

Probabilistic Logit Models

By Alfredo A Romero

Draft: November, 2008

1 Introduction

In the traditional binomial and multinomial models, researchers have to specify at least two assumptions before estimating the model. The first assumption involves the probability distribution of the error term. Such an assumption would determine the functional form of the so-called link function, that is, the function that computes the probability of selecting a particular choice by an individual. The second assumption involves assuming a functional form for the argument of the link function, that is, how the explanatory variables will enter the link function. Whether the explanatory variables in the argument enter linearly in the link function or not relies completely in the specification suggested by the theory or in the cunning of the researcher. None of the previous two assumptions, however, is the result of a careful and systematic study of the data. Additionally, there is no guideline on how to assess the statistical adequacy of the proposed model afar from standard likelihood ratio tests. In the present document, we introduce an alternative approach to the specification and estimation of binomial and multinomial models, the probabilistic reduction approach. Through this approach, we are able to show that the logit link function is a sufficient condition for the existence of a joint distribution of all the variables involved in the model, the dependent and independent variables, that allows the specification of the models as an orthogonal decomposition of systematic and unsystematic information. The approach provides a set of probabilistic conditions that would have to be satisfied

by the models in order to achieve the status of statistically adequate and thus the use of the estimation results in statistical inference. Furthermore, this set of probabilistic conditions is testable through a Neyman-Pearson framework using the observed data and not the unobserved error. But the most important result is that the approach provides the modeler with a functional form for the argument of the link function that is warranted by the probabilistic structure of the data. This specification follows naturally from the definition of a proper joint distribution, the probabilistic reduction of the observed data, and the study of the conditional distribution of the conditioning set.

The rest of this document is organized as follows: section 2 presents the binomial and multinomial models from the traditional error specification approach, section 3 introduces the probabilistic reduction approach and derives the binomial and multinomial models, section 4 assesses the performance of both methodologies via simulation analysis while section 5 present a real data application. Finally, section 6 summarizes this paper.

2 Error Specification Approach

The error specification approach for the estimation of binary and multinomial models has firmly solidified in economics for its use in random utility models. The specification is achieved by imposing a latent variable structure in the proposed models. This is warranted under the belief that whereas individual utility cannot be observed, the choices that individuals make can be observed. From this perspective, once a choice has been made, it is possible to assume that an unobserved threshold of the utility rendered by the argument of the utility function has been crossed. Probability theory is then used to establish the likelihood of the decision crossing that threshold. While this threshold is also unobservable, it is still possible to model it using the latent

variables approach.

Maddala (1983) use Y_i^* to denote this latent continuous unobserved random variable, assuming that Y_i^* can be described by the following regression function:

$$Y_i^* = \beta' \mathbf{x}_i + u_i$$

In the usual interpretation of this kind of regression models, the response (yet unobserved) variable Y_i^* is a function of a set of explanatory variables \mathbf{x}_i with some probabilistic error attached to the functional relationship, u_i . Usually, this error term will be assumed to be independent and identically distributed with the additional condition of zero expected value, i.e., $E(u_i) = 0$. The interpretation in economics of the previous regression equation is that the functional relationship represents the difference in utility between two alternative outcomes for a consumer having to make a particular decision.

It is possible to generalize the previous case to more than one decision. In this respect, the response variable is not limited to only two values for it can be binary or polychotomous. For the case with more than two alternatives, the difference in utility represents the difference in utility with respect to a base (normalized) decision.

Under the premise that Y_i^* is unobserved, it is possible to observe whether the individual decided to consume a good or not with respect to the normalized decision. In the simplest case, when only one good is under consideration, the variable becomes a binary variable and illustrates whether a decision to consume was made or not. Thus, if consumption is observed, $Y_i = 1$, which implies that $Y_i^* > 0$; and $Y_i = 0$ otherwise. Because of the binary nature of the variable in this simplest case, a set of assumptions about the error allows the modeler to determine the frequency or probability of each outcome for a set of T responses. Thus, the frequency of the observed outcome is equal to,

$$\mathbf{P}(Y_i^* > 0) = \mathbf{P}(Y_i = 1) = \mathbf{P}(\beta' \mathbf{x}_i + u_i > 0) = \mathbf{P}(u_i > -\beta' \mathbf{x}_i) = 1 - G(-\beta' \mathbf{x}_i) = G(\beta' \mathbf{x}_i).$$

where the last equality arises from a symmetry assumption.

It is clear than to be able to estimate the preceding probabilities, at least two assumptions have to made. The first one regarding the distribution of the error term, which in turn would provide the functional form for the cumulative distribution function $G(\cdot)$. A second assumption, not less important, is the functional form of the argument of the cumulative density function. Notice that the preceding equation implies a linear combination of the \mathbf{x}' s that might not be granted by the data or the theory. It is possible then to rewrite the previous probability as $\mathbf{P}(Y_i^* > 0) = \mathbf{P}(Y_i = 1) = 1 - G(h(\mathbf{x}_i))$.

The selection of the functional form of the argument is not a trivial one. The researcher is confronted with the simultaneous selection of two assumptions: what is the functional form for $G(\cdot)$? which will be given by the probabilistic assumption regarding the error term, and what is the functional form for the argument of $h(\cdot)$? The researcher faces these two questions with no a priori probabilistic information provided. The indirect implication of these assumptions is that the particular probability distribution specified for the error term will determine the functional form of the regression equation.

The assumptions that the researcher makes for the error generates several specifications. In general, if the error term is assumed to be distributed extreme value, then $G(\cdot)$ is the logistic cumulative distribution function and the resulting model is the so-called logit binary model. Similarly, if the error term is assumed to be normally distributed, then the $G(\cdot)$ of the cumulative distribution function is the normal cumulative distribution function and the resulting model is the so-called probit binary model. Additional specifications can be obtained by imposing different distributional assumptions on the error, like the Gumbel Model, the Complementary Log-Log Model

(Greene, 2006), etcetera¹. The obvious caveat is that in order to get a specific functional form for the choice of probabilities, one has to make an assumption about the distribution of the stochastic term, about which we generally have no a priori knowledge (Cosslett, 1983).

In the more general case, where the individual is facing a set of J different choices, the dependent variable becomes a multinomial choice. In this kind of models, the individual is making a single decision amongst two or more alternatives which are unordered in nature. Notice that these unordered choice models can similarly be motivated by random utility models (Greene, 2006). In the usual approach, the i th consumer faces J choices. We can assume that the utility of choice j is given by the following function,

$$U_{i,j} = z'_{ij} + \epsilon_{ij}$$

Then, if we observe that the consumer made the choice j , we assume that U_{ij} is the maximum among the J utilities. Hence, the statistical model is driven by the probability that choice j is made, which is implicitly given by

$$P(U_{ij} > U_{ik}) \text{ for all other } k \neq j.$$

Suppose that there are m categories. Let P_1, P_2, \dots, P_m be the probabilities associated with these m categories. Then the idea is to express these probabilities in binary form. Let $\frac{P_1}{P_1 + P_m} = F(\beta_1^\top \mathbf{x})$, $\frac{P_2}{P_2 + P_m} = F(\beta_2^\top \mathbf{x})$, and $\frac{P_{m-1}}{P_{m-1} + P_m} = F(\beta_{m-1}^\top \mathbf{x})$. These imply that the odds ratio between the alternatives is nothing but a nonlinear combination of the cumulative distribution function of the set of explanatory variables \mathbf{x} , that is

$$\frac{P_j}{P_m} = \frac{F(\beta_j^\top \mathbf{x})}{1 - F(\beta_j^\top \mathbf{x})} = G(\beta_j^\top \mathbf{x}), \quad j = 1, 2, \dots, m - 1.$$

¹These last two models do not assume symmetry of the errors.

But since $\sum_{j=1}^{m-1} \frac{P_j}{P_m} = \frac{1 - P_m}{P_m} = \frac{1}{P_m} - 1$, we have that $P_m = \left[1 + \sum_{j=1}^{m-1} G(\beta_j^\top \mathbf{x}) \right]^{-1}$,

and hence, $P_j = \frac{G(\beta_j^\top \mathbf{x})}{1 + \sum_{j=1}^{m-1} G(\beta_j^\top \mathbf{x})}$ (Greene, 2006).

