“New Developments in Econometrics”

Lecture 8

Discrete Choice Models

Guido Imbens

Cemmap Lectures, UCL, June 2009

Outline

1. Introduction

2. Multinomial and Conditional Logit Models

3. Independence of Irrelevant Alternatives

4. Models without IIA

5. Berry-Levinsohn-Pakes

6. Models with Multiple Unobserved Choice Characteristics

7. Hedonic Models

1. Introduction

Various versions of multinomial logit models developed by McFadden in 70's.

In IO applications with substantial number of choices IIA property found to be particularly unattractive because of unrealistic implications for substitution patterns.

Random effects approach is more appealing generalization than either nested logit or unrestricted multinomial probit

Generalization by BLP to allow for endogenous choice characteristics, unobserved choice characteristics, using only aggregate choice data.

2. Multinomial and Conditional Logit Models

Models for discrete choice with more than two choices.

The choice Y_i takes on non-negative, unordered integer values between zero and J.

Examples are travel modes (bus/train/car), employment status (employed/unemployed/out-of-the-laborforce), car choices (suv, sedan, pickup truck, convertible, minivan).

We wish to model the distribution of Y in terms of covariates individual-specific, choice-invariant covariates Z_i (e.g., age) choice (and possibly individual) specific covariates X_{ij}.

2.A Multinomial Logit

Individual-specific covariates only.

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

for choices $j = 1, \ldots, J$ and for the first choice:

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

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.

2.B Conditional Logit

Suppose all covariates vary by choice (and possibly also by individual). The conditional logit model specifies:

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

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

Also easy to estimate.

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

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

and define the parameter vector as $\beta = (\gamma_1', \ldots, \gamma_J')'$. Then

$$\Pr(Y_i = j | Z_i) = \frac{\exp(Z_i' \gamma_j)}{1 + \sum_{k=1}^{J} \exp(Z_i' \gamma_k)}$$

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

2.D Link with Utility Maximization

Utility, for individual i, associated with choice j, is

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

i choose option j if choice j provides the highest level of utility

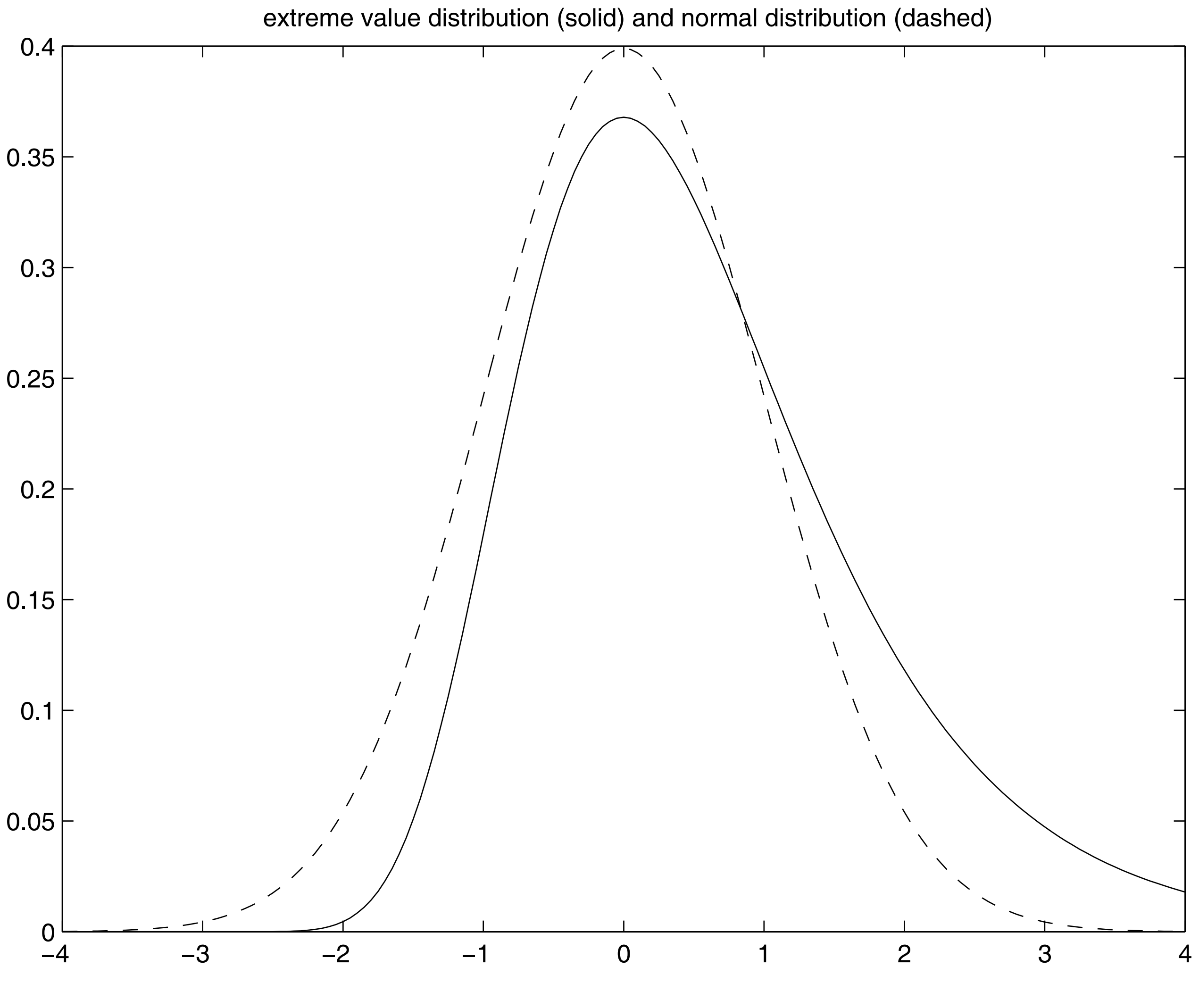
$$Y_i = j \quad \text{if } U_{ij} \geq U_{il} \text{ for all } l = 0, \ldots, J,$$

