMETER VALUE θ_1	SAMPLE SIZE	PROBABILITY OF EXISTENCE OF MAXIMUM LIKELIHOOD ESTIMATOR	CONDITIONAL EXPECTATION OF MAXIMUM LIKELIHOOD ESTIMATOR	PERCENT BIAS	CONDITIONAL VARIANCE OF MAXIMUM LIKELIHOOD ESTIMATOR	CALCULATED VARIANCE OF MAXIMUM LIKELIHOOD ESTIMATOR	PERCENT BIAS	LINEAR PROBABILITY MODEL ESTIMATOR	PERCENT BIAS
.01	5	.93748	.00828	-17.2	.75018	1.21529	62.0	.01000	*
	10	.99804	.01103	10.3	.49988	.50943	1.9	.01000	*
	20	1.0	.01057	5.7	.22422	.22314	- .5	.01000	*
	30	1.0	.01036	3.6	.14334	.14308	*	.01000	*
	50	1.0	.01021	2.1	.08343	.08338	*	.01000	*
	75	1.0	.01014	1.4	.05482	.05481	*	.01000	*
	100	1.0	.01010	1.0	.04083	.04082	*	.01000	*
.1	5	.93594	.08272	-17.3	.74726	1.21643	67.0	.09992	- .1
	10	.99782	.11023	10.2	.50044	.51128	2.2	.09992	- .1
	20	1.0	.10575	5.8	.22503	.22387	- .5	.09992	- .1
	30	1.0	.10364	3.6	.14379	.14349	*	.09992	- .1
	50	1.0	.10211	2.1	.08366	.08360	*	.09992	- .1
	75	1.0	.10138	1.4	.05497	.05495	*	.09992	- .1
	100	1.0	.10103	1.0	.04093	.04093	*	.09992	- .1
.5	5	.89888	.40054	-20.0	.68164	1.24250	83.0	.48984	- 2.0
	10	.99121	.54433	8.9	.50916	.55472	8.9	.48984	- 2.0
	20	.99992	.53020	6.0	.24536	.24215	- 1.3	.48984	- 2.0
	30	1.0	.51912	3.8	.15515	.15391	- .8	.48984	- 2.0
	50	1.0	.51101	2.2	.08955	.08922	*	.48984	- 2.0
	75	1.0	.50720	1.4	.05865	.05852	*	.48984	- 2.0
	100	1.0	.50535	1.1	.04361	.04354	*	.48984	- 2.0
1.0	5	.78978	.73234	-26.8	.52531	1.30757	148.0	.92423	- 7.6
	10	.95639	1.04040	4.0	.49435	.67376	36.5	.92423	- 7.6
	20	.99610	1.06643	6.6	.30950	.30826	- .4	.92423	- 7.6
	30	.99992	1.04437	4.4	.19662	.19146	- 2.6	.92423	- 7.6
	50	1.0	1.02525	2.5	.11041	.10881	- 1.5	.92423	- 7.6
	75	1.0	1.01641	1.6	.07145	.07083	- .8	.92423	- 7.6
	100	1.0	1.01216	1.2	.05285	.05252	- .6	.92423	- 7.6
2.0	5	.46885	1.12824	-44.5	.23655	1.44030	510.0	1.5232	-23.8
	10	.71897	1.71841	-14.1	.30827	.95333	209.0	1.5232	-23.8
	20	.92102	2.04023	2.0	.37726	.60123	59.4	1.5232	-23.8
	30	.97780	2.09760	4.9	.35233	.41411	17.6	1.5232	-23.8
	50	.99825	2.08151	4.0	.23948	.23167	- 3.3	1.5232	-23.8
	75	.99993	2.05384	2.2	.15052	.14352	- 4.7	1.5232	-23.8
	100	1.0	2.03925	2.0	.10754	.10383	- 3.5	1.5232	-23.8
3.0	5	.21568	1.28953	-57.1	.09262	1.51321	1655.0	1.8103	-39.7
	10	.38464	2.01670	-32.8	.13505	1.11987	830.0	1.8103	-39.7
	20	.62158	2.59216	-13.6	.21698	.88161	307.0	1.8103	-39.7
	30	.76721	2.84486	- 5.2	.27726	.74378	162.0	1.8103	-39.7
	50	.91191	3.04643	1.5	.33452	.54907	64.4	1.8103	-39.7
	75	.97385	3.10492	3.5	.32443	.38745	19.4	1.8103	-39.7
	100	.99224	3.10337	3.4	.27741	.28459	2.5	1.8103	-39.7

