


LIBRARY
OF THE
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY



Digitized by the Internet Archive
in 2011 with funding from
Boston Library Consortium Member Libraries

<http://www.archive.org/details/conditionalprobi00haus>

HB31
.M415
No.173



**working paper
department
of economics**

A CONDITIONAL PROBIT MODEL FOR QUALITATIVE CHOICE:
DISCRETE DECISIONS RECOGNIZING
INTERDEPENDENCE AND HETEROGENOUS PREFERENCES

Jerry A. Hausman (MIT)

David A. Wise (Harvard)

Number 173

April 1976

**massachusetts
institute of
technology**

**50 memorial drive
cambridge, mass.02139**



A CONDITIONAL PROBIT MODEL FOR QUALITATIVE CHOICE:
DISCRETE DECISIONS RECOGNIZING
INTERDEPENDENCE AND HETEROGENOUS PREFERENCES

Jerry A. Hausman (MIT)

David A. Wise (Harvard)

Number 173

April 1976

The views expressed here are the responsibility of the authors and do not reflect those of the Department of Economics or the Massachusetts Institute of Technology.

The authors wish to thank Gary Chamberlain, Zvi Griliches, Charles Manski, and Dan McFadden for helpful suggestions. An earlier version of this paper was presented at the Third World Congress of the Econometric Society, August, 1975.

HR31
M415
NO. 173



M.I.T. LIBRARIES
JUN 26 1976

1. Introduction and Background

The social sciences attempt to explain and predict the behavior of individuals. In practice, this often requires that they predict individual decisions or choices. In many situations, choices are made over a continuum of possibilities; for example, "how much" to spend or how much to work. But in many other situations, choices are made from a limited number of possibilities or alternatives; the possible alternatives are "discrete" or "quantal." Indeed, many decisions made in the public sector could be considered to be informed only if knowledge of the determinants of discrete choices by individuals were available. Examples of these kinds of choice are whether or not to work, where to live, where to work, mode of transportation, and size of family. Knowledge of the determinants of such decisions is important to the policy maker in designing, for example, income maintenance programs, urban renewal projects, medical education programs, public transportation networks, and child care facilities.

Logit and probit analysis are the most widely used methods for estimating the relationship between choices on the one hand and attributes of alternatives and individual decision makers on the other in binary choice, or two alternative, situations (e.g., Cox [1970]). In multiple alternative situations the most widely used method is a generalization of logit analysis, often called conditional logit analysis. Professor McFadden has developed qualitative choice models based on the conditional logit specification to a high degree of sophistication. He first applied the model to the choice of urban freeway routes by state highway departments [1975] and since has done extensive investigation of transit mode choice by individuals [1974]. Others have applied the same model to college

choice, plant location, occupational choice, and the choice of fuel for electric power generators. Conditional logit analysis has been preferred over other theoretical possibilities primarily because of computational simplicity, a distinct advantage. The primary disadvantage of the functional form providing the basis of conditional logit is a property termed the "independence of irrelevant alternatives." This restriction of the model is quite unrealistic in many situations. To date, attempts to correct for this shortcoming have been on an ad hoc basis and not generally applicable.

This paper proposes a computationally feasible method of estimation not constrained by the "independence" restriction and which allows for a much richer range of human behavior than does the conditional logit approach. An important characteristic of the model is the explicit allowance for variation in tastes across individuals for the attributes of alternatives. This gain in realism, though, is at the expense of computational simplicity. To date, application of the model is limited to choice situations with four or five alternatives. The example in this paper uses only three. However, the increasing capacity of new generations of computer facilities may be expected to broaden the applicability of our approach.

The general problem that we are dealing with may be formulated as follows. Consider an individual who faces J alternatives and must choose one of them. Let the probability that he chooses the j^{th} alternative be P_j , where $\sum_{j=1}^J P_j = 1$. Let the outcome be represented by a vector $Y = (y_1, y_2, \dots, y_J)$, where y_j is either zero or one, and $\sum_{j=1}^J y_j = 1$. Then the probability that the first alternative is chosen is given by the probability that $Y = (1, 0, \dots, 0)$, where the probability of any Y

is given by $P_1^{y_1} P_2^{y_2} \dots P_J^{y_J}$. For N identical individuals indexed by i , the likelihood that $Y_1 = (y_{11}, y_{12}, \dots, y_{1J}), \dots, Y_i = (y_{i1}, y_{i2}, \dots, y_{iJ})$, etc.... is given by the following likelihood function,

$$e^L = \prod_{i=1}^N P_1^{y_{i1}} P_2^{y_{i2}} \dots P_J^{y_{iJ}}.$$

If individuals face different numbers of alternatives, J must be indexed by i , and if the i^{th} individual is faced with R_i repetitions of the same choice situation, then $\sum_{j=1}^{J_i} y_{ij} = R_i$, and the likelihood function is given by,

$$e^L = \prod_{i=1}^N y_{i1}^{y_{i1}} y_{i2}^{y_{i2}} \dots y_{iJ_i}^{y_{iJ_i}} P_1^{y_{i1}} P_2^{y_{i2}} \dots P_{J_i}^{y_{iJ_i}}.$$

A common statistical problem is to find the values of the P_j that maximize the value of this likelihood function. A more general problem is to allow the selection probabilities to be dependent on attributes of the alternatives in the choice set and on attributes of the individual making the choice. That is, the probability that the i^{th} individual chooses the j^{th} alternative is given by $P_{ij} = P(X_{ij}, a_i)$, where X_{ij} is a vector of attributes of the j^{th} alternative faced by individual i and a_i is a vector of characteristics of the i^{th} individual. Conditional probit analysis differs from conditional logit analysis in the stochastic specification of the probabilities P_{ij} . The probit specification is based on the multivariate normal distribution, while the logit formulation rests on the univariate extreme value distribution. In turn, it is useful to relate the selection probabilities, given the attributes of the alternatives and characteristics of decision makers, to underlying theories of

consumer choice. While both models can be related to the idea of the representative individual (explained below), we will see that the different stochastic formulations imply quite different theories of individual behavior and, in fact, lead to quite different predictions of selection probabilities in some important choice situations. Even though both models are likely to "fit the data" well -- analogous to the similar results obtained from logit and probit analysis in the binary case;¹ the predicted effect of the introduction of a new alternative based on one model is likely to differ substantially from that of the other. This possibility is investigated in the last sections of the paper.

Section 2 examines the general specifications of qualitative choice models. The deterministic theory of the representative individual is discussed and then a stochastic theory is formulated from which the choice probabilities are derived. In Section 3 specific parametric distributions are developed and conditional probit and conditional logit models are discussed. Section 4 deals with maximum likelihood estimation of the unknown parameters in the probit model and the formulation of statistics to compare different model specifications. An empirical example of transportation mode choice for commuters is analyzed in Section 5. Important differences between the conditional probit and conditional logit models are found. In Section 6 artificial data are used to compare forecasts based on the two models when a new transit mode is introduced. One of the important uses of conditional logit models has been in this situation;

1. We will see below, however, that our model will lead to a covariance term, for each choice situation, that depends on the attributes of the alternatives being compared -- the choice set.

thus, the comparative forecasts of the two specifications should be of interest. Again, important differences are found. Finally, the treatment of the "red-bus, blue-bus" problem by logit and probit models is discussed in Section 7.

2. A Model of Individual Choice

While economic theory provides a well-determined axiomatic theory of individual choice, use of this theory in econometrics is not always straightforward. Even when observations on individual choices are available, two problems remain. The investigator observes and measures only some portion of the factors that determine individual decisions. There are unobserved attributes of the alternatives in the choice sets faced by decision makers and unobserved attributes of the decision makers themselves. Also, the investigator usually lacks repeated observations on choices made by any given individual, in particular, under changing conditions. The usual situation in economics is that data is collected for many individuals, but with only one (random) observation for each. The following information is typically available: the observed attributes of the alternatives in the choice sets faced by individuals, their observed attributes, and their choices. In qualitative choice situations with appropriate sampling techniques each trial is assumed to be a single drawing from an independent but not identical multinomial distribution. The task of the empirical investigator is to construct a model of individual behavior that is consistent with estimation of the probabilities in the multinomial distribution. The estimation procedure can use only observed data; but a very important aspect of any such model is the treatment of the unobserved determinants of individual behavior.

A common procedure used in both economic theory and econometrics is to assume the existence of a "representative" or "average" individual who is assumed to have tastes equal to the average over all decision maker's with given observed attributes. Suppose the representative individual i faces alternatives X_{ij} ($j = 1, \dots, J$), where X_{ij} is a vector of the observed

characteristics of alternative j , and he is described by a vector of observed attributes a_i . Then this representative person is assumed to have a utility function \bar{U} defined over alternatives X , often assumed linear in parameters, such that,

$$(2.1) \quad \bar{U}_{ij} = \bar{U}(X_{ij}, a_i) = Z_{ij}\beta,$$

where Z_{ij} is a vector of arithmetic combinations of the elements of X_{ij} and a_i , and β is a vector of parameters. Note, the further assumption has been made that $\bar{U}(\cdot)$ or, equivalently, β is common to the entire population. This assumption is made necessary by the lack of individual repetitions. If β is assumed constant only over subsets of the entire population, the sample would be partitioned according to observed characteristics and different utility functions would be estimated for each.

