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

Chapter 3. Logit

- In this chapter, we use the general notation from Chapter 2 and add a specific distribution for unobserved utility. A decision maker, labeled n, faced J alternatives. The utility that the decision maker obtains from utility j is decomposed into (1) a part labeled V_{nj} that is known by the researcher up to some parameters, and (2) an unknown part ε_{nj} that is treated by the researcher as random: $U_{nj} = V_{nj} + \varepsilon_{nj}$ for any j.

3.1. Choice probabilities

- The logit model is obtained by assuming that each ε_{nj} is independently, identically distributed extreme value:

  $$f(\varepsilon_{nj}) = \mathrm{e}^{-\varepsilon_{nj}} \mathrm{e}^{-\mathrm{e}^{-\varepsilon_{nj}}}$$

  and

  $$F(\varepsilon_{nj}) = \mathrm{e}^{-\mathrm{e}^{-\varepsilon_{nj}}},$$

  with a variance of $\pi^2/6$.
- The difference between two extreme value variables is distributed logistic. That is, if ε_{nj} and ε_{ni} are i.i.d. extreme value, then $\varepsilon^*_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ follows the logistic distribution

  $$F\left(\varepsilon^*_{nji}\right) = \frac{\mathrm{e}^{\varepsilon^*_{nji}}}{1 + \mathrm{e}^{\varepsilon^*_{nji}}}.$$

- It is better to understand the logit model as an ideal scenario that the research has specified V_{nj} sufficiently that the remaining unobserved portion of utility is essentially "white noise."
- With the above assumptions, we can derive a succinct closed-form expression:

  $$P_{ni} = \frac{\mathrm{e}^{V_{ni}}}{\sum_j \mathrm{e}^{V_{nj}}}.$$

  Further, if the representative utility is specified to be linear in parameters: $V_{nj} = \beta' x_{nj}$, the logit probabilities become

  $$P_{ni} = \frac{\mathrm{e}^{\beta' x_{ni}}}{\sum_j \mathrm{e}^{\beta' x_{nj}}}.$$

- Several desirable properties of the logit probabilities:
  - P_{ni} is necessarily between zero and one. The probability is never exactly zero and is one only if the choice set consists of a single alternative.
  - The choice probabilities of all alternatives sum to one. The decision maker necessarily chooses one of the alternatives.
  - The relation of the logit probability to representative utility is sigmoid.
- Consider a binary case with alternatives 1 and 2. Then we have the probability

  $$P_{n1} = \frac{\mathrm{e}^{V_{n1}}}{\mathrm{e}^{V_{n1}} + \mathrm{e}^{V_{n2}}} = \frac{1}{1 + \mathrm{e}^{V_{n2} - V_{n1}}}.$$

If only demographics of the decision makers, s_n, enter the model, and the coefficients of these demographic variables are normalized to zero for the first alternatives, the probability of the first alternative is $P_{n1} = 1/(1 + \mathrm{e}^{\alpha' s_n})$, which is the form that is used in most textbooks and computer manuals for binary logit.

3.2. The scale parameter

- In general, utility can be expressed as $U^*_{nj} = V_{nj} + \varepsilon^*_{nj}$, where the unobserved portion has variance $\sigma^2 \times (\pi^2/6)$. Since the scale of utility is irrelevant to behavior, utility can be divided by σ without changing behavior. Utility becomes $U_{nj} = V_{nj}/\sigma + \varepsilon_{nj}$ where $\varepsilon_{nj} = \varepsilon^*_{nj}/\sigma$ and now the unobserved portion has variance $\pi^2/6$.
- Now the choice probability is

$$P_{ni} = \frac{\mathrm{e}^{V_{ni}/\sigma}}{\sum_j \mathrm{e}^{V_{nj}/\sigma}}.$$

  If V_{nj} is linear in parameters with coefficient β^*, the choice probabilities become

$$P_{ni} = \frac{\mathrm{e}^{(\beta^*/\sigma)' x_{ni}}}{\sum_j \mathrm{e}^{(\beta^*/\sigma)' x_{nj}}}.$$

  Each of the coefficients is scaled by $1/\sigma$. The parameter σ is called the *scalar parameter*, because it scales the coefficients to reflect the variance of the unobserved portion of utility.
- Only the ratio $\beta = \beta^*/\sigma$ can be estimated, and usually the model is expressed in its scaled form (i.e., β is estimated). For interpretation it is useful to recognize that these estimated parameters are actually estimates of the "original" coefficients divided by the scale parameter.
- Consider two different markets with independent decision making. If we assume that original coefficients (i.e., β^*) are the same, then the difference in the estimated coefficients in these two markets reflects the difference in the variances of the unobserved portions in these two markets. Similar consideration applies to two different data sets.