*Negligible

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 $\hat{\theta}_1 = \log \dfrac{S_1}{R-S_1}$ and $\hat{\theta}_2 = \log\left(\dfrac{S_2}{R-S_2}\right) - \log\left(\dfrac{S_1}{R-S_1}\right)$. Figure 3 gives the conditional expectations of these estimators and the expectations of the calculated variances 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 ten, 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 ten percent, 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 probabilties 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 non-existence of the the maximum likelihood estimator are acceptably small

FIGURE 3

Sampling Properties of the Maximum Likelihood Estimator, Example 2

	SAMPLE SIZE					
	10	20	40	80	140	200
First Parameter						
True value	.10000	.10000	.10000	.10000	.10000	.10000
Expect. of M.L. estimate	.08272	.11023	.10575	.10267	.10148	.10103
Per cent bias of est.	-17.28364	10.22808	5.74896	2.66732	1.48170	1.02609
True variance of M.L. est.	.74726	.50044	.22503	.10576	.05902	.04094
Calc. variance of M.L. est.	.80200	.40100	.20050	.10025	.05729	.04010
Per cent bias of var. est.	7.32580	-19.87098	-10.90107	-5.21060	-2.93154	-2.04035
Linear prob. model est.	.09992	.09992	.09992	.09992	.09992	.09992
Per cent bias of linear est.	-.08325	-.08325	-.08325	-.08325	-.08325	-.08325
Second Parameter						
True value	-.50000	-.50000	-.50000	-.50000	-.50000	-.50000
Expect. of M.L. estimate	-.40695	-.54780	-.52948	-.51365	-.50757	-.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	.46265	.21631	.12049	.08353
Calc. variance of M.L. ext.	1.63443	.81722	.40861	.20430	.11675	.08172
Per cent bias of var. est.	12.53225	-18.87888	-11.68067	-5.54934	-3.11136	-2.16302
Linear prob. model est.	-.49467	-.49467	-.49467	-.49467	-.49467	-.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.	.85421	.99181	.99996	1.00000	1.00000	1.00000
First Parameter						
True value	.10000	.10000	.10000	.10000	.10000	.10000
Expect. of M.L. estimate	.08272	.11023	.10575	.10267	.10148	.10103
Per cent bias of est.	-17.28364	10.22808	5.74896	2.66723	1.48170	1.02609
True variance of M.L. est.	.74726	.50044	.22503	.10576	.05902	.04094
Calc. variance of M.L. est.	.80200	.40100	.20050	.10025	.05729	.04010
Per cent bias of var. est.	7.32580	-19.87098	-10.90107	-5.21060	-2.93154	-2.04035
Linear prob. model est.	.09992	.09992	.09992	.09992	.09992	.09992
Per cent bias of linear est.	-.08325	-.08325	-.08325	-.08325	-.08325	-.08325
Second Parameter						
True value	.10000	.10000	.10000	.10000	.10000	.10000
Expect. of M.L. estimate	.08204	.10990	.10582	.10270	.10150	.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	.45254	.21246	.11852	.08220
Calc. variance of M.L. ext.	1.61003	.80501	.40251	.20125	.11500	.08050
Per cent bias of var. est.	8.36201	-19.69933	-11.05610	-5.27618	-2.96650	-2.06423
Linear prob. model est.	.09942	.09942	.09942	.09942	.09942	.09942
Per cent bias of linear est.	-.58076	-.58076	-.58076	-.58076	-.58076	-.58076
Prob. of existence of M.L. est.	.87160	.99496	.99999	1.00000	1.00000	1.00000

FIGURE 3 (continued)

