What's New in Econometrics   **NBER, Summer 2007**

Lecture 11, Wednesday, Aug 1st, 9.00-10.30am

Discrete Choice Models

1. INTRODUCTION

In this lecture we discuss multinomial discrete choice models. The modern literature on these models goes back to the work by Daniel McFadden in the seventies and eighties, (McFadden, 1973, 1981, 1982, 1984). In the nineties these models received much attention in the Industrial Organization literature, starting with Berry (1994), Berry, Levinsohn, Pakes (1995, BLP), and Goldberg (1995). In the IO literature the applications focused on demand for differentiated products, in settings with relatively large numbers of products, some of them close substitutes. In these settings a key feature of the conditional logit model, namely the Independence of Irrelevant Alternatives (IIA), was viewed as particularly unattractive. Three approaches have been used to deal with this. Goldberg (1995) used nested logit models to avoid the IIA property. McCulloch and Rossi (1994), and McCulloch, Polson and Rossi (2000) studied multinomial probit models with relatively unrestricted covariance matrices for the unobserved components. BLP, McFadden and Train (2000) and Berry, Levinsohn and Pakes (2004) uses random effects or mixed logit models, in BLP in combination with unobserved choice characteristics and using methods that allow for estimation using only aggregate choice data. The BLP approach has been very influential in the subsequent empirical IO literature.

Here we discuss these models. We argue that the random effects approach to avoid IIA is indeed very attractive, both substantively and computationally, compared to the nested logit or unrestricted multinomial probit models. In addition to the use of random effects to avoid the IIA property, the inclusion in the BLP methodology of unobserved choice characteristics, and the ability to estimate the models with market share rather than individual level data makes their methods very flexible and widely applicable. We discuss extensions to the BLP set up allowing multiple unobserved choice characteristics, and the richness required for these

models to rationalize general choice data based on utility maximization. We also discuss the potential benefits of using Bayesian methods.

2. Multinomial and Conditional Logit Models

First we briefly review the multinomial and conditional logit models.

2.1 Multinomial Logit Models

We focus on models for discrete choice with more than two choices. We assume that the outcome of interest, the choice Y_i takes on non-negative, un-ordered integer values between zero and J; $Y_i \in \{0, 1, \ldots, J\}$. Unlike the ordered case there is no particular meaning to the ordering. Examples are travel modes (bus/train/car), employment status (employed/unemployed/out-of-the-laborforce), car choices (suv, sedan, pickup truck, convertible, minivan), and many others.

We wish to model the distribution of Y in terms of covariates. In some cases we will distinguish between covariates Z_i that vary by units (individuals or firms), and covariates that vary by choice (and possibly by individual), X_{ij}. Examples of the first type include individual characteristics such as age or education. An example of the second type is the cost associated with the choice, for example the cost of commuting by bus/train/car, or the price of a product, or the speed of a computer chip. This distinction is important from the substantive side of the problem. McFadden developed the interpretation of these models through utility maximizing choice behavior. In that case we may be willing to put restrictions on the way covariates affect utilities: characteristics of a particular choice should affect the utility of that choice, but not the utilities of other choices.

The strategy is to develop a model for the conditional probability of choice j given the covariates. Suppose we only have individual-specific covariates, and the model is $\Pr(Y_i = j | Z_i = z) = P_j(z; \theta)$. Then the log likelihood function is

$$L(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{J} 1\{Y_i = j\} \cdot \ln P_j(Z_i; \theta).$$

A natural extension of the binary logit model is to model the response probability as

$$\Pr(Y_i = j | Z_i = z) = \frac{\exp(z'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for choices $j = 1, \ldots, J$ and

$$\Pr(Y_i = 0 | Z_i = z) = \frac{1}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for the first choice. The γ_l here are choice-specific parameters. This multinomial logit model leads to a very well-behaved likelihood function, and it is easy to estimate using standard optimization techniques. Interestingly, it can be viewed as a special case of the following conditional logit.

2.2 CONDITIONAL LOGIT MODELS

Suppose all covariates vary by choice (and possibly also by individual, but that is not essential here). Then McFadden proposed the conditional logit model:

$$\Pr(Y_i = j | X_{i0}, \ldots, X_{iJ}) = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)},$$

for $j = 0, \ldots, J$. Now the parameter vector β is common to all choices, and the covariates are choice-specific.

The multinomial logit model can be viewed as a special case of the conditional logit model. Suppose we have a vector of individual characteristics Z_i of dimension K, and J vectors of coefficients γ_j, each of dimension K. Then define for choice j, $j = 1, \ldots, J$, the vector of covariates X_{ij} as the vector of dimension $K \times J$, with all elements equal to zero other than the elements $K \times (j-1) + 1$ to $K \times j$ which are equal to Z_i:

$$X_{i1} = \begin{pmatrix} Z_i \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad X_{ij} = \begin{pmatrix} 0 \\ \vdots \\ Z_i \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad X_{iJ} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ Z_i \end{pmatrix}, \text{ and } X_{i0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and define the common parameter vector β, of dimension $K \cdot J$, as

$$\beta = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_J \end{pmatrix}.$$

Then

$$\Pr(Y_i = j | Z_i) = \frac{\exp(Z_i'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)} = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = j | X_{i0}, \ldots, X_{iJ}),$$

for $j = 1, \ldots, J$, and

$$\Pr(Y_i = 0 | Z_i) = \frac{1}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)} = \frac{\exp(X_{i0}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = 0 | X_{i0}, \ldots, X_{iJ}).$$

2.3 LINK WITH UTILITY MAXIMIZATION

McFadden motivates the conditional logit model by extending the single latent index model to multiple choices. Suppose that the utility, for individual i, associated with choice j, is

$$U_{ij} = X_{ij}'\beta + \varepsilon_{ij}. \tag{1}$$

Furthermore, let individual i choose option j (so that $Y_i = j$) if choice j provides the highest level of utility, or

$$Y_i = j \ \text{ if } U_{ij} \geq U_{il} \text{ for all } l = 0, \ldots, J,$$

(ties have probability zero because of the continuity of the distribution for ε).