Now suppose that the ε_{ij} are independent accross choices and individuals and have type I extreme value distributions.

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \qquad f(\epsilon) = \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)).$$

(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.)

Then the choice Y_i follows the conditional logit model.


extreme value distribution (solid) and normal distribution (dashed)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4
−3
−2
−1
0
1
2
3
4

3. Independence of Irrelevant Alternatives

The main problem with the conditional logit is the property of Independence of Irrelevant Alternative (IIA).

The conditional probability of choosing j given 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.

Also unattractive implications for marginal probabilities for new choices.

Although multinomial and conditional logit models may fit well, they are not necessarily attractive as behavior/structural models. because they generates unrealistic substitution patterns.

Suppose that individuals have the choice out of three restaurants, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B). Suppose we have two characteristics, price and quality

price	$P_C = 95$, $P_L = 80$, $P_B = 5$,
quality	$Q_C = 10$, $Q_L = 9$, $Q_B = 2$
market share	$S_C = 0.10$, $S_L = 0.25$, $S_B = 0.65$.

These numbers are roughly consistent with a conditional logit model where the utility associated with individual i and restaurant j is

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

Now suppose that we raise the price at Lalime's to 1000 (or raise it to infinity, corresponding to taking it out of business).

The conditional logit model predicts that the market share for Lalime's gets divided by Chez Panisse and the Bongo Burger, proportional to their original market share, and thus $\tilde{S}_C = 0.13$ and $\tilde{S}_B = 0.87$: most of the individuals who would have gone to Lalime's will now dine (if that is the right term) at the Bongo Burger.

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, implying $\tilde{S}_C \approx 0.35$ and $\tilde{S}_B \approx 0.65$.

Recall the latent utility set up with the utility

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

In the conditional logit model we assume independent extreme value ϵ_{ij}. The independence 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 similar choices (like Chez Panisse and LaLime's), we should allow the latent utilities for these choices to be similar.

4. Models without IIA

Here we discuss 3 ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the ϵ_{ij}.

The first is the nested logit model where the researcher groups together sets of choices. This allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation across 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 vary between individuals, generating positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

Nested Logit Models

Partition the set of choices $\{0, 1, \ldots, J\}$ into S sets $B_1, \ldots, B_S$

Now let the conditional probability of choice j given that your choice is in the set 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 choice in the set B_s is

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

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

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

and we are back in the conditional logit model.

The implied 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 restaurant example by having two nests, the first $B_1 = \{\text{Chez Panisse}, \text{LaLime's}\}$, and the second one $B_2 = \{\text{Bongoburger}\}$.

Estimation of Nested Logit Models

Maximization of the likelihood function is difficult.

An easier alternative is to use the nesting structure. Within a nest we have a conditional logit model with coefficients β/ρ_s. Estimates these as $\widehat{\beta/\rho_s}$.

Then the probability of a particular set B_s can be used to estimate ρ_s through

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

where the "inclusive values" are

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

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** which choices are potentially close substitutes, with the data being used to estimate the amount of correlation.

Researcher would have to choose nest for new good to estimate market share.

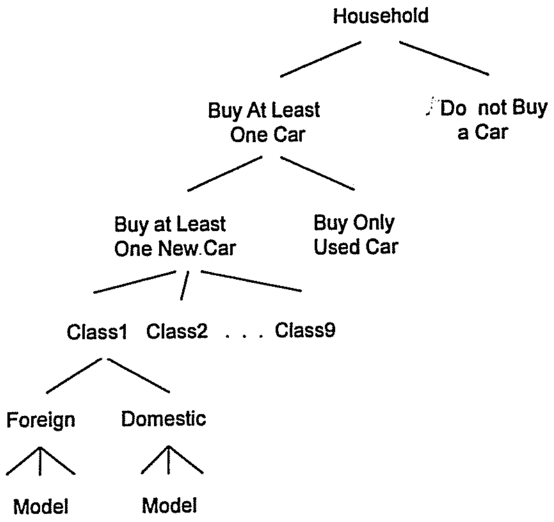


FIGURE 1.—Automobile choice model.