Once a functional form representing the behavior of the average individual is given, a stochastic theory is used to describe unobserved components that differentiate a particular individual from the average. That is, the deterministic model, equation (2.1), is assumed to represent average (e.g., mean) behavior, and a nondeterministic part to represent (random) deviations from this average. A convenient parametrization of the random utility of alternative j to person i is then

$$(2.2) \quad U_{ij} = \bar{U}(X_{ij}, a_i) + \varepsilon(X_{ij}, a_i) = Z_{ij}\beta + \varepsilon_{ij},$$

where ε is a random variable. Two possible explanations for the stochastic term may be given. The first is that individuals behave randomly, perhaps due to random firing of neurons; so that faced repeatedly with the same alternative set, the same individual makes different choices. A more attractive explanation is to assume that given the observed data (X_{ij}, a_i) ,

a stochastic distribution is induced by unobserved data in each trial of the experiment. That is, there are unobserved characteristics of the decision maker (or random preferences) and unobserved attributes of the alternatives. We will discuss this possibility in some detail below.

Given the specification of the utility function U_{ij} , the individual is assumed to choose the alternative that maximizes his utility. Suppose individual i faces three choices, $J = 3$. The probability that he chooses the first alternative is,

$$(2.3) \quad P_{i1} = \text{pr}[U_{i1} > U_{i2} \text{ and } U_{i1} > U_{i3}] \\ = \text{pr}[\epsilon_{i2} < \bar{U}_{i1} - \bar{U}_{i2} + \epsilon_{i1} \text{ and } \epsilon_{i3} < \bar{U}_{i1} - \bar{U}_{i3} + \epsilon_{i1}]$$

Similar expressions are obtained for P_{i2} and P_{i3} . It is clear that the P_{ij} are well defined probabilities once we choose a joint density function for the ϵ_{ij} . Let $f(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}) = f_i(\epsilon)$ be this density function and let $F(k_{i1}, k_{i2}, k_{i3})$ be the corresponding distribution function. Then the probability that person i chooses alternative 1 is

$$(2.4) \quad P_{i1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\bar{U}_{i1} - \bar{U}_{i2} + \epsilon_{i1}} \int_{-\infty}^{\bar{U}_{i1} - \bar{U}_{i3} + \epsilon_{i1}} f(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}) d\epsilon_{i3} d\epsilon_{i2} d\epsilon_{i1} \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\bar{U}_{i,12} + \epsilon_{i1}} \int_{-\infty}^{\bar{U}_{i,13} + \epsilon_{i1}} f(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}) d\epsilon_{i3} d\epsilon_{i2} d\epsilon_{i1} \\ = \int_{-\infty}^{\infty} F_1(\epsilon_{i1}, \bar{U}_{i,12} + \epsilon_{i1}, \bar{U}_{i,13} + \epsilon_{i1}) d\epsilon_{i1}$$

where $\bar{U}_{i,jj'}$ is the difference in utility of alternatives j and j' to the representative individual and $F_1 = \partial F / \partial k_{i1}$. It is sometimes more convenient to look at equations (2.3) and (2.4) in differenced form.

(The subscript i will be dropped and will be used only where needed to prevent confusion.) Then the probability of choosing alternative 1 is,

$$(2.5) \quad P_1 = \text{pr}[\eta_{21} < \bar{U}_{12} \text{ and } \eta_{31} < \bar{U}_{13}],$$

where $\eta_{jj'} = \epsilon_j - \epsilon_{j'}$. This change of variables will induce a new joint density $g_1(\eta_{21}, \eta_{31})$ that depends on which probability is being considered (i.e., $g_1(\cdot) \neq g_2(\cdot)$). The new density $g_j(\cdot)$ is easily derived from the density $f(\epsilon)$ by a linear transformation with Jacobian equal to unity, then from (2.5),

$$(2.6) \quad P_1 = \int_{-\infty}^{\bar{U}_{12}} \int_{-\infty}^{\bar{U}_{13}} g_1(\eta_{21}, \eta_{31}) d\eta_{21} d\eta_{23}$$

Two important points to note are that this transformation reduces the order of integration by one, and because only subtraction is involved in going from $f(\cdot)$ to $g_j(\cdot)$, distributions which are closed under subtraction or are transformed into mathematically convenient distributions by subtraction may be desirable candidates for $f(\cdot)$.

The specification of the density function $f(\epsilon)$ will complete the formulation of the model of individual choice. It then remains to estimate the unknown parameters of \bar{U}_{ij} as well as any unknown parameters of $f(\epsilon)$. While mathematical convenience of estimation must be an important consideration in choosing the density function f , because equation (2.6) contains a $J-1$ dimension integral, a reasonable stochastic theory, represented by $f(\epsilon)$, is essential for a model that implies acceptable behavioral characteristics of individuals. In the next section two stochastic parametrizations are discussed which lead to convenient expressions for the basic probability equation (2.6).

3. Probit and Logit Models of Stochastic Choice

Beginning with a general formulation, we will discuss first the conditional probit model; then the familiar conditional logit specification and its primary disadvantage. The focus of the latter discussion will be on the parametric specification of the covariance matrix of the joint density function $f(\epsilon)$. The goal is to develop a parametrization with reasonable behavioral implications that is also computationally feasible, and that overcomes the main shortcoming of the logit model. An important property of the proposed random coefficients parametrization is explicit allowance for a distribution of tastes among decision makers in the population. For purposes of exposition, we will consider only three alternatives.

Following the discussion above, we assume that the values of the three alternatives to the i^{th} individual can be represented by,

$$(3.1a) \quad U(X_1, a_i) = \bar{U}(X_1, a_i) + \epsilon(X_1, a_i) = \bar{U}_{i1} + \epsilon_{i1}$$

$$(3.1b) \quad U(X_2, a_i) = \bar{U}(X_2, a_i) + \epsilon(X_2, a_i) = \bar{U}_{i2} + \epsilon_{i2}$$

$$(3.1c) \quad U(X_3, a_i) = \bar{U}(X_3, a_i) + \epsilon(X_3, a_i) = \bar{U}_{i3} + \epsilon_{i3}$$

We may assume, as is usual, that $E(\epsilon_{ij}) = 0$, because any nonzero term would be absorbed in the mean function \bar{U}_{ij} .

a. The conditional probit model rests on the assumption that the ϵ_j in equation (3.1) have a multivariate normal distribution. The normal distribution provides a good approximation to many multivariate distributions, and has the advantage that $\eta_{jj} = \epsilon_j - \bar{\epsilon}_j$, is also distributed normally. Suppose then that $f_i(\epsilon)$ is multivariate normal with covariance matrix given by,

$$(3.2) \quad \Sigma_i = \begin{bmatrix} \sigma_{i,1}^2 & & \\ \sigma_{i,12} & \sigma_{i,2}^2 & \\ \sigma_{i,13} & \sigma_{i,23} & \sigma_{i,3}^2 \end{bmatrix} .$$

Consider the probability of selecting the first alternative. The covariance matrix for $\eta_{21} = \epsilon_2 - \epsilon_1$ and $\eta_{31} = \epsilon_3 - \epsilon_1$, with density function $g_1(\eta_{21}, \eta_{31})$, is given by

$$(3.3) \quad \Omega_1 = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & & \\ \sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23} & \sigma_1^2 + \sigma_3^2 - 2\sigma_{13} & \\ & & \end{bmatrix} = \begin{bmatrix} \omega_{1,11} & & \\ \omega_{1,12} & \omega_{1,22} & \\ & & \end{bmatrix} ,$$

where the index i has been suppressed. Note that g and Ω are subscripted according to the alternative whose choice probability is being referenced. Then the probability that the first alternative is chosen is given by,

$$(3.4) \quad P_1 = \int_{-\infty}^{\bar{u}_{12}/\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}} \int_{-\infty}^{\bar{u}_{13}/\sqrt{\sigma_1^2 + \sigma_3^2 - 2\sigma_{13}}} b_1(\eta_{21}, \eta_{31}; r_1) d\eta_{21} d\eta_{31}$$

where b_1 is a standardized bivariate normal distribution with correlation coefficient $r_1 = \omega_{1,12}/\sqrt{\omega_{1,11}\omega_{1,22}} = (\sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23})/$

$\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})(\sigma_1^2 + \sigma_3^2 - 2\sigma_{13})}$. A further transformation of variables allows (3.9) to be written as

$$(3.5) \quad P_1 = \int_{-\infty}^{\bar{u}_{12}/\sqrt{\omega_{1,11}}} \phi(\lambda) \Phi \left[\frac{\bar{u}_{13}/\sqrt{\omega_{1,22}(1-r_1^2)} - \lambda r_1/\sqrt{1-r_1^2}}{\sqrt{1-r_1^2}} \right] d\lambda,$$

where ϕ is a unit normal density function and Φ is a standardized normal cumulative distribution function. The probabilities P_2 and P_3 are similarly calculated. The stochastic specification is complete given a parametrization

of the covariance terms, $\sigma_{i,kk'}$, in (3.2). With no more than one observation for each individual, the covariance can be estimated only through parametrization.