Now suppose that the ε_{ij} are independent accross choices and individuals and have type I extreme value distributions. Then the choice Y_i follows the conditional logit model. The type I extreme value distribution has cumulative distribution function

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \qquad \text{and pdf } \ f(\epsilon) = \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)).$$

This distribution has a unique mode at zero, a mean equal to 0.58, and a a second moment of 1.99 and a variance of 1.65. See Figure 1 for the probability density function and the comparison with the normal density. Note the assymmetry of the distribution.

Given the extreme value distribution the probability of choice 0 is
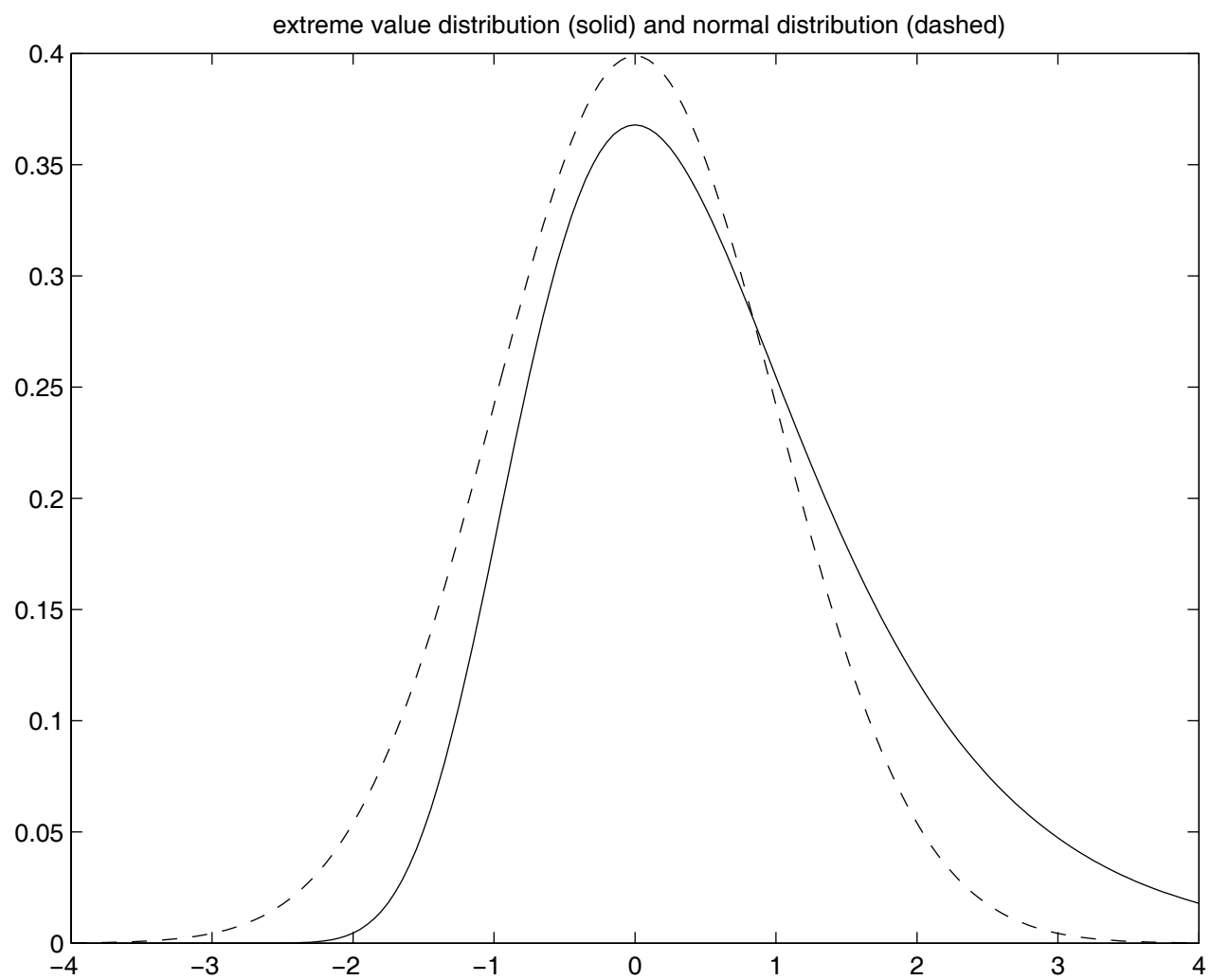
$$\Pr(Y_i = 0|X_{i0}, \ldots, X_{iJ}) = \Pr(U_{i0} > U_{i1}, \ldots, U_{i0} > U_{iJ})$$

$$= \Pr(\varepsilon_{i0} + X'_{i0}\beta - X'_{i1}\beta > \varepsilon_{i1}, \ldots, \varepsilon_{i0} + X'_{i0}\beta - X'_{iJ}\beta > \varepsilon_{iJ})$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\varepsilon_{i0}+X'_{i0}\beta-X'_{i1}\beta} \cdots \int_{-\infty}^{\varepsilon_{i0}+X'_{iJ}\beta-X'_{iJ}\beta} f(\varepsilon_{i0}) \ldots f(\varepsilon_{iJ}) d\varepsilon_{iJ} \ldots, d\varepsilon_{i0}$$

$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{0i}) \exp(-\exp(-\varepsilon_{0i}) \cdot \exp(-\exp(-\varepsilon_{i0} - X'_{i0}\beta + X'_{i1}\beta)) \cdots$$

$$\times \exp(-\exp(-\varepsilon_{i0} - X'_{i0}\beta + X'_{iJ}\beta)) d\varepsilon_{i0}$$