3.3. Power and limitations of logit

- Logit can represent systematic taste variation but not random taste variation.
  - Consider households' choice among makes and models of cars to buy. Suppose there are two attributes observed: purchase price PP_j and inches of shoulder room SR_j for make/model j. The value of households place on these two attributes varies over households, and so the utility is

$$U_{nj} = \alpha_n \mathrm{SR}_j + \beta_n \mathrm{PP}_j + \varepsilon_{nj},$$

    where α_n and β_n are parameters specific to household n. Suppose that such heterogeneity is determined by some observable demographics. For example, if

$$\alpha_n = \rho M_n,$$

    where M_n is the number of members, and

$$\beta_n = \theta / I_n,$$

    where I_n refers to income, we have

$$U_{nj} = \rho\,(M_n \mathrm{SR}_j) + \theta\,(\mathrm{PP}_j / I_n) + \varepsilon_{nj}.$$

This means we can use interaction terms between alternative attributes and household demographics to depict systematic taste variation.
    - Similarly, we can incorporate higher-order polynomials to depict non-linear relationship between attributes and utility. In conclusion, when tastes vary systematically in the population in relation to observed variables, the variation can be incorporated into logit models.
    - However, the limitation of the logit model arises when we attempt to allow tastes to vary with respect to unobserved variables or purely randomly. For example, the coefficient contains a random term (like random-coefficients model discussed in chapter of mixed logit). Suppose that $\alpha_n = \rho M_n + \mu_n$ where μ_n is unobserved, then the utility is specified as

$$U_{nj} = \rho\left(M_n \mathrm{SR}_j\right) + \theta\left(\mathrm{PP}_j / I_n\right) + \tilde{\varepsilon}_{nj},$$

      where $\tilde{\varepsilon}_{nj} = \varepsilon_{nj} + \mu_n \mathrm{SR}_j$. As all new unobserved terms contain the same μ_n, they cannot possibly be distributed independently. Therefore, if taste variation is at least partly random, logit is a misspecification.
    - As an approximation, logit might be able to capture the average tastes fairly well even when tastes are random, since the logit formula seems to be fairly robust to misspecifications. But there is no guarantee. Additionally, logit does not provide information of the distribution of tastes around the average, which can be important in many situations. To incorporate random taste variation, a probit or mixed logit model can be used instead.
- The logit model implies proportional substitution across alternatives, given the researcher's specification of representative utility. To capture more flexible forms of substitution, other models are needed.
    - For any two alternatives i and k, the ratio of the logit probabilities is

$$\frac{P_{ni}}{P_{nk}} = \mathrm{e}^{V_{ni} - V_{nk}},$$

      which does not depend on any alternatives other than i and k. This ratio is said to be independent from *irrelevant* alternatives. In other words, the logit model exhibits the *independence from irrelevant alternatives*, or IIA.
    - By definition, the elasticity of P_{ni} with respect to a variable that enters the representative utility of alternative $j \neq i$ is

$$E_{iz_{nj}} = -z_{nj} P_{nj} \frac{\partial V_{nj}}{\partial z_{nj}}$$

      or

$$E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$$

      if the representative utility is linear in attributes, where z_{nj} is the attributes of alternative j and β_z is its coefficient. This cross-elasticity is the same for all i since i does not enter the formula. In other words, if one alternatives' probabilities drops by ten percent, then all the other alternatives' probabilities (except j) also drop by ten percent. This pattern of substitution, which can be called *proportionate shifting*, is a manifestation of the IIA property. Indeed, the pattern of proportionate shifting is intuitive because IIA ensures the ratio of probabilities between alternatives (except j) unchanged.
    - However, the property of IIA seems to be inappropriate in some choice situations like the red-bus-blue-bus problem. In some cases, when introducing a new alternative (e.g., a new bus line), we expect the probabilities of some alternatives (e.g., a old bus line) decrease more and those of others (e.g., walking) decrease less. This results in different probability ratios between original