Of course, as mentioned in Section 2, a reasonable assumption may be that the σ_{ij} are independent. In that case Ω_1 , for example, has a particularly simple form given by,

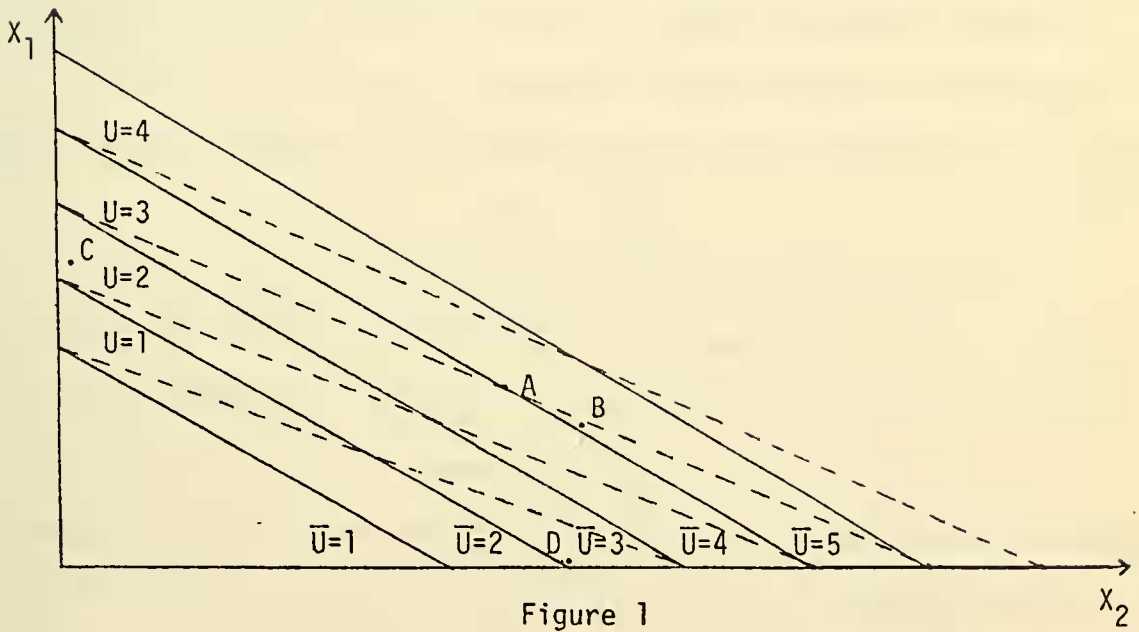
$$(3.6) \quad \Omega_1 = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_3^2 \end{bmatrix}.$$

If the variances are assumed to be equal across alternatives, then the Ω_j are identical for all j . And, because the variance terms can only be determined up to a scale factor, we can set them equal to one. The matrix Ω then has twos on the diagonal and ones on the off-diagonal. This case is sometimes referred to as the equi-correlated case. The independence assumption eases the computation burden of evaluating the integrals in equations (3.4) or (3.5) considerably. However, computational convenience is only one criteria for choosing Σ ; thus, other specifications are also considered. The problem then becomes one of choosing "good" parametrizations. We will argue below that some rather simple functional forms imply quite plausible behavioral assumptions about individual decision makers.

Consider again an individual facing a set of alternatives, each described by a vector of measured characteristics X_{ij} . As described above, we consider the "worth" to him of each alternative to be composed of two parts; the "average" worth of an alternative with measured characteristics X_{ij} , plus a deviation from this average, an "error" term. The average is the average over all alternatives with measured characteristics X ;

and, over the group of all decision makers, from which a particular individual is selected at random. The deviation is thus assumed to be a function of two factors: unobserved characteristics of the alternative together with a deviation in the tastes of a given individual from average tastes, those of the "representative" individual. We will argue that it may not be reasonable to assume that these deviations or errors are uncorrelated across alternatives in the choice set for a given decision maker. Indeed, we will argue that the degree of correlation between any two errors might be expected to depend on how "close" the corresponding alternatives are in measured characteristics. We will first try to motivate this idea in a heuristic manner. Then we will discuss possible metrics for measuring "closeness." In particular we will propose a general parametrization of the covariance matrix, simple cases of which are easily seen to capture the idea of closeness.

For purposes of exposition, let us assume first that all relevant characteristics of alternatives are measured; there are no unobserved attributes. Then the deviation of the utility of any individual from representative utility is due only to differences in tastes across the population of decision makers. Assume that the preferences, \bar{U} , of the average or representative individual over characteristics X_1 and X_2 are represented by the solid lines in figure 1. The preferences of an individual U are represented by the dashed lines. This individual is assumed to have an "unusually weak" taste for characteristic X_2 , and thus for the alternative indicated by point A on the graph. He is likely also to have a "weak" taste for any point such as B that is "close" to A. Knowing his preferences for A, however, may tell us much less about his valuation of alternatives like C or D that are relatively "far" from A.



Because we don't know the "shape" of any individual's preferences --that is, we don't know how any individual's preferences differ from the average -- our ability to predict the relationship between any two deviations decreases as the "distance" between them in attribute space increases.

Now assume that all decision makers have identical tastes; but that not all characteristics of alternatives are observed. That is, the "error" results from the values of unmeasured attributes; that for a given alternative are the same for all individuals, but vary from one alternative to the other. This is true even if measured attributes of two alternatives, for example, are the same; there may be many values of unobserved ones. We may expect unobserved attributes to be closer together if observed attributes are closer than if they are distant from each other.

A reasonable argument is based on the assumption that the set of all relevant (to the decision maker) attributes of alternatives has a multivariate distribution, say normal. If we assume in addition that the covariances between observed and unobserved attributes are not all zero, the expected value of unobserved attributes depends on the values of observed attributes. In fact, the expected values of unobserved attributes will be closer together, the nearer are observed characteristics. This can be seen by considering the expected value of unobserved, given observed attributes, when both groups are jointly normal.

But does this imply that deviations from representative utility are closer together the closer are observed characteristics? Recall that representative utility, $\bar{U}(X_j)$, is the expected value of $U(X_j)$, given observed attributes. Unobserved attributes are "included" in \bar{U} . The relationships between deviations from $\bar{U}(X)$ and $\bar{U}(Y)$ should not depend

on how close X and Y are. In fact, assuming that the rules of random sampling are followed, the covariance of $\varepsilon(X)$ and $\varepsilon(Y)$ is zero.¹

Unobserved attributes, however, should be expected to affect the correlation between deviations when tastes vary across individuals.

We will propose a rather general random utility formulation of the model that captures the essence of these heuristic ideas. Special cases of the formulation are then discussed. For convenience of exposition, we assume that there are only two measured attributes, X_1 and X_2 . The analysis can easily be extended to more.

Let,

$$\begin{aligned} (3.7) \quad U(X) &= \bar{U}(X,a) + \varepsilon(X,a) \\ &= (\bar{\beta}_1 + \beta_1)X_1 + (\bar{\beta}_2 + \beta_2)X_2 + \gamma \\ &= \bar{\beta}_1 X_1 + \bar{\beta}_2 X_2 + \beta_1 X_1 + \beta_2 X_2 + \gamma. \end{aligned}$$

In this specification, $\bar{U} = \bar{\beta}_1 X_1 + \bar{\beta}_2 X_2$, $\varepsilon(X,a) = \beta_1 X_1 + \beta_2 X_2 + \gamma$, and

1. More formally, let X and Y be the observed attributes of two alternatives and let X^C and Y^C be unobserved. Assume the observed and unobserved attributes have a joint multivariate distribution (e.g., normal).

Then

$$\bar{U}(X) = EU(X) = EU(X, X^C | X) = \int U(X, X^C) dX^C,$$

and the covariance between $\varepsilon(X)$ and $\varepsilon(Y)$ by,

$$\begin{aligned} \text{Cov}[\varepsilon(X), \varepsilon(Y)] &= E[U(X) - EU(X, X^C | X)][U(Y) - EU(Y, Y^C | Y)] = \\ &E[U(X) \cdot U(Y)] - EU(X, X^C | X) \cdot EU(Y, Y^C | Y) = 0, \end{aligned}$$

since $E[U(X) \cdot U(Y)] = EU(X, X^C | X) \cdot EU(Y, Y^C | Y)$.

β_1 , β_2 , and γ are assumed to be uncorrelated random terms. The random variables β_1 and β_2 may be thought of as random taste parameters representing the effects of unobserved attributes of individuals. The term γ may be considered to represent "purely" random components of utility -- unobserved characteristics of alternatives, or purely random behavior on the part of individuals, for example.