$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{0i}) \exp\left[-\exp(-\varepsilon_{0i}) - \exp(-\varepsilon_{i0} - X'_{i0}\beta + X'_{i1}\beta)) \cdots\right.$$

$$\left. - \exp(-\varepsilon_{i0} - X'_{i0}\beta + X'_{iJ}\beta)\right] d\varepsilon_{i0}$$

$$= \frac{\exp(X'_{i0}\beta)}{\sum_{j=0}^{J} \exp(X'_{j0}\beta)}.$$

To see the different steps in this derivation note that

$$\int_{-\infty}^{c} \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)) d\epsilon = F(c) = \exp(-\exp(-c)),$$

for the extreme value distribution. Also,

$$\int_{-\infty}^{\infty} \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon - c)) d\epsilon$$

extreme value distribution (solid) and normal distribution (dashed)

$$= \int_{-\infty}^{\infty} \exp(-\eta + c) \cdot \exp(-\exp(-\eta)) d\eta$$

$$= \exp(c) \cdot \int_{-\infty}^{\infty} \exp(-\eta) \cdot \exp(-\exp(-\eta)) d\eta = \exp(c),$$

by change of variables, which we apply with

$$c = -\ln\left(1 + \exp(X'_{i1}\beta - X'_{i0}\beta) + \ldots + \exp(X'_{iJ}\beta - X'_{i0}\beta)\right).$$

3. Independence of Irrelevant Alternatives

The main problem with the conditional logit is the property of Independence of Irrelevant Alternative (IIA). Consider the conditional probability of choosing j given that you choose either j or l:

$$\Pr(Y_i = j | Y_i \in \{j, l\}) = \frac{\Pr(Y_i = j)}{\Pr(Y_i = j) + \Pr(Y_i = l)} = \frac{\exp(X'_{ij}\beta)}{\exp(X'_{ij}\beta) + \exp(X'_{il}\beta)}.$$

This probability does not depend on the characteristics X_{im} of alternatives m other than j and l. This is sometimes unattractive. The traditional example is McFadden's famous blue bus/red bus example. Suppose there are initially three choices: commuting by car, by red bus or by blue bus. It would seem reasonable be to assume that people have a preference over cars versus buses, but are indifferent between red versus blue buses. One could capture this by assuming that

$$U_{i,\text{redbus}} = U_{i,\text{bluebus}},$$

with the choice between the blue and red bus being random. So, to be explicit, suppose that $X_{i,\text{bluebus}} = X_{i,\text{redbus}} = X_{i,\text{bus}}$. Then suppose that the probability of commuting by bus is

$$\Pr(Y_i = \text{bus}) = \Pr(Y_i = \text{redbus or bluebus}) = \frac{\exp(X'_{i,\text{bus}}\beta)}{\exp(X'_{i,\text{bus}}\beta) + \exp(X'_{i,\text{car}}\beta)},$$

and the probability of choosing a red bus or blue bus, conditional on choosing a bus, is

$$\Pr(Y_i = \text{redbus}|Y_i = \text{bus}) = \frac{1}{2}.$$

That would imply that the conditional probability of commuting by car, given that one commutes by car or red bus, would differ from the same conditional probability if there is no blue bus. Presumably taking away the blue bus choice would lead all the current blue bus users to shift to the red bus, and not to cars.

The conditional logit model does not allow for this type of substitution pattern. Another way of stating the problems with the conditional logit model is to say that it generates unrealistic substitution patterns. Let us make that argument more specific. Suppose that individuals have the choice out of three Berkeley restaurants, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B). Suppose the two characteristics of the restaurants are price with $P_C = 95$, $P_L = 80$, and $P_B = 5$, and quality, with $Q_C = 10$, $Q_L = 9$, and $Q_B = 2$. Suppose that market shares for the three restaurants are $S_C = 0.10$, $S_L = 0.25$, and $S_B = 0.65$. These numbers are roughly consistent with a conditional logit model where the utility associated with individual i and restaurant j is

$$U_{ij} = -0.2 \cdot P_j + 2 \cdot Q_j + \epsilon_{ij},$$

with independent extreme value ϵ_{ij}, and individuals go to the restaurant with the highest utility. Now suppose that we raise the price at Lalime's to 1000 (or raise it to infinity, corresponding to taking it out of business). In that case the prediction of the conditional logit model is that the market shares for Chez Panisse and the Bongo Burger go to $\tilde{S}_C = 0.13$ and $\tilde{S}_B = 0.87$. That seems implausible. The people who were planning to go to Lalime's would appear to be more likely to go to Chez Panisse if Lalime's is closed than to go to the Bongo Burger, and so one would expect $\tilde{S}_C \approx 0.35$ and $\tilde{S}_B \approx 0.65$. The model on the other hand predicts that most of the individuals who would have gone to Lalime's will now dine (if that is the right term) at the Bongo Burger.

Recall the latent utility set up with the utility for individual i and choice j equal to

$$U_{ij} = X'_{ij}\beta + \epsilon_{ij}. \tag{2}$$

In the conditional logit model we assume independent ϵ_{ij} with extreme value distributions. This is essentially what creates the IIA property. (This is not completely correct, because other distributions for the unobserved, say with normal errors, we would not get IIA exactly, but something pretty close to it.) The solution is to allow in some fashion for correlation between the unobserved components in the latent utility representation. In particular, with a choice set that contains multiple versions of essentially the same choice (like the red bus or the blue bus), we should allow the latent utilities for these choices to be identical, or at least very close. In order to achieve this the unobserved components of the latent utilities would have to be highly correlated for those choices. This can be done in a number of ways.