SAMPLE SIZE

	10	20	40	80	140	200
First Parameter						
True value	.10000	.10000	.10000	.10000	.10000	.10000
Expect. of M.L. estimate	.08272	.11023	.10575	.10267	.10148	.10103
Per cent bias of est.	-17.28364	10.22808	5.74896	2.66732	1.48170	1.02609
True variance of M.L. est.	.74726	.50044	.22503	.10576	.05902	.04094
Calc. variance of M.L. est.	.80200	.40100	.20050	.10025	.05729	.04010
Per cent bias of var. est.	7.32580	-19.87098	-10.90107	-5.21060	-2.93154	-2.04035
Linear prob. model est.	.09992	.09992	.09992	.09992	.09992	.09992
Per cent bias of linear est.	-.08325	-.08325	-.08325	-.08325	-.08325	-.08325
Second Parameter						
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	.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	.98412	.55049	.25613	.14019	.09663
Calc. variance of M.L. ext.	1.86941	.93470	.46735	.23368	.13353	.09347
Per cent bias of var. est.	50.93327	-5.02084	-15.10210	-8.76667	-4.75197	-3.26642
Linear prob. model est.	.90112	.90112	.90112	.90112	.90112	.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.	.71254	.94144	.99680	.99999	1.00000	1.00000
First Parameter						
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	.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.	.52531	.49435	.30950	.14142	.07686	.05285
Calc. variance of M.L. est.	1.01723	.50862	.25431	.12715	.07266	.05086
Per cent bias of var. est.	93.64346	2.88570	-17.83370	-10.08782	-5.46810	-3.76027
Linear prob. model est.	.92423	.92423	.92423	.92423	.92423	.92423
Per cent bias of linear est.	-7.57657	-7.57657	-7.57657	-7.57657	-7.57657	-7.57657
Second Parameter						
True value	-.50000	-.50000	-.50000	-.50000	-.50000	-.50000
Expect. of M.L. estimate	-.33180	-.49607	-.53624	-.51826	-.50991	-.50681
Per cent bias of est.	-33.64020	-.78652	7.24721	3.65173	1.98177	1.36160
True variance of M.L. est.	1.20695	1.00351	.55487	.25493	.13986	.09646
Calc. variance of M.L. ext.	1.86828	.93414	.46707	.23354	.13345	.09341
Per cent bias of var. est.	54.79349	-6.91277	-15.82285	-8.39383	-4.58190	-3.15818
Linear prob. model est.	-.43440	-.43440	-.43440	-.43440	-.43440	-.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.	.70992	.94799	.99802	1.00000	1.00000	1.00000

FIGURE 3 (continued)

SAMPLE SIZE

	10	20	40	80	140	200
First Parameter						
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	.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.	.52531	.49435	.30950	.14142	.07686	.05285
Calc. variance of M.L. est.	1.01723	.50862	.25431	.12715	.07266	.05086
Per cent bias of var. est.	93.64346	2.88570	−17.83370	−10.08782	−5.46810	−3.76027
Linear prob. model est.	.92423	.92423	.92423	.92423	.92423	.92423
Per cent bias of linear est.	−7.57657	−7.57657	−7.57657	−7.57657	−7.57657	−7.57657
Second Parameter						
True value	.10000	.10000	.10000	.10000	.10000	.10000
Expect. of M.L. estimate	.05511	.08855	.10715	.10465	.10248	.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	.97802	.63496	.29179	.15804	.10854
Calc. variance of M.L. ext.	2.08464	1.04232	.52116	.26058	.14890	.10423
Per cent bias of var. est.	105.05601	6.57426	−17.92244	−10.69591	−5.78006	−3.96928
Linear prob. model est.	.07681	.07681	.07681	.07681	.07681	.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.	.60127	.90235	.99491	.99999	1.00000	1.00000
First Parameter						
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	.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.	.52531	.49435	.30950	.14142	.07686	.05285
Calc. variance of M.L. est.	1.01723	.50862	.25431	.12715	.07266	.05086
Per cent bias of var. est.	93.64346	2.88570	−17.83370	−10.08782	−5.46810	−3.76027
Linear prob. model est.	.92423	.92423	.92423	.92423	.92423	.92423
Per cent bias of linear est.	−7.57657	−7.57657	−7.57657	−7.57657	−7.57657	−7.57657
Second Parameter						
True value	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Expect. of M.L. estimate	.39590	.67801	.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.	.76186	.80262	.68676	.43638	.24022	.16047
Calc. variance of M.L. ext.	2.92211	1.46106	.73053	.36526	.20872	.14611
Per cent bias of var. est.	283.54796	82.03495	6.37315	−16.29632	−13.11233	−8.94926
Linear prob. model est.	.59895	.59895	.59895	.59895	.59895	.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.	.37108	.68762	.91927	.99376	.99986	1.00000

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 extreme and non-extreme selection probabilties. 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 non-linear 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 Professors L. Miller and R. Radner (1970), and the results have been used to forecast the effects of changing educational policy on college enrollment. Professor M. 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 Charles River Associates (1972).[15]

---

15. The results below are reproduced with the permission of Charles River Associates and the Federal Highway Administration, U.S. Dept. of Transportation.

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.[16] The results of each analysis are summarized in turn.