Note that the taste parameters β_1 and β_2 are assumed to be uncorrelated. An obvious reason for this is the saving in computation that it allows. There is, however, a more fundamental rationalization. In some sense we describe alternatives by their attributes to allow explicit description of why one alternative may be preferred to another. If the tastes of individuals for one "attribute" are correlated with those for another, it must be that the two attributes have something "in common." If this common component is in fact identifiable, then we would like to isolate it by explicit consideration of it as a separate attribute. This would presumably eliminate correlation between the new attributes -- now more precisely defined. In this sense, precise definition of attributes, if successful, would lead to defined attributes for which individual tastes are uncorrelated.

The variance of the error term corresponding to the j^{th} alternative faced by the i^{th} individual is given by,

$$(3.8) \quad \text{Var}(\epsilon_{ij}) = \sigma_{ij}^2 = \sigma_{\beta_1}^2 X_{1ij}^2 + \sigma_{\beta_2}^2 X_{2ij}^2 + \sigma_{\gamma ij}^2,$$

where $\sigma_{\beta_1}^2$ represents the variance in tastes across individuals relative to the measured characteristics, X_1 , etc... The covariance between the error terms corresponding to two alternatives, j and j' , faced by the

i^{th} individual is given by,

$$(3.9) \quad \text{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_{\beta_1}^2 X_{1ij} X_{1ij'} + \sigma_{\beta_2}^2 X_{2ij} X_{2ij'}$$

Because the variance of ε is identified only up to some arbitrary multiple, we can fix one of the variances, $\sigma_{\beta_1}^2$, $\sigma_{\beta_2}^2$, or σ_{γ}^2 , at an arbitrary value. We have elected to set $\sigma_{\gamma_{ij}}^2 = 1$. We must estimate only $\sigma_{\beta_1}^2$ and $\sigma_{\beta_2}^2$. In general then the covariance matrix Σ is of the form,

$$(3.10) \quad \Sigma_i = \begin{matrix} \sum_k \sigma_{\beta_k}^2 X_{kil}^2 + \sigma_{\gamma_{il}}^2 & & & \\ \sum_k \sigma_{\beta_k}^2 X_{kil} X_{ki2} & \sum_k \sigma_{\beta_k}^2 X_{ki2}^2 + \sigma_{\gamma_{i2}}^2 & & \\ \sum_k \sigma_{\beta_k}^2 X_{kil} X_{ki3} & \sum_k \sigma_{\beta_k}^2 X_{ki2} X_{ki3} & \sum_k \sigma_{\beta_k}^2 X_{ki3}^2 + \sigma_{\gamma_{i3}}^2 & \end{matrix}$$

where the summation is over all measured characteristics.¹ Note that the γ_{ij} are assumed to be independent across alternatives faced by a given individual, as well as across individuals.

If tastes do not vary across individuals -- that is, if $\sigma_{\beta_k}^2 = 0$ for all k -- and we assume that the $\sigma_{\gamma_{ij}}^2 = 1$ for all j , then,

$$(3.11) \quad \Sigma_i = \begin{matrix} & & 1 & \\ & 0 & 1 & \\ & & & \\ 0 & 0 & 1 & \end{matrix},$$

1. This specification thus allows an "alternative set" effect as used by McFadden []. The logit and independent probit assume this effect to be zero.

and,

$$(3.12) \quad \Omega_j = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix},$$

for all choices j . This is the independence case.

We mentioned above that the variance in the values, or utilities, that different individuals assign to any particular alternative, or alternatives with the same measured characteristics, can be thought of as resulting from two factors -- differences in tastes across individuals, and unobserved characteristics of the alternative. (This ignores the possibility of purely random behavior.) Some idea of the relative importance of these two factors can be had by comparing the estimates of the normalized $\sigma_{\beta_k}^2$ with the normalized value of σ_γ^2 , fixed at 1.

The number of parameters to estimate can be reduced and the model simplified by constraining $\sigma_{\beta_k}^2$ to equal some constant variance σ_β^2 for all characteristics k . For this simplification to be at all reasonable certainly requires that the variables X_k be normalized, since the units in which they are measured is completely arbitrary. We experimented with this constrained model after normalizing measures on the X_k by dividing them by their respective sample standard deviations (determined from measures across all alternatives and individual decision makers in the sample).

We can constrain the covariance specification even further by assuming that there are no unmeasured characteristics that affect individual decisions and setting $\sigma_\gamma^2 = 0$. That is, we assume not only that the variance in tastes is the same for all measured attributes of alternatives, but

also that all the randomness in utility results from variation in tastes. This formulation in fact allows a straightforward intuitive feeling for the properties of the more general model. The relationship between this specification and the loose idea of the correlation of errors depending on "closeness" in attribute space is easily seen. In this case, $U(X)$ is given by,

$$(3.13) \quad U(X) = \bar{\beta}_1 X_1 + \bar{\beta}_2 X_2 + \beta_1 X_1 + \beta_2 X_2,$$

where $\bar{U}(X,a) = \bar{\beta}_1 X_1 + \bar{\beta}_2 X_2$, and $\epsilon(X,a) = \beta_1 X_1 + \beta_2 X_2$. If β_1 and β_2 have equal variances and are uncorrelated, the covariance between any two errors, say for the alternatives X and Y , is given by: $\text{Cov}[\epsilon(X), \epsilon(Y)] = \sigma^2(X_1 Y_1 + X_2 Y_2)$. The correlation between the two is given by,

$$(3.14) \quad \rho_{XY} = \frac{\sigma^2(X_1 Y_1 + X_2 Y_2)}{\sqrt{\sigma^2(X_1^2 + X_2^2)} \sqrt{\sigma^2(Y_1^2 + Y_2^2)}} = \frac{X_1 Y_1 + X_2 Y_2}{\|X\| \|Y\|} = \cos(X, Y).$$

This formulation assumes that if A and B have the same measured characteristics, a decision maker will treat them as identical; the deviation of his valuation of A from that of the representative individual will equal the deviation in his valuation of B . (See figure 2.) If there were no unmeasured characteristics of alternatives, we would want precisely this property. Identical alternatives are treated identically by a given decision maker.¹ (We will see below that under this formulation, adding

1. It also has the property that if A and B are orthogonal, or at right angles to one another, so that $A_1 B_1 + A_2 B_2 = 0$; the corresponding deviations are assumed to be uncorrelated. Finally, if two alternatives are in the same direction, but different "distances" from the origin, like A and D , they are assumed to have the same correlation as alternatives closer together, like A and C , or two alternatives A .

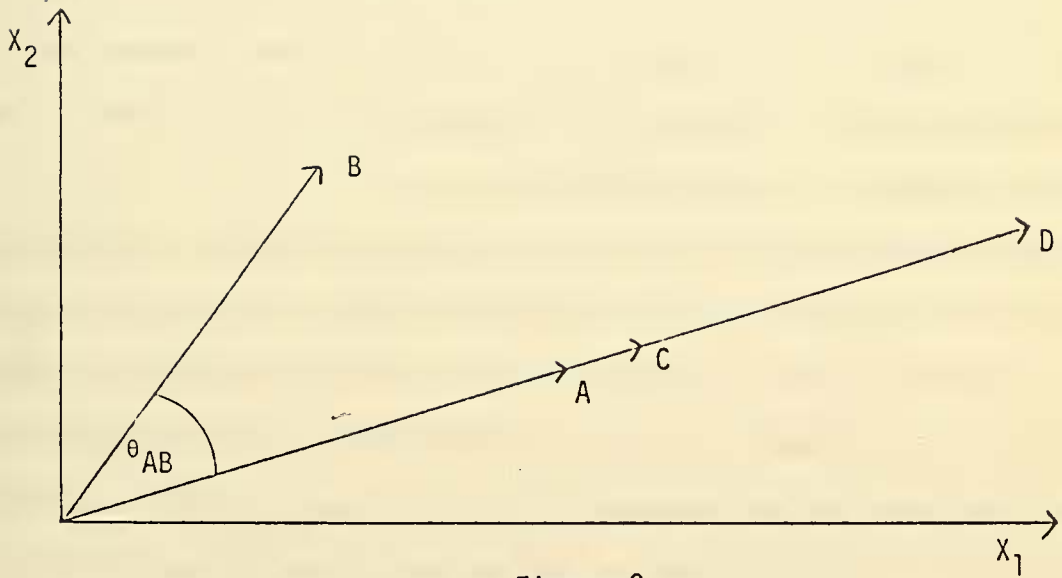


Figure 2

an alternative identical to an existing one will not change the predicted probability of choosing other alternatives. This represents the absence of the unwanted "independence of irrelevant alternatives", property; however, it is an extreme case, stronger than we would like to impose.)

In summation, we will experiment with three parametrizations of the covariance matrix. The first constrains all off-diagonal elements to be zero. We call this the independent probit case. The second assumes that off diagonal elements are given by (3.9). We refer to it as covariance probit. And, third, we will use an intermediate parametrization that constrains the taste variation parameters to be equal. The more flexible parametrizations correspond to letting the data "choose" the degree of association of ε_{ij} and ε_{ij} , conditional on how "close" the observed alternatives are in attribute space.

