Economics 499 J.M. Wooldridge
Spring 2020

Problem Set #2
60 Points
Due: **10/09/2020**

Instructions: Answer each question on your choice of paper, but be sure to staple the sheets together and to write your name on the front. Point totals are in parentheses. Within a question, each part receives equal weight. You should attach a printout of your Stata commands and results. But it is very important that you answer the questions by providing any relevant numbers from the Stata output. Act as if the grader will not even look at the output (which may or may not be true).

1. (30 points) Use the data in TUTOR_GPA.DTA to answer the following questions.

(i) How many students are in the data set? What percentage of students participated in the tutoring program?

(ii) The outcome variable of interest is *grad*, a binary variable equal to one of the student graduated from college. What percentage of students graduated from college?

(iii) Run the simple regression of *grad* on *tutor*. What is the estimated difference in the probability of graduating based on participating and not participating in the tutoring program. Is it statistically significant?

(iv) To the regression in part (ii) add the variables *sat* and *hsperc* along with squares in both variables and the interaction *sat* • *hsperc*. Now what is the estimated tutoring effect? How does it compare with part (iii)? What would you report as a 95% confidence interval?

(v) Now use a logit model for *grad*, using the same explanatory variables in part (iv). What is the estimated average treatment effect of *tutor* on the graduation probability?

(vi) Using the same variables in part (iv), use separate linear regression to estimate the ATE. Use the `teffects` command. Find $\hat{\tau}_{ate}$ and $\hat{\tau}_{att}$. Summarize what you find, including statistical significance.

(vii) Replace part (vi) but replace linear RA with logistic RA. Describe your findings. [Ignore the error message that Stata gives you; as far as I can tell, there is no problem.]

(viii) Estimate a logit model for the treatment variable, *tutor*, using the same explanatory variables in part (iv). Obtain the fitted probabilities, $\hat{p}_i$. How many observations have $\hat{p}_i < 0.1$? How many observations have $\hat{p}_i > 0.9$?

(ix) Keep only the observations with $0.1 \leq \hat{p}_i \leq 0.9$. Restimate the propensity score model in part (viii). Now how many observations have $\hat{p}_i < 0.1$?

(x) Use the smaller data set created in part (ix) and repeat part (vii). What happens to $\hat{\tau}_{ate}$ and $\hat{\tau}_{att}$? Does dropping the observations have any costs?

2. (30 points) Use the data in TUTOR_RD.DTA to answer the following queestions. The data are on a tutoring program for students entering college. Students participated in the program if and only if their high school GPA was less than or equal to 2.8. The response variable is college grade point average.

(i) Explain the underlying counterfactuals in evaluating the program. What are the two potential outcomes?

(ii) How many students participated in the tutoring program? How many students have a *hsgpa* of exactly 2.8?

(iii) Find the average college GPA for students participating in the program, and compare it to the average college GPA for those who did not. Is the difference statistically significant? What do you conclude?

(iv) Using all of the data, run the regression of *colgpa* on *tutor* and *hsgpa*. Now what is the estimated effect of the tutoring program? Is it statistically significant?

(v) Draw a graph associated with the estimates from part (iv). Remember that the "treatment" rule is of the form $x_i \leq c$.

(vi) Now consider the Stata command

```
reg colgpa tutor hsgpa i.tutor#c.hsgpa, robust
```

Explain what this command is doing. What happens to the coefficient on *tutor*? Why?

(vii) Use the following commands:

```
gen hsgpa_2p8 = hsgpa - 2.8
gen tutor_hsgpa_2p8 = tutor*hsgpa_2p8
reg colgap tutor hsgpa tutor_hsgpa_2p8
```

Now what is the coefficient on *tutor*? Explain why it is much closer to the estimate in part (iv) than in part (vi).

(viii) In the regression from part (vii), restrict the estimation to $2.5 < hsgpa < 3.1$. What happens to the estimated effect of the tutoring program? What about the precision of the estimate?

(ix) Now use the following sequence of commands:

```
ssc install rdrobust
rdrobust colgap hsgpa, c(2.805) kernel(uniform) bwselect(ik)
```

Why is the estimated effect negative? (Hint: `rdrobust` takes the treatment rule as being $x_i \geq c$.) Removing the negative sign, are the results similar to those in part (viii)? How many observations are used on either side of the cutoff?

(x) Obtain and save a graph using the commands

```
rdplot colgpa hsgpa, c(2.805) numbinl(40) numbinr(40)
graph save tutor_rd_1
```

Include the graph and describe what you see.