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 Figure 4. 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

---

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

FIGURE 4

Conditional Logit Model of Shopping Mode Choice
Dependent Variable = 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)*	-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)	-----	-----

* 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 which would be replicated in a random sample of the population.

a value of walk time of $5.46 per hour and a value of in-vehicle time of
$ .95 per hour in Model 1. These values are in close accord 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 pre-
school children. The inclusive price is defined from the shopping mode
choice Model 1 to be

[Inclusive Price] = 0.0654[Auto in-vehicle time] + 4.11[Auto operating cost] .

The measure of destination attractiveness is taken to be the retail employ-
ment in the zone as a percentage of total retail employment in the region.
Because the alternative destinations are unranked and vary from one observ-
ation 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 Figure 5. The two independent variables above

FIGURE 5

Conditional Logit Model of Shopping Destination Choice

Dependent Variable = 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 = inclusive price of trip times number of pre-school children	----	-0.521 (0.343)

are both found to be 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 non-trip 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 interaction variable. The estimates are given in Figure 6. Model 7 predicts the actual decision with probability .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.[16] These

---

16. Such tables could also be generated by aggregating over individuals for a random sample of the population, a procedure which 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.

FIGURE 6

Conditional Logit Model of Shopping Trip Frequency

Dependent Variable = 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 of 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)

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" which could occur in attempting to infer response elasticities from data on behavior of heterogeneous groups. Thus, this modelling approach has the potential of providing accurate forecasts of the response of shopping travel demand to policy variables, in a framework which 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 analogue 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 one 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_m-th 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 $1 \leq i_m, j_m \leq J_{n_m}$ 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 $M_{\ell-1}$ containing $\ell - 1$ indices which satisfy the required property. Suppose there does not exist an index m_ℓ such that $M_\ell = M_{\ell-1} \cup \{m_\ell\}$ has the desired property. Then, for all $m > m_{\ell-1}$ and $1 \leq i, j \leq J_{n_m}$, 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 $M_{\ell-1}$. 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 $M = \overset{\infty}{\underset{\ell=1}{\cup}} M_\ell$ 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

$$z_{i_m n_m} - z_{j_m n_m} \quad \text{for} \quad m \in M^q .$$

Then, W^q is non-singular 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 $m \in M^q$ if a_m^q is positive. Suppose this event occurs, but Axiom 6 fails, and let $\gamma \neq 0$ be such that $S_{in}(z_{jn} - z_{in})\gamma \leq 0$ for all i, j, n . Then, $W^{q-1}\gamma$ is a non-negative vector. Since W^{q-1} is non-singular, 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

$$(42) \qquad P_{in} \geq 1/J_* e^{2M|\theta|} \equiv P_* > 0 \quad ,$$

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 $q \leq 2q'$ is at least $1 - (1 - P_*^{2K})^{q'}$. This last probability approaches one as $q' \to +\infty$, proving the lemma. Q.E.D.

<u>Lemma 6</u>. Suppose Axioms 1-4 and 7 hold, θ° 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

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

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, $|z_{in}| \leqq M$ for a positive scalar M . Differentiation

of equation (16) yields the bounds

$$
\begin{array}{ll}
(43) &
\begin{array}{l}
|\partial \log P_{in}/\partial \theta| \leq 2M \quad , \\[4pt]
|\partial^2 \log P_{in}/\partial \theta \partial \theta'| \leq 4M^2 \quad , \\[4pt]
|\partial^3 \log P_{in}/\partial \theta \partial \theta' \partial \theta_k| \leq 8M^3 \quad ,
\end{array}
\end{array}
$$

uniform in θ .

Let m be a serial index of trials and repetitions, and let s_{im} equal

one if alternative i is chosen at this observation and zero otherwise, for

$1 \leqq i \leqq J_{n_m}$. Define a sequence of independent random variables

$$(44) \qquad \lambda^m(\theta) = \sum_{i=1}^{J_{n_m}} s_{im} \log P_{in}(\theta) \quad .$$

Then

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

with C_q a constant independent of θ , is the log likelihood function.

From (43), the derivatives of λ^m , denoted $\partial \lambda^m/\partial \theta = \lambda_\theta^m$, etc., satisfy the

uniform bounds $|\lambda_\theta^m| \leqq 2M$, $|\lambda_{\theta\theta'}^m| \leqq 4M^2$, and $|\lambda_{\theta\theta'\theta_k}^m| \leqq 8M^3$. Further,

$$(46) \qquad E \, \lambda_\theta^m(\theta^0) = \left[\sum_{i=1}^{J_{n_m}} \partial P_{in_m}/\partial \theta \right]_{\theta^0} = 0 \quad .$$

