Structural Models of Utility Maximization

Tyler Ransom

University of Oklahoma, Dept. of Economics

April 17, 2018

Outline

Today's plan

1. Describe static discrete choice models
2. How do they fit in with other data science models we've talked about in this class?
3. Derive logit/probit probabilities from intermediate microeconomic theory
4. Go through examples of how to estimate
5. How discrete choice models relate to sample selection bias

Note: These slides are based on the introductory lecture of a PhD course taught at Duke University by Peter Arcidiacono, and are used with permission. That course is based on Kenneth Train's book *Discrete Choice Methods with Simulation*, which is freely available here (PDF).

Outline

What are discrete choice models?

- Discrete choice models are one of the workhorses of structural economics
- 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

Why use discrete choice models?

- Provides link between human optimization behavior and economic theory
- Parameters of these models map directly to economic theory
- Parameter values can quantify a particular policy
- Can be used to form counterfactual predictions (e.g. by adjusting certain parameter values)
- Allows a research to quantify “tastes”

Why not use discrete choice models?

- They're not the best predictive models
  - Trade-off between out-of-sample prediction and counterfactual prediction
- You don't want to form counterfactual predictions, you just want to be able to predict handwritten digits
- You aren't interested in economic theory
- The math really scares you
- You don't like making assumptions
  - e.g. that decision-makers are rational

Example of a discrete choice model

- Cities in the Bay Area are interested in how the introduction of rideshare services will impact ridership on Bay Area Rapid Transit (BART)
- Questions that cities need to know the answers to:
    - Is rideshare a substitute for public transit or a complement?
    - How inelastic is demand for BART? Should fares be $\uparrow$ or $\downarrow$?
    - Should BART services be scaled up to compete with rideshares?
    - Will the influx of rideshare vehicles increase traffic congestion / pollution?
- Each of these questions requires making a counterfactual prediction
- In particular, need a way to make such a prediction confidently and in a way that is easy to understand

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**

Example illustrating these properties

- In San Francisco, people can commute to work by the following (and *only* the following) methods:
  - Drive a personal vehicle (incl. motorcycle)
  - Carpool in a personal vehicle
  - Use taxi/rideshare service (incl. Uber, Lyft, UberPool, LyftLine, etc.)
  - BART (bus, train, or both)
  - Bicycle
  - Walk

Outline

Mathematically representing utility

Let d_i indicate the choice individual (or decision-maker) i makes where $d_i \in \{1, \cdots, J\}$. Individuals choose d to maximize their utility, U. U generally is written as:

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

where:

1. u_{ij} relates observed factors to the utility individual i receives from choosing option j
2. ε_{ij} are unobserved to the researcher but observed to the individual
3. $d_{ij} = 1$ if $u_{ij} + \varepsilon_{ij} > u_{ij'} + \varepsilon_{ij'}$ for all $j' \neq j$

Breakdown of the assumptions

- Examples of what's in ε
  - Person's mental state when making the decision
  - Choices of friends or relatives (maybe, depends on the data)
  - $\vdots$
  - Anything else about the person that is not in our data
- Reasonable to assume additive separability?
  - This is a big assumption: that there are no interactive effects between unobservable and observable factors
  - This results in linear separation regions and may be too restrictive
  - For now, go with it, and remember that there are no free lunches

Probabilistic choice

With the ε's unobserved, we must consider choices as probabilistic instead of certain. The Probability that i chooses alternative j is:

$$\begin{aligned} P_{ij} &= \Pr(u_{ij} + \varepsilon_{ij} > u_{ij'} + \varepsilon_{ij'} \ \forall\ j' \neq j) && (2) \\ &= \Pr(\varepsilon_{ij'} - \varepsilon_{ij} < u_{ij} - u_{ij'} \ \forall\ j' \neq j) && (3) \\ &= \int_{\varepsilon} I(\varepsilon_{ij'} - \varepsilon_{ij} < u_{ij} - u_{ij'} \ \forall\ j' \neq j) f(\varepsilon) d\varepsilon && (4) \end{aligned}$$

Transformations of utility

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:

1. Add a constant to the utility of all options (utility is relative to one of the options, only differences in utility matter)
2. Multiply by a positive number (need to scale something, generally the variance of the ε's)

This is just like in consumer choice theory: utility is ordinal, and so is invariant to the above two transformations

Variables

Suppose we have:

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

Since only differences in utility matter:

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

- Thus, we cannot tell whether men are happier than women, but can tell whether men have a preference for a particular option over another.
- We can only obtain **differenced** coefficient estimates on X's, and can obtain an estimate of a coefficient that is constant across choices only if the variable it is multiplying varies by choice.

