Economics 311 Fall 2017
Problem Set #7 Due: Wednesday, November 22

This problem set again asks you to use Stata to perform some regressions and to interpret the results. Assemble your answers *along with all relevant Stata output* into a Word document and send it to me as an email attachment in .docx or .pdf format.

This project uses the dataset nels.dta, which is drawn from the National Education Longitudinal Study of 1988. It has 6,649 observations and 14 variables, defined below:

psechoice	= 1 if first postsecondary education was no college
	= 2 if first postsecondary education was a 2-year college
	= 3 if first postsecondary education was a 4-year college
hscath	= 1 if catholic high school graduate
hsrural	= 1 if high school rural
grades	= average grade in math, English and social studies on 13 point scale with 1 = highest
faminc	= gross 1991 family income (in $1000)
famsiz	= number of family members
parsome	= 1 if most educated parent had some college, but not a 4-year degree
parcoll	= 1 if most educated parent graduated from college or had an advanced degree
female	= 1 if female
asian	= 1 if asian
hispan	= 1 if hispan
black	= 1 if black
grants	= 1 if student had a grant/scholarship or fellowship when attending college
loans	= 1 if student received a loan while attending college

The summary statistics for these variables are:

Variable	Obs	Mean	Std. Dev.	Min	Max
psechoice	6,649	2.30952	.7958798	1	3
hscath	6,649	.0774553	.2673325	0	1
hsrural	6,649	.3263649	.4689178	0	1
grades	6,649	6.435658	2.261568	1.15	12.67
faminc	6,649	50.79523	40.60194	0	250
famsiz	6,649	4.270717	1.342331	1	10
parsome	6,649	.4620244	.4985933	0	1
parcoll	6,649	.3286208	.4697471	0	1
female	6,649	.5095503	.4999464	0	1
asian	6,649	.0851256	.2790895	0	1
hispan	6,649	.0891863	.285034	0	1
black	6,649	.0917431	.2886847	0	1
grants	6,649	.338096	.473097	0	1
loans	6,649	.1998797	.3999398	0	1

The dependent variable of interest in this project is *psechoice*, which reflects the person's decision to attend college and whether to attend a 2-year or 4-year college. The other variables are assumed to be exogenous and are prospective explanatory variables.

For this project, you are to use the following variables (only) to explain college choice using several estimation methods: *grades, parcoll, faminc, female, black*. Do not include other variables in the regressions (unless you do the optional part at the end).

1. Linear probability model

For this part (and the next two) we will ignore the distinction between 2-year and 4-year colleges and simply focus on whether the person attended *any* college. Construct an appropriate dummy variable with gen college = psechoice > 1. This variable will be 1 when the condition is true (*psechoice* = 2 or 3) and zero if it is false (*psechoice* = 1).

- Use OLS to estimate the linear probability model with *college* as the dependent variable and the five selected independent variables.
- Interpret the results of the *t* tests in the regression table.
- Interpret the meaning of each estimated coefficient, such as "If *X* changes by this, the effect on *Y* is that."
- Use the predict command to create a series *olspred* that has the fitted values for this regression. What is the interpretation of these values? Use summarize to get the sample statistics for *olspred*. Are all of the predicted values between 0 and 1? Why is this an issue?

2. Logit

Now we want to re-estimate the model using binomial logit, which Studenmund introduces in section 13.2. There are two alternative commands to estimate a logit regression in Stata. If you use the command logit, you get the actual estimated regression coefficients, which have an obscure direct interpretation. If you instead use logistic, Stata transforms the estimated coefficients into effects on the "log-odds ratio," or proportional effects on the odds ratio itself. (In the column where the estimated coefficients are usually found, Stata produces a column headed "Odds Ratio.") The odds ratio is the ratio Probability of $Y = 1$/Probability of $Y = 0$. In this problem that is the Probability of attending college/Probability of not attending college. So if someone has a 0.5 probability of attending college (and 0.5 probability of not attending), the odds ratio will be 0.5/0.5 = 1. If the probability of attending is 0.75 (so the probability of not attending is 1 – 0.75 = 0.25), then the odds ratio is 0.75/0.25 = 3.