Clearly, the observations for the dependent variable \mathbf{Y}_t are arising from a multinomial distribution with the probabilities given by the previous two equations. Again, in line with the binary case, for the link function $G(\cdot)$, any proper *CDF* could be used. Additionally, there is no a priori guidance to decide whether the explanatory variables enter the model linearly or not, that is, there is no functional form for $h(\cdot)$ suggested by probability.

In a direct analogy with the binomial case, different assumptions about the error term will lead to different functional forms for the link function. From the computational point of view, however, the logistic link function has proved to be the easiest to handle. If this functional form is chosen, then $G(h(\beta_j^\top \mathbf{x}))$ is nothing but $\exp(h(\beta_j^\top \mathbf{x}))$.

Thus, we can rewrite the previous set of equations as,

$$P_j = \frac{e^{h(\beta_j^\top \mathbf{x})}}{D}, \quad j = 1, 2, \dots, m-1, \quad \text{and} \quad P_m = \frac{1}{D}$$

$$\text{where } D = 1 + \sum_{k=1}^{m-1} e^{h(\beta_k^\top \mathbf{x})}$$

For the multinomial logit specification, let \mathbf{Y}_i be a random variable that indicates the choice made. McFadden (1974) has shown that if the J disturbances are independent and identically distributed with Gumbel Type I Extreme Value distribution, that is, $F(\epsilon_{ij}) = \exp(-\exp(-\epsilon_{ij}))$, then

$$P(Y_i = j) = \frac{\exp(h(\mathbf{x}'_{ij}\theta))}{\sum_{j=1}^J \exp(h(\mathbf{x}'_{ij}\theta))}$$

Arguably, the odds ratio in the multinomial logit model are independent of the other alternatives. The property of the logit model whereby $\frac{P_{ij}}{P_{im}}$ is independent of the remaining probabilities is called the independence from irrelevant alternatives. The

independence assumption follows from the initial assumptions that the disturbances are independent and homoskedastic.

The multinomial models have had several extensions, especially regarding the condition of independence of irrelevant alternatives. Luce (1959) derived the multinomial logit model starting from the assumption of irrelevant alternatives. McFadden showed (1973) that a necessary and sufficient condition for the Luce model to hold the property of irrelevant alternatives is that the disturbances be independently and identically distributed with the extreme-value distribution. The primary drawback of the model is that when the assumption does not hold, the model predicts too high a joint probability distribution of selection for two alternatives that are in fact perceived as similar rather than independent by the individual.

This problem can be solved to some extent by assuming that the errors are multivariate normal. This assumption gives birth to the multinomial probit model MNP. This model is attributed to Thurstone (1972). Given that the residuals follow a multivariate normal distribution, computationally speaking, the multinomial probit model can be applied only for a small number of alternatives, because the computations involve evaluating multiple integrals. The advantage to the model is that the specification of the variance-covariance matrix of the errors allows the direct imposition of restrictions about the independence of the errors.

Additional, more appealing alternatives have been proposed. In this alternative models, the individual undertakes a series of choices or choice nodes that take care of the dependence of alternatives directly. One of this models is the the elimination by aspects model (EBA). The EBA model views choice as a covert sequential process. It is assumed that each alternative is described by a set of aspects, or characteristics, and that at each stage of the process an aspect is selected from the ones included in the available alternatives, with a probability that is proportional to its weight. This process eliminates all the alternatives that do not contain the selected aspect,

and the selection continues until a single alternative remains. Aspects that are common to all the alternatives under consideration do not affect the choice probabilities. Unfortunately, similar to the multinomial probit model, the EBA model becomes computationally infeasible for large choice sets. If there are n aspects, the total number of subsets one must consider is $2^n - 2$, and this can get very large very quickly. As another caveat, it does not have the latent-variable characterization that we have given to the multinomial logit and multinomial probit models (McFadden, 1982).

One final model considered is the hierarchical elimination by aspects model (Tversky and Sattah, 1979), HEBA. This model eliminates the problem of the choice set by assigning a hierarchical structure to the choices. The reduction is the result of assigning a tree structure to the EBA model. Under this framework, the weight of the probability is directly proportional to the link in the tree and eliminates all the alternatives that do not include that link. The same process is applied to each selected branch until one alternative remains.

Notice, however, that in all the preceding models, how we obtain the probabilities depends on the probabilistic structure assigned to the error term. Whether the choices made by the individual are simultaneous or sequential, it is this assumption about the error term what determines the functional form for the so-called link function. Similarly, notice that there is no guideline on how the independent variables of the conditioning set, the argument of the link function, should enter the model. A serious flaw is that if the assumption concerning the stochastic error term is wrong, not only in the binary case but also in the polychotomous case, the estimable model obtained will be misspecified. Unfortunately, the error term in equation is unobservable and its distribution unverifiable!

The problems of the error specification approach have just recently started being considered. In the current econometric literature, it is commonly believed that the choice of the distribution of the error is not transcendental in the binomial and

multinomial models. The usual solution to the problem of what specification form to select is simply to be ‘agnostic’ about how the covariates enter the regression equation (Johnston and Dinardo, 1997). The rationale behind this assumption is that most models seem to produce similar answers in most empirical applications. This similarity of the estimators has been considered evidence of robustness of the assumptions about the error term. The common way to approach the model selection problem is to “stick with what is convenient in a particular application, making certain that one’s inference does not depend unduly on the particular choices,” (Johnston and Dinardo, 1997). Furthermore, usual tests to discriminate between the models, like the χ^2 - test of twice the difference between the two log-likelihood functions, have very little power and in practice the difference is rarely large enough to discriminate between competing models.

3 Probabilistic Reduction Approach

In both the binomial and the multinomial models, two sets of assumptions, most of the time done separately, have to be made before estimating the models. The first one has to deal with the nature of the error term. The decision about the probability density of the error provides the researcher with the functional form for the link function. After this decision has taken place, the researcher has to make another assumption, this time regarding the functional form for the argument of the link function. Clearly, if one or both of the decisions are incorrect, any statistical inferences generated from the estimated model would be misleading. The Probabilistic Reduction (PR) Approach (see below) can be used to shed some light in terms of what functional form to select for the error and what functional form to select for the argument of the link function in a holistic manner. More importantly, the approach allows both assumptions to be

put to the test not by using the unobservable disturbances but rather the observable data.

A crucial first step is then to make probabilistic assumptions about the data. Consider the framework of Fahrmeir and Tutz (1994). They state that the binary logit and probit models are the expected value of the conditional distribution of Y_i given X_i . That is, they argue that the models are estimating $E(Y_i|\mathbf{X}_i = \mathbf{x}_i)$. Clearly, the functional form of the binary model being estimated is determined by the probability density function of the conditional distribution aforementioned, $f(Y_i|\mathbf{X}_i; \psi_1)$. Extending these results to the multinomial case, the multinomial regression model amounts to estimating the expected value of the conditional distribution of \mathbf{Y}_i given a set of explanatory variable \mathbf{X}_i , where \mathbf{Y}_i can take J different values, that is, $E(f(\mathbf{Y}_i|\mathbf{X}_i; \psi_1))$, where ψ_1 is a set of parameters, $\{\mathbf{Y}_1, \dots, \mathbf{Y}_N\}$ is a stochastic process where \mathbf{Y}_i is distributed multinomial and $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$, the conditioning set, is a stochastic vector process with joint density $f(\mathbf{X}_i; \psi_2)$ and $E(X_{k,i}^2) < \infty$ for $k = 1, \dots, K$.

Finding this conditional density is then a matter of finding the correct set of probabilistic conditions that reduces the joint distribution of all the variables, dependent and independent, involved in the model. De Finetti's representation theorem can be used as a formal way of reducing the joint distribution of all possible random variables involved into simplified products of distributions by imposing certain probabilistic assumptions. This decomposition provide us with a formal and intuitive mechanism for constructing statistical models, with the added benefit of identifying the underlying probabilistic assumptions of the statistical model being examined. Additionally, these probabilistic assumptions can be put to the test to determine their adequacy.