Define

$$
(47) \qquad
\begin{aligned}
\Omega_m &= -E \, \lambda_{\theta\theta'}^m(\theta^0) \\
&= \sum_{i=1}^{J_{n_m}} (z_{in_m} - \bar{z}_{n_m})' \, P_{in_m} (z_{in_m} - \bar{z}_{n_m}) \quad .
\end{aligned}
$$

Then by Axiom 7,

$$
(48) \qquad \lim_{q \to \infty} \frac{1}{q} \sum_{m=1}^{q} \Omega_m = \Omega \quad .
$$

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

Given a small positive scalar ε , choose $\delta = \min \{\varepsilon, \beta/4(1 + 4KM^3)\}$. The weak law of large numbers [Feller (1966), Vol. I, p. 238] implies that there exists $q_1 \geqq q_0$ such that for $q \geqq q_1$, $L_\theta^q(\theta^0) = \sum_{m=1}^{q} \lambda_\theta^m (\theta^0)$ satisfies

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

with probability at least $1 - \varepsilon$.

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

$$
(50) \qquad \frac{1}{q} L_{\theta_k}^q (\theta) - \frac{1}{q} L_{\theta_k \theta}^q(\theta^0)(\theta - \theta^0)
$$

$$
= \frac{1}{q} L_{\theta_k}^q (\theta^0) + \frac{1}{2} \frac{1}{q} (\theta - \theta^0)' L_{\theta_k \theta \theta'}^q(\tilde{\theta})(\theta - \theta^0) \quad ,
$$

where $\tilde{\theta}$ lies between θ and θ^0 . Consider $q \geqq q_1$ and θ satisfying $(\theta - \theta^0)'(\theta - \theta^0) = \delta^2$, and suppose (49) holds. Then

$$
(51) \qquad \left| \frac{1}{q} L_{\theta_k}^q (\theta) - \frac{1}{q} L_{\theta_k \theta}^q(\theta^0)(\theta - \theta^0) \right| \leqq \delta^2(1 + 4K^2 M^3) \leqq \delta \beta/4 \quad .
$$

Hence,

$$
(52) \qquad \frac{1}{q} \left| (\theta - \theta^0)' L_\theta^q(\theta) - (\theta - \theta^0)' L_{\theta \theta'}(\theta^0)(\theta - \theta^0) \right| \leqq \delta^2 \beta/4 \quad .
$$

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

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

Hence, at each point in 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 $1 - \epsilon$, 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

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

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

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

equation (54) can be written

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

Then,

$$(57) \qquad \underset{q\to\infty}{\text{plim}}\, D_q = I - 4M^3\Omega^{-1/2}A\Omega^{-1/2}\underset{q\to\infty}{\text{plim}}\,|\hat{\theta}^q-\theta^0| = I \quad.$$

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

$$(58) \qquad \frac{1}{q}\,\Omega^{-1/2}L^q_{\theta}(\theta^0) = \frac{1}{q}\sum_{m=1}^{q}\Omega^{-1/2}\lambda^m_{\theta}(\theta^0) \quad.$$

But the independent random variables $\Omega^{-1/2}\lambda^m_{\theta}(\theta^0)$ satisfy the Lindeberg-Levy theorem [Feller (1966), Vol. II, pp. 256-258], implying that (58) is asymptotic standard normal. This proves the lemma. Q.E.D.

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, p. 230-236). Rao gives several asymptotically equivalent forms for the test.

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

$$(59) \qquad D^0_{in} = (S_{in} - R_n P^0_{in}) / \sqrt{R_n P^0_{in}} \quad ,$$

$$(60) \qquad D^1_{in} = \sum_{m=1}^{N} \sum_{j=1}^{J_m} \sqrt{r_n r_m P^0_{in} P^0_{jm}} \, (z_{in} - \bar{z}_n) \Omega^{-1} (z_{jm} - \bar{z}_m)' D^0_{jm} \quad ,$$

where $P^0_{in} = P_{in}(\theta^0)$. Then D^0_{in} has a multivariate distribution with $E \, D^0_{in} = 0$ and $E \, D^0_{in} D^0_{jm} = \delta_{nm}(\delta_{ij} - \sqrt{P^0_{in} P^0_{jn}})$, implying $E \, D^1_{in} = 0$,

$$(61) \qquad E \, D^0_{in} D^1_{jm} = \sqrt{r_n r_m P^0_{in} P^0_{jm}} \, (z_{in} \, \bar{z}_n) \, \Omega^{-1} (z_{jm} - \bar{z}_m) \quad ,$$

and asymptotically

$$(62) \qquad E \, D^1_{in} D^1_{jm} = \sqrt{r_n r_m P^0_{in} P^0_{jm}} \, (z_{in} - \bar{z}_n) \, \Omega^{-1} (z_{jm} - \bar{z}_m) \quad .$$