Number of error terms

Similar to socio-demographic characteristics, there are restrictions on the number of error terms. Recall that he probability i will choose j is given by:

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

where the integral is J-dimensional.

Number of error terms (cont'd)

But we can rewrite the last line as $J-1$ dimensional integral over the differenced ε's:

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

Note that this means one dimension of $f(\varepsilon)$ is not identified and must therefore be normalized.

Derivation of Logit Probability

Consider the case when the choice set is $\{1, 2\}$. The Type 1 extreme value cdf for ε_2 is:

$$F(\varepsilon_2) = e^{-e^{(-\varepsilon_2)}}$$

To get the probability of choosing 1, substitute in for ε_2 with $\varepsilon_1 + u_1 - u_2$:

$$Pr(d_1 = 1|\varepsilon_1) = e^{-e^{-(\varepsilon_1+u_1-u_2)}}$$

But ε_1 is unobserved so we need to integrate it out (see Appendix to these slides if you want the math steps)

Derivation of Logit Probability

In the end, we can show that, for any model where there are two choice alternatives and ε is drawn from the Type 1 extreme value distribution,

$$P_{i1} = \frac{\exp(u_{i1} - u_{i2})}{1 + \exp(u_{i1} - u_{i2})}, P_{i2} = \frac{1}{1 + \exp(u_{i1} - u_{i2})}$$

Suppose we have a data set with N observations. The log likelihood function we maximize is then:

$$\ell(\beta, \gamma) = \sum_{i=1}^{N} (d_{i1} = 1)(u_{i1} - u_{i2}) - \ln\left(1 + \exp(u_{i1} - u_{i2})\right)$$

Derivation of Probit Probability

In the probit model, we assume that ε is Normally distributed. So for a binary choice we have:

$$P_{i1} = \Phi\left(u_{i1} - u_{i2}\right), P_{i2} = 1 - \Phi\left(u_{i1} - u_{i2}\right)$$

where $\Phi\left(\cdot\right)$ is the standard normal cdf
The log likelihood function we maximize is then:

$$\ell(\beta,\gamma) = \sum_{i=1}^{N}(d_{i1} = 1)\ln\left(\Phi\left(u_{i1} - u_{i2}\right)\right) + (d_{i2} = 1)\ln\left(1 - \Phi\left(u_{i1} - u_{i2}\right)\right)$$

Pros & Cons of Logit & Probit

Logit model:

- Has a much simpler objective function
- Is by far most popular
- ... but has more restrictive assumptions about how people substitute choices
- (this is known as the Independence of Irrelevant Alternatives or IIA assumption)

Probit model:

- Much more difficult to estimate
- ... but can accommodate more realistic choice patterns

Outline

Estimation in R

The R function `glm` is the easiest way to estimate a binomial logit or probit model:

```
library(mlogit)
data(Heating) # load data on residential heating choice in CA
levels(Heating$depvar) <- c("gas","gas","elec","elec","elec")
estim <- glm(depvar ~ income+agehed+rooms+region,
             family=binomial(link='logit'),data=Heating))
print(summary(estim))
```

Interpreting the coefficients

Estimated coefficients using the code in the previous slide:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.599142   0.252032  -2.377   0.0174 *
income        0.015579   0.027905   0.558   0.5766
agehed       -0.006535   0.003342  -1.955   0.0505 .
rooms         0.024291   0.026916   0.902   0.3668
regionscostl -0.053096   0.126665  -0.419   0.6751
regionmountn  0.041827   0.169787   0.246   0.8054
regionncostl -0.219136   0.137692  -1.591   0.1115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the coefficients

- Positive coefficients ⇒ household more likely to choose the non-baseline alternative (in this case: electric)
  - Whatever the first `level` of the factor dependent variable is will be the "baseline" alternative
- Negative coefficients imply the reverse
- Coefficients **not** linked to changes in probability of choosing the alternative (since probability is a nonlinear function of X)