We are going to follow Spanos' (1986, 1999) definition of a statistical model and define it as a set of probabilistic assumptions that adequately capture the systematic information in the observed data in a parsimonious and efficient way. This definition is the first step of the PR Approach. In the approach, the construction of a statistical

model begins with the observed data. The observed data, $(\mathbf{y}_1, \dots, \mathbf{y}_N, \mathbf{x}_1, \dots, \mathbf{x}_N)' = (\mathbf{z}_1, \dots, \mathbf{z}_N)'$, can be considered one particular realization of the vector stochastic process, $\{\mathbf{Z}_i, i = 1, \dots, N\}$. Following Haavelmo (1944), the entire probabilistic information contained in the vector \mathbf{Z}_t is captured by the joint probability distribution of this process, the Haavelmo Distribution. In this fashion, we can represent this joint density by $f(\mathbf{Z}_1, \dots, \mathbf{Z}_N; \phi)$, for all $(\mathbf{z}_1, \dots, \mathbf{z}_N)' \in \mathbf{R}_Z^{(K+1) \times N}$.

A weaker version of De Finetti's representation theorem allows us to decompose or 'reduce' the Haavelmo distribution by imposing testable probabilistic assumptions from three broad categories:

(D) Distribution (M) Dependence (H) Heterogeneity

By imposing probabilistic assumptions from these broad categories, the modeler is essentially partitioning the space of all possible statistical models into a family of operational models. With these assumptions, we can derive a plethora of statistical models. As an illustration, consider for instance the derivation of the normal linear regression model starting from the Haavelmo distribution $f(\mathbf{Z}_1, \dots, \mathbf{Z}_N; \phi)$.

$$\begin{aligned} f(\mathbf{Z}_1, \dots, \mathbf{Z}_N; \phi) &\stackrel{I}{=} \prod_{i=1}^N f_i(\mathbf{Z}_i; \phi_1) \stackrel{ID}{=} \prod_{i=1}^N f(\mathbf{Z}_i; \phi) = \prod_{i=1}^N f(Y_i | \mathbf{X}_i; \psi_1) \cdot f(\mathbf{X}_i; \psi_2) \stackrel{N}{=} \\ &\stackrel{N}{=} \prod_{i=1}^N f(Y_i | \mathbf{X}_i; \psi_1), \end{aligned}$$

where we can ignore the marginal distribution of \mathbf{X}_i by the normality assumption.

Clearly, the imposition of the independent and identical distribution conditions provides the modeler with a method for defining a proper statistical model. The conditional distribution $f(Y_i | \mathbf{X}_i; \psi_1)$ allows the modeler to define a Statistical Generating Mechanism, SGM, which is viewed as an idealized representation of the true underlying data generating process. In a regression function framework, the SGM is given by $Y_i = E(Y_i | \mathbf{X}_i = \mathbf{x}_i) + u_i$, where $E(\cdot)$ represents the systematic component

of the data and u_i represents the unsystematic component of the data. Note that the functional form of $E(\cdot)$ is dependent upon the distribution assumptions plus the rest of the reduction assumptions. Additionally, the probabilistic structure of the error will be derived from the probabilistic structure of the data and not the other way around.

To obtain the probabilistic reduction specification for a multinomial regression model, we are going to start by recoding the polychotomous dependent variable. We can consider that the vector C_t represents whether a choice has been made at each $t \in T$. At each t , we can consider the vector of choices as $\{C_t; t = 1, \dots, T\}$. This stochastic vector C_t takes values $\{0, 1, 2, 3, \dots, J\}$, where J is the number of different choices; J is constant throughout $t \in T$ and the ordering of the choices is irrelevant. We can recode the variable C_t using $Y_{r,t}$ for $r = 1, \dots, R$. In this particular case,

$$Y_{r,i} = \begin{cases} 1 & \text{if } C_i = r \\ 0 & \text{otherwise} \end{cases}.$$

For every t , we have that $\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{R,t})'$. This indicator vector would take the value of zero at every R except for the one when the decision is made, the r^{th} position. In this fashion, if $C_t = r$, then $\mathbf{Y}_t = (0, \dots, 1, \dots, 0)' = \mathbf{e}_r$. Assuming that all the choices are independent, then the probability of $C_t = r$ and the probability of $Y_{r,t} = 1$ are identical. That is, $\mathbf{P}(C_t = r) = \mathbf{P}(Y_{r,t} = 1) = p_r$. The joint probability of the vector C_t (and/or the vindicator vector Y_t) is a multinomial distribution with $\mathbf{Y}_t \sim M(\mathbf{p}, 1)$ and $\mathbf{p} = (p_1, \dots, p_R)$.

In its general form, the multinomial distribution is given by (Fahrmeir and Tutz, 1994),

$$f(\mathbf{Y}_t; \mathbf{p}) = p_1^{Y_{1,t}} \cdot p_2^{Y_{2,t}} \cdot \dots \cdot p_R^{Y_{R,t}} \cdot \left(1 - \sum_{r=1}^R p_r\right)^{1 - \sum_{r=1}^R Y_{r,t}},$$

with $\sum_{r=1}^R p_r = 1$, and $\sum_{r=1}^R Y_{r,t} = 1$. Notice that when J is equal to one, then $f(Y_i; p) = p^{Y_i}(1-p)^{1-Y_i}$.

Under the assumption of independence, the multinomial distributed variable \mathbf{Y}_t will be completely characterized by the vector or probabilities \mathbf{p} . In this case, $E(\mathbf{Y}_t | \mathbf{X}_t = \mathbf{x}_t) = E(\mathbf{Y}_t) = \mathbf{p}$. In the more interesting case, however, the vector of probabilities would be heterogeneous with respect to \mathbf{X}_t in a specific form, that is, $\mathbf{p} = G(h(\mathbf{X}_t))$. Thus, it is a matter of obtaining a proper functional form for $h(\cdot)$, and in turn for $G(\cdot)$, what will capture the systematic information of the heterogeneous nature of \mathbf{p} through the variates \mathbf{X}_t . We can call this endeavor the “pass the bucket approach.” The purpose of modeling will be to find the functional form that captures the mean (and variance) heterogeneity of the multinomial variable \mathbf{Y}_t . The probabilistic reduction approach would prove key in the determination of the functional form of the conditional moments’ dependence of \mathbf{Y}_t on \mathbf{X}_t .

Let us consider the joint probability density of the vector stochastic process of $\{\mathbf{Z}_t := \{\mathbf{Y}_t, \mathbf{X}_t\}, t = 1, \dots, T\}$, that is $f(\mathbf{Y}_1, \dots, \mathbf{Y}_N, \mathbf{X}_1, \dots, \mathbf{X}_N; \phi)$ where ϕ is a relevant set of parameters. As we have argued above, by imposing several probabilistic conditions, it is possible to reduce the previous joint probability density (Haavelmo distribution) into an operational model to establish statistical relationships between the vector of choices and the conditioning set.

