Notes on *Discrete Choice Methods with Simulation* (Train, 2009)

Chapter 2. Properties of discrete choice models

2.1. The choice set

- Three characteristics of the choice set in discrete choice models:
  1. The alternatives must be *mutually exclusive* from the decision maker's perspective.
  2. The choice set must be *exhaustive*, in that all possible alternatives are included.
  3. The number of alternatives must be finite.
- By re-defining the alternatives, the first and second characteristics are actually not restrictive. Besides examining the choice of "which," discrete models can be and have been used to examine choices of "how much" by defining the alternatives with quantities (e.g., buying no car, buying 1 car, buying 2 cars, and buying more cars). Therefore, the third characteristic is restrictive when considering decision making related to continuous alternatives (e.g., money saving decisions).

2.2. Derivation of choice probabilities

- The choice probabilities can be derived from random utility models (RUMs).
- Consider a decision maker n facing a choice among J alternatives. The utility obtains from alternative j is defined as $U_{nj}, j = 1, \cdots, J$. The decision maker chooses alternative i if and only if $U_{ni} > U_{nj}$ ($\forall j \neq i$).
- There are some observable attributes of the alternatives faced by the decision maker $x_{nj}, j = 1, \cdots, J$, and some observable attributes of the decision maker s_n, so a *representative utility* function can be specified to relate these observable factors to utility, denoted by $V_{nj} = V(x_{nj}, s_n)$. Now taking into account the unobservable factors, utility is decomposed as $U_{nj} = V_{nj} + \varepsilon_{nj}$.
- The researcher does not know $\varepsilon_{nj}, j = 1, \cdots, J$, and therefore treats them as random. The joint density is denoted as $f(\varepsilon)$, where $\varepsilon_n = [\varepsilon_{n1}, \cdots, \varepsilon_{nJ}]'$. Therefore, the probability that decision maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= \Pr\left\{V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj},\ \forall j \neq i\right\} \\ &= \Pr\left\{\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj},\ \forall j \neq i\right\} \\ &= \int_{\varepsilon} \mathbb{I}\left[\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj},\ \forall j \neq i\right] f\left(\varepsilon_n\right) \mathrm{d}\varepsilon_n. \end{aligned}$$

  Consider a population of people who face the same observed utility $V_{nj}, j = 1, \cdots, J$, as person n but have different values of the unobservables. The density $f(\varepsilon_n)$ is the distribution of the unobserved portion of utility within the population of people who face the same observed portion of utility. Under this interpretation, the probability P_{ni} is the share of people who choose alternative i within the population of people who face the same observed utility for each alternative as person n.
- Different choice models are derived under different specifications of the density of unobserved factors $f(\varepsilon_n)$:
  - *Logit:* i.i.d. extreme distribution.
  - *GEV:* extreme distribution allowing for any correlation.
  - *Probits:* normal distribution allowing for any covariance matrix.
  - *Mixed logit:* the unobserved factors can be decomposed into a part that contains all correlation and heteroskedasticity, and another part that is i.i.d. extreme value.
  - ...

2.3. Identification of choice models

- Several aspects of the behavioral decision process affect the specification and estimation of any discrete choice model. The issues can be summarized in two statements:
  1. Only differences in utility matter.
  2. The scale of utility is arbitrary.
- **Alternative-specific constants.** Consider two specifications of utility (with different constant term, i.e., $k_j^0 \neq k_j^1$):

$$U_{nj}^0 = x_{nj}'\beta + k_j^0 + \varepsilon_{nj},$$
$$U_{nj}^1 = x_{nj}'\beta + k_j^1 + \varepsilon_{nj},$$

  where $k_i^0 - k_j^0 = d_{ij} = k_i^1 - k_j^1$ ($\forall i, j$). These two specifications give the same probabilities

$$P_{ni} = \int_{\varepsilon} \mathbb{I}\left[\varepsilon_{nj} - \varepsilon_{ni} > x_{ni}'\beta - x_{nj}'\beta + d_{ij},\ \forall j \neq i\right] f(\varepsilon_n) \mathrm{d}\varepsilon_n.$$