4. Models without Independence of Irrelevant Alternatives

Here we discuss three ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the unobserved components of the latent utility. All of these originate in some form or another in McFadden's work (e.g., McFadden, 1981, 1982, 1984). The first is the nested logit model where the researcher groups together sets of choices. In the simple version with a single layer of nests this allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation between the unobserved components of choices in different nests. Second, the unrestricted multinomial probit model with no restrictions on the covariance between unobserved components, beyond normalizations. Third, the mixed or random coefficients logit where the marginal utilities associated with choice characteristics are allowed to vary between individuals. This generates positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

4.1 Nested Logit

One way to induce correlation between the choices is through nesting them. Suppose the

set of choices $\{0, 1, \ldots, J\}$ can be partitioned into S sets $B_1, \ldots, B_S$, so that the full set of choices can be written as

$$\{0, 1, \ldots, J\} = \cup_{s=1}^{S} B_s.$$

Let Z_s be set-specific characteristics. (It may be that the set of set specific variables is empty, or just a vector of indicators, with Z_s an S-vector of zeros with a one for the sth element.) Now let the conditional probability of choice j given that your choice is in the set B_s, or $Y_i \in B_s$ be equal to

$$\Pr(Y_i = j | X_i, Y_i \in B_s) = \frac{\exp(\rho_s^{-1} X_{ij}' \beta)}{\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}' \beta)},$$

for $j \in B_s$, and zero otherwise. In addition suppose the marginal probability of a chocie in the set B_s is

$$\Pr(Y_i \in B_s | X_i) = \frac{\exp(Z_s' \alpha) \left(\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}' \beta) \right)^{\rho_s}}{\sum_{t=1}^{S} \exp(Z_t' \alpha) \left(\sum_{l \in B_t} \exp(\rho_t^{-1} X_{il}' \beta) \right)^{\rho_s}}.$$

If we fix $\rho_s = 1$ for all s, then

$$\Pr(Y_i = j | X_i) = \frac{\exp(X_{ij}' \beta + Z_s' \alpha)}{\sum_{t=1}^{S} \sum_{l \in B_t} \exp(X_{il}' \beta + Z_t \alpha)},$$

and we are back in the conditional logit model.

In general this model corresponds to individuals choosing the option with the highest utility, where the utility of choice j in set B_s for individual i is

$$U_{ij} = X_{ij}' \beta + Z_s' \alpha + \epsilon_{ij},$$

where the joint distribution function of the ϵ_{ij} is

$$F(\epsilon_{i0}, \ldots, \epsilon_{iJ}) = \exp\left(-\sum_{s=1}^{S} \left(\sum_{j \in B_s} \exp\left(-\rho_s^{-1} \epsilon_{ij} \right) \right)^{\rho_s} \right).$$

Within the sets the correlation coefficient for the ϵ_{ij} is approximately equal to $1-\rho$. Between the sets the ϵ_{ij} are independent.

The nested logit model could capture the blue bus/red bus example by having two nests, the first $B_1 = \{\text{redbus}, \text{bluebus}\}$, and the second one $B_2 = \{\text{car}\}$.

How do you estimate these models? One approach is to construct the log likelihood and directly maximize it. That is complicated, especially since the log likelihood function is not concave, but it is not impossible. An easier alternative is to directly use the nesting structure. Within a nest we have a conditional logit model with coefficients β/ρ_s. Hence we can directly estimate β/ρ_s using the concavity of the conditional logit model. Denote these estimates of β/ρ_s by $\widehat{\beta/\rho_s}$. Then the probability of a particular set B_s can be used to estimate ρ_s and α through

$$\Pr(Y_i \in B_s | X_i) = \frac{\exp(Z_s'\alpha)\left(\sum_{l \in B_s} \exp(X_{il}'\widehat{\beta/\rho_s})\right)^{\rho_s}}{\sum_{t=1}^{S} \exp(Z_t'\alpha)\left(\sum_{l \in B_t} \exp(X_{il}'\widehat{\beta/\rho_t})\right)^{\rho_s}} = \frac{\exp(Z_s'\alpha + \rho_s \hat{W}_s)}{\sum_{t=1}^{S} \exp(Z_t'\alpha + \rho_t \hat{W}_t)},$$

where

$$\hat{W}_s = \ln\left(\sum_{l \in B_s} \exp(X_{il}'\widehat{\beta/\rho_s})\right),$$

known as the "inclusive values". Hence we have another conditional logit model back that is easily estimable. These two-step estimators are not efficient. The variance/covariance matrix is provided in McFadden (1981).

These models can be extended to many layers of nests. See for an impressive example of a complex model with four layers of multiple nests Goldberg (1995). Figure 2 shows the nests in the Goldberg application. The key concern with the nested logit models is that results may be sensitive to the specification of the nest structure. The researcher chooses the choices that are potentially close, with the data being used to estimate the amount of correlation. In contrast, in the random effects models, choices can only be close if they are close in terms of observed choice characteristics, with the data being used to estimate the

from: PINELOPI KOUJIANOU GOLDBERG   (1995)

Household

Buy At Least
One Car

Do not Buy
a Car

Buy at Least
One New Car

Buy Only
Used Car

Class1   Class2   . . .   Class9
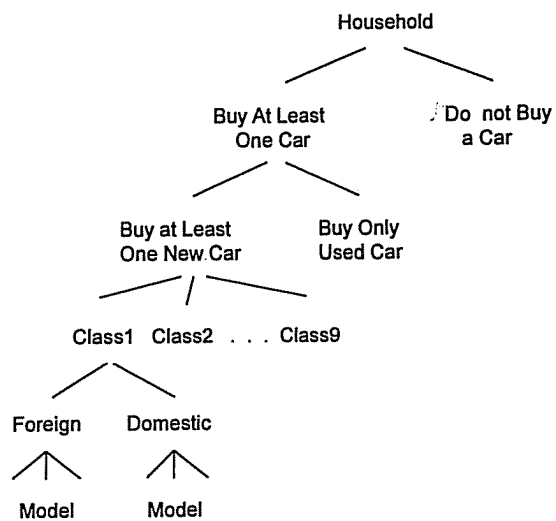
Foreign   Domestic

Model   Model

FIGURE 1.—Automobile choice model.

relative importance of the various choice characteristics. In that sense the nested logit model can be more flexible, allowing the researcher to group together choices that are far apart in terms of observed choice characteristics, but it is more demanding in requiring the researcher to make these decisions *a priori*.

