Lecture 1: Discrete Choice.

Greene(19.1-19.8)

Sam Kortum - Fall 2002

This Draft: October 23, 2002.

1 Discrete Choice Models

Individuals $i = 1, ..., I$ choose one from the set of alternatives $j = 0, 1, ..., J$. If individual i chooses j then we observe a vector y_i with all elements zero except $y_{ij} = 1$. Individual i gets utility U_{ij} by choosing j, where

$$U_{ij} = g(x_{ij})\varepsilon_{ij}$$

The individual knows x_{ij} and ε_{ij} when making his choice of

$$\arg\max_j \{U_{ij}\}$$

The econometrician observes x_{ij}, but knows only the distribution from which the ε_{ij} are drawn. Hence to the econometrician y_i is the realization of a random vector Y_i.

This model has very wide application, its uses are expanding exponentially. It has been applied to choice of a career, voting choice, choice of a car, and choice of breakfast cereal. It can be nested within a complicated dynamic programming problems. It's use in microeconometrics was pioneered by Daniel McFadden.

Assuming the ε_{ij} are independent across individuals, Y_i is distributed Multinomial $(1, P(x_i))$, where $P_j(x_i) = \Pr[Y_{ij} = 1|x_i]$. (Note that x_i contains any relevant observations about individual i, any information about that individual with respect to any possible choice, and any relevant observations on the choices themselves.) The likelihood function is

$$L(\beta) = \Pr[Y_1 = y_{1,}..., Y_I = y_I|x, \beta] = \prod_{i=1}^{I}\prod_{j=0}^{J} P_j(x_i;\beta)^{y_{ij}}$$

where β is the vector of parameters determining g and the distribution of ε. Note that the parameters are hidden in the choice probabilities $P_j(x_i;\beta)$. The maximum likelihood estimate of the parameter vector is simply the $\arg\max$ of $L(\beta)$. Maximizing the likelihood is a standard computational problem. The interesting part is the expressions for the choice probabilities. We want these probabilities to capture the economics of the problem while not being too hard to compute.

1.1 Extreme Value Distributions

If we assume that the ε_{ij} are i.i.d. with an extreme value distribution, the choice probabilities are very simple. To understand why, it is useful to review the theory of extremal distributions. Keep in mind that the distribution of the maximimum of n draws from $F(x)$ has a distribution $[F(x)]^n$.

Definition (Billingsley): A distribution F is extremal if, for some distribution G, the distribution $[G(a_n x + b_n)]^n$converges in distribution to $F(x)$.

Definition (Billingsley): Distribution functions F and G are of the same type if there exist constants $a > 0$ and b such that $F(ax + b) = G(x)$ for all x.

An extremal distribution is one for which $F(x)$ and $[F(x)]^m$ are of the same type. Note that for large n, $[F(x)]^m$ is close to $G(a_n x + b_n)^{nm}$ while $F(x)$ is close to $G(a_{nm} x + b_{nm})^{nm}$. Roughly speaking, extremal distributions are closed under max (like Poisson is closed under sum and Normal distributions are closed under averaging).

Theorem (Billingsley): The class of extremal distributions is exactly distributions having one of the following 3 types:

$$\begin{aligned} F_1(x) &= e^{-e^{-x}} \\ F_2(x) &= e^{-x^{-\alpha}} \qquad x \geq 0 \\ F_3(x) &= e^{-(-x)^{\alpha}} \qquad x \leq 0 \end{aligned}$$

Note that if X has a distribution of type F_2 then $\ln X$ has a distribution of type F_1 and $-X^{-1}$ has a distribution of type F_3.