alternatives. Why logit is not suitable in this case? This is because the unobserved components of different bus lines are indeed correlated.
    - There are several statistical methods to test the IIA assumption.
- If unobserved factors are independent over time in repeated choice situations, then logit can capture the dynamics of repeated choice, including state dependence. However, logit cannot handle situations where unobserved factors are correlated over time.
    - If the unobserved factors that affect decision makers are independent over the repeated choices, then logit can be used to examine panel data in the same way as purely cross-sectional data. Any dynamics related to observed factors that enter the decision process, such as state dependence (by which the person's past choices influence their current choices) or lagged response to changes in attributes, can be accommodated. However, dynamics associated with unobserved factors cannot be handled, since the unobserved factors are assumed to be unrelated over choices.

3.4. Consumer surplus

- By definition, a person's consumer surplus is the utility that the person receives in the choice situation. The decision maker chooses the alternative that provides the greatest utility. Consumer surplus is therefore $\mathrm{CS}_n = (1/\alpha_n)\max_j\{U_{nj}\}$, where α_n is the marginal utility of income. The division by α_n translates utility into dollars.
- The researcher does not observe U_{nj} and therefore cannot use this expression to calculate the decision maker's consumer surplus. Instead, the researcher observes V_{nj} and knows the distribution of the remaining portion of utility. With this information, the research is able to calculate the expected consumer surplus:

  $$\mathbb{E}\left[\mathrm{CS}_n\right] = \frac{1}{\alpha_n}\mathbb{E}\left[\max_j\left\{V_{nj}+\varepsilon_{nj}\right\}\right].$$

  If each ε_{nj} is i.i.d. extreme value and utility is linear in income (so that α_n is a constant with respect to income), then this expectation becomes

  $$\mathbb{E}\left[\mathrm{CS}_n\right] = \frac{1}{\alpha_n}\ln\left(\sum_{j=1}^{J}\mathrm{e}^{V_{nj}}\right) + C,$$

  where C is an unknown constant that represents the fact that the absolute level of utility cannot be measured. Aside from the division and additional of constants, expected consumer surplus in a logit model is simply the log of the denominator of the choice probability. It is often called the *log-sum term*.
- Under the standard interpretation for the distribution of errors, $\mathbb{E}[\mathrm{CS}_n]$ is the average consumer surplus in the subpopulation of people who have the same representative utilities as person n. The total consumer surplus in the population is calculated as a weighted sum of $\mathbb{E}[\mathrm{CS}_n]$ over a sample of decision makers.
- The change in consumer surplus that results from a change in the alternatives and/or the choice set is

  $$\Delta\mathbb{E}[\mathrm{CS}_n] = \frac{1}{\alpha_n}\left[\ln\left(\sum_{j=1}^{J^1}\mathrm{e}^{V_{nk}^1}\right) - \ln\left(\sum_{j=1}^{J^0}\mathrm{e}^{V_{nk}^0}\right)\right].$$

3.5. Estimation

3.5.1. Exogenous sample

- Consider first the situation in which the sample is exogenously drawn and the explanatory variables are independent of the unobserved component of utility.
- A sample of N decision makers is obtained. The probability of person n choosing the alternative that he was actually observed to choose can be expressed as $\Pi_i (P_{ni})^{y_{ni}}$ where $y_{ni} = 1$ if person n chose i and zero otherwise. Assuming that each decision maker's choice is independent of that of other decision makers, the probability of each person in the sample choosing the alternative that he was observed actually to choose is

  $$\mathcal{L}(\beta) = \prod_{n=1}^{N} \prod_{i} (P_{ni})^{y_{ni}}.$$

  The log-likelihood function is then

  $$\mathrm{LL}(\beta) = \sum_{n=1}^{N} \sum_{i} y_{ni} \ln P_{ni}.$$

  We estimate the parameters by solving the following first-order condition:

  $$\frac{\mathrm{dLL}(\beta)}{\mathrm{d}\beta} = 0$$

- Let the representative utility be linear in parameters: $V_{nj} = \beta' x_{nj}$, then it can be shown that the first-order condition becomes

  $$\sum_{n} \sum_{i} (y_{ni} - P_{ni}) x_{ni} = 0.$$

  McFadden (1974) shows that $\mathrm{LL}(\beta)$ is globally concave for linear-in-parameters utility, and many statistical packages are available for estimation of these models.
- Rearranging the dividing both sides by N, we have

  $$\frac{1}{N} \sum_{n} \sum_{i} y_{ni} x_{ni} = \frac{1}{N} \sum_{n} \sum_{i} P_{ni} x_{ni}.$$