4.2 MULTINOMIAL PROBIT

A second possibility is to directly free up the covariance matrix of the error terms. This is more natural to do in the multinomial probit case. See McCulloch and Rossi (1994) McCulloch, Polson, and Rossi (2000) for general discussion.

We specify:

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X'_{i0}\beta + \epsilon_{i0} \\ X'_{i1}\beta + \epsilon_{i1} \\ \vdots \\ X'_{iJ}\beta + \epsilon_{iJ} \end{pmatrix},$$

with

$$\epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} \Big| X_i \sim \mathcal{N}(0, \Omega),$$

for some relatively unrestricted $(J+1) \times (J+1)$ covariance matrix Ω. We do need some normalizations on Ω beyond symmetry. Recall that in the binary choice case (which corresponds to $J = 1$) there were no free parameters in the distribution of ϵ, which implies three restrictions on the symmetric matrix Ω.

In principle we can derive the probability for each choice given the covariates, construct the likelihood function based on that, and maximize it using an optimization algorithm like Davidon-Fletcher-Powell (Gill, Murray, and Wright, 1981) or something similar. In practice this is very difficult with $J \geq 3$. Evaluating the probabilities involves calculating a third order integral involving normal densities. This is difficult to to using standard integration methods. There are two alternatives.

There is a substantial literature on simulation methods for computing estimates in these models. See for an early example Manski and Lerman (1981), general studies McFadden (1989), and Pakes and Pollard (1989), and Hajivassiliou and Ruud (1994) for a review. Geweke, Keane, and Runkle (1994) and Hajivasilliou and McFadden (1990) proposed a way of calculating the probabilities in the multinomial probit models that allowed researchers to deal with substantially larger choice sets. A simple attempt to estimate the probabilities would be to draw the ϵ_i from a multivariate normal distribution and calculate the probability of choice j as the number of times choice j corresponded to the highest utility. This does not work well in cases with many (more than four) choices. The Geweke-Hajivasilliou-Keane (GHK) simulator uses a more complicated procedure that draws sequentially and combines the draws with the calculation of univariate normal integrals so that the resulting probabilities are smooth in the parameters.

From a Bayesian perspective drawing from the posterior distribution of β and Ω is straightforward. The key is setting up the vector of unobserved random variables as

$$\theta = (\beta, \Omega, U_{i0}, \ldots, U_{iJ}),$$

and defining the most convenient partition of this vector. Suppose we know the latent utilities U_i for all individuals. Then the normality makes this a standard linear model problem, and we can sample sequentially from $\beta|\Omega$ and $\Omega|\beta$ given the appropriate conjugate prior distributions (normal for β and inverse Wishart for Ω). Given the parameters drawing from the unobserved utilities can be done sequentially: for each unobserved utility given the others we would have to draw from a truncated normal distribution, which is straightforward. See McCulloch, Polson, and Rossi (2000) for details.

The attraction of this approach is that there are no restrictions on which choices are close. In contrast, in the nested logit approach the researcher specifies which choices are potentially close, and in the random effects approach only choices that are close in terms of observed choice characteristics can be close. The difficulty, however, with the unrestricted multinomial probit approach is that with a reasonable number of chocies this frees up a

large number of parameters (all elements in the $(J+1) \times (J+1)$ dimensional covariance matrix of latent utilities, minus some that are fixed by normalizations.) Estimating all these covariance parameters precisely, based on only first choice data (as opposed to data where we know for each individual additional orderings, e.g., first and second choices), is difficult with the sample sizes typically available.

4.3 Random Coefficient (Mixed) Logit (or Probit)

A third possibility to get around the IIA property is to allow for unobserved heterogeneity in the slope coefficients. This is a very natural idea. Why do we fundamentally think that if Lalime's price goes up, the individuals who were planning to go Lalime's go to Chez Panisse instead, rather than to the Bongo Burger? The reason is that we think individuals who have a taste for Lalime's are likely to have a taste for close substitute in terms of observable characteristics, Chez Panisse as well, rather than for the Bongo Burger.

We can model this by allowing the marginal utilities to vary at the individual level:

$$U_{ij} = X'_{ij}\beta_i + \epsilon_{ij},$$

where the ϵ_{ij} are again independent of everything else, and of each other, either extreme value, or normal. We can also write this as

$$U_{ij} = X'_{ij}\overline{\beta} + \nu_{ij},$$

where

$$\nu_{ij} = \epsilon +_{ij} + X_{ij} \cdot (\beta_i - \overline{\beta}),$$

which is no longer independent across choices. The key ingredient is the vector of individual specific taste parameters β_i. See for a general discussion of such models and their properties in approximating general choice patterns McFadden and Train (2000). One possibility is to

assume the existence of a finite number of types of individuals, similar to the mixture models used by Heckman and Singer (1984) in duration settings:

$$\beta_i \in \{b_0, b_1, \ldots, b_K\},$$

with

$$\Pr(\beta_i = b_k | Z_i) = p_k, \qquad \text{or} \quad \Pr(\beta_i = b_k | Z_i) = \frac{\exp(Z_i'\gamma_k)}{1 + \sum_{l=1}^{K} \exp(Z_i'\gamma_l)}.$$

Here the taste parameters take on a finite number of values, and we have a finite mixture. We can use either Gibbs sampling with the indicator of which mixture an observations belongs to as an unobserved random variable, or use the EM algorithm (Dempster, Laird, and Rubin, 1977).