By imposing independence and identical distribution of the joint density, it is possible to decompose the Haavelmo distribution into a product of individual distributions,

$$f(\mathbf{Z}_1, \dots, \mathbf{Z}_N; \phi) \stackrel{I}{=} \prod_{t=1}^T f_t(\mathbf{Z}_t; \phi_t) \stackrel{ID}{=} \prod_{t=1}^T f(\mathbf{Z}_T; \phi),$$

Notice that in order to establish the statistical relationship between the vector \mathbf{Y}_t and the vector \mathbf{X}_t , it is necessary to take this decomposition one step further, that

is,

$$f(\mathbf{Z}_1, \dots, \mathbf{Z}_N; \phi) \stackrel{I}{=} \prod_{t=1}^T f_t(\mathbf{Z}_t; \phi_t) \stackrel{ID}{=} \prod_{t=1}^T f(\mathbf{Z}_t; \phi) = \prod_{t=1}^T f(\mathbf{Y}_t | \mathbf{X}_t; \psi_1) \cdot f(\mathbf{X}_t; \psi_2)$$

Clearly, the adequacy of a model that captures the heterogeneity in the conditional distribution of \mathbf{Y}_t given \mathbf{X}_t is granted by the existence of the marginal distribution of the conditioning set. It is important, however, that this decomposition is indeed warranted and it is the result of a proper joint density for \mathbf{Z}_t . If the joint density of \mathbf{Z}_t is indeed a proper probability density, then it will be the case that (Arnold et al. 1999),

$$f(\mathbf{Y}_t | \mathbf{X}_t; \psi_1) \cdot f(\mathbf{X}_t; \psi_2) = f(\mathbf{X}_t | \mathbf{Y}_t; \eta_1) \cdot f(\mathbf{Y}_t; \mathbf{p}) = f(\mathbf{Y}_t, \mathbf{X}_t; \phi)$$

where $f(\mathbf{Y}_t; \mathbf{p})$ is the unconditional multinomial distribution of \mathbf{Y}_t .

By looking at the odds ratio for a particular choice versus the base alternative, it is possible to establish the properness of the distribution. That is,

$$\frac{f(\mathbf{X}_t | \mathbf{Y}_t = \mathbf{e}_r; \eta_1)}{f(\mathbf{X}_t | \mathbf{Y}_t = \mathbf{0}; \eta_1)} \cdot \frac{f(\mathbf{Y}_t = \mathbf{e}_r; p)}{f(\mathbf{Y}_t = \mathbf{0}; p)} = \frac{f(\mathbf{Y}_t = \mathbf{e}_r | \mathbf{X}_t; \psi_1)}{f(\mathbf{Y}_t = \mathbf{0} | \mathbf{X}_t; \psi_1)} \cdot \frac{f(\mathbf{X}_t; \psi_2)}{f(\mathbf{X}_t; \psi_2)}$$

Recall from the “pass the bucket approach” that the vector or probabilities of \mathbf{Y}_t can be rewritten as a function of the random vector \mathbf{X}_t . Thus, we can rewrite the conditional distribution of \mathbf{Y}_t given \mathbf{X}_t using the joint distribution of a multinomially distributed random variable, that is,

$$f(\mathbf{Y}_t | \mathbf{X}_t; \psi_1) = \left(\prod_{r=1}^R g_r(\mathbf{X}_t; \psi_1)^{Y_{r,t}} \right) \left(1 - \sum_{r=1}^R g_r(\mathbf{X}_t; \psi_1) \right)^{1 - \sum_{r=1}^R Y_{r,t}}$$

The last, and of course, most important step is still finding the particular functional form for the link function $G(\cdot)$. We can rewrite the conditional distribution of \mathbf{Y} given \mathbf{X} in the previous equation to obtain,

$$\frac{f(\mathbf{X}_t|\mathbf{Y}_t = \mathbf{e}_r; \eta_1)}{f(\mathbf{X}_t|\mathbf{Y}_t = \mathbf{0}; \eta_1)} \cdot \frac{\pi_r}{\pi_0} = \frac{g_r(\mathbf{X}_t; \psi_1)}{1 - \sum_{r=1}^R g_r(\mathbf{X}_t; \psi_1)} \cdot \frac{f(\mathbf{X}_t; \psi_2)}{f(\mathbf{X}_t; \psi_2)},$$

where we are rewriting the unconditional densities of \mathbf{Y}_t as $\pi_r = f(\mathbf{Y}_t = \mathbf{e}_r; \mathbf{p})$ and $\pi_0 = f(\mathbf{Y}_t = \mathbf{0}; \mathbf{p})$. Also, note that the unconditional marginal density of \mathbf{X}_t will fall out of the equation. By doing this, we are ensuring that the joint distribution of \mathbf{Z}_t continues being a proper joint density.

Now it is possible to solve for $g_r(\mathbf{X}_i; \psi_1)$. Using the transformation suggested by Kay and Little (1987), $x = \exp\{\ln(x)\}$, we get,

$$g_r(\mathbf{X}_t; \psi_1) = \frac{\exp\{h_r(\mathbf{X}_t; \eta_{1,r})\}}{1 + \sum_{r=1}^R \exp\{h_r(\mathbf{X}_t; \eta_{1,r})\}},$$

where $h_r(\mathbf{X}_t; \eta_{1,r}) = \ln\left(\frac{f(\mathbf{X}_t|\mathbf{Y}_t = \mathbf{e}_r; \eta_1)}{f(\mathbf{X}_t|\mathbf{Y}_t = \mathbf{0}; \eta_1)}\right) + \kappa_r$, and $\kappa_r = \ln\left(\frac{\pi_r}{\pi_0}\right)$.

Note that the binomial model is a special case where $J = 1$. For the binomial model, the previous link function is,

$$g(\mathbf{X}_i; \psi_1) = \frac{\pi_1 \cdot f(\mathbf{X}_i|Y_i = 1; \eta_1)}{\pi_0 \cdot f(\mathbf{X}_i|Y_i = 0; \eta_1) + \pi_1 \cdot f(\mathbf{X}_i|Y_i = 1; \eta_1)}$$

Using the same transformation, $x = \exp\{\ln(x)\}$, and rearranging terms, we get,

$$g(\mathbf{X}_i; \psi_1) = \frac{\exp\{h(\mathbf{X}_i; \eta_1)\}}{1 + \exp\{h(\mathbf{X}_i; \eta_1)\}}$$

where $h(\mathbf{X}_i; \eta_1) = \ln\left(\frac{f(\mathbf{X}_i|Y_i = 1; \eta_1)}{f(\mathbf{X}_i|Y_i = 0; \eta_1)}\right) + \kappa$ and $\kappa = \ln(\pi_1) - \ln(\pi_0)$.

There are several things to be mentioned here. First, note that the composite function $g(\cdot)$, the so-called link function, represents the logistic cumulative density function for the index function $h(\cdot)$. Thus, the binomial logit and the multinomial logit arise naturally from the probabilistic structure of the data. We can then claim

that a logit link function is a sufficient condition for the appropriateness of the joint density of all the variables involved in the model. Note however, that the probabilistic structure of the conditioning set will provide us with the correct functional form for the argument of the link function (see below).

This point to notice is perhaps the most crucial difference with the traditional logit specification. In the logit model, the terms in the $h(\cdot)$ equation might enter linearly with no guidance provided by the probabilistic structure of the data whereas in the present model, the Probabilistic Logit Model, P-Logit, the functional form for the argument of $G(\cdot)$ will be determined by the conditional distribution of \mathbf{X}_i given \mathbf{Y}_i , with the obvious result that the conditional distribution might not necessarily lead to linear specifications if the joint conditional distribution is not part of the symmetric elliptical family of distributions (see below and Fang et al., 1989).

Consider the binomial case, we can now decompose the previous reduction of the stochastic process to obtain an operational model,

$$Y_i = E(Y_i|\mathbf{X}_i = x_i) + u_i = g(\mathbf{x}_i; \psi_1) + u_i = \frac{\exp\{h(\mathbf{X}_i; \eta_1)\}}{1 + \exp\{h(\mathbf{X}_i; \eta_1)\}} + u_i.$$

In the multinomial case, we would have,

$$\mathbf{Y}_i = E(\mathbf{Y}_i|\mathbf{X}_i = x_i) = g_r(\mathbf{x}_i; \psi_1) = \frac{\exp\{h_r(\mathbf{x}_i; \eta_{i,r})\}}{1 + \sum_{r=1}^R \exp\{h_r(\mathbf{x}_i; \eta_{i,r})\}}$$

Since the functional forms of both $g(\cdot)$ and $h(\cdot)$ are simultaneously dependent upon the functional form of $f(\mathbf{X}_i|\mathbf{Y}_i; \eta_1)$ and in turn the joint distribution of \mathbf{Y}_i and \mathbf{X}_i , thus, a crucial in-between step in formulating the model would be the assessment of the conditional distributions of the vector \mathbf{X} .

