Static Discrete Choice Models

Gabriel Petrini

September 3st, 2020

Contents

1 Pre-class reading: train2009

1.1 Reading Quiz

1. The most widely used discrete choice model is:
   - GEV
   - probit
   - logit
   - mixed logit
2. List the three properties that the choice set in a discrete choice model must satisfy
3. Describe the most common interpretation that is given to the distribution of (epsilon)

4. Assess the validity of the following claim: utility-maximizing behavior can be expressed by a random utility model (RUM), but not all RUMs express utility-maximizing behavior

5. Why do you think it is called a Random Utility Model?

2 What are discrete choice models?

Discrete choice models are one of the workhorses of structural economics and are deeply tied to economic theory:

- utility maximization
- revealed preference
- Used to model "utility" (broadly defined), for example:
  - consumer product purchase decisions
  - firm market entry decisions
  - investment decisions

2.1 Properties of discrete choice models

1. Agents choose from among a **finite** set of alternatives (called the choice set)
2. Alternatives in choice set are **mutually exclusive**
3. Choice set is **exhaustive**

2.2 Notation

Let d_i indicate the choice individual i makes where $d_i \in \{1, \ldots, J\}$.

- Individuals choose d to maximize their utility, U, which generally is written as:

$$U_{ij} = u_{ij} + \epsilon_{ij} \tag{1}$$

where:

- u_{ij} relates observed factors to the utility individual i receives from choosing option j
- ϵ_{ij} are unobserved to the econometrician but observed to the individual

$$d_{ij} = 1 \text{ if } u_{ij} + \epsilon_{ij} > u_{ij'} + \epsilon_{ij'} \text{ for all } j' \neq j \tag{2}$$

2.3 Probabilities

With the ϵ's unobserved, the probability of i making choice j is given by:

$$\begin{aligned} P_{ij} &= \Pr(u_{ij} + \epsilon_{ij} > u_{ij'} + \epsilon_{ij'} \ \forall\ j' \neq j) \\ &= \Pr(\epsilon_{ij'} - \epsilon_{ij} < u_{ij} - u_{ij'} \ \forall\ j' \neq j) \\ &= \int_{\epsilon} I(\epsilon_{ij'} - \epsilon_{ij} < u_{ij} - u_{ij'} \ \forall\ j' \neq j) f(\epsilon) d\epsilon \end{aligned}$$

- From the researcher perspective, the choice is probabilistic
- There are some assumptions about the distribution of $f(\epsilon)$ that is needed to be made

Note that, regardless of what distributional assumptions are made on the ϵ's, the probability of choosing a particular option does not change when we:

- Add a constant to the utility of all options (i.e. **only differences in utility matter**)
- Multiply by a positive number (need to **scale something**; e.g. the variance of the ϵ's)

2.4 Variables

Suppose we have (observable by the econometrician):

$$\begin{aligned} u_{i1} &= \alpha Male_i + \beta_1 X_i + \gamma Z_1 \\ u_{i2} &= \alpha Male_i + \beta_2 X_i + \gamma Z_2 \end{aligned}$$

Since only differences in utility matter:

$$u_{i1} - u_{i2} = (\beta_1 - \beta_2) X_i + \gamma (Z_1 - Z_2)$$

- We can't tell whether men are happier than women, but can tell whether they more strongly prefer one option
  - αMale_i drops out (is not observable)
- We can only obtain differenced coefficient estimates on X's
- We can only obtain an estimate of a coefficient that is constant across choices if its corresponding variable varies by choice