  This implies that:
  - With J alternatives, at most $J-1$ alternative-specific constants can enter the model, with one of the constants normalized to zero. It is irrelevant which constant is normalized to zero: the other constants are interpreted as being relative to whichever one is set to zero;
  - If we are interested in effects of the attributes that do not vary over alternatives (e.g., attributes of the decision maker like income, which could have different margin effects on utility of different alternatives), we can only estimate the differences in marginal effects rather than the absolute marginal effects.
  - The number of independent error terms is actually $J-1$:

$$\begin{aligned} P_{ni} &= \int \mathbb{I}\left[\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj},\ \forall j \neq i\right] f(\varepsilon_n) \mathrm{d}\varepsilon_n \\ &= \int \mathbb{I}\left[\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj},\ \forall j \neq i\right] g(\tilde{\varepsilon}_{ni}) \mathrm{d}\tilde{\varepsilon}_{ni}, \end{aligned}$$

    where $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$, for any $j = 1, \cdots, J$.
- **The overall scale of utility is irrelevant.** The two specifications are equivalent no matter how utility is scaled by $\lambda > 0$:

$$U_{nj}^1 = V_{nj} + \varepsilon_{nj},$$
$$U_{nj}^\lambda = \lambda V_{nj} + \lambda\varepsilon_{nj},\ \forall \lambda > 0.$$

  The standard way to normalized the scale of utility is to normalize the variance of the error terms.
  - **Error terms are assumed to be i.i.d.** Suppose that the original model is $U_{nj}^0 = x_{nj}'\beta + \varepsilon_{nj}^0$ where the variance of the error terms is $\mathrm{Var}[\varepsilon_{nj}^0] = \sigma^2$, we can normalize the original model as $U_{nj}^1 = x_{nj}'(\beta/\sigma) + \varepsilon_{nj}^1$ with $\mathrm{Var}[\varepsilon_{nj}^1] = 1$. The new coefficients β/σ reflect, therefore, the effect of the observable variables relative to the standard deviation of the unobservable factors.
  - **Heteroskedastic errors.** Suppose that the original models are

$$\begin{aligned} U_{A,nj}^0 &= x_{nj}'\beta + \varepsilon_{A,nj}^0 \\ U_{B,nj}^0 &= x_{nj}'\beta + \varepsilon_{B,nj}^0 \end{aligned},$$

    with $\mathrm{Var}[\varepsilon_{A,nj}^0] \neq \mathrm{Var}[\varepsilon_{B,nj}^0]$. Define the ratio of variances of unobservable factors in regions(/data sets/time/other factors) B and A is $k_{BA} \equiv \mathrm{Var}[\varepsilon_{B,nj}^0]/\mathrm{Var}[\varepsilon_{A,nj}^0]$, then we can re-

scale the model in region B to

$$U^1_{A,nj} = x'_{nj}\beta + \varepsilon^1_{nj}$$
$$U^1_{B,nj} = x'_{nj}(\beta/\sqrt{k_{BA}}) + \varepsilon^1_{nj},$$

where $\mathrm{Var}[\varepsilon^1_{nj}] = \mathrm{Var}[\varepsilon^0_{A,nj}] = \mathrm{Var}[\varepsilon^0_{B,nj}/\sqrt{k_{BA}}]$ and returns to the homoskedastic case. The parameter k_{BA}, which is often called the scale parameter, is estimated along with β. The estimated value $\hat{k}_{BA}$ of k_{BA} tells the researcher the variance of unobservable factors in B relative to that in A. (Remarks: here A and B refer to different markets with independent decision making.)

- **Correlated errors.**

2.4. Aggregation

- If an alternative gives different representative utilities to different kinds of people (which yields different probabilities P_{ni}), we cannot simply aggregate/average the explanatory variables since discrete choice models are not linear in explanatory variables. Aggregate outcome variables can be obtained consistently from discrete choice models in two ways: *sample enumeration* or *segmentation*.
- **Sample enumeration.** A consistent estimate of the total number of decision makers in the population who choose alternative i, labeled $\hat{N}_i$, is simply the weighted sum of the individual probabilities:

$$\hat{N}_i = \sum_n w_n P_{ni},$$

where w_n represents the number of decision makers similar to him in the population. For samples based on exogeneous factors, this weight is the reciprocal of the probability that the decision maker was selected into the sample. If the sample is purely random, then w_n is the same for all n; and if the sample is stratified random, then w_n is the same for all n within a stratum.

- **Segmentation.** If the total number of different types of decision makers (called *segments*) is small (e.g., education level and gender), we can estimate aggregate outcomes without utilizing a sample of decision makers because the number of people in each segments is available. Since the probability within each segment is identical, we can estimate the probability for each segment P_{si} and then aggregate based the weights of segments:

$$\hat{N}_i = \sum_s w_s P_{si},$$

where w_s is the number of decision makers in segment s.