  That is, the maximum likelihood estimates of β are those that make the predicted average of each explanatory variable equal to the observed average in the sample.
- Note that the difference between a person's actual choice y_{ni} and the probability of that choice P_{ni} is a modeling error, or residual. Therefore, another interpretation is that the maximum likelihood estimates make the sample covariance of the residuals with explanatory variables zero. Under this interpretation, the estimates can be motivated as providing a sample analog to the moment condition that explanatory variables are uncorrelated with the modeling errors.

3.5.2. Estimation on a subset of alternatives

- In some situations, the number of alternatives facing the decision maker is so large that estimating model parameters is very expensive or even impossible. With a logit model, estimation can be performed on a subset of alternatives without inducing inconsistency. For example, a researcher examining a choice situation that involves 100 alternatives can estimate on a subset of 10 alternatives for each sampled decision maker, with the person's chosen alternative included as well as 9 alternatives randomly selected from the remaining 99.
- Suppose that the researcher has used some specific method to select alternatives into the subset that is used in estimation for each sampled decision maker. Denote the full set of alternatives as F and a subset of alternatives as K. Our goal is to derive a formula for the probability that the person chooses

alternative i conditional on the researcher selecting subset K for him, i.e., $P_n(i \mid K)$. Using Bayes Theorem, we have

$$\begin{aligned} P_n\left(i \mid K\right) &= \frac{P_{ni} P_n\left(K \mid i\right)}{\sum_j P_{nj} P_n\left(K \mid j\right)} \\ &= \frac{\mathrm{e}^{V_{ni}} P_n\left(K \mid i\right)}{\sum_j \mathrm{e}^{V_{nj}} P_n\left(K \mid j\right)} \\ &= \frac{\mathrm{e}^{V_{ni}} P_n\left(K \mid i\right)}{\sum_{j \in K} \mathrm{e}^{V_{nj}} P_n\left(K \mid j\right)}. \end{aligned}$$

- Further, suppose that the researcher has designed the selection procedure so that $P_n(K \mid j)$ is the same for all $j \in K$ (called "uniform conditional property"). Then, the conditional probability becomes

  $$P_n\left(i \mid K\right) = \frac{\mathrm{e}^{V_{ni}}}{\sum_{j \in K} \mathrm{e}^{V_{nj}}},$$

  which is simply the logit formula for a person who faces the alternatives in subset K. The conditional log-likelihood function under the unform conditional property is

  $$\mathrm{CLL}\left(\beta\right) = \sum_{n=1}^{N} \sum_{i \in K_n} y_{ni} \ln \frac{\mathrm{e}^{V_{ni}}}{\sum_{j \in K_n} \mathrm{e}^{V_{nj}}}.$$

3.6. Goodness of fit and hypothesis testing

- A statistic called the *likelihood ratio index* is often used with discrete choice models to measure how well the models fit the data:

  $$\rho = 1 - \frac{\mathrm{LL}(\hat{\beta})}{\mathrm{LL}(0)},$$

  where $\mathrm{LL}(0)$ typically refers to having no model at all. As $\mathrm{LL}(0) \leq \mathrm{LL}(\hat{\beta}) \leq 0$, $\rho \to 1$ when the model is perfectly fitted, and $\rho \to 0$ when the model does not work at all.
- Consider a null hypothesis H that can be expressed as constraints on the values of the parameters. Let $\mathcal{L}(\hat{\beta}^H)$ be the constrained maximum value of the likelihood function and $\mathcal{L}(\hat{\beta})$ be the unconstrained maximum value of the likelihood function. Define the ratio of likelihoods, $R = \mathcal{L}(\hat{\beta}^H)/\mathcal{L}(\hat{\beta})$, then the test statistic defined as $-2 \ln R$ is distributed chi-squared with degrees of freedom equal to the number of restrictions implied by the null hypothesis.