For this aim, several conditional distributions are provided in the statistics literature (see Arnold, Castillo, and Sarabia, 1999). Kay and Little provides us with

several conditional probability distributions that give birth to functional forms that warrant linearity in the specification. For instance, if the conditional distribution of the vector \mathbf{X}_i is a multivariate normal distribution with homogeneous covariance matrix, then,

$$g(\mathbf{x}_i; \psi_1) = \frac{\exp\left(-\beta_0 - \sum_{k=1}^K \beta_k x_{k,i}\right)}{1 + \exp\left(-\beta_0 - \sum_{k=1}^K \beta_k x_{k,i}\right)}$$

Notice that in this particular case, the regression model will be,

$$Y_i = E(Y_i | \mathbf{X}_i = x_i) + u_i = g(\mathbf{x}_i; \psi_1) + u_i = \frac{\exp\left(-\beta_0 - \sum_{k=1}^K \beta_k x_{k,i}\right)}{1 + \exp\left(-\beta_0 - \sum_{k=1}^K \beta_k x_{k,i}\right)} + u_i$$

In practice, however, we can rarely assume joint conditional normality of the explanatory variables. A solution to the problem can be obtained following Arnold et al. (1999), achieved by decomposing the conditioning set into a product of simpler conditional density functions. Of course, in order to obtain the right partitioning the condition of independence in the sub-densities would have to hold.

As an illustration, let us consider that case where the variables in the conditioning set are independent from each other conditional on Y_i . In this case, we would have, $f(\mathbf{X}_i, \eta_{1,j}) = \prod_{k=1}^K f(X_{k,i}; \eta_{1,k,j})$, so that the index function $h(\cdot)$ becomes
$$h(\mathbf{x}_i, \eta_1) = \sum_{k=1}^K \ln \left(\frac{f(X_{k,i}; \eta_{1,k,1})}{f(X_{k,i}; \eta_{1,k,0})} \right) + \kappa.$$
 If it is the case that some or none of the explanatory variables are independent conditional on Y_i , then the use of sequential conditioning can be of help.

Thus, contingent upon the distribution of the conditioning set with respect to \mathbf{Y}_i , we can specify the family of multinomial logit models as follows,

SGM:	$\mathbf{Y}_{i,r} = g_r(\mathbf{x}_i; \psi_1) + u_i, i = 1, \dots, T$
Multinomial	$(\mathbf{Y}_{i,r} \mathbf{X}_i = \mathbf{x}_i) \sim M(g(\mathbf{X}_i; \psi_1), 1)$
Non-Linearity	$E(\mathbf{Y}_{i,r} \mathbf{X}_i = \mathbf{x}_i) = g_r(\mathbf{x}_i; \psi_1) = \frac{\exp\{h_r(\mathbf{x}_i; \eta_{i,r})\}}{1 + \sum_{r=1}^R \exp\{h_r(\mathbf{x}_i; \eta_{i,r})\}}$
	where $h_r(\mathbf{x}_i; \eta_{i,r}) = \ln \left[\frac{f(\mathbf{X}_i Y_i = \mathbf{e}_r; \eta_1)}{f(\mathbf{X}_i Y_i = 0; \eta_1)} \right] + \kappa_r$ and $\psi_1 = G(\eta_1)$
Heteroskedasticity	$Var(\mathbf{Y}_i \mathbf{X}_i = \mathbf{x}_i) = diag(g(\mathbf{x}_i; \psi_1) - g(\mathbf{x}_i; \psi_1) g(\mathbf{x}_i; \psi_1)'),$
	where $g(\mathbf{x}_i; \psi_1) = (g_1(\mathbf{x}_i; \psi_1), \dots, g_R(\mathbf{x}_i; \psi_1))$
Homogeneity	$\psi_1 = G(\eta_1)$ is not a function of $i = 1, \dots, N$.
Independence	$\{\mathbf{Y}_{i,r} \mathbf{X}_i = \mathbf{x}_i, i = 1, \dots, N\}$ is an independent process.

4 Empirical Analysis: Simulation Based

In order to compare the performance of the models, we created three different simulations assuming different probability distributions of the conditional joint distribution of the explanatory variables: A normally distributed joint conditional distributed data, a gamma distributed joint conditional distributed data, and a normal-beta joint conditional distributed data. The only relevant model for comparison purposes would be the logit function with a linear specification of the explanatory variables. To make the comparisons meaningful, we also compared the marginal probabilities obtained from the average individual in the sample under the P-Logit model and the traditional Logit model.

4.1 Normally Distributed Data

For the normally distributed data, we created a sample of N observations from a bivariate conditional normal distribution, letting as many values for $Y_t = 1$ as needed to achieve a pre-specified p parameter, representing the unconditional mean of Y_t .

Consider the process $\{X_i, i = 1, \dots, N\}$, where $X_i \sim N(\mu, \sigma^2)$. Then, the conditional distribution of X_t given Y_t is given by,

$$f(X_i|Y_i; \eta_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (X_i - \alpha_0 - \alpha_1 Y_i^2) \right\}.$$

Substituting into $h(x_i; \eta_1)$ we get, $h(x_i; \eta_1) = -\frac{(\alpha_1^2 + 2\alpha_0\alpha_1)}{2\sigma^2} + \frac{\alpha_1}{\sigma^2}x_i + \kappa$, and thus

$$g(\mathbf{X}_i; \psi_1) = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}},$$

where $\beta_1 = \frac{\alpha_1}{\sigma^2}$, $\beta_0 = -\frac{(\alpha_1^2 + 2\alpha_0\alpha_1)}{2\sigma^2} + \frac{\alpha_1}{\sigma^2}x_i + \kappa$, and $\kappa = \ln\left(\frac{\pi_1}{\pi_0}\right)$.

We estimated different sample sizes, from 50 to a 1000 observations each, using the following pre-specified parameters for the simulation, $p = 0.6$, $\alpha_0 = 2$, $\alpha_1 = 1$, and $\sigma^2 = 1$.

	P-L Mean	Z-stat	Logit Mean	Z-stat
True Value	-2.095			
T=50	-2.3032	-2.2217	-2.3032	-2.2217
T=100	-2.1988	-3.1938	-2.1988	-3.1938
T=500	-2.1132	-7.1874	-2.1132	-7.1874
T=1000	-2.1013	-10.1652	-2.1013	-10.1652
True Value	1.000			
T=50	1.0913	2.7354	1.0913	2.7354
T=100	1.0458	3.9396	1.0458	3.9396
T=500	1.0082	8.8960	1.0082	8.8960
T=1000	1.0031	12.5921	1.0031	12.5921

Clearly, the estimates are identical in both models since the logit and the p-logit models both have their explanatory variables entering linearly in the specification of the logistic function. Obviously, the marginal probabilities obtained from the average individual would have to be also identical. For additional comparison and as a robustness check, we also added the marginal effect from the average individual using the probit model.

	P-L Model	Logit Model	Probit Model
True Value P=0.235			
T=50	0.2540	0.2540	0.2471
T=500	0.2364	0.2364	0.2293
T=1000	0.2353	0.2353	0.2283

4.2 Gamma Distributed Data

For the gamma distributed data, we created a sample of N observations from a heterogeneous gamma distribution, letting the change in the heterogeneity to be a function of the values of Y_t . We let as many values for $Y_t = 1$ as needed to achieve a pre-specified p parameter, representing the unconditional mean of Y_t . In this case, the simulation is specified below.

Consider the process $\{X_i, i = 1, \dots, N\}$, where $X_i \sim \Gamma(\gamma, \alpha)$. Then, the conditional distribution of X_t given Y_t is given by,

$$f(X_i|Y_i; \eta_1) = \frac{1}{\gamma_j \Gamma(\alpha_j)} \left(\frac{X_i}{\gamma_j} \right)^{\alpha_j - 1} \exp \left\{ -\frac{X_i}{\gamma_j} \right\}.$$

Notice that the source of the heterogeneity will be directly related to the shape parameter of the gamma distribution.