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

$\sqrt{R_n} \, (P^0_{in} - P_{in}) / \sqrt{P^0_{in}}$ and $\sqrt{r_n P^0_{in}} \, (z_{in} - \bar{z}_n)(\sum_{n=1}^{N} R_n)^{1/2} (\theta^0 - \hat{\theta})$ differ from $-D^1_{in}$ 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, $D^0_{in} - D^1_{in}$ has an asymptotic distribution with mean zero and covariance

$$\Lambda_{in,jm} = \delta_{nm}[\delta_{ij} - \sqrt{P^0_{in} P^0_{jm}} \,]$$

$$(63) \qquad\qquad - \sqrt{r_n r_m P^0_{in} P^0_{jm}} \, (z_{in} - \bar{z}_n) \, \Omega^{-1} (z_{jm} - \bar{z}_m) \quad .$$

Write

$$(64) \qquad D_{in} = \frac{S_{in} - R_n P^0_{in}}{\sqrt{R_n P_{in}}} + \frac{\sqrt{R_n}(P^0_{in} - P_{in})}{\sqrt{P_{in}}} \qquad .$$

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

$$(65) \qquad \Lambda = I - \sum_{m=0}^{N} q_m q'_m \qquad ,$$

where

$$(q_0)_{in} = \sqrt{r_n P^0_{in}} \, (z_{in} - \bar{z}_n) \, \Omega^{-1/2} \qquad ,$$

$$(q_m)_{in} = \delta_{mn} \sqrt{P^0_{in}} \qquad , \qquad m = 1,\ldots,N \qquad .$$

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

Next consider the linear transformations

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

where $\alpha_n = (1 - \sqrt{P_{1n}})/(1 - P_{1n})$. Then, Y_{in} has the same asymptotic distribution as the random variable Y^0_{in} formed by replacing P_{in} with P^0_{in} in (66), and the latter random variable has asymptotic moments $E \, Y^0_{in} = 0$ and

$$\Gamma_{in,jm} = E \, Y^0_{in} Y^0_{jm} = E \, D_{in} D_{jm} - \sqrt{P^0_{jm}} \, \alpha_m E \, D_{in} D_{1m}$$

$$- \sqrt{P^0_{in}} \, \alpha_n E \, D_{1n} D_{jm} + \sqrt{P^0_{in} P^0_{jm}} \, \alpha_n \alpha_m E \, D_{1n} D_{1m}$$

$$(67) \qquad\qquad\qquad = \delta_{nm} \delta_{ij} - q'_{in} q_{jm} \qquad ,$$

with $q'_{in} = \sqrt{r_n P^0_{in}} \, [(z_{in} - \bar{z}_n) + \dfrac{P^0_{1n} - \sqrt{P^0_{1n}}}{1 - P^0_{1n}} (z_{1n} - \bar{z}_n)] \, \Omega^{-1/2}$. Then,

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

$$\sum_{n=1}^{N} \sum_{i=2}^{J_n} Y_{in}^2 = \sum_{n=1}^{N} \sum_{i=1}^{J_n} D_{in}^2 \quad \text{has an asymptotic chi-square distribution}$$

with N^* degrees of freedom.

REFERENCES

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

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

[3] Allouche, J. (1972), "Approach to Probability Distribution of Value of Walking Time and Pedestrian Circulation Models", *Highway Research Record* 392 pp. 121-133.

[4] Amemiya, T. (1972), "Bivariate Probit Analysis: Minimum Chi-Square Methods", Department of Economics, Stanford University, unpublished.

[5] 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", Institute for Mathematical Studies in the Social Sciences, Tech. Report No. 75.

[6] Antle, C., L. Klimko, and W. Harkness (1970), "Confidence intervals for the parameters of the logistic distribution", *Biometrika* 57, pp. 397-

[7] Atkinson, A. (1972), "A Test of the Linear Logistic and Bradley-Terry Models", *Biometrika*, pp. 37-42.

[8] Berkson, J. (1955), "Maximum Likelihood and Minimum Chi-Square Estimates of the Logistic Function", *Journal of the American Statistical Association*, pp. 130-162.

[9] Berkson, J. (1951), "Why I prefer Logits to Probits", *Biometics*, pp. 327-339.

[10] Bloch, D. and G. Watson (1967), "A Bayesian Study of the Multinomial Distribution", *Annals of Mathematical Statistics* 38, pp. 1423-1435.

