Tuesday, October 12. Unordered Choice Models

    I.   Motivating the Model
    II.   Multinomial Logit
    III.   Independence of Irrelevant Alternatives
    IV.   Multinomial Probit

I. Motivating the Model (see Greene, 21.7)

A. When to Use an Unordered Choice Model. In the ordered probit model, we could give a clear objective ordering of the values of the dependent variable from less to more of the concept that we were measuring. When your dependent variable can take on more than two possible values, but you cannot say whether value A denotes more of the concept than value B (and B more than C) for all observation, you have an "unordered choice" model. "Choice" gets into the name because these models are usually used to analyze the choices of individuals over things like candidates, consumer products, or, in the canonical example, transportation options. But the dependent variable does not have to be a choice; it can be just about anything that takes on three or more categories that cannot be given any universal order.

B. The Random Utility Model. This is a behavior model in which individuals chose the option that gives them the most utility, but sometimes they screw up. For the ith individual with j choices, the utility of the jth choice can be given by:

$$U_{ij} = z_{ij}\beta + e_{ij}$$

Here their utility is a function of some set of choice- and individual-specific independent variables z (we'll get to them later), a set of weights β that do not vary across individuals or choices, and an error term. (Note that we are specifying a linear systematic component to the model in this step, because the covariates are linearly related to utility.) Because this individual's choice reveals their preferences, we know that if individual i chose j, then the utility they (think they) get from j is great than their utility from all other options k. To build a pdf for this observation, we need to find the probability that $U_{ij} > U_{ik}$ for all k not equal to j.

$$\begin{aligned}
Pr(Y_i = j) &= Pr(U_{ij} > U_{ik}) \\
&= Pr(z_{ij}\beta + e_{ij} > z_{ik}\beta + e_{ik}) \\
&= Pr[(z_{ij} - z_{ik})\beta > e_{ik} - e_{ij})]
\end{aligned}$$

Here comes the hand waving. To find this probability, we want to find the integral of this expression up to where $z_{ij} - z_{ik} = e_{ik} - e_{ij}$. So we are going to make an assumption about the stochastic distributions of these error terms (the stochastic component of our likelihood model) basic in order to make the difference between these error terms pretty. I know, I know, I promised that all of these models would make an assumption about the distribution of y around its predicted value that would be substantively meaningful and all, but instead we

are going to make an assumption about how e varies around zero over many experiments that is mathematically tractable.

So here's the assumption: the error for each observation is distributed according to the Gumbel distribution (also called the "type I extreme value distribution," but 4 out of 5 methodologists think "Gumbel" is more fun to say). It turns out that if two Gumbel variables are independent, their difference is distributed logistic, and the cdf of a logistic is the logit. So when we make the assumption that our errors are independent and identically distributed as Gumbels, the integral up to where the difference between them is equal to something can be given by the logit function. The following pdf gives the probability that the utility from choice j is greater than the utility from all of the other choices:

$$\Pr(Y_i = j) = \frac{e^{z_{ij}\beta}}{\sum_{k=0}^{J} e^{z_{ik}\beta}}$$

II. Multinomial Logit

A. What are the Covariates? The above model is the basis of economist Dan McFadden's (1973) "conditional logit" model, which he used predict individuals' choices of transportation options. But what are these zs? They can be partitioned into:
    i. x_{ij}, the potentially individual-specific *attributes* of the choices, and
    ii. w_i, the *characteristics* of the individuals. But now notice that

$$\Pr(Y_i = j) = \frac{e^{x_{ij}\beta + \alpha_i w_i}}{\sum_{k=0}^{J} e^{x_{ik}\beta + \alpha_i w_i}} = \frac{e^{x_{ij}\beta} e^{\alpha_i w_i}}{\sum_{k=0}^{J} e^{x_{ik}\beta} e^{\alpha_i w_i}}$$

and $e^{\alpha w}$ the terms drop out! That creates an estimation problem using the basic model. So to estimate the effects of some individual characteristic, create j-1 dummy variables for choices an interact each of them with an attribute variable. This lifts a few restrictions on your model, and stops the terms from dropping out by tying them to a term that depends upon the choice.

Normally, an analysis of a nominal variable where choice is determined by characteristics of individual choosers is called "unconditional multiple logit," while choice that is also determined by individual's perceptions of the attributes of the choices is referred to as "multinomial logit." Some authors such as Greene reserve the term "conditional logit" to use for cases in which choice is determined only by attributes of the choice. Neal Beck's notes show that they are mathematically identical, but that the interpretation is different.