Substituting into $h(x_i; \eta_1)$ we get, $h(x_i; \eta_1) = -\ln \frac{\frac{1}{\gamma_1 \Gamma(\alpha_1)} \left(\frac{X_i}{\gamma_1}\right)^{\alpha_1-1} \exp\left\{-\frac{X_i}{\gamma_1}\right\}}{\frac{1}{\gamma_0 \Gamma(\alpha_0)} \left(\frac{X_i}{\gamma_0}\right)^{\alpha_0-1} \exp\left\{-\frac{X_i}{\gamma_0}\right\}} + \kappa,$

and thus

$$g(\mathbf{X}_i; \psi_1) = \frac{\exp\{\beta_0 + \beta_1 x_i + \beta_2 \ln(x_i)\}}{1 + \exp\{\beta_0 + \beta_1 x_i + \beta_2 \ln(x_i)\}},$$

where $\beta_2 = \alpha_1 - \alpha_0$, $\beta_1 = \frac{1}{\gamma_2} - \frac{1}{\gamma_1}$,

$\beta_0 = \ln \frac{1}{\gamma_1} \Gamma(\alpha_1) - \alpha_1 \ln(\gamma_1) + \ln(\gamma_1) - \ln \frac{1}{\gamma_2} \Gamma(\alpha_2) - \ln \gamma_2 + \alpha_2 \ln \gamma_2 + \kappa$, and $\kappa = \ln \left(\frac{\pi_1}{\pi_0}\right)$.

Note that heterogeneous gamma distributed explanatory variables enables the use of linear terms in the specification but with additional sources of non-linearities in the variables. A one variable model needs the variable in levels and a logarithmic transformation of the variable to completely capture the systematic information in the data.

For the simulations, the implicit parametrization used is $p = 0.60$, $\beta_0 = 1.38$, $\beta_1 = -0.88$, and $\beta_2 = 2.5$. The results are given in the following table,

	P-L Mean	Z-stat	Logit Mean	Z-stat
True Value	1.38			
T=50	1.5334	2.1835*	.7446	1.6738
T=100	1.4955	3.1495**	.7575	2.4974*
T=500	1.3970	7.0031**	.7352	5.6517**
T=1000	1.3912	9.9224**	.7331	8.0022**
True Value	-0.8888			
T=50	-1.0480	-2.4726*	-.0867	-1.0062
T=100	-.9918	-3.4873**	-.0915	-1.5003
T=500	-.9050	-7.7655**	-0.0907	-3.4795**
T=1000	-0.8982	-10.9866**	-.0906	-4.9416**
True Value	2.500			
T=50	2.9422	2.5151*		
T=100	2.7548	3.5575**		
T=500	2.5447	7.9597**		
T=1000	2.5257	11.2532**		

Clearly, both models produce very different results. One of the most prominent differences is that a researcher using the traditional logit model would find no evidence of a statistical relationship between X_t and Y_t at the sample sizes $T = 50$ and $T = 100$. Thus, the addition of the logarithm of the variables has improved the efficiency of the estimators. Note also that the estimated coefficient using the traditional logit does not converge to the true parameter even in relatively large sample. This estimator suffers from bias due to the the omitted variables problem.

It could be the case, however, that although biased, the estimators obtained via the traditional logit still produce relatively close marginal effects. To find this out, we computed the marginal effects obtained from the traditional logit model, the probit model, and the P-logit model for the average individual. The results are presented in the following table,

	P-L Model	Logit Model	Probit Model
True Value P=-0.03			
T=50	-.0344	-.0207	-.0209
T=500	-0.0314	-0.0218	-0.0217
T=1000	-0.0312	-.0217	-0.0217

There are several issues to mention. One, the traditional logit model overestimates the true marginal effect for the average individual even in relatively large samples. Second, the claim that the traditional models seem to produce similar answers in most empirical applications holds true, both the traditional logit and the probit produced very close marginal probabilities, and they both are off by almost 30 percent! It is then possible to argue that the bias of the estimators is transferred to the computation of the marginal probabilities even for large samples.

4.3 Normal-Beta Distributed Data

For the third simulation we created a set of three variables with a more complicated probabilistic structure. The set of three variables, X_{1t} , X_{2t} and X_{3t} have the following statistical properties: X_{2t} and X_{3t} are jointly conditionally distributed, given Y_t , but independently distributed from X_{3t} . Their joint conditional distribution is bivariate normal with a homogeneous variance-covariance matrix. Additionally, X_{3t} is heterogeneously distributed beta. The source of the heterogeneity for the beta distributed variable will be in both the location parameter and the shape parameter and would be directly related to Y_t . We let as many values for $Y_t = 1$ as needed to achieve a pre-specified p parameter, representing the unconditional mean of Y_t , and created

a sample of N observations. In this case, the simulation was specified as follows,

$$f(X_{1,i}|X_{2,i}, Y_i; \eta_1) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_1} (X_{1,i} - \alpha_0 - \alpha_1 X_{2,i} - \alpha_2 Y_i)^2\right\},$$

$$f(X_{2,i}|Y_i; \eta_1) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_2} (X_{2,i} - \gamma_0 - \gamma_1 Y_i)^2\right\},$$

$$f(X_{3,i}|Y_i; \eta_1) = \frac{X_{3,i}^{\delta_j-1} (1 - X_{3,i})^{\eta_j-1}}{B[\delta_j, \eta_j]},$$

Substituting into $h(x_i; \eta_1)$ and then into $g(\cdot)$ we get

$$g(\mathbf{X}_i; \psi_1) = \frac{\exp\{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 \ln(x_{3,i}) + \beta_4 \ln(1 - x_{3,i})\}}{1 + \exp\{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 \ln(x_{3,i}) + \beta_4 \ln(1 - x_{3,i})\}},$$

where $\beta_4 = \eta_1 - \eta_0$, $\beta_3 = \delta_1 - \delta_0$, $\beta_2 = \frac{\gamma_1}{\sigma_2^2} - \frac{\alpha_1 \alpha_2}{\sigma_1^2}$, $\beta_1 = \frac{\alpha_2}{\sigma_2^2}$,

$$\beta_0 = \ln\left(\frac{B[\delta_0, \delta_1]}{B[\eta_1 - \eta_0]}\right) - \frac{(2\alpha_0 \alpha_2 + \alpha_2)^2}{2\sigma_1^2} - \frac{2\gamma_0 \gamma_1 - \gamma_1^2}{2\sigma_2^2} + \kappa, \text{ and } \kappa = \ln\left(\frac{\pi_1}{\pi_0}\right).$$

Note at this point the richer specification obtained. The normally distributed variables enter linearly in the specification, as we would expect, but the heterogeneous beta distributed variable enters the specification with two transformations, the first one being the logarithm of the value taken by the variable at time t and the second being the logarithm of the results of one minus the value taken by the variable at time t . The traditional logit specification would miss both transformations.