2.5 Number of Error Terms

Similar to the X's, there are restrictions on the number of error terms

- This is because only differences in utility matter

Recall that he probability i will choose j is given by:

$$\begin{aligned} P_{ij} &= \Pr(u_{ij} + \epsilon_{ij} > u_{ij'} + \epsilon_{ij'} \ \forall\ j' \neq j) \\ &= \Pr(\epsilon_{ij'} - \epsilon_{ij} < u_{ij} - u_{ij'} \ \forall\ j' \neq j) \\ &= \int_{\epsilon} I(\epsilon_{ij'} - \epsilon_{ij} < u_{ij} - u_{ij'} \ \forall\ j' \neq j) f(\epsilon) d\epsilon \end{aligned} \tag{3}$$

where the integral is J dimensional

Rewriting the last line of (3) as a $J-1$ dimensional integral over the differenced ϵ's:

$$P_{ij} = \int_{\tilde{\epsilon}} I(\tilde{\epsilon}_{ij'} < \tilde{u}_{ij'} \ \forall\ j' \neq j) g(\tilde{\epsilon}) d\tilde{\epsilon} \tag{4}$$

- This means one dimension of $f(\epsilon)$ is not identified and must therefore be normalized
- Arises from only differences in utility mattering (**location normalization**)
- The scale of utility also doesn't matter (**scale normalization**)
  - The scale normalization implies we must place restrictions on the variance of ϵ's
  - This mean that is impossible to compare across estimations

2.5.1 More on the scale normalization

The need to normalize scale means that we can never estimate the variance of $F(\tilde{\epsilon})$

- This contrasts with linear regression models, where we can easily estimate MSE
- The scale normalization means our β's and γ's are implicitly divided by an unknown variance term:

$$\begin{aligned}
u_{i1} - u_{i2} &= (\beta_1 - \beta_2)X_i + \gamma(Z_1 - Z_2) \\
&= \tilde{\beta}X_i + \gamma\tilde{Z} \\
&= \frac{\beta^*}{\sigma}X_i + \frac{\gamma^*}{\sigma}\tilde{Z}
\end{aligned}$$

- $\tilde{\beta}$ is what we estimate, but we will never know β^* because utility is scale-invariant

3 Logit Derivation

3.1 Where does the logit formula come from?

Consider a binary choice set $\{1, 2\}$. The Type 1 extreme value CDF for ϵ_2 is:

$$F(\epsilon_2) = e^{-e^{(-\epsilon_2)}}$$

To get the probability of choosing 1, substitute in for ϵ_2 with $\epsilon_1 + u_1 - u_2$:

$$\Pr(d_1 = 1|\epsilon_1) = e^{-e^{-(\epsilon_1 + u_1 - u_2)}} \tag{5}$$

- But ϵ_1 is unobserved so we need to integrate it out

Taking the integral over what is random (ϵ_1):

$$\begin{aligned}
\Pr(d_1 = 1) &= \int_{-\infty}^{\infty} \overbrace{\left(e^{-e^{-(\epsilon_1 + u_1 - u_2)}}\right)}^{p.d.f} \overbrace{f(\epsilon_1)}^{\text{dist}} d\epsilon_1 \\
&= \int_{-\infty}^{\infty} \left(e^{-e^{-(\epsilon_1 + u_1 - u_2)}}\right) e^{-\epsilon_1} e^{-e^{-\epsilon_1}} d\epsilon_1 \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\epsilon_1} - e^{-(\epsilon_1 + u_1 - u_2)}\right) e^{-\epsilon_1} d\epsilon_1 \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\epsilon_1}\left[1 + e^{u_2 - u_1}\right]\right) e^{-\epsilon_1} d\epsilon_1
\end{aligned}$$

- We can simplify by U-substitution where $t = \exp(-\epsilon_1)$ and $dt = -\exp(-\epsilon_1)d\epsilon_1$

- And adjusting the bounds of integration accordingly, $\exp(-\infty) = 0$ and $\exp(\infty) = \infty$

Substituting in then yields:

$$\begin{aligned}
\Pr(d_1 = 1) &= \int_{\infty}^{0} \exp\left(-t\left[1 + e^{(u_2 - u_1)}\right]\right)(-dt) \\
&= \int_{0}^{\infty} \exp\left(-t\left[1 + e^{(u_2 - u_1)}\right]\right) dt \\
&= \left.\frac{\exp\left(-t\left[1 + e^{(u_2 - u_1)}\right]\right)}{-\left[1 + e^{(u_2 - u_1)}\right]}\right|_{0}^{\infty} \\
&= 0 - \frac{1}{-\left[1 + e^{(u_2 - u_1)}\right]} \\
&= \frac{\exp(u_1)}{\exp(u_1) + \exp(u_2)}
\end{aligned}$$