[11] Block, H. and J. Marschak (1960), "Random Orderings and Stochastic Theories of Respons" in I. Olkin, ed., *Contributions to Probability and Statistics*, Stanford University Press.

[12] Bock, R. (1969), "Estinating Multinomial Response Relations" in R. Bose, ed., Contributions to Statistics and Probability: Essays in Memory of S. N. Roy, University of North Carolina Press.

[13] Boskin, M. (1972), "A Conditional Logit Model of Occupational Choice", Department of Economics, Stanford University, unpublished.

[13a] Charles River Associates, Inc. (1972), A Disaggregated Behavioral Model of Urban Travel Demand, Contract No. FH-11-7566 of the Federal Highway Administration, U. S. Department of Transportation.

[14] Cox, D. (1970), Analysis of Binary Data, Methuen.

[15] Cox, D. and E. Snell (1971), "On Test Statistics Calculated from Residuals", Biometrika 58, pp. 589-594.

[16] Cox, D. (1958), "The Regression Analysis of Binary Sequences", Journal of the Royal Statistical Society, Series B, pp. 215-242.

[17] Cox, D. (1966), "Some Procedures Connected with the Logistic Qualitative Response Curve", F. David, ed., Research Papers in Statistics, Wiley, pp. 55-71.

[18] Cox, D. and E. Snell (1968), "A General Definition of Residuals", Journal of the Royal Statistical Society, Series B 30, pp. 248-265.

[19] Ergun, G. (1971), "Development of a Downtown Parking Model", Highway Research Record 369, pp. 118-134.

[20] Feller, W. (1966), An Introduction to Probability Theory and Its Applications, Vol. 2, Wiley.

[21] Finney, D. (1964), Probit Analysis, Cambridge University Press.

[22] Fisher, J. (1962), "An Analysis of Consumer Goods Expenditures in 1957", Review of Economics and Statistics, pp. 64-71.

[23] Gart, J. and J. Zweifel (1967), "On the Bias of Various Estimators of the Logit and its Variance", Biometrika 54, pp. 181-187.

[24] Gilbert, E. (1968), "On Discrimination Using Qualitative Variables", _Journal of the American Statistical Association_, pp. 1399-1412.

[25] Goodman, L. (1972), "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables", _American Sociological Review_ 37, pp. 28-46.

[26] Goodman, L. (1970), "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications", _Journal of the American Statistical Association_ 65, pp. 226-256.

[27] Grizzle, J. (1962), "Asymptotic Power of Tests of Linear Hypotheses Using the Probit and Logit Transformations", _Journal of the American Statistical Association_, pp. 877-894.

[28] Grizzle, J. (1971), "Multivariate Logit Analysis", _Biometrics_ 27, pp. 1057-1062.

[29] Gupta, S., A. Qureishi, and B. Shah (1967), "Best Linear Unbiased Estimators of the Parameters of the Logistic Distribution Using Order Statistics", _Technometrics_ 9, pp. 43-56.

[30] Gurland, J., I. Lee, and P. Dolan (1960), "Polychotomous quantal response in biological assay", _Biometrics_ 16, pp. 382-398.

[30a] Haldane, J. (1955), "The Estimation and Significance of the logarithm of a ratio of frequencies," _Annals of Human Genetics_ 20, pp. 309-311.

[31] Harter, J. and A. Moore (1967), "Maximum likelihood estimation, from censored samples, of the parameters of a logistic distribution", _Journal of the American Statistical Association_ 62, pp. 675-683.

[32] Kendall, M. and A. Stuart (1967), _The Advanced Theory of Statistics_ Vol. 2, Hafner.

[33] Korbel, J. (1966), "Labor Force Entry and Attachment of Young People", _Journal of the American Statistical Association_, pp. 117-127.

[34] Lave, C. (1968), _Modal Choice in Urban Transportation: A Behavioral Approach_, Department of Economics, Stanford University, Ph.D dissertation.

[35] Lee, T. (1963), "Demand for Housing: A Cross-Section Analysis",
      Review of Economics and Statistics, pp. 190-196.

[36] Leonard, T. (1972), "Bayesian Methods for Multinomial Data", _American_
      College Testing Program, Technical Bulletin No. 4.

[37] Lisco, T. (1967), _The Value of Commuters' Travel Time: A Study in_
      Urban Transportation, Department of Economics, University of Chicago,
      Ph.D. dissertation.

[38] Luce, R. (1959), _Individual Choice Behavior_, Wiley.