The implicit parametrization selected for this simulation are $\beta_0 = -5.5$, $\beta_1 = 0.7$, $\beta_2 = 1$, $\beta_3 = -3$, and $\beta_4 = 1$. The results are given by the following table,

		P-L Mean	Z-stat	Logit Mean	Z-stat
β_0	True Value	-5.5000			
	T=50	-9.1009	-1.4193	-0.7652	-0.1217
	T=100	-6.0950	-1.9414*	0.8250	0.3385
	T=500	-5.6348	-4.6140***	0.6245	0.6615
	T=1000	-5.5734	-6.5280***	0.6281	0.9555
β_1	True Value	.7000			
	T=50	1.0982	1.5000	0.9721	1.5927
	T=100	0.7998	2.3118**	0.7860	2.3410**
	T=500	0.7241	5.4501***	0.7222	5.4836***
	T=1000	0.7153	7.7348***	0.7144	7.7768***
β_2	True Value	1.0000			
	T=50	1.6430	1.9114*	1.4960	1.9664**
	T=100	1.1424	2.7043***	1.1128	2.7316***
	T=500	1.0239	6.3416***	1.0212	6.3818***
	T=1000	1.0118	9.0010***	1.0105	9.0524***
β_3	True Value	-3.0000			
	T=50	-2.8407	-0.5960	-11.9537	-2.8396***
	T=100	-3.4776	-1.1757	-10.2613	-4.1868***
	T=500	-3.1110	-2.8563***	-9.2913	-9.7498***
	T=1000	-3.0509	-4.0430***	-9.2078	-13.8599***
β_4	True Value	1.0000			
	T=50	2.0974	1.1036		
	T=100	1.2557	1.1767		
	T=500	1.0344	2.6569***		
	T=1000	1.0218	3.7803***		

Not surprisingly, the normally distributed variables, both in the P-logit and in the traditional logit model converge relatively quickly to the true parameter and have similar estimated values. The discrepancy lies on the estimated coefficients for X_{3t} . Notice how in the traditional logit specification the bias with respect to the true parameter increases with the sample size along with the precision of the estimator! To

corroborate the claim that the bias is transferred to the computation of the marginal probabilities we computed the marginal effect for the average individual with respect to X_{3t} . The results are presented in the following table,

	P-L Model	Logit Model	Probit Model
True Value P=-1.35			
T=50	-1.5048	-1.6195	-1.7062
T=500	-1.4014	-1.6943	-1.6949
T=1000	-1.3915	-1.6904	-1.6883

Again, notice that the traditional logit and the probit models produce similar marginal probabilities, as expected, that differ systematically from the true probabilities even for relatively large samples. For this particular case, the marginal probabilities for the average individual with respect to X_{3t} are off by almost 25 percent for the logit and the probit versus a mere 3 percent for the P-logit model.

5 Empirical Analysis: Real Data

5.1 Data and Specification

For an empirical application, we used data on travel choices made by 210 individuals originally gathered by Greene and Hensher (1995) and closely studied in Greene (2006). The data is used for the estimation of a model of travel mode choice for travel between Sydney and Melbourne, Australia. The data contains 840 observations on choice among four alternatives, AIR, TRAIN, BUS, and CAR. The attributes used for their example were choice-specific constants and two continuous measures: GC, a measure of the generalized cost of the travel, constructed from measures of in-vehicle

cost, and a wage-like measure times the amount of time spent traveling; TTME, the travel time; and HINC, household income. Although they estimated a nested logit, the first two branches of their model specifies a logit model for the decision of traveling by AIR versus traveling by GROUND.

The specification of the model is then given by,

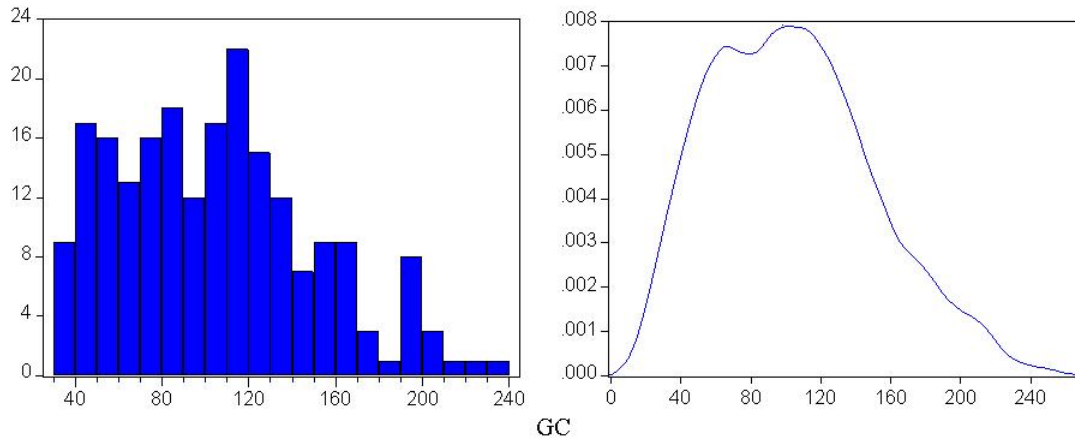
$$Y_i = E(Y_i|\mathbf{X}_i = x_i) + u_i = g(\mathbf{x}_i; \psi_1) + u_i$$

$$\text{where } Y_i = \begin{cases} 1 & \text{if } Y_i = \textit{Air} \\ 0 & \text{if } Y_i = \textit{Ground} \end{cases}, \text{ and } \mathbf{X}_i = \{GC_i, TTME_i, HINC_i\}.$$

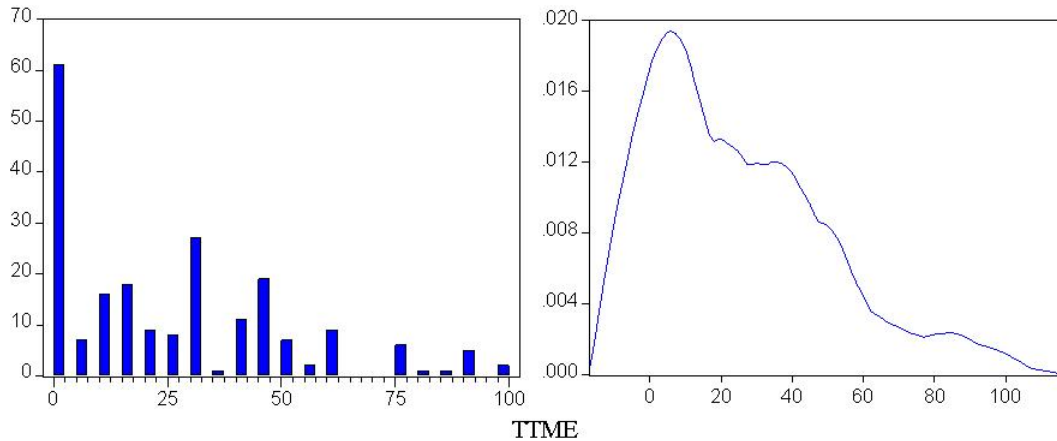
We will estimate the model from the traditional logit specification testing for additional sources of non-linearities and re-specifying accordingly. This way, we can bypass the assessment and estimation of the joint conditional distribution of the the conditioning set and still estimate a statistically adequate model that captures the systematic information of the data in a satisfactory way. The use of histograms and empirical distribution tests, however, would prove extremely useful at suggesting the appropriate form of non-linearities to include in the specification.

5.2 Analysis of the Conditioning Set

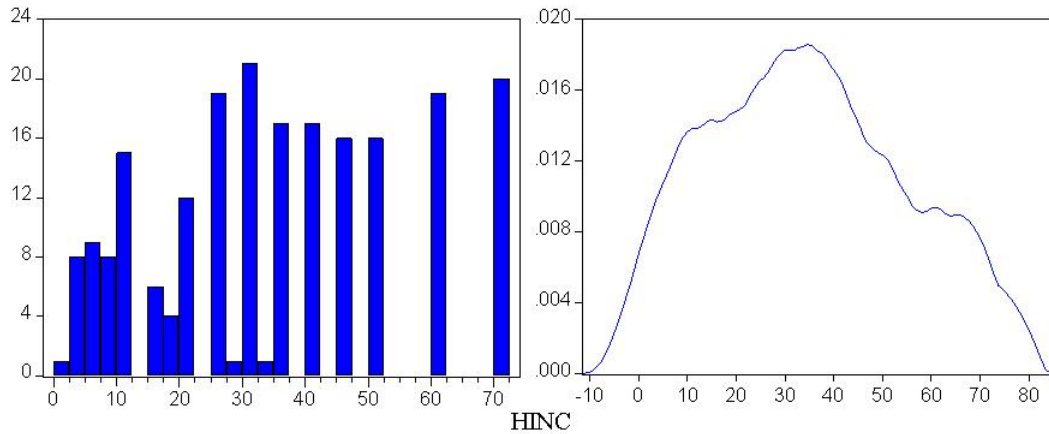
The following figures show the histograms and the kernel density graphs (using Epanechnikov kernels) for the three explanatory variables. Several specifications were proposed per variable after conducting empirical distribution tests.



