Lecture Notes 7/24/2012

Discrete Choice

We have $J+1$ mutually-exclusive choices indexed by $j=0,\ldots J$. We only observe the choice that was made, not the utility arising from each choice. Model:

$$y_{ij}=x_i'\beta_j+\varepsilon_{ij}$$

in vector form:

$$y_j=X\beta_j+\varepsilon_j$$

We want to estimate β because it has some sort of economic significance: e.g. demand for good j, utility of working in occupation j, utility of living in neighborhood j; etc.

Estimation procedure for recovering β depends on what assumptions we make on the J-dimensional error term, ε_i.

Binomial Logit

Let's assume that $\varepsilon_{ij}\overset{iid}{\sim}$Type I Extreme Value (also sometimes referred to as the Gumbel distribution). The CDF for this distribution is $F(x)=e^{-e^{-x}}$. It is noteworthy that the difference in two TIEV random variables is distributed logistic; i.e. if X and Y are both distributed Gumbel, then $X-Y$ is distributed Logistic with CDF $F(x)=\frac{1}{1+e^{-x}}=\frac{e^x}{1+e^x}$. Now let's write out our system of equations:

$$\begin{aligned} y_{i0} &= X_i\beta_0+\varepsilon_{i0} \\ y_{i1} &= X_i\beta_1+\varepsilon_{i1} \end{aligned}$$

where $\varepsilon_{ij}\overset{iid}{\sim}$Type I Extreme Value. Now let's take differences between the two equations:

$$\begin{aligned} y_{i1}-y_{i0} &= X_i(\beta_1-\beta_0)+\varepsilon_{i1}-\varepsilon_{i0} \qquad (1) \\ \tilde{y}_i &= X_i\tilde{\beta}+\tilde{\varepsilon}_i \end{aligned}$$

where $\tilde{\varepsilon}_i$ is now distributed iid Logistic. The choice probabilities are now

$$\begin{aligned}
\Pr(\tilde{y}_i = 1) = P_{i1} &= \frac{\exp\left(X_i\tilde{\beta}\right)}{1+\exp\left(X_i\tilde{\beta}\right)} \\
&= \frac{\exp\left(X_i\left(\beta_1-\beta_0\right)\right)}{1+\exp\left(X_i\left(\beta_1-\beta_0\right)\right)} \\
\Pr(\tilde{y}_i = 0) = P_{i0} &= \frac{1}{1+\exp\left(X_i\tilde{\beta}\right)} \\
&= \frac{1}{1+\exp\left(X_i\left(\beta_1-\beta_0\right)\right)}.
\end{aligned}$$

Now let's multiply and divide by $\exp(X_i\beta_0)$:

$$\begin{aligned}
P_{i1} &= \frac{\exp\left(X_i\left(\beta_1-\beta_0\right)\right)\exp\left(X_i\beta_0\right)}{\left[1+\exp\left(X_i\left(\beta_1-\beta_0\right)\right)\right]\exp\left(X_i\beta_0\right)} \\
&= \frac{\exp\left(X_i\beta_1-X_i\beta_0\right)\exp\left(X_i\beta_0\right)}{\exp\left(X_i\beta_0\right)+\exp\left(X_i\beta_1-X_i\beta_0\right)\exp\left(X_i\beta_0\right)}.
\end{aligned}$$

Now we can rewrite this formula using the property of exponentials that $e^{x-y} = \frac{e^x}{e^y}$:

$$P_{i1} = \frac{\frac{\exp(X_i\beta_1)}{\exp(X_i\beta_0)}\exp\left(X_i\beta_0\right)}{\exp\left(X_i\beta_0\right)+\frac{\exp(X_i\beta_1)}{\exp(X_i\beta_0)}\exp\left(X_i\beta_0\right)}$$

and simplifying, we get

$$P_{i1} = \frac{\exp\left(X_i\beta_1\right)}{\exp\left(X_i\beta_0\right)+\exp\left(X_i\beta_1\right)}.$$

We can do something similar for the formula of P_{i0} to get

$$P_{i0} = \frac{\exp\left(X_i\beta_0\right)}{\exp\left(X_i\beta_0\right)+\exp\left(X_i\beta_1\right)}.$$

In general, for any multinomial choice model where $\varepsilon_{ij} \overset{iid}{\sim}$Type I Extreme Value, the formula for the probability of making choice j is

$$P_{ij} = \frac{\exp\left(X_i\beta_j\right)}{\sum_k \exp\left(X_i\beta_k\right)}. \tag{2}$$

For a more rigorous (i.e. an actual) proof of this, see Train pp. 36, 74-75.