Forming predictions

To get predicted probabilities for each observation in the data:

```r
Heating$predLogit <- predict(estim, newdata = Heating, type = "response")
print(summary(Heating$predLogit))
```

Estimating a probit model

For the probit model, we repeat the same code, except change the "link" function from "logit" to "probit"

```
estim2 <- glm(depvar ~ income+agehed+rooms+region,
             family=binomial(link='probit'),data=Heating))
print(summary(estim2))
Heating$predProbit <- predict(estim2, newdata = Heating, type = "response
    ")
print(summary(Heating$predProbit))
```

A simple counterfactual simulation

- We talked a lot about doing counterfactual comparisons, but how do we *actually* do it?
- Let's show how to do this on a previous example. Suppose that we introduce a policy that makes richer people more likely to use electric heating.
- Mathematically, what does this look like?
- It would correspond to an increase in the parameter in front of *income* in our regression

A simple counterfactual simulation

- Suppose the coefficient increased by a factor of 4. What is the new share of gas vs. electricity usage?

```
estim$coefficients["income"] <- 4*estim$coefficients["income"]
Heating$predLogitCfl <- predict(estim, newdata = Heating, type = "
    response")
print(summary(Heating$predLogitCfl))
```

This policy would increase electric usage by 7 percentage points (from 22% to 29%)

Outline

Discrete choice models and sample selection bias

- Discrete choice models are common tools used to evaluate sample selection bias
- Why? Because variables that are MNAR can be thought of as following a utility-maximizing process
- Examples:
    - Suppose you want to know what the returns to schooling are, but you only observe wages for those who currently hold jobs
    - As a result, your estimate of the returns to schooling might be invalidated by the non-randomness of the sample of people who are currently working
    - How to get around this? Use a discrete choice model (This was the problem we ran into in PS7, if you recall)

Heckman selection correction

The Heckman selection model specifies two equations:

$$u_i = \beta x_i + \nu$$
$$y_i = \gamma z_i + \varepsilon$$

The first equation is a utility maximization problem, determining if the person is in the labor force.
The second equation is the log wage equation, where y_i is only observed for people who are in the labor force.
To solve the model, one needs to use the so-called “Heckit” model, which involves adding a correction term in the wage equation which accounts for the fact that workers are not randomly selected.

Derivation of Logit Probability

$$
\begin{aligned}
Pr(d_1 = 1) &= \int_{-\infty}^{\infty} \left(e^{-e^{-(\varepsilon_1 + u_1 - u_2)}} \right) f(\varepsilon_1) d\varepsilon_1 \\
&= \int_{-\infty}^{\infty} \left(e^{-e^{-(\varepsilon_1 + u_1 - u_2)}} \right) e^{-\varepsilon_1} e^{-e^{-\varepsilon_1}} d\varepsilon_1 \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\varepsilon_1} - e^{-(\varepsilon_1 + u_1 - u_2)} \right) e^{-\varepsilon_1} d\varepsilon_1 \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\varepsilon_1} \left[1 + e^{u_2 - u_1} \right] \right) e^{-\varepsilon_1} d\varepsilon_1
\end{aligned}
$$

Derivation of Logit Probability

Now need to do the substitution rule where $t = \exp(-\varepsilon_1)$ and $dt = -\exp(-\varepsilon_1)d\varepsilon_1$.

Note that we need to do the same transformation of the bounds as we do to ε_1 to get t. Namely, $\exp(-\infty) = 0$ and $\exp(\infty) = \infty$.

Derivation of Logit Probability

Substituting in then yields:

$$
\begin{aligned}
Pr(d_1 = 1) &= \int_{\infty}^{0} \exp\left(-t\left[1 + e^{(u_2 - u_1)}\right]\right)(-dt) \\
&= \int_{0}^{\infty} \exp\left(-t\left[1 + e^{(u_2 - u_1)}\right]\right) dt \\
&= \left.\frac{\exp\left(-t\left[1+e^{(u_2-u_1)}\right]\right)}{-\left[1+e^{(u_2-u_1)}\right]}\right|_0^{\infty} \\
&= 0 - \frac{1}{-\left[1 + e^{(u_2-u_1)}\right]} = \frac{\exp(u_1)}{\exp(u_1) + \exp(u_2)}
\end{aligned}
$$