1.2 Multinomial Logit

Since I set up utility with multiplicative shocks ε, I want to use F_2 which is closed under multiplication (it is more typical, and equivalent, to set up additive shocks and to use F_1.)

$$\Pr[\varepsilon_{ij} \leq \varepsilon] = F(\varepsilon) = e^{-T\varepsilon^{-\theta}}$$

Thus

$$\Pr\left[U_{ij} \leq u\right] = F(u/g(x_{ij})) = e^{-Tg(x_{ij})^{\theta}u^{-\theta}} = G_{ij}(u)$$

The choice probability is therefore

$$\begin{aligned} P_j(x_i) &= \Pr[U_{ij} > \max_{k \neq j}\{U_{ik}\}] \\ &= \int_0^{\infty} \Pr[\max_{k \neq j}\{U_{ik}\} \leq u] dG_{ij}(u) \end{aligned}$$

Since

$$\Pr[\max_{k \neq j}\{U_{ik}\} \leq u] = \prod_{k \neq j} G_{ik}(u)$$

we get,

$$P_j(x_i) = \int_0^{\infty} e^{-T\left[\sum_{k \neq j} g(x_{ik})^{\theta}\right]u^{-\theta}} dG_{ij}(u) = \frac{g(x_{ij})^{\theta}}{\sum_{k=0}^{J} g(x_{ik})^{\theta}}$$

Note that T is not identified.

Without taking a stand on g, we already see an important property of the choice probabilities. The ratio of two choice probabilities does not depend on anything having to do with any of the other choices:

$$\frac{P_j(x_i)}{P_k(x_i)} = \frac{g(x_{ij})^{\theta}}{g(x_{ik})^{\theta}}$$

This property of the model is called Independence of Irrelevant Alternatives. There is nothing wrong with it, but in some applications it may be an implausible restriction to impose a priori. Suppose that you are deciding between buying a red apple, a green apple, or a donut. The odds of buying a green apple relative to a donut is likely to be a function of whether the red apples are being given away free. If you're going to get an apple you will choose the free one, but you still might want a donut instead. Introspection suggests IIA is inappropriate (atleast as an a priori restriction) in this situation.

A standard assumption in applications is that $g(x_{ij}) = e^{x'_{ij}\gamma}$. In that case it is clear at best only $\alpha = \theta\gamma$ is identified. Furthermore, if some of the elements of x_i are common to all the choices j then we cannot identify common coefficients on these elements. In other words, if $x'_{ij}\alpha = x'_{ij}\beta + \rho_i$ then ρ_i is not identified. Finally, without loss of generality we can measure all of the x's relative to the 0 choice. Taking into account these limits on identification, the choice probabilities become:

$$P_j(x_i) = \frac{e^{(x_{ij}-x_{i0})'\beta}}{1 + \sum_{k=1}^{J} e^{(x_{ik}-x_{i0})'\beta}}$$

We have derived the multinomial logit model.

Theorem(Ruud): The log likelihood function is globally concave in the parameters β.

1.3 Binary Choice

If we set $J = 1$ then there is a binary choice between alternatives 0 and 1. Without ambiguity we can let $y_i = 1$ if the choice is 1 (0 otherwise). The multinomial logit model then reduces to the logit model:

$$E\left[Y_i|x_i\right] = P_1(x_i) = \frac{e^{x'_i\beta}}{1 + e^{x'_i\beta}} = \Lambda(x'_i\beta)$$

where we redefine $x_i = x_{i1} - x_{i0}$. Note that $\Lambda(-\infty) = 0$, $\Lambda'(x) > 0$, $\Lambda(\infty) = 1$, i.e. Λ is just a distribution function (the logistic).

Why not generalize to any distribution $F(x)$ in place of $\Lambda(x)$? The typical derivation is in terms of a latent variable or index function

$$y_i^* = x'_i\beta - \epsilon_i$$

Although y^* is not observed, $Y_i = 1$ if and only if $y_i^* > 0$. The random variable ϵ_i is drawn from F. Thus

$$\begin{aligned} E\left[Y_i|x_i\right] &= \Pr[y_i^* > 0] = \Pr[-\epsilon_i > -x'_i\beta] \\ &= \Pr[\epsilon_i \le x'_i\beta] = F(x'_i\beta). \end{aligned}$$

For $F(x) = \Lambda(x)$ we get the Logit model. For $F(x) = \Phi(x)$ (cumulative Normal) we get the Probit model. For $F(x) = x$ we get the linear probability model. The later can be justified as a uniform distribution except that the constraints on x are typically not imposed. By ignoring these constraints we can simply run an OLS regression. The linear probability model is very helpful to get a quick feel for the data.