These three parametrizations of the covariance matrix are all generalizations of the "probit" model used often in economic analysis. To date only the independent probit model has been used in the binary choice case where its properties are rather similar to the more commonly used logit model because the distribution functions on which the models are based are similar except in the extreme tails. However, with three or more alternatives the behavior of the logit and covariance probit models is apt to differ since the logit model is based on binary comparisons while the covariance probit model is based on an n-way comparison with interdependent stochastic terms. In particular, predicted effects of the introduction of a new alternative are likely to differ substantially between the two models. But before comparing results from the two models we will review briefly the relevant aspects of the logit model.

b. The conditional logit specification is based on the assumption that the ϵ_j in (3.1) are independently and identically distributed with extreme value density functions (Type I extreme value -- Johnson and Kotz, pp. 272 ff),

$$(3.15) \quad f(\epsilon_j) = e^{-\epsilon_j} e^{-e^{-\epsilon_j}}$$

and distribution functions,

$$(3.16) \quad F(k_j) = \text{pr}(\epsilon_j \leq k_j) = e^{-e^{-k_j}}.$$

It is the limiting distribution (as $n \rightarrow \infty$) of the greatest value of n independent and identically distributed random variables. While it is difficult to argue that the extreme value distribution is a particularly good representation of the stochastic nature of the ϵ_j , it turns out to be extremely convenient mathematically. The difference between any two random variables with this distribution [e.g., $\eta_{jj'} = \epsilon_{ij} - \epsilon_{ij'}$, in equations (2.5) and (2.6)] has a logistic distribution function, that gives rise to the binary logit model. For example, if only two alternatives are available, the probability that the first is chosen is given by,

$$(3.17) \quad P_1 = \frac{e^{\bar{U}_1}}{e^{\bar{U}_1} + e^{\bar{U}_2}} = \frac{1}{1 + e^{\bar{U}_2 - \bar{U}_1}}.$$

The probability that the first is chosen from three alternatives is given by,

$$(3.18) \quad P_1 = \frac{e^{\bar{U}_1}}{e^{\bar{U}_1} + e^{\bar{U}_2} + e^{\bar{U}_3}} = \frac{1}{1 + e^{\bar{U}_2 - \bar{U}_1} + e^{\bar{U}_3 - \bar{U}_1}},$$

which appears as a straightforward extension of the binary case. This simple form arises because the relevant probabilities in equation (2.4) are independent, as well as having convenient functional forms.¹

That is, $F_1(k_1, \bar{U}_{12} + k_1, \bar{U}_{13} + k_1) = f(k_1) \cdot \Pr(\epsilon_2 \leq \bar{U}_{12} + k_1) \cdot \Pr(\epsilon_3 \leq \bar{U}_{13} + k_1) = f(k_1) \cdot F(\bar{U}_{12} + k_1) \cdot F(\bar{U}_{13} + k_1)$. The integration in (2.4) is essentially taking a weighted (by $f(\cdot)$) average over the values of ϵ_1 , of the product of binary comparisons, where the value of ϵ_1 is fixed in each. Another way to see that the model assumes that only binary comparisons need be made is to rewrite (3.18) as the inverse of the sum of binary odds. That is P_1 can be written as,

$$(3.19) \quad P_1 = 1 / \left(\frac{e^{\bar{U}_1}}{e^{\bar{U}_1}} + \frac{e^{\bar{U}_2}}{e^{\bar{U}_1}} + \frac{e^{\bar{U}_3}}{e^{\bar{U}_1}} \right).$$

(The extension to a greater number of alternatives is straightforward.)

In fact, we can let any alternative be a "basis" for the set of alternatives, and then write any probability P_j in terms of binary comparisons with it as,

1. McFadden [1973] has shown that a necessary and sufficient condition for the random utility model with independent and identically distributed errors to yield the conditional logit or "strict utility" model, is that the errors have extreme value distributions.

$$(3.20) \quad P_j = \frac{e^{\bar{U}_j - \bar{U}_1}}{e^{\bar{U}_1 - \bar{U}_1} + e^{\bar{U}_2 - \bar{U}_1} + e^{\bar{U}_3 - \bar{U}_1}} .$$

Thus all choices may be assumed to result from binary comparisons with a basis alternative. Well defined probabilities are obtained by appropriate transformation of these "comparisons" (differences), $e^{\bar{U}_j - \bar{U}_1}$, and normalization. We emphasize the binary comparison aspect of this model because it is integrally related to its primary shortcoming. This very powerful simplification brings with it very restrictive assumptions on individual behavior. As Luce and Suppes [1965], Marshak [1960], and McFadden [1973] have pointed out, the relative odds of alternative j being chosen over alternative j' is independent of the number, or attributes, of other alternatives in the set. This so called independence of irrelevant alternatives assumption follows directly from equation (3.18).

While for many problems the logit choice model is adequate, for some problems which contain alternatives that are close substitutes for each other the specification is too restrictive. For example, consider an individual with a choice of two residence locations, say Florida and Vermont. Assume that he likes the sun in Florida; but he likes equally the beautiful fall and winter skiing possibilities in Vermont. This results in a 50-50 chance that he will choose Florida over Vermont; $P_{\text{Florida}}/P_{\text{Vermont}} = 1$. Now assume that his alternative set is expanded to include New Hampshire, which we assume to be identical to Vermont in skiing opportunities and fall beauty; $U(\text{Vermont}) = U(\text{New Hampshire})$. We would expect that the individual would still choose Florida with probability .5 and would choose Vermont or New Hampshire with probability .5

and each of them with probability .25. Contrary to this expectation, the conditional logit functional form constrains the odds of choosing Florida over Vermont to remain at 1. The probability of choosing Florida, as well as the probability of choosing Vermont, falls. The model predicts that each state will be chosen with probability $1/3$. In the empirical application used later there are three alternatives for commuting to work: drive alone, car pool, or bus; the characteristics of the first two alternatives are similar. Yet, if we had started only with the drive alone-bus split and wanted to predict the effect of car-pooling, the relative odds of the original choices would be constrained to remain the same, while it seems likely that much more substitution exists between driving alone and car pooling than between taking a bus and car pooling. These restrictions essentially result from the assumption of independent errors in (3.1). The goal of the conditional probit model is to allow relaxation of these restrictions.

The probit and logit models have been specified in terms of the theory of a representative individual and a stochastic theory of the distribution of "deviations" from the representative individual. On prior grounds it is difficult to choose between them because a more general specification is gained at the expense of computational convenience. After discussing the estimation procedure for the probit model in the next section, an empirical example is used to demonstrate differences between probit and logit models in estimation and prediction. We might expect the independent probit and the logit models to have similar properties and, in fact, they lead to almost identical empirical results. Both assume independence and after normalizing the variances, the distributions that form the basis of the

models -- independent normal and extreme value respectively -- are quite similar. The independent probit is introduced to allow direct comparison (nested hypothesis testing) with the covariance probit. Although we can only make "precise" comparisons between the two probit models, the fact that the independent probit and the logit models give almost identical results allows us in practice to make implicit comparisons between the logit and the covariance probit models.

4. Estimation

Given a random sample of individuals, the unknown parameters are estimated by maximum likelihood. A sample (without repetitions) may be thought of as N independent drawings from a multinomial distribution with log-likelihood function,

$$(4.1) \quad L = K + \sum_{i=1}^N \sum_{j=1}^J y_{ij} \log p_{ij}$$

where $y_{ij} = 1$ if person i chooses alternative j , and $y_{ij} = 0$ otherwise. Both the probit and logit likelihood functions have this same general form, but have different specifications of the probabilities p_{ij} . Estimation of the logit model (equation 3.4) is discussed at length by MacFadden [1973] and will not be described here. In the case of three alternatives, the relevant probabilities for the probit model are given by equations corresponding to (3.9) or (3.10) with the bivariate distributions having the covariance matrices,

$$(4.2) \quad \begin{aligned} \Omega_1 &= \begin{bmatrix} \sigma_{\gamma 1}^2 + \sigma_{\gamma 2}^2 + \sum_{i=1}^K \sigma_{\beta_i}^2 (X_{1i} - X_{2i})^2 & & \\ \sigma_{\gamma 1}^2 + \sum_{i=1}^K \sigma_{\beta_i}^2 (X_{1i} - X_{2i})(X_{1i} - X_{3i}) & \sigma_{\gamma 1}^2 + \sigma_{\gamma 3}^2 + \sum_{i=1}^K \sigma_{\beta_i}^2 (X_{1i} - X_{3i})^2 & \\ \sigma_{\gamma 2}^2 + \sum_{i=1}^K \sigma_{\beta_i}^2 (X_{2i} - X_{1i})(X_{2i} - X_{3i}) & \sigma_{\gamma 2}^2 + \sigma_{\gamma 3}^2 + \sum_{i=1}^K \sigma_{\beta_i}^2 (X_{2i} - X_{3i})^2 & \end{bmatrix} \\ \Omega_2 &= \begin{bmatrix} \sigma_{\gamma 1}^2 + \sigma_{\gamma 2}^2 + \sum_i \sigma_{\beta_i}^2 (X_{2i} - X_{1i})^2 & & \\ \sigma_{\gamma 2}^2 + \sum_i \sigma_{\beta_i}^2 (X_{2i} - X_{1i})(X_{2i} - X_{3i}) & \sigma_{\gamma 2}^2 + \sigma_{\gamma 3}^2 + \sum_i \sigma_{\beta_i}^2 (X_{2i} - X_{3i})^2 & \\ \sigma_{\gamma 1}^2 + \sigma_{\gamma 3}^2 + \sum_i \sigma_{\beta_i}^2 (X_{3i} - X_{1i})^2 & & \\ \sigma_{\gamma 3}^2 + \sum_i \sigma_{\beta_i}^2 (X_{3i} - X_{1i})(X_{3i} - X_{2i}) & \sigma_{\gamma 2}^2 + \sigma_{\gamma 3}^2 + \sum_i \sigma_{\beta_i}^2 (X_{3i} - X_{2i})^2 & \end{bmatrix} \\ \Omega_3 &= \end{aligned}$$