Alternatively we could specify

$$\beta_i | Z_i \sim \mathcal{N}(Z_i'\gamma, \Sigma),$$

where we use a normal (continuous) mixture of taste parameters. Just evaluating the likelihood function would be very difficult in this setting if there is a large number of choices. This would involve integrating out the random coefficients which could be very computationally intensive. See McFadden and Train (2000). Using Gibbs sampling with the unobserved β_i as additional unobserved random variables may be an effective way of doing inference.

5. BERRY-LEVINSOHN-PAKES

Here we consider again random effects logit models. BLP extended these models to allow for unobserved product characteristics, endogeneity of choice characteristics, and developed methods that allowed for consistent estimation without individual level choice data. Their approach has been widely used in Industrial Organization, where it is used to model demand for differentiated products, often in settings with a large number of products. See Nevo (2000) and Ackerberg, Benkard, Berry, and Pakes (2005) for reviews and references.

Compared to the earlier examples we have looked at there is an emphasis in this study, and those that followed it, on the large number of goods and the potential endogeneity of some of the product characteristics. (Typically one of the regressors is the price of the good.) In addition the procedure only requires market level data. We do not need individual level purchase data, just market shares and estimates of the distribution of individual characteristics by market. In practice we need a fair amount of variation in these things to estimate the parameters well, but in principle this is less demanding in terms of data required. On the other hand, we do need data by market, where before we just needed individual purchases in a single market (although to identify price effects we would need variation in prices by individuals in that case).

The data have three dimensions: products, indexed by $j = 0, \ldots, J$, markets, $t = 1, \ldots, T$, and individuals, $i = 1, \ldots, N_t$. We only observe one purchase per individual. The large sample approximations are based on large N and T, and fixed J.

Let us go back to the random coefficients model, now with each utility indexed by individual, product and market:

$$U_{ijt} = \beta_i' X_{jt} + \zeta_{jt} + \epsilon_{ijt}.$$

The ζ_{jt} is a unobserved product characteristic. This component is allowed to vary by market and product. It can include product and market dummies (for example, we can have $\zeta_{jt} = \zeta_j + \zeta_t$). Unlike the observed product characteristics this unobserved characteristic does not have a individual-specific coefficient. The inclusion of this component allows the model to rationalize any pattern of market shares. The observed product characteristics may include endogenous characteristics like the price.

The ϵ_{ijt} unobserved components have extreme value distributions, independent across all individuals i, products j, and markets t.

The random coefficients β_i, with dimension equal to that of the observable characteristics X_{jt}, say K, are assumed to be related to individual observable characteristics. We postulate

the following linear form:

$$\beta_i = \beta + Z_i'\Gamma + \eta_i,$$

with

$$\eta_i|Z_i \sim \mathcal{N}(0, \Sigma).$$

So if the dimension of Z_i is $L \times 1$, then Γ is a $L \times K$ matrix. The Z_i are normalized to have mean zero, so that the β's are the average marginal utilities. The normality assumption is not necessary, and unlikely to be important. Other distributional assumptions can be substituted.

BLP developed an approach to estimate models of this type that does not require individual level data. Instead it exploits aggregate (market level) data in combination with estimates of the distribution of Z_i. Specifically the data consist of estimated shares $\hat{s}_{tj}$ for each choice j in each market t, combined with observations from the marginal distribution of individual characteristics (the Z_i's) for each market, often from representative data sets such as the CPS.

First write the latent utilities as

$$U_{ijt} = \delta_{jt} + \nu_{ijt} + \epsilon_{ijt},$$

where

$$\delta_{jt} = \beta' X_{jt} + \zeta_{jt}, \qquad \text{and} \quad \nu_{ijt} = (Z_i'\Gamma + \eta_i)' X_{jt}.$$

Now consider for fixed Γ and Σ and δ_{jt} the implied market share for product j in market t, s_{jt}. This can be calculated analytically in simple cases. For example with $\Gamma_{jt} = 0$ and $\Sigma = 0$, the market share is a very simple function of the δ_{jt}:

$$s_{jt}(\delta_{jt}, \Gamma = 0, \Sigma = 0) = \frac{\exp(\delta_{jt})}{\sum_{l=0}^{J} \exp(\delta_{lt})}.$$

More generally, this is a more complex relationship. We can always calculate the implied market share by simulation: draw from the distribution of Z_i in market t, draw from the distribution of η_i, and calculate the implied purchase probability (or even simulate the implied purchase by also drawing from the distribution of ϵ_{ijt}). Do that repeatedly and you will be able to calculate the market share for this product/market. Call the vector function obtained by stacking these functions for all products and markets $s(\delta, \Gamma, \Sigma)$.

Next, fix only Γ and Σ. For each value of δ_{jt} we can find the implied market share. Now find the vector of δ_{jt} such that the implied market shares are equal to the observed market shares $\hat{s}_{jt}$ for all j, t. BLP suggest using the following algorithm. Given a starting value for δ_{jt}^0, use the updating formula:

$$\delta_{jt}^{k+1} = \delta_{jt}^k + \ln s_{jt} - \ln s_{jt}(\delta^k, \Gamma, \Sigma).$$

BLP show this is a contraction mapping, and so it defines a function $\delta(s, \Gamma, \Sigma)$ expressing the δ as a function of observed market shares, and parameters Γ and Σ. In order to implement this, one needs to approximate the implied market shares accurately for each iteration in the contraction mapping, and then you will need to do this repeatedly to get the contraction mapping to converge.