Normalizations

Recall that, for the classic logit formula,

$$P_{i0} = \frac{1}{1+\exp\left(X_i\beta_1\right)}, \tag{3}$$

not

$$P_{i0} = \frac{\exp(X_i\beta_0)}{\exp(X_i\beta_0) + \exp(X_i\beta_1)}. \tag{4}$$

Note that (3) is equivalent to (4) if we set $\beta_0 = 0$. Indeed, we can only ever identify $\beta_1 - \beta_0$ (see (1) for intuition). Therefore, we need to set the scale of this difference, typically by setting $\beta_0 = 0$. We choose this because we still want to be able to interpret the β_1 parameters in a meaningful way, and when $\beta_0 = 0$ then $\tilde{\beta} = \beta_1$.

Scaling our parameters in this way is known as making a *location normalization* or setting the *level* of utility. In general, we also need to normalize the *scale* of the model by setting the variance of one of the ε_{ij} to be 1, for example. However, in the logit model this is not necessary because the standardized Type I Extreme Value distribution already has the variance scaled (where variance is equal to $\pi^2/6$).

In the multinomial logit, we need to set one of the β vectors to be zero (usually this is β_0 or β_J).

Why do we need to normalize?

If we tried doing MLE on the logit model without normalizing the location (i.e. setting $\beta_0 = 0$), we would find that Matlab would be able to arbitrarily set β_0 to any value and β_1 to any other value, with the only constraint being that $P_{i0} + P_{i1} = 1$. We would quickly discover that the MLE would never converge, because any combination of β_0 and β_1 arbitrarily satisfies the model. Hence, if we don't normalize the location, then we can't separately estimate β_0 and β_1. This is what is meant by *identification*.

Interpreting in the face of normalizations

Recall that interpretation of coefficients β_1 in the classic logit model is always relative to the option that is $y = 0$. For example, if we have a model where y is 1 if the individual chooses to retire and 0 if the individual chooses to stay in the workforce, then β_1 is always interpreted relative to β_0. In other words, if we estimated a positive coefficient on the variable *male*, we would conclude that males are *more likely to retire* than females. That's the most we can say about our model estimates.

Probit

With probit, we assume that ε_i is distributed $N(0,\Sigma)$, where ε_i is a $J+1$-dimensional vector of choice-specific error terms. Analytical derivation of the multinomial probit choice probabilities is impossible because there is no closed form solution for them. The general formula for a multinomial probit choice probability is a $J+1$-dimensional integral:

$$P_{ij} = \int \cdots \int I\left[X_i\beta_j + \varepsilon_{ij} > X_i\beta_k + \varepsilon_{ik} \, \forall k \neq j\right] \phi(\varepsilon_i)\, d\varepsilon_i \tag{5}$$

where $I[\cdot]$ is an indicator function, and $\phi(\varepsilon_i)$ is the PDF of the multivariate normal distribution with mean vector 0 and covariance matrix Σ.

$$\phi(\varepsilon_i) = \frac{1}{(2\pi)^{(J+1)/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\varepsilon_i'\Sigma^{-1}\varepsilon_i\right).$$

Just as with the logit, one of the β_k needs to be set equal to 0, and additionally one of the variances of ε_k needs to be set to 1.

Maximum Likelihood

Maximum likelihood for discrete choice models is fairly straightforward. Recall that for the binomial case, we used the PDF of the Bernoulli distribution as our functional form for the likelihood: $f(y;p) = p^y (1-p)^{1-y}$. In a Bernoulli distribution, y can only take on values $\{0,1\}$.

In the multinomial case, we use a "multivariate" Bernoulli, which is called the Categorical distribution. In this distribution, y can take on values $\{0,1,\ldots J\}$. The PDF of this distribution is $f(y;p_0,\ldots,p_J) = p_0^{I[y=0]} p_1^{I[y=1]} \cdots p_J^{I[y=J]}$, where $I[y=J]$ is an indicator function that is 1 if $y=J$ and 0 otherwise.

The likelihood function for a multinomial choice model is then

$$\begin{aligned}
\mathscr{L}(y,X;\beta) &= \prod_i p_{i0}^{d_{i0}} p_{i1}^{d_{i1}} \cdots p_{iJ}^{d_{iJ}} \\
&= \prod_i \prod_j p_{ij}^{d_{ij}} \\
\log(\mathscr{L}(y,X;\beta)) = \ell(y,X;\beta) &= \sum_i \sum_j d_{ij} \log\left(p_{ij}\right)
\end{aligned}$$

where $d_{ij} = I[y_i = j]$.

The formula for p_{ij} is then either (2) if Gumbel distribution is assumed, or the ugly integral formula (5) if Normal distribution is assumed.