Multinomial Probit with Unrestricted Covariance Matrix

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.

We specify:

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X_{i0}'\beta + \epsilon_{i0} \\ X_{i1}'\beta + \epsilon_{i1} \\ \vdots \\ X_{iJ}'\beta + \epsilon_{iJ} \end{pmatrix} \qquad \epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} |X_i \sim \mathcal{N}(0, \Omega),$$

for some relatively unrestricted $(J+1) \times (J+1)$ covariance matrix Ω (beyond normalizations).

Direct maximization of the log likelihood function is infeasible for more than 3-4 choices.

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.

The Geweke-Hajivasilliou-Keane (GHK) simulator uses a more complicated procedure that draws $\epsilon_{i1}, \ldots, \epsilon_{iJ}$ sequentially and combines the draws with the calculation of univariate normal integrals.

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.

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.

Merits of Unrestriced Multinomial Probit

The attraction of this approach is that there are no restrictions on which choices are close substitutes.

The difficulty, however, with the unrestricted multinomial probit approach is that with a reasonable number of choices there are a large number of parameters: all elements in the $(J+1) \times (J+1)$ dimensional Ω minus some normalizations and symmetry restrictions.

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.

Prediction for new good would require specifying correlations with all other goods.

Random Effects Models

A third possibility to get around the IIA property is to allow for unobserved heterogeneity in the slope coefficients.

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? One argument 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:

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

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.

One possibility to implement this is to assume the existence of a finite number of types of individuals, similar to the finite 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.

Alternatively we could specify

$$\beta_i|Z_i \sim \mathcal{N}(\beta + Z_i'\Gamma, \Sigma),$$

where we use a normal (continuous) mixture of taste parameters.

Using simulation methods or Gibbs sampling with the unobserved β_i as additional unobserved random variables may be an effective way of doing inference.

The models with random coefficients can generate more realistic predictions for new choices (predictions will be dependent on presence of similar choices)

5. Berry-Levinsohn-Pakes

BLP extended the random effects logit models to allow for

1. unobserved product characteristics,
2. endogeneity of choice characteristics,
3. estimation with only aggregate choice data
4. with large numbers of choices.

Their approach has been widely used in Industrial Organization, where it is used to model demand for differentiated products.

The utility is 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.

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

The random coefficients β_i are related to individual observable characteristics:

$$\beta_i = \beta + Z_i'\Gamma + \eta_i, \quad \text{with} \quad \eta_i | Z_i \sim \mathcal{N}(0, \Sigma).$$

The data consist of

- estimated shares $\widehat{s}_{tj}$ for each choice j in each market t,
- 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

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

Now consider for fixed Γ, 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 which we may need to calculate by simulation of choices.

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 all implied market shares are equal to the observed market shares $\widehat{s}_{jt}$.

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 s, and parameters Γ and Σ.

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 these residuals are equal to the unobserved product characteristic ζ_{jt}.

Now we can use GMM given instruments that are orthogonal to these residuals, typically things like characteristics of other products by the same firm, or average characteristics by competing products.

This step is where the method is most challenging. Finding values of the parameters that set the average moments closest to zero can be difficult.

Let us see what this does if we 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.

6. Models with Multiple Unobserved Choice Characteristics

The BLP approach can allow only for a single unobserved choice characteristic. This is essential for their estimation strategy with aggregate data.

With individual level data one may be able to establish the presence of two unobserved product characteristics (invariant across markets). Elrod and Keane (1995), Goettler and Shachar (2001), and Athey and Imbens (2007) study such models.

These models can be viewed as freeing up the covariance matrix of unobserved components relative to the random coefficients model, but using a factor structure instead of a fully unrestricted covariance matrix as in the multinomial probit.

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:

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

Even in the case with all choice characteristics exogenous, maximum likelihood estimation would be difficult (multiple modes). Bayesian methods, and in particular markov-chain-monte-carlo methods are more effective tools for conducting inference in these settings.

7. Hedonic Models

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 small number of unobserved product characteristics and unobserved variation in taste parameters to generate stochastic choices.

Why can it still be useful to include such an ϵ_{ij}?

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 due to a **single** mismeasured observation.

Thus, one might wish to generalize the model to be more robust. One possibility is to related the observed choice Y_i to the optimal choice Y_i^*:

$$\Pr(Y_i = y | Y_i^*, X_i, \nu_i, Z_1, \ldots, Z_J, \zeta_1, \ldots, \zeta_J)$$

$$= \begin{cases} 1 - \delta & \text{if } Y = Y_i^*, \\ \delta/(J-1) & \text{if } Y \neq Y_i^*. \end{cases}$$

This nests the pure characteristics model (by setting $\delta = 0$), without the extreme sensitivity.

However, if the optimal choice Y_i^* is not observed, all of the remaining choices are equally likely.

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.

Second, this model approximately nests the pure characteristics model 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.