Notice that the histogram and density graphs for GC weakly resemble a normally distributed variable. Several empirical distribution tests suggested additionally the Logistic, the Lognormal, and the Gamma distributions as potential candidates. Consequently, we decided to test the following non-linear transformations for GC : GC^2 , $\ln(GC)$, and $\ln(GC)^2$.



With respect to $TTME$, we also tested several empirical distributions. The results were similar to those of the variable GC , so we proposed the testing of the same non-linear transformations.



The distribution of the variable $HINC$ proved more difficult to asses. The empirical distribution tests were unable to discriminate between the distributions observed in the previous two variables plus the χ^2 , the Uniform, and the Weibull distributions. Thus, we tested the same set of non-linear transformations for $HINC$ plus $HINC^\alpha$, where α is an additional parameter to be estimated.

5.3 Empirical Results

5.3.1 Naive Logit Specification

We started the analysis using a simple naive logit specification where the explanatory variables enter the $h(\cdot)$ function linearly. The results are presented in the following table,

Variable	Coefficient	S.E.	z-statistic	p-value
C	-3.9686	0.7032	-5.6432	0.0000
GC	-0.0043	0.0047	-0.9139	0.3607
TTME	0.0660	0.0103	6.3821	0.0000
HINC	0.0416	0.0104	4.0004	0.0001

Notice that under the naive logit specification, the estimator for GC is non-statistically different from zero. We evaluated the predictive ability of the model using a naive expectation-prediction table choosing 0.50 as the cutoff point. The results are presented in the following table,

Expectation-Prediction		
Y=0	Y=1	Total Y
93.42	62.07	84.76

A researcher may be tempted to drop the variable GC from the model for lack of statistical significance. The results from that Modified Naive Logit Specification are presented here for comparison,

Variable	Coefficient	S.E.	z-statistic	p-value
C	-4.3241	.6017	-7.1855	.0000
TTME	.0630	.0096	6.5547	.0000
HINC	.0409	.0103	3.9530	.0001

The predictive ability of the model is analogously given by,

Expectation-Prediction		
Y=0	Y=1	Total Y
92.11	60.33	83.33

Notice that dropping the variable deteriorates the explanatory power of the model in every category.

5.3.2 P-Logit Specification

To obtain the statistically adequate model, we tested the naive logit specification for additional sources of heterogeneities using the non-linear transformations of all the variables involved. The process of specification, misspecification testing, and re-specification yielded the following model,

Variable	Coefficient	S.E.	z-statistic	p-value
C	-125.7032	35.3727	-3.5536	0.0004
$\ln(GC)$	52.8744	15.2329	3.4710	0.0005
$\ln(GC)^2$	-5.7731	1.6436	-3.5123	0.0004
$TTME$	0.1192	0.0279	4.2750	0.0000
$TTME^2$	-0.0006	0.0003	-2.2355	0.0254
$HINC$	0.0494	0.0116	4.2568	0.0000

Notice that after the variable GC enters non-linearly into the function $h(\cdot)$, its estimators become statistically different from zero at the usual significance levels. We were also able to incorporate additional sources of non-linearity in the $TTME$ variables. Note also that it is the study of the data, and not an a priori theoretical restriction, that has suggested the implicit testing of decreasing returns to scale in

$\ln(GC)$ and $TTME$, variables whose respective estimator for their squared value have a negative coefficient. Finally, note that the overall predictive power of the model has increased in practically every category: ten percent more accurate prediction with respect to $Y_t = 1$ and an overall increase in explanatory power of almost two percent.

Expectation-Prediction		
Y=0	Y=1	Total Y
92.76	70.69	86.67

5.3.3 Misspecification Testing

Before proposing the P-logit model as a model that adequately captures all the systematic information contained in the conditioning set, we conducted a final misspecification test for additional sources of non-linearities. The test takes the form,

$$y_t = \alpha_0 + \alpha_1 \hat{y}_t + \alpha_2 \hat{y}_t^2 + u_t$$

where \hat{y}_t^2 can be substituted by any other powered variable or non-linear transformation. Similarly, we can test for additional sources of heterogeneity in the mean process with the following specification,

$$y_t = \gamma_0 + \gamma_1 \hat{y}_t + \gamma_2 t + u_t$$

where in both of the previous cases, α_2 and γ_2 are expected to be equal to zero. We tested simultaneously additional sources of heterogeneity and additional non-linearities using the estimated residuals. The following table presents the results,

Variable	Coefficient	S.E.	t-statistic	p-value
C	0.0466	0.1372	.3400	0.7342
\widehat{MODE}	-0.2092	0.9282	-0.2253	0.8219
\widehat{MODE}^2	2.8148	2.2007	1.2790	0.2023
\widehat{MODE}^3	-1.7328	1.5188	-1.1409	0.2552
$\ln(\widehat{MODE})$	0.0127	0.0217	0.5851	0.5591
TREND	0.0005	0.0003	1.5196	0.1302

Clearly, all additional sources of heterogeneity and nonlinearity are not statistically different from zero. We conducted a F-test to test the assumptions that all variables are statistically not different from zero. We did not find any evidence against this hypothesis ($F = 1.1509$, $F_{5\%}(4, 204) = 2.4159$).

6 Conclusion

Unless certain probabilistic conditions are met, the use of the traditional logit specification with the argument of the link function in linear form might not be warranted by the data and might lead the researcher to invalid inferences.

The problem is that these conditions exist and are imposed in the specification whether the researcher is aware of them or not. An alternative approach to modeling binomial and multinomial models is the probabilistic reduction approach. The approach specifies the models in a holistic manner, taking into account not only the probabilistic structure of the conditioning set but also the set of probabilistic reductions in the observed data that would warrant an statistically adequate model. Under the approach, the researcher is not only aware of the set of conditions that have to be satisfied by the data but he or she is also provided with a testing framework for the conditions. Through a Neyman-Pearson testing framework, the modeler has control over the error probabilities if the inferences drawn from the models.

7 References

1. Arnold, B. C., E. Castillo and J. M. Sarabia (1999), *Conditional Specification of Statistical Models*, New York: Springer Verlag.
2. Cosslett, S.R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51.
3. Fahrmeir, L. and G. Tutz (1994), *Multivariate Statistical Modeling Based on Generalized Linear Models*, New York: Springer-Verlag.
4. Fang, Kai-Tai, Samuel Kotz and Kai Wang Ng (1989), *Symmetric Multivariate and Related Distributions*, Chapman & Hall/CRC.
5. Greene, William H. (2006), *Econometric Analysis*, Pearson Prentice Hall.
6. Greene, W. and D. Hensher (1997), *Multinomial Logit and Discrete Choice Models*, in Greene, W., *LIMPDEP Version 7.0 User's Manual*, Revised, Econometrics Software Inc., NY.
7. Haavelmo, T. (1994), "The Probability Approach to Econometrics," *Econometrica* 12.
8. Johnston, Jack and John DiNardo (1997), *Econometric Methods*, McGraw Hill.
9. Kay, R. and S. Little (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, 74.
10. Luce, R. D. (1959), *Individual Choice Behavior: A Theoretical Analysis*, New York: Wiley.
11. Maddala, G. S. (1985), *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press.
12. McFadden, D. (1973), *Conditional Logit Analysis of Qualitative Choice Behavior*, in P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic.
13. McFadden, D. (1974), "The measurement of Urban Travel Demand," *Journal of Public Economics* 3.
14. McFadden, D. (1982), *Econometric Models of Probabilistic Choice*, in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data: With Econometric Applications*, Cambridge, Mass.: M. I. T. Press.
15. Spanos, A. (1986), *Statistical Foundations of Econometrics Modeling*, Cambridge, UK: Cambridge University Press.
16. Spanos, A. (1999), *Probability Theory and Statistical Inference: Econometrics Modeling with Observational Data*, Cambridge UK: Cambridge University Press.

17. Thurstone, L. (1927), "A Law of Comparative Judgment," *Psychological Review* 34.
18. Tversky, A., and S. Sattath (1979), "Preference Trees," *Psychology Review* 86.