Qualitative and Limited Dependent Variable Models

Background for Problem Set #3

There are many examples that involve discrete dependent variables rather than continuous dependent variables and that are not binary, i.e. have more than two choices. For example, individuals or firms such as banks can make decisions that are “either-or” in nature. One example of this is an individual’s decision about what type of transportation to use, a personal car or public transportation. Another example of this is when a bank decides whether or not to lend money. Typically such choices are represented by a binary variable that has the value 1 (one) if the outcome is yes or chosen and 0 (zero) otherwise.

Remember that the linear regression model can be represented as:

When we use a model to explain a “choice variable y” we call this the **linear probability model**. That is, what we seek from this model is the probability density functions for the left-hand and right-hand sides of the equation. There are a number of difficulties applying this model to “choice” or binary problems. To overcome these difficulties we use the probit or logit models. In addition, when we have more than two alternative choices we use multinomial logit models.

# Multinomial Logit

In the probit and logit models already presented decision makers choose between two alternatives. This is not always the case. There are many times where choices involve more than two alternatives, e.g choosing a laundry detergent or choosing a major in college, etc. These are also cases where there is no natural ordering involved. In these cases multinomial logit is used. I will leave it up to you to go through the math involved. There are many references available online[[1]](#footnote-1).

Suffice to say that in the end the probability of a choice is given by a modified sigmoid function that depends on the number of choices available. For example, if there are three alternatives the probability of choice 1 is given by:

And the probabilities of the other alternatives, i.e. 2 and 3, can be obtained using analogous equations. To evaluate which alternative is chosen we might select which is the maximum of the estimated probabilities. Or, another very useful measure is the probability ratio, e.g.

Let’s consider an example of a multinomial logit model.

The National Education Longitudinal Study of 1988 (NELS:88) was the first nationally representative longitudinal study of eighth-grade students in public and private schools in the U.S. It was sponsored by the National Center for Education Statistics. In 1988, some 25,000 eighth-graders and their parents, teachers, and principals were surveyed. In 1990, these same students (who were then mostly 10th-graders and some dropouts) and their teachers, and principals were surveyed again. In 1992, the second follow-up survey was conducted of students, mostly in the 12th grade, but dropouts, parents, teachers, school administrators, and high school transcripts were also surveyed. The third follow-up was in 1994, after most students had graduated.[[2]](#footnote-2)

The nels\_small.gdt dataset is a subset of the total data, namely those who stayed in the panel of data through the third follow-up. On this group we have complete data on the individuals and their households, high school grades, and test scores, as well as their post-secondary education choices. In the file nels\_small.dat we have 1,000 observations on students who chose, upon graduating from high school, either no college (PSECHOICE = 1), a two-year college (PSECHOICE = 2), or a four-year college (PSECHOICE = 3). For illustration purposes we focus on the explanatory variable GRADES, which is an index ranging from 1.0 (highest level, A+ grade) to 13.0 (lowest level, F grade) and represents combined performance in English, math, and social studies. NOTE, this scale runs counter to the typical scale where higher performance is associated with higher values, e.g. F=0 and A=4.0 or 5.0 depending. This means that a one positive unit change in grade represents poorer performance rather than better performance.

# Conditional Logit

The objective here is to understand the factors that lead a consumer to choose one alternative over another. To do this a model is built for the probability that individual chooses alternative , i.e.

There is considerable math here I won’t go into. Suffice to say that in the end you will have an expression for a probability given the “likelihood” of the related estimated coefficients. For question 2.c.ii. of this assignment, the probabilities of the multinomial logit model are:

where:

and for the nels\_small dataset:

Once the multinomial model has been built, then you will have the estimated coefficients to complete the required computations.

Problem Set #3

1. **Descriptive Statistics** Consider the data in the the nels\_small dataset. (As usual, you can use gretl, or the script provided for gretl, or any other suitable program/code. Note that the numbers computed by the script provided might not exactly match what the questions in the assignment ask, one-to-one. However, all the information needed to answer the questions in the assignment is there, somewhere.)
   1. How many observations are in the dataset?
   2. How many variables are in the dataset?
   3. What is the dependent variable in this dataset?
   4. Is this variable binary?
   5. Given the observations in this dataset, what percentage of these students decided to pursue a college degree at either a 2-year college or a 4-year college?
   6. Given the observations in this dataset, what percentage of these students decided to pursue a college degree at a 4-year college?
   7. Given the observations in this dataset, what percentage of these students decided not to pursue a college degree?
   8. Of those students who decided to go onto college,
      1. what was the average grade?
      2. What was the minimum grade?
      3. What was the maximum grade?
      4. In terms of their average grade, did these students do better or worse than the overall average student?
      5. What percentage of students going to college were female?
      6. What percentage of students going to college were black?
      7. What percentage of students going to college were black females?
      8. What was the average grades for black females going onto college?
         1. Was this average grade higher than, lower than, or equal to the overall average grade for all students going to college?
2. **Modelling** Follow the directions below to build models and interpret their results.
   1. Given the dependent and independent variables in the nels\_small dataset, what type of model would be best for this data?
      1. linear
      2. Probit
      3. Logit
      4. Multinomial logit
   2. Build a model for this data using psechoice as the dependent variable and grades as the independent variable.
      1. Which psechoice level is the baseline?
      2. Are the intercept and independent variable both statistically significant at all levels?
      3. In terms of the number of cases predicted, did the model correctly predict more than, less than or a number equal to the total number of cases?
      4. What is the predicted probability that a student in the 5th percentile will go to a 2-year college?
      5. What is the predicted probability that a student in the 50th percentile will go to a 2-year college?
      6. Remember that an increase in the value of the variable grades actually indicates worse performance, not better! If the grades increase by 1 unit for a student in the 95th percentile how much more likely is it that he or she will go to a 4-year college?
      7. Remember that an increase in the value of the variable grades actually indicates worse performance, not better! If the grades increase by 1 unit for a student in the 50th percentile how much more likely is it that he or she will not go to college?
3. Build a model for this data using psechoice as the dependent variable and grades faminc, female, and black as independent variables.
   1. The appropriate model for this case is:
      * + 1. Linear
4. Linear
5. Probit
6. Logit
7. Multinomial logit
8. Some other model
   1. The intercept and estimated coefficients for all variables at all levels are all statistically significant. True/**False**
   2. Using the estimated coefficients from the model, what is the estimated probability that a white male student with median values of GRADES and FAMINC will attend a 4-year college? (Hint: I will provide the formula you need below. They actually come from the derivation of maximum likelihood estimates.)

where:

* 1. Compute the probability ratio that a white male student with median values of GRADES and FAMINC will attend a 4-year college rather than not attend any college.

(Again, use the formulae provided for you. Eliminating the math from this problem, the formula to find this answer is:

1. For example, Regression Models for Categorical and Limited Dependent Variables by J. Scott Long or Quantitative Models in Marketing Research by Philip Hans Franses and Richard Paap. A more advanced text is Econometric Analysis: 6th Edition by William Greene. [↑](#footnote-ref-1)
2. The study and data are summarized in National Education Longitudinal Study: 1988–1994, Descriptive Summary Report With an Essay on Access and Choice in Post-Secondary Education, by Allen Sanderson, Bernard Dugoni, Kenneth Rasinski, and John Taylor, C. Dennis Carroll project officer, NCES 96-175, National Center for Education Statistics, March 1996. [↑](#footnote-ref-2)