B. Again with the Hand Waving. Another problem with the above model is that if you add some vector q to the vector of coefficients for any choice, you will get the same probabilities for each choice as you would have if you hadn't added q (because all of the q terms drop out of the model). Trust me, or rather, trust Greene (p. 721). To fix this "indeterminacy," we anchor the model with the assumption that the βs for choice 0 are all equal to 0. Restating the model with the βs subscripted (allowing the relationship between

an individual's characteristics and her probability of making a certain choice to vary for each choice), we have

$$\Pr(Y_i = j) = \frac{e^{x_{ij}\beta_j}}{\sum\limits_{k=0}^{J} e^{x_{ik}\beta_k}} \ where \ j = 0,1...J$$

$$= \frac{e^{x_{ij}\beta}}{1 + \sum\limits_{k=1}^{J} e^{x_{ik}\beta}} \ where \ j = 0,2...J, \beta_0 = 0$$

Now it is easy to see that when J =1 (when there are only choices 0 and 1), this expression reduces to the pdf for a logit. In Stata, you are going to have to leave one category (one candidate, one party, etc.) as the baseline case in your mlogit model. When you do that, you are deciding which choice will be your β_0. The coefficients that you estimate are going to tell you how individual characteristics make choice j more (or less) likely, compared to this baseline category. Multiplying one of an individual's characteristics by the estimated coefficient for that characteristic that Stata spits out at you will give you the log odds ratio of making choice j compared to making the baseline choice.

$$x_i\beta_j = \ln[\Pr_{ij}/\ \Pr_{i0}]$$

III. Independence of Irrelevant Alternatives

Life is good when we can justify using the multinomial logit model. But we can only justify using it if we can make the independence of irrelevant alternatives (IIA) assumption. Remember that we assumed that the errors in our random utility model were independent across choices? That's the math version of IIA, but here are a couple of substantive ways to think about it. (Read more on pp. 71-76 of Alvarez and Nagler or p. 724 of Greene, who calls it I from IA).

First, IIA means that the odds ratio above doesn't shift when alternative choices are added. Suppose you looked at the two candidates for mayor of San Diego a couple of weeks ago and said that you were 60/40 for Dick Murphy over Ron Roberts, two old white Republican guys. Then liberal Democrat Donna Frye jumps into the race. Now as a liberal Democrat I'm really excited about Frye, and there's a 75% chance that I'll vote for her, a 15% chance that I'll vote for Murphy, and 10% chance for Ron Roberts. My choice has changed, but notice that adding this new candidate didn't affect the odds ratio of my chances of voting for the other two candidates (1.5 in each case). So for me, and for this added choice, at least, the IIA assumption holds. But what if next week John Moores, another conservative rich white old guy, hops into the race? Maybe I'd be 75% for Frye, 10% for Murphy, 10% for Roberts, and 5% for Moores. Now, my odds ratio has shifted, and the IIA assumption fails. Alvarez and Nagler argue that IIA generally fails when you have spatial voting and, for instance, two conservative candidates and one liberal.

Second, when you have a nominal variable that is not really a choice, the IIA assumption says that there are no factors affecting the probability that an observation falls into one category that: a. are left unexplained in your model of this outcome, and b. also influence the chances that an observation ends up in a different category.

Hausman and McFadden propose a way to test whether the IIA assumption holds. You can estimate a model that omits the subset of choices that you think are irrelevant, and then check to see if this systematically changes your parameter estimates for the other choices. You hope that your βs are invariant to the set of choices available. I have not seen this test used in political science. Typically, people discuss the implications of the IIA assumption for their research question, making a substantive argument for whether or not IIA holds.

IV. Multinomial Probit

A. **What if you are really sure that IIA does not hold?** Another maximum likelihood model that relaxes the assumption that the error terms are independent across choices (and thus relaxes IIA) is the multinomial probit model. Instead of assuming that the errors are independent Gumbels, you assume that they are distributed according to a multivariate Normal distribution. The MVN distribution has as many variables as there are J choices minus one. With two choices, it reduces to a univariate probit model.

B. **The limits of multinomial probit**. With three choices, you have a bivariate Normal distribution. This is somewhat hard to actually compute, but not terribly. Alvarez and Nagler wrote a simulation program to take draws from this distribution, but it is not canned in Stata yet. If you have four choices, you've got draws from a three dimensional multivariate Normal, and as far as I know this is too hard for our computers. Alternative ML models with slightly different substantive assumptions are the generalized extreme value model and nested logit.