If the reported coefficient in the Odds Ratio column is 1.0, then the variable has no effect: an increase of one unit in the variable multiplies the odds ratio of the dependent variable by one, not changing it at all. Thus, an estimate of one for the logistic effect is equivalent to an

estimate of zero for the regression coefficient: no effect. If the reported value is, for example, 1.5 then a one-unit increase in the regressor multiplies the odds ratio by a factor of 1.5, increasing it by 50%. If the reported coefficient is 0.6, then an increase of one unit in the regressor multiplies the odds ratio by 0.6, reducing it to 60% of its previous value (or lowering it by 40%). So an estimated coefficient in this column that is *larger than one* means that the regressor *increases* the probability that the dependent variable is one and an estimated coefficient that is *less than one* means that the regressor *decreases* the probability that the dependent variable is one.

- Use the logistic command to estimate the model using the binomial logit model.
- Interpret the results of the *z* tests in the table and compare them to the results of the linear probability model from the previous exercise. (*z* is used here because the coefficients in the logit model are asymptotically normal, but do not have a *t* distribution.)
- Interpret the meaning of each estimated "coefficient" in the regression table, again as "If *X* changes by this, the effect on *Y* is that."

3. Probit

An alternative to logit that many prefer is probit, which substitutes the cumulative normal probability distribution for the cumulative logistic. As with logit, Stata offers two commands that differ only in the form in which the coefficients are reported. The probit command reports the actual coefficients, which, again, are not directly informative. To get something that we humans can understand, use dprobit. In dprobit, the column in the regression table where the coefficients would usually be is headed "dF/dx." The reported coefficients here are the estimated "marginal effect" of a one-unit change in the regressor on the probability that the dependent variable is one. As Studenmund discusses, this effect varies depending on where the observation is in the distribution; Stata evaluates the effects at the means of all variables.

A reported marginal effect of zero means no effect (as with normal regression coefficients), a reported marginal effect of 0.2 means that a one-unit increase in the regressor increases the probability that the dependent variable is one (that the person attended some college in this example) by 0.2. So if *X* is a dummy variable, changing the dummy from 0 to 1 would raise the probability of attending college by 0.2 for a person with average values for all variables. If the reported marginal effect is negative, then an increase in the regressor lowers the probability of attending college.

- Use the dprobit command to estimate the effects using the probit model.
- Interpret the results of the *z* tests and compare them to the results of the linear probability model and of logit.

- Interpret the meaning of each estimated marginal effect in the table in the same manner that you have done for the other models.

4. Ordered probit

As discussed in class, ordered probit is a method for estimating models in which the dependent variable has more than two outcomes that can be ordered, but that are not measured on an "interval" scale. (In other words, 3 is not necessarily bigger than 2 by the same amount that 2 is bigger than 1.) The original variable *psechoice* has this property: Choosing a 4-year college is "more than" choosing a 2-year college, which is "more than" choosing not to go to college at all.

Stata's oprobit command estimates by ordered probit. The coefficients reported by oprobit tell how much a one-unit change in the regressor advances the underlying "index" for the dependent variable up the scale from lowest outcome to highest. The "/cut" values reported at the bottom of the table are the "cut-points" or threshold values between each of the levels of the dependent variable.

- Estimate the model by ordered probit.
- Interpret the results of the *z* tests and compare them to your previous results.
- Interpret the signs of the coefficients (but not their magnitudes)
- Use the five coefficients reported to calculate the index value for a person with *grades* = 10, parents who have not completed college, *faminc* = 20, and who is neither female nor black. (The constant term is normalized to zero.) Does the computed index fall below /cut1 so that the prediction is no college, between /cut1 and /cut2 so the prediction is 2-year college, or above /cut2 so the prediction is 4-year college?
- Now compute the index for another person with the same grades, income, sex, and race, but who has a parent who finished college. Does your prediction change?

5. Conclusions

What do you conclude about the effect of these variables on college choice? Are your results consistent across estimation methods?

6. Optional (extra credit and extra fun): Choose one preferred estimation method and include some or all of the other variables from the data set in your regression. Interpret the results and compare them to the ones above from the restricted variable list.