Note that does require that each market share is accurately estimated. If all we have is an estimated market share, then even if this is unbiased, the procedures will not necessarily work. In that case the log of the estimated share is not unbiased for the log of the true share. In practice the precision of the estimated market share is so much higher than that of the other parameters that this is unlikely to matter.

Given this function $\delta(s, \Gamma, \Sigma)$ define the residuals

$$\omega_{jt} = \delta_{jt}(s, \Gamma, \Sigma) - \beta' X_{jt}.$$

At the true values of the parameters and the true market shares this is equal to the unobserved product characteristic ζ_{jt}.

Now we can use GMM or instrumental variable methods. We assume that the unobserved product characteristics are uncorrelated with observed product characteristics (other than typically price). This is not sufficient since the observed product characteristics enter directly into the model. We need more instruments, and typically use things like characteristics of other products by the same firm, or average characteristics by competing products. The general GMM machinery will also give us the standard errors for this procedure. This is where the method is most challenging. Finding values of the parameters that set the average moments closest to zero can be difficult.

It is instructive to see what this approach does if we in fact have, and know we have, a conditional logit model with fixed coefficients. In that case $\Gamma = 0$, and $\Sigma = 0$. Then we can invert the market share equation to get the market specific unobserved choice-characteristics

$$\delta_{jt} = \ln s_{jt} - \ln s_{0t},$$

where we set $\delta_{0t} = 0$. (this is typically the outside good, whose average utility is normalized to zero). The residual is

$$\zeta_{jt} = \delta_{jt} - \beta' X_{jt} = \ln s_{jt} - \ln s_{0t} - \beta' X_{jt}.$$

With a set of instruments W_{jt}, we run the regression

$$\ln s_{jt} - \ln s_{0t} = \beta' X_{jt} + \epsilon_{jt},$$

using W_{jt} as instrument for X_{jt}, using as the observational unit the market share for product j in market t.

So here the technique is very transparent. It amounts to transforming the market shares to something linear in the coefficients so we can use two-stage-least-squares. More generally the transformation is going to be much more difficult with the random coefficients implying that there is no analytic solution. Computationally these things can get very complicated. Note however that we can estimate these models now without having individual level data,

and that at the same time we can get a fairly flexible model for the substitution patterns. At the same time you would expect to need a lot of structure to get the parameters precisely estimated just as in the other models. Of course if you compare the current model to the nested logit model you can impose such structure by imposing restrictions on the covariance matrix.

Comparisons of the models are difficult. Obviously if the structure imposed is correct it helps, but we typically do not know what the truth is, so we cannot conclude which one is better on the basis of the data typically available.

6. Models with Multiple Unobserved Choice Characteristics

The BLP approach allows for a single unobserved choice characteristic. This is essential for their estimation strategy that requires only market share data, and exploits a one-to-one relationship between market-specific unobserved product characteristics and market shares given other parameters and covariates. With individual level data one may be able to, and wish to allow for, multiple unobserved product characteristics. Elrod and Keane (1995), Goettler and Shachar (2001), and Athey and Imbens (2007), among others, study such models, in all cases with the unobserved choice characteristics constant across markets. Athey and Imbens model the latent utility for individual i in market t for choice j as

$$U_{ijt} = X'_{it}\beta_i + \zeta'_j\gamma_i + \epsilon_{ijt},$$

with the individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Delta Z_i, \Omega).$$

Even in the case with all choice characteristics exogenous, maximum likelihood estimation would be difficult. Athey and Imbens show that Bayesian methods, and in particular markov-chain-monte-carlo methods are effective tools for conducting inference in these settings.

7. Hedonic Models and the Motivation for a Choice and Individual Specific Error Term

Recently researchers have reconsidered using pure characteristics models for discrete choices, that is models with no idiosyncratic error ϵ_{ij}, instead relying solely on the presence of a few unobserved product characteristics and unobserved variation in taste parameters to generate stochastic choices. Such an error term is the only source of stochastic variation in the original multinomial choice models with only observed choice and individual characteristics, but in models with unobserved choice and individual characteristics their presence needs more motivation. Athey and Imbens (2007) discuss two arguments for including the additive error term.

First, the pure characteristics model can be extremely sensitive to measurement error, because it can predict zero market shares for some products. Consider a case where choices are generated by a pure characteristics model that implies that a particular choice j has zero market share. Now suppose that there is a single unit i for whom we observe, due to measurement error, the choice $Y_i = j$. Irrespective of the number of correctly measured observations available that were generated by the pure characteristics model, the estimates of the latent utility function will not be close to the true values corresponding to the pure characteristics model due to the single mismeasured observation. Such extreme sensitivity puts a lot of emphasis on the correct specification of the model and the absence of measurement error, and is undesirable in most settings.

Thus, one might wish to generalize the model to be robust against small amounts of measurement error of this type. One possibility is to define the optimal choice Y_i^* as the choice that maximizes the utility and assume that the observed choice Y_i is equal to the optimal choice Y_i^* with probability $1 - \delta$, and with probability $\delta/(J-1)$ any of the other choices is observed:

$$\Pr(Y_i = y | Y_i^*, X_i, \nu_i, Z_1, \ldots, Z_J, \zeta_1, \ldots, \zeta_J) = \begin{cases} 1 - \delta & \text{if } Y = Y_i^*, \\ \delta/(J-1) & \text{if } Y \neq Y_i^*. \end{cases}$$

This nests the pure characteristics model (by setting $\delta = 0$), without having the disad-

vantages of extreme sensitivity to mismeasured choices that the pure characteristics model has. If the true choices are generated by the pure characteristics model the presence of a single mismeasured observation will not prevent the researcher from estimating the true utility function. However, this specific generalization of the pure characteristics model has an unattractive feature: if the optimal choice Y_i^* is not observed, all of the remaining choices are equally likely. One might expect that choices with utilities closer to the optimal one are more likely to be observed conditional on the optimal choice not being observed.