Consider our model from before:

$$u_{i1} - u_{i2} = (\beta_1 - \beta_2)X_i + \gamma(Z_1 - Z_2)$$

- We observe X_i, Z_1, Z_2, and d_i
- Assuming $\epsilon_1, \epsilon_2 \overset{iid}{\sim} T1EV$ gives the likelihood of choosing 1 and 2 respectively as:

$$\begin{aligned}
P_{i1} &= \frac{\exp(u_{i1} - u_{i2})}{1 + \exp(u_{i1} - u_{i2})} \\
P_{i2} &= \frac{1}{1 + \exp(u_{i1} - u_{i2})}
\end{aligned}$$

- Note: if $\epsilon_1, \epsilon_2 \overset{iid}{\sim} T1EV$ then $\tilde{\epsilon}_1 \sim Logistic$, where $\tilde{\epsilon}_1 := \epsilon_1 - \epsilon_2$

3.2 Likelihood function

We can view the event $d_i = j$ as a weighted coin flip

- This gives us a random variable that follows the Bernoulli distribution
- Supposing our sample is of size N, the likelihood function would then be

$$\begin{aligned}
\mathcal{L}(X, Z; \beta, \gamma) &= \prod_{i=1}^{N} P_{i1}^{d_{i1}} P_{i2}^{d_{i2}} \\
&= \prod_{i=1}^{N} P_{i1}^{d_{i1}} \left[1 - P_{i1}\right]^{(1 - d_{i1})} \qquad (6)
\end{aligned}$$

where P_{i1} and P_{i2} are both functions of X, Z, β, γ
For many reasons, it's better to maximize the log likelihood function

- Taking the log of (6) gives

$$
\begin{aligned}
\ell\left(X, Z ; \beta, \gamma\right) =& \sum_{i=1}^{N} d_{i1} \log P_{i1} + \left(1 - d_{i1}\right) \log\left(1 - P_{i1}\right) \\
=& \sum_{i=1}^{N} \sum_{j=1}^{2} d_{ij} \log P_{ij} \\
=& \sum_{i=1}^{N} d_{i1} \left[\log\left(\exp(u_{i1} - u_{i2})\right) - \log\left(1 + \exp(u_{i1} - u_{i2})\right)\right] + \\
& \left(1 - d_{i1}\right)\left[\log\left(1\right) - \log\left(1 + \exp(u_{i1} - u_{i2})\right)\right] \\
=& \sum_{i=1}^{N} d_{i1}\left[u_{i1} - u_{i2}\right] - \log\left(1 + \exp(u_{i1} - u_{i2})\right)
\end{aligned}
\tag{7}
$$

3.3 Multinomial Logit Estimation

Adding more choices with i.i.d. Type I extreme value errors yields the **multinomial logit**

- Normalizing with respect to alternative J we have (for $j \in \{1, \ldots, J-1\}$)

$$
u_{ij} - u_{iJ} = (\beta_j - \beta_J) X_i + \gamma (Z_j - Z_J) \tag{8}
$$

We observe $X_i, Z_1, \ldots, Z_J$, and d_i. The likelihood of choosing j and J respectively is:

$$
P_{ij} = \frac{\exp(u_{ij} - u_{iJ})}{1 + \sum_{j'=1}^{J-1} \exp(u_{ij'} - u_{iJ})}, \qquad P_{iJ} = \frac{1}{1 + \sum_{j'=1}^{J-1} \exp(u_{ij'} - u_J)} \tag{9}
$$