The derivatives of (4.1) with respect to β_k yields,

$$(4.3) \quad \frac{\partial L}{\partial \beta_k} = \sum_{i=1}^N \sum_{j=1}^J \frac{y_{ij}}{P_{ij}} \frac{\partial P_{ij}}{\partial \beta_k} .$$

The derivatives of the P_{ij} with respect to β_k have a simple form comprised of standardized normal densities and distributions. For example, let us rewrite equation (3.3) as

$$(4.4) \quad P_1 = \int_{-\infty}^{\tilde{U}_{12}} \int_{-\infty}^{\tilde{U}_{13}} b_1(\eta_{21}, \eta_{31}; r_1) d\eta_{21} d\eta_{31}$$

where $\tilde{U}_{12} = \bar{U}_{12} / \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} = (Z_{i1} - Z_{i2}) \beta / \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} = \tilde{Z}_{12}\beta$ and likewise for \tilde{U}_{13} . Then the derivative has the formula

$$(4.5) \quad \frac{\partial P_1}{\partial \beta_k} = \phi(\tilde{Z}_{12}\beta) \Phi \frac{(Z_{13}\beta - r_1 Z_{12}\beta)}{\sqrt{1-r_1^2}} \tilde{Z}_{12}^k + \phi(\tilde{Z}_{13}\beta) \Phi \frac{\tilde{Z}_{12}\beta - r_1 \tilde{Z}_{13}\beta}{\sqrt{1-r_1^2}} \tilde{Z}_{13}^k .$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are standard normal density and distribution functions respectively. Thus, in the three alternative case the gradient involves only univariate normal densities and distributions that are easily evaluated on a computer. To obtain likelihood values requires evaluation of bivariate normal distributions. This is done using a modification of an algorithm first introduced by Owen [1956]. Each additional alternative past three increases the order of the integrals in the derivatives by one. Thus computation with many alternatives may be prohibitively expensive. To

date, the specification has been used for three and four alternatives and costs have been moderate. Significant cost reductions do accrue to careful programming of the maximization routine. In the case of the unknown σ_i^2 entering the covariance matrix, the corresponding derivative again has a simple form. For example, the derivative of P, with respect to σ_i^2 is given by,

$$\begin{aligned}
 \frac{\partial P_1}{\partial \sigma_{\beta_i}^2} = & -\phi(\tilde{Z}_{12}^\beta) \phi \frac{\tilde{Z}_{13}^\beta - r_1 \tilde{Z}_{12}^\beta}{\sqrt{1 - r_1^2}} \cdot \frac{\tilde{Z}_{12}^\beta}{2\omega_{11}} \cdot \frac{\partial \omega_{1,11}}{\partial \sigma_{\beta_i}^2} \\
 (4.6) \quad & -\phi(\tilde{Z}_{13}^\beta) \phi \frac{\tilde{Z}_{12}^\beta - r_1 \tilde{Z}_{13}^\beta}{\sqrt{1 - r_1^2}} \cdot \frac{\tilde{Z}_{13}^\beta}{2\omega_{22}} \cdot \frac{\partial \omega_{1,22}}{\partial \sigma_{\beta_i}^2} \\
 & + \phi(\tilde{Z}_{13}^\beta) \phi \frac{\tilde{Z}_{12}^\beta - r_1 \tilde{Z}_{13}^\beta}{\sqrt{1 - r_1^2}} \cdot \frac{1}{\sqrt{1 - r_1^2}} \cdot \frac{\partial r_1}{\partial \sigma_{\beta_i}^2},
 \end{aligned}$$

where the last term may be written as,

$$b_1(\tilde{Z}_{12}^\beta, \tilde{Z}_{13}^\beta; r_1) \frac{1}{\sqrt{\omega_{1,11}\omega_{1,22}}} \cdot \frac{\partial \omega_{1,12}}{\partial \sigma_{\beta_i}^2} - \frac{r_1}{2\omega_{1,11}} \frac{\partial \omega_{1,11}}{\partial \sigma_{\beta_i}^2} - \frac{r_1}{2\omega_{1,22}} \frac{\partial \omega_{1,22}}{\partial \sigma_{\beta_i}^2}.$$

The method used to maximize the likelihood function is that proposed by Berndt, Hall, Hall, and Hausman [1974]. It requires only first derivatives, each iteration is guaranteed to increase the value of the likelihood function and, given an additional requirement likely to be satisfied (equation 2.1 of Berndt et. al.), will converge to a stationary point. If a global maximum of the likelihood function L^* is assured, then under the usual regularity conditions (see Cox and Hinckley [1974]) the maximum likelihood estimates will be consistent and asymptotically normal. The asymptotic

covariance matrix of the maximum likelihood estimates is equal to the covariance matrix of the gradient of the likelihood function evaluated at the maximum $\theta^* = (\beta, \sigma^2)^*$. The expression,

$$(4.7) \quad Q(\theta) = \sum_{i=1}^N \frac{\partial(\sum_j y_{ij} \log P_{ij})}{\partial \theta} \frac{\partial(\sum_j y_{ij} \log P_{ij})}{\partial \theta} ',$$

has the same expectation in the limit as the covariance matrix of the gradient and thus provides a consistent estimate of the inverse covariance matrix of the parameters.

Given the covariance matrix of the estimates, large sample tests of coefficient values can be made in the usual way. The diagonal elements of the inverse of the covariance matrix of the gradient provide consistent estimates of the variances of the unknown parameters. Tests on model specification can also be constructed. One interesting test of the model specification might be that $\theta = (\beta, \alpha) = 0$. Then P_j in equation (4.4) equals $\frac{1}{3}$ and the log likelihood function (4.1) takes the value

$$(4.8) \quad L' = N \log \frac{1}{3} = N \log \frac{1}{3} .$$

Hypothesis testing then follows the classical likelihood procedure with $-2(L'-L^*) \sim \chi_{2K+2}^2$, where K is the dimension of β . For the independent covariance specification where $\omega_i = 1$ and the $\sigma_i^2 = 0$ are assumed zero, the appropriate test statistic is $-2(L'-L^*) \sim \chi_K^2$.

Unfortunately, in trying to test the probit against the logit specification a problem arises. Although both model specifications are intended to estimate the same multinomial probabilities in the likelihood function (4.1), neither model is a "nested" special case of the other. Thus,

classical likelihood ratio tests cannot be applied. While the two different likelihood values give some indication of how successful the respective models are with respect to the sample, no easy distributional theory can be developed to choose between the specifications. However, since the logit specification gives such similar results to the identity probit specification which is a vested case of the more general covariance specification, the relative likelihood values might be used in an "approximate" χ^2 test.

To test the two different classes of models a measure of fit against the observed frequencies can be constructed. Let,

$$(4.9) \quad Z = \sum_{i=1}^N \sum_{j=1}^J \frac{(y_{ij} - p_{ij}(\hat{\theta}))^2}{p_{ij}(\hat{\theta})} .$$

Then as $\sum_{i=1}^N \sum_{j=1}^J p_{ij}$ becomes infinite for each j , Z approaches the χ^2 distribution with $N \cdot J - (K+1) - N$ (because probabilities add to 1) degrees of freedom.

An alternative interpretation of the measure is that under random sampling each individual in the sample is given the same weight and the proportion of decision makers selecting the j^{th} alternative is estimated to be,

$$(4.10) \quad \hat{p}_j = \frac{1}{N} \sum_{i=1}^N p_{ij}(\hat{\theta}) .$$

We note that while predicted frequencies from both the logit and probit models can be compared to the observed frequencies, this does not provide a formal test of one model specification against the other. An associated nondistribution test is to compare the Z statistics from the three models.

Alternatively, the \hat{p}_j can be compared to the observed \bar{p}_j where $p_{ij} = 1$ if individual i makes choice j and is zero otherwise. Since given correct model specification the estimated empirical distribution function converges to the underlying population distribution function, comparing the estimated \hat{p}_j to the sample \bar{p}_j provides some guide to relative model performance.

In the next section both probit and logit are estimates, based on transportation mode choice data, are presented. The models are compared on the basis of both the formal and informal tests.