[39] Luce, R. and P. Suppes (1965), "Preference, Utility, and Subjective
      Probability" in R. Luce, R. Bush and E. Galanter,eds., _Handbook of_
      Mathematical Psychology, Vol. 3, Wiley.

[40] 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, Department of Economics, University of
      California at Berkeley.

[41] McFadden, D. (1968), "The Revealed Preferences of a Government Bureau-
      cracy", Department of Economics, University of California at Berkeley,
      unpublished.

[42] McGillivray, R. (1969), _Binary Choice of Transport Mode in the San_
      Francisco Bay Area, Department of Economics, University of California at
      Berkeley, Ph.D. dissertation.

[43] McGillivray, R. (1970), "Demand and Choice Models of Modal Split",
      Journal of Transport Economics and Policy 4, pp. 192-207.

[44] Marschak, J. (1960), "Binary Choice Constraints on Random Utility
      Indicators" in K. Arrow, ed., _Stanford Symposium on Mathematical Methods_
      in the Social Sciences, Stanford University Press.

[45] Miller, L. and R. Radner (1970), "Demand and Supply in U. S. Higher Education". American Economic Review 60.

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

[47] Moses, L., R. Beals, and M. Levy (1967), "Rationality and Migration in Ghana", Review of Economics and Statistics, pp. 480-486.

[48] Quandt, R. and K. Young (1969), "Cross Sectional Travel Demand Models: Estimation and Tests", Journal of Regional Science 9, pp. 201-214.

[49] Quandt, R. and W. Baumol (1966), "The Demand for Abstract Transport Modes: Theory and Measurement", Journal of Regional Science 6, pp. 13-26.

[50] Quandt, R. (1970), The Demand for Travel, Heath.

[51] Quandt, R. (1968), "Estimation of Modal Splits", Transportation Research 2, pp. 41-50.

[52] Quarmby, D. (1967), "Choice of Travel Mode for the Journey to Work: Some Findings", Journal of Transport Economics and Policy.

[53] Rao, C. (1965), Linear Statistical Inference and its Applications, Wiley.

[54] Rassam, P., R. Ellis, and J. Bennett (1971), "The n-Dimensional Logit Model: Development and Application", Highway Research Record 369, pp. 135-147.

[55] Reichman, S. and P. Stopher (1971), "Disaggregate Stochastic Models of Travel-Mode Choice", Highway Research Record 369, pp. 91-103.

[56] Stopher, P. and T. Lisco (1970), "Modelling Travel Demand: A Disaggregate Behavioral Approach - Issues and Applications", Transportation Research Forum Proc., pp. 195-214.

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

[58] Stopher, P. (1969), "A Probability Model of Travel Mode Choice for the Work Journey", Highway Research Record 283, pp. 57-65.

[59] Talvitie, A. (1972), "Comparison of Probabilistic Modal-Choice Models: Estimation Methods and System Inputs", Highway Research Record 392, pp. 111-120.

[60] Theil, H. (1969), "A Multinomial Extension of the Linear Logit Model", International Economic Review 10, pp. 251-259.

[61] Theil, H. (1970), "On the Estimation of Relationships involving Qualitative Variables", American Journal of Sociology 76, pp. 103-154.

[62] Theil, H. (1971), Principles of Econometrics, Wiley.

[63] Thomas, T. and G. Thompson (1971), "Value of Time Saved by Trip Purpose", Highway Research Record 369, pp. 104-113.

[64] Thurstone, L. (1927), "A Law of Comparative Judgment", Psychological Review 34, pp. 273-286.

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

[66] Tversky, A. and J. Russo (1969), "Substitutability and Similarity in Binary Choices", Journal of Mathematical Psychology 6, pp. 1-12.

[67] Uhler, R. (1968), "The Demand for Housing: An Inverse Probability Approach", Review of Economics and Statistics, pp. 129-134.

[68] Walker, F. (1968), "Determinants of auto scrappage", Review of Economics and Statistics, pp. 503-506.

[69] Walker, S. and D. Duncan (1967), "Estimation of the Probability of an Event as a Function of Several Independent Variables", Biometrika, pp. 167-179.

[70] Warner, S. (1967), Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice, Northwestern University Press.

[72] Zellner, A. and T. Lee (1965), "Joint Estimation of Relationships
      Involving Discrete Random Variables", Econometrica 33, pp. 382-394.

[73] Zinnes, J. (1969), "Scaling", Annual Review of Psychology 20,
      pp. 447-478.

[74] Grizzle, J., C. Starmer, G. Koch (1969), "Analysis of Categorical Data
      by Linear Models," Biometrics 25, pp. 498-504.