The log likelihood function we maximize is then:

$$
\ell(X, Z; \beta, \gamma) = \sum_{i=1}^{N} \left[\sum_{j=1}^{J-1} (d_{ij} = 1)(u_{ij} - u_{iJ})\right] - \ln\left(1 + \sum_{j'=1}^{J-1} \exp(u_{ij'} - u_{iJ})\right) \tag{10}
$$

4 Independence of Irrelevant Alternatives (IIA)

One of the properties of the multinomial logit model is **IIA**

- P_{ij}/P_{ik} does not depend upon what other alternatives are available:

$$
\begin{aligned}
\frac{P_{ij}}{P_{ik}} &= \frac{e^{u_{ij}} / \sum_{j'} e^{u_{ij'}}}{e^{u_{ik}} / \sum_{j'} e^{u_{ij'}}} \\
&= \frac{e^{u_{ij}}}{e^{u_{ik}}} \\
&= e^{u_{ij} - u_{ik}}
\end{aligned}
$$

4.1 Advantage of IIA

IIA can simplify estimation. Instead of using as our likelihood

$$P_{ij} = \frac{\exp(u_{ij})}{\sum_{j'}^{J} \exp(u_{ij'})}, \tag{11}$$

we can use the conditional likelihood $P_i(j|j \in K)$ where $K < J$.

- The log likelihood function is then (**beggGray1984**):

$$\begin{aligned} L(\beta, \gamma | d_i \in K) = \sum_{i=1}^{N} \left[\sum_{j=1}^{K-1} (d_{ij} = 1)(u_{ij} - u_{iK}) \right] \\ - \ln \left(1 + \sum_{j'=1}^{K} \exp(u_{ij'} - u_{iK}) \right) \end{aligned}$$

4.2 Disadvantage of IIA

Most famously illustrated by the "red bus/blue bus problem"

- Consider a commuter with the choice set $\{\text{ride a blue-colored bus}, \text{drive a car}\}$
- Now add a red-colored bus to the choice set
- Assume that the only difference in utility between a red bus and a blue bus is in ϵ
- This will **double** the probability of taking a bus
- Why? $P(\text{blue bus})/P(\text{car})$ does not depend upon whether the red bus is available

5 Expected Utility

It is possible to move from the estimates of the utility function to expected utility

- (or at least differences in expected utility)
- Individual i is going to choose the best alternative
- Thus, expected utility from the best choice, V_i, is given by:

$$V_i = E \max_j (u_{ij} + \epsilon_{ij})$$

where the expectation is over all possible values of ϵ_{ij}

For the multinomial logit, this has a closed form:

$$V_i = \ln \left(\sum_{j=1}^{J} \exp u_{ij} \right) + C \tag{12}$$

where C is Euler's constant (a.k.a. Euler-Mascheroni constant)

- We will use this later when we discuss dynamic discrete choice models

5.1 Alternative expression for expected utility

Note that we can also express V_i as:

$$
\begin{aligned}
V_i &= \ln\left(\sum_{j=1}^{J} \frac{\exp(u_{iJ})\exp(u_{ij})}{\exp(u_{iJ})}\right) + C \\
&= \ln\left(\sum_{j=1}^{J} \exp(u_{ij} - u_{iJ})\right) + u_{iJ} + C \\
&= \ln\left(1 + \sum_{j=1}^{J-1} \exp(u_{ij} - u_{iJ})\right) + u_{iJ} + C
\end{aligned}
$$

- This representation will become useful later in the course

5.2 From Expected Utility to Consumer Surplus

We may want to transform utility into dollars to get consumer surplus

- We need something in the utility function (such as price) that is measured in dollars
- Suppose $u_{ij} = \beta_j X_i + \gamma Z_j - \delta p_j$
- The coefficient on price, δ then gives the utils-to-dollars conversion:

$$
E(CS_i) = \frac{1}{\delta}\left[\ln\left(\sum_{j=1}^{J} \exp u_{ij}\right) + C\right] \tag{13}
$$

- We can calculate the change in consumer surplus after a policy change as $E(CS_{i2}) - E(CS_{i1})$ where the C's cancel out