5. Empirical Example: Transit Mode Choice

Disaggregate models of transit mode choice are widely used to analyze factors that determine the type of transportation individuals use. The models are important in answering two types of questions: patronage of a new transit mode (e.g. new subways in San Francisco or Washington) and effects on patronage of changes in existing modes (e.g. introduction of off-peak fares in Boston). To date, almost all such models have been based on a conditional logit specification. The major weakness of such a specification is that when a new mode is introduced or the characteristics of an existing mode change, all predicted probabilities of choice are constrained to change by the same proportion. This result follows from the independence of irrelevant alternatives assumption. While on the micro level of the individual, this property seems undesirable, it is not clear that aggregate forecasts will be seriously wrong. Therefore, both probit and logit specifications are used to estimate mode choice for commuters to the central business district (CBD) of Washington, and in the next section a forecasting example is discussed.

Three alternative transit modes are available in this example.¹ The first mode is driving alone, the second is car pooling, and the third is public transit (bus). The model analyzes the worker's choice of travel mode from his home to his work place in the CBD. As the model of the representative individual postulates, two types of factors are important:

1. We would like to thank Professors M.E. Ben-Akiva, F. Koppelman, and S.R. Lerman for kindly providing us with the survey data used in this empirical study.

characteristics of the alternatives, x_{ij} , and attributes of decision makers, a_i .

The data used in the study are from the Washington Council of Governments Home Interview Survey described in [4]. The specification is similar to that used by Koppelman and Watchnatada [1975]. Three mode characteristics, and one personal attribute are used. The mode factors are cost of trip divided by income (PINC), in-vehicle travel time (INTIME), and out-of-vehicle travel time (OUTTIME). Cost and travel time are typically found to be the most important determinants of transit mode choice.

Three models are initially estimated. The first two are probit models corresponding to independence and a covariance specification where the alternative specific variances are set to one ($\sigma_{\gamma i}^2 = 1$). Lastly, conditional logit estimates are obtained for comparison purposes. One hundred observations are used with the cost of computing being quite small (under \$10 in all cases). Parameter estimates are presented in table 1. The parametric estimates of the probit model are roughly similar and accord with prior expectations with respect to sign. Note that the mean estimates are quite different for the probit estimates depending on the specification of the covariance matrix.

The hypothesis that $\theta = 0$, [$L' = -109.89$ -- equation (4.7)] is rejected at the 1% level by all the models using a χ^2 variate with 3 or 6 degrees of freedom. Another important test is a "saturated" model specification referred to as the presence of "alternative specific effects" by McFadden ([1973], p. 114) or as the "pure mode preference effect" ([1973], p. 131). This test entails inclusion of a constant for the different choices representing choice characteristics which have been

Table 1: Parameter Estimates, CBD Transit Mode Choice.

Variable	Probit Estimates (Standard Errors)		Logit (Standard Errors)
	Identity Probit	Covariance Probit	
L*(θ)	-103.0	-99.4	-102.9
1. PINC	-.411 (.135)	-1.05 (.369)	-.531 (.186)
2. INTIME	-.0549 (.0151)	-.0651 (.0416)	-.0713 (.0210)
3. OUTTIME	-.0884 (.0723)	-.0813 (.0729)	-.132 (.132)
Covariance Parameters			
4. PINC, $\sigma_{\beta_1}^2$		3.07 (3.88)	
5. INTIME, $\sigma_{\beta_2}^2$.0331 (.105)	
6. OUTTIME, $\sigma_{\beta_3}^2$		2.13 (5.15)	
Degrees of Freedom	97	94	97

left out of the model specification, e.g., a "convenience factor" for driving alone versus car pooling or taking transit. Besides providing a check for correct model specification, this test is important since forecasting the effects of the introduction of a new choice or altering the characteristics of an existing mode are impossible if alternative specific effects are present. The two saturated models which are tested are first to include alternative specific constants for the second and third choices in the identity probit specification (for choice one the alternative specific effect is normalized at zero) and second in the covariance probit model to have two alternative specific effects and to estimate $\sigma_{\gamma 2}^2$ and $\sigma_{\gamma 3}^2$ while normalizing $\sigma_{\gamma 1}^2 = 1$. Neither saturated model provides a significant improvement over the corresponding unsaturated model at the 10% level by a likelihood ratio test although some evidence is present that there may be a specific effect for the transit choice. Thus the model specifications would be appropriate to use in a forecasting situation.

<u>Variable</u>	<u>Identity Probit</u>	<u>Covariance Probit</u>
L*(θ)	-101.2	-98.9
1. PINC	-.107 (.230)	-.187 (.722)
2. INTIME	-.0379 (.0282)	-.0216 (.0567)
3. OUTTIME	-.139 (.073)	-.246 (.877)
4. Alt. Effect 2	-.033 (.496)	-.047 (1.11)
5. Alt. Effect 3	.483 (.403)	.361 (.460)
<hr/>		
Covariance Parameters		
6. PINC, σ_1^2		.098 (.529)
7. INTIME, σ_2^2		.0054 (.024)
8. OUTTIME, σ_3^2		.204 (.782)
9. Alt. Effect 2, $\sigma_{\gamma 1}^2$		2.02 (3.46)
10. Alt. Effect 3, $\sigma_{\gamma 2}^2$		1.001 (8.42)
<hr/>		
Degrees of Freedom	95	90
Unsaturated LR Statistic Against Unsaturated Model	3.60	1.00

The two unsaturated probit specifications can be tested against each other in two ways. First a likelihood ratio test can be constructed. Two times the likelihood ratio is distributed as χ^2 with three degrees of freedom. This statistic has the value of 7.2 which is significant at about the 7% level. The more general specification seems to provide evidence of considerable variation in tastes for the first and third mode attributes. A second test of the covariance specification as well as the logit specification is to calculate the predicted sample frequencies as discussed in Section 4.

Estimated Sample Frequency Distribution

	<u>Mode 1</u>	<u>Mode 2</u>	<u>Mode 3</u>	<u>χ^2</u>
Independent Probit	.362	.218	.421	1.67
Covariance Probit	.346	.197	.457	.281
Logit	.363	.218	.419	1.73
Sample	.34	.18	.48	

Two findings should be noted. All model specifications do a good job with the hypothesis of the predicted and empirical distributions not being significantly different not rejected in all cases. Also, as expected, the independent probit and logit specifications give virtually identical population forecasts. The covariance probit specification, however, does better than either of the other models. Not only is the χ^2 statistic lower, but also it never misses the sample frequency by more than .02. Given the wide variation in transit mode choice, these results appear promising for further development of the model.

6. Forecasting Example: Introduction of a New Transit Mode

Disaggregate mode choice models are often used to forecast patronage for a potentially new transit mode. Transit modes are described in terms of their characteristics x_{ij} . If a new mode is introduced its characteristic vector $x_{i,J+1}$ along with the individual attributes a_i will be used to form the representative utility $\bar{U}_{i,J+1} = Z_{i,J+1}\beta$. This utility will then be compared with the utility of the other modes \bar{U}_{ij} ($j = 1, J$ and $i = 1, N$) and the probabilities of use by each individual will be predicted by the stochastic model.

To ascertain if important differences between probit and logit forecasts might be expected, a new mode was "created" and resulting effects were predicted with the different specifications. The "new" mode is intended to correspond roughly to a new subway mode. Cost divided by income (PINC) is set at a mean value higher than the bus mode, the mean of in-vehicle travel time (INTIME) is assumed to be lower than the bus mean, and out-of-vehicle time (OUTTIME) is set at a higher mean. Two types of experiments were carried out. In the first, all x_{ij} for the new mode were set at the mean values. In the second, not reported on here, a random number generator assigns the x_{ij} randomly according to normal or rectangular distributions. Of course, in any actual forecast situation the design characteristics of the new mode would be used to set $x_{i,J+1}$.

For the probit models the probabilities of taking the new mode follow from equation (2.6) and are given by,

$$(6.1) \quad \tilde{p}_4 = \int_{-\infty}^{\bar{U}_{41}} \int_{-\infty}^{\bar{U}_{42}} \int_{-\infty}^{\bar{U}_{43}} h_1(\eta_{14}, \eta_{24}, \eta_{34}) d\eta_{14} d\eta_{24} d\eta_{34},$$

where $h_1(\cdot)$ is a trivariate normal density. For existing modes (say the first mode) the probability in equation (2.6) changes with the addition of a new dimension of integration. The new probability, \tilde{P}_1 , will be less than the former probability P_1 , but no fixed relationship between \tilde{P}_1 and the old P_1 can be ascertained. It depends in a complex way on both the \bar{U}_{jj} , and the covariance matrix of the normal distribution. On the other hand, the new probability according to the logit specification is

$$(6.2) \quad \tilde{p}_4 = \frac{e^{(Z_4 - Z_1)\beta}}{\sum_j e^{(Z_j - Z_1)\beta}} .$$

The old probabilities can be seen to change from equation (3.12) only in the addition of a new term in the denominator so that $\frac{\tilde{p}_1}{\tilde{p}_2} = \frac{p_1}{p_2}$, the independence of irrelevant alternatives assumption. As discussed above, many people find it unreasonable that for representative individuals the same proportion will change from driving alone to the subway mode as will switch from the bus mode. However, it is important to realize that this assumption is a micro one and may not have an adverse effect on macro (population) predictions.