An alternative modification of the pure characteristics model is based on adding an idiosyncratic error term to the utility function. This model will have the feature that, conditional on the optimal choice not being observed, a close-to-optimal choice is more likely than a far-from-optimal choice. Suppose the true utility is U_{ij}^* but individuals base their choice on the maximum of mismeasured version of this utility:

$$U_{ij} = U_{ij}^* + \epsilon_{ij},$$

with an extreme value ϵ_{ij}, independent across choices and individuals. The ϵ_{ij} here can be interpreted as an error in the calculation of the utility associated with a particular choice. This model does not directly nest the pure characteristics model, since the idiosyncratic error term has a fixed variance. However, it approximately nests it in the following sense. If the data are generated by the pure characteristics model with the utility function $g(x, \nu, z, \zeta)$, then the model with the utility function $\lambda \cdot g(x, \nu, z, \zeta) + \epsilon_{ij}$ leads, for sufficiently large λ, to choice probabilities that are arbitrarily close to the true choice probabilities (e.g., Berry and Pakes, 2007).

Hence, even if the data were generated by a pure characteristics model, one does not lose much by using a model with an additive idiosyncratic error term, and one gains a substantial amount of robustness to measurement or optimization error.

References

ACKERBERG, D., L. BENKARD, S. BERRY, AND A. PAKES, (2005), "Econometric Tools for Analyzing Market Outcomes," forthcoming, *Handbook of Econometrics*, Vol 5, Heckman and Leamer (eds.)

AMEMIYA, T., AND F. NOLD, (19750, "A Modified Logit Model," *Review of Economics and Statistics*, Vol 57(2), 255-257.

ATHEY, S., AND G. IMBENS, (2007), "Discrete Choice Models with Multiple Unobserved Product Characteristics," *International Economic Review*, forthcoming.

BAJARI, P., AND L. BENKARD, (2004), "Demand Estimation with Heterogenous Consumers and Unobserved Product Characteristics: A Hedonic Approach," Stanford Business School.

BERRY, S., (1994), "Estimating Discrete-Choice Models of Product Differentiation,"*RAND Journal of Economics*, Vol. 25, 242-262.

BERRY, S., J. LEVINSOHN, AND A. PAKES, (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, Vol. 63, 841-890.

BERRY, S., J. LEVINSOHN, AND A. PAKES (2004), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, Vol 112(1), 68-105.

BERRY, S., O. LINTON, AND A. PAKES, (2004), "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems ", *Review of Economic Studies*, Vol. 71, 613-654.

BERRY, S., AND A. PAKES, (2007), "The Pure Characteristics Discrete Choice Model of Differentiated Products Demand,"*International Economic Review*, forthcoming.

DEMPSTER, A., N. LAIRD, AND D. RUBIN, (1974), "Maximum Likelihood from Incomplete Data via the EM Algorithm", (with discussion), *Journal of the Royal Statistical*

Society, Series B, Vol. 39, 1-38.

ELROD, T., AND M. KEANE, (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data, "*Journal of Marketing Research*, Vol. XXXII, 1-16.

GEWEKE, J., M. KEANE, AND D. RUNKLE, (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 76, No 4, 609-632.

GILL, P., W. MURRAY, AND M. WRIGHT, (1981), *Practical Optimization*, Harcourt Brace and Company, London

GOETTLER, J., AND R. SHACHAR (2001), "Spatial Competition in the Network Television Industry," *RAND Journal of Economics*, Vol. 32(4), 624-656.

GOLDBERG, P., (1995), "Product Differentiation and Oligopoly in International Markets: The Case of the Automobile Industry," *Econometrica*, 63, 891-951.

HAJIVASSILIOU, V., AND P. RUUD, (1994), "Classical Estimation Methods for LDV Models Using Simulation," in Engle and McFadden 9eds.), *Handbook of Econometrics*, Vol 4, Chapter 40, Elseviers.

HAJIVASSILIOU, V., AND D. MCFADDEN, (1990, "The method of simulated scores,", with application to models of external debt crises," unpublished manuscript, Department of Economics, Yale University.

HECKMAN, J., AND B. SINGER, (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2).

MANSKI, C., AND S. LERMAN,, (1981) "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds.), 305-319, MIT Press, Cambridge, MA.

MCCULLOCH, R., AND P. ROSSI, (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model, "*Journal of Econometrics* 64 207-240.

McCulloch, R., N. Polson, and P. Rossi, (2000) "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters, "*Journal of Econometrics* 99, 173-193.

McFadden, D., (1973), "Conditional Logit Analysis of Qualitative Choice Behavior "in P. Zarembka (ed), *Frontiers in Econometrics* Academic Press, New York 105-142.

McFadden, D., (1981) "Econometrica Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds.), 198-272, MIT Press, Cambridge, MA.

McFadden, D., (1982), "Qualitative Response Models," in Hildenbrand (ed.), *Advances in Econometrics*, Econometric Society Monographs, Cambridge University Press.

McFadden, D., (1984), "Econometric Analysis of Qualitative Response Models,"in Griliches and Intriligator (eds), *Handbook of Econometrics*, Vol. 2, 1395- 1457, Amsterdam.

McFadden, D., (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57(5), 995-1026.

McFadden, D., and K. Train, (2000), "Mxed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15(5), 447-470.

Nevo, A. (2000), "A Practioner's Guide to Estimation of Random-Coefficient Logit Models of Demand," *Journal of Economics & Management Science,* Vol. 9, No. 4, 513-548.

Pakes, A., and D. Pollard, (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027-1057.