Given the new alternative the macro forecasts that a transit planner might use in designing the new mode are:

Aggregate Probabilities of Transit Choice

	<u>Mode 1</u>	<u>Mode 2</u>	<u>Mode 3</u>	<u>Mode 4</u>
Independent Probit	.335	.200	.391	.074
Covariance Probit	.255	.181	.377	.187
Logit	.334	.200	.386	.080

The forecasts again demonstrate that the logit and identity probit specifications give very similar results. While the independence of irrelevant alternative property holds only at the level of individual probabilities, note that the logit specification forecast a ridership of 8% for the new mode with the three existing modes losing ridership of from 7.9% to 8.3%. Thus the independence property holds approximately at the macro level in this example. The covariance probit forecasts differ in two ways. First, a much greater ridership of 18.7% is forecast for the new mode. Also, many of the new riders are from the existing transit mode and from those people who currently drive alone. Current mode two which is car pooling has a decline of less than half that of the other two modes. Thus a differential response of the three existing modes to the new mode is found. While this experiment cannot be validated since artificial data is used, it does demonstrate that the different specifications may lead to different forecasts. Furthermore, both the logit and independent probit models suffer from the disadvantage that more people are forecast to take existing mode two than currently do so in the presence of three choices. These forecasts seem counterintuitive in the presence of an enlarged choice set.

7. Conditional Probit Specification and Random Utility Models

The stochastic specification of the utility function U_{ij} in equation (2.2) is sometimes called the random utility model. Block and Marschak [1960] and Luce and Suppes [1965] discuss the model at length, and recently Manski [1975] has given an interesting theoretical discussion of a special case of this model -- the independent and identically distributed random utility model. By this terminology is meant that the stochastic terms are I.I.D. and thus independent of x_{ij} and a_i . The logit specification, being a particular case of this specification, is often criticized on the following grounds.¹ (McFadden has called the problem the "red-bus blue-bus problem".) Suppose in a transit choice problem the original choice set consists of driving alone or taking a (red) bus. Next, an additional alternative is added identical in all characteristics to the red bus, except its color is blue. From equation (3.4) it can be seen that the logit specification will "correctly" forecast equal probability of use of the two buses; but, unfortunately, the decreased proportion of car drivers must exactly equal the decreased proportion of red bus riders. Thus, the odds of the first two choices must remain identical, so if the original probabilities were 2/3 car and 1/3 bus, the new probabilities are 1/2 for car and 1/4 for each type of bus. This counterintuitive result has led to many attempts at "correction" of the logit specification to remove the property.

Note that the independent probit specification of equation (3.5) has a similar undesirable property. The original probability of driving is

1. G. Debreu, to the best of our knowledge, was the first to point out this property, although his example is less prosaic than ours.

$$(7.1) \quad P_1 = \int_{-\infty}^{\bar{U}_{12}/\sqrt{2}} \phi(\eta_{21}) d\eta_{21}$$

where $\phi(\cdot)$ is the standard normal density. When the blue bus is added, the probability of driving falls to

$$(7.2) \quad \tilde{P}_1 = \int_{-\infty}^{\bar{U}_{12}/\sqrt{2}} \int_{-\infty}^{\bar{U}_{13}/\sqrt{2}} b(\eta_{21}, \eta_{31}; 1/2) d\eta_{12} d\eta_{13}$$

where $\bar{U}_{12} = \bar{U}_{13}$ since all characteristics are the same. While the exact relationship between \tilde{P}_1 and P_1 depends on \bar{U}_{12} and follows no simple formula, as does the logit formula, the result is still counterintuitive. The independent probit specification would not be appropriate in such situations.

On the other hand, the covariance probit specification offers a solution to the problem. Originally, the probability of driving is

$$(7.3) \quad P_1 = \int_{-\infty}^{\bar{U}_{12}/(2+\Sigma\sigma_i^2 X_{12}^2)^{1/2}} \phi(\eta_{21}) d\eta_{21}.$$

With the addition of an identical alternative the probability is

$$(7.4) \quad \tilde{P}_1 = \int_{-\infty}^{\bar{U}_{12}/(2+\Sigma\sigma_i^2 X_{12}^2)^{1/2}} \int_{-\infty}^{\bar{U}_{12}/(2+\Sigma\sigma_i^2 X_{12}^2)^{1/2}} b(\eta_{21}, \eta_{31}; 1) d\eta_{21} d\eta_{31}$$

The correlation is unity since in equation (3.8) $\omega_{1,11} = \omega_{1,12} = \omega_{1,22}$ if the natural assumption is made that the unobserved characteristics are the same for the red and blue buses and have covariance equal to their respective variances. Because the limits of integration in equations (7.3) and (7.4) are identical, $P_1 = \tilde{P}_1$. The second integration is equivalent to the first integration along the line $\eta_{21} = \eta_{31}$. Thus, the covariance specification completely reproduces what our intuition desires. The probability of

driving remains the same while $\tilde{P}_2 = \tilde{P}_3 = 1/2 P_2$.¹ Thus the model specification seems satisfactory in the "red-bus blue-bus problem".

1. Actually, the probabilities \tilde{P}_2 and \tilde{P}_3 are degenerate in one dimension, so we adopt the convention of setting them equal.

References

1. Berndt, E.K., B.H. Hall, R.E. Hall, and J.A. Hausman, "Estimation and Inference in Nonlinear Structural Models," Annals of Economic and Social Measurement, March 1974, pp. 653-665.
2. Block, H.D. and J. Marschak, "Random Orderings and Stochastic Theories of Responses," in I. Olkin, et. al. (eds.), Contributions to Probability and Statistics, Stanford, pp. 97-132, 1960.
3. Bock, R.D. and L.V. Jones, The Measurement and Prediction of Judgement and Choice, Holden-Day, San Francisco, 1968.
4. Cambridge Systematics, Inc., The Effect of Transit Service on Automobile Ownership, U.S. D.O.T. Contract DOT-OS-30056, 1974.
5. Cox, D., The Analysis of Binary Data, Methuen, London, 1970.
6. Cox, D. and D.V. Hinckley, Theoretical Statistics, London, 1974.
7. Johnson, N. and S. Kotz, Continuous Univariate Distributions - 1, Houghton Mifflin, Boston, 1970.
8. Koppelman, F. and T. Watanatada, "Disaggregate Three-Mode Choice Model for Aggregate Forecast Testing," Unpublished Mimeograph, MIT, 1975.
9. Luce, R.D. and P. Suppes, "Preferences, Utility and Subjective Probability," in Luce et. al. (eds.), Handbook of Mathematical Psychology, III, Wiley, New York, pp. 249-441, 1965.
10. Manski, C., "The Structure of Random Utility Models," School of Urban and Public Affairs, Carnegie-Mellon University, Working Paper, 1975.

11. Marschak, J., "Binary Choice Constraints on Random Utility Indicators,"
in K. Arrow (ed.), Stanford Symposium on Mathematical Methods in
the Social Sciences, Stanford, 1960.
12. McFadden, D., "Conditional Logit Analysis of Qualitative Choice Behavior,"
in P. Zarembka (ed.), Frontiers in Econometrics, Academic Press,
New York, 1973.
13. McFadden, D., "The Measurement of Urban Travel Demand," Journal of Public
Economics, 26, 1974.
14. McFadden, D., "The Revealed Preferences of a Government Bureaucracy:
Theory," Bell Journal, Autumn 1975.
15. McFadden, D., "The Revealed Preferences of a Government Bureaucracy:
Evidence," Bell Journal, Spring 1976.
16. Owen, D., "Tables for Computing Bivariate Normal Probabilities," Annals
of Mathematical Statistics, 27, 1956.

Date Due

3/29/78

Lib-26-67

JY01 '88

NOV 24 1988

JAN 10 1989
DEC 9

JUN 10 1995

JUN 11 1995

JUN 15 1995

DEC 9 2004

HB31.M415 no.169
Fisher, Frankl/Quantity constraints, s
726267 D*BKS 00019890



3 9080 000 646 353

T-J5 E19 w no.170
Temin, Peter. /Lessons for the present
726484 D*BKS 00023136



3 9080 000 689 445

= 51

HB31.M415 no. 171
Joskow, Paul L/Regulatory activities b
727644 D*BKS 00048046



3 9080 000 990 058

HB31.M415 no. 172
Bhagwati, Jagd/Optimal policy interven
727641 D*BKS 00048047



3 9080 000 990 074

HB31.M415 no.173
Hausman, Jerry/A conditional probit mo
727640 D*BKS 00048049



3 9080 000 990 108

HB31.M415 no. 174
Carlton, Denni/Vertical integration in
727637 D*BKS 00048050



3 9080 000 990 124

HB31.M415 no. 175
Fisher, Frankl/On donor sovereignty an
727636 D*BKS 00048056



3 9080 000 990 231

