Niclas Knecht (niclas.knecht@u-bordeaux.fr)
Econometrics
Spring 2024

Problem Set 4

Problem 1 (Stata)

Use the data in **td4_lf.dta** to investigate the determinants of labour force participation among married women during 1975:

$$inlf = \beta_0 + \beta_1\, nwifeinc + \beta_2\, educ + \beta_3\, exper + \beta_4\, exper^2 + \beta_5\, age + \beta_6\, kidslt6 + \beta_7\, kidsge6 + u,$$

where *inlf* is a dummy equal to one if the woman reports working for a wage outside the home at some point during the year, and zero otherwise, *nwifeinc* is husband's earnings (measured in thousands of dollars), *educ* years of education, *exper* past years of labour market experience, *kidslt6* is the number of children less than six years old, and *kidsge6* is the number of kids between 6 and 18 years of age.

a) Estimate the model using LPM. What is the effect of one more small child (*kidslt6*) on the probability of labour force participation?

b) Check if all fitted values are strictly between zero and one.

c) Estimate the same model using logit. Compare your results to LPM.

d) Take a woman with $nwifeinc = 20.13$, $educ = 12.3$, $exper = 10.6$, and $age = 42.5$ — which are roughly the sample averages and $kidsge6 = 1$. What is the estimated effect on the probability of working in going from zero to one small child? What would be the effect of going from one child to two small children?

e) Repeat c) and d) using probit.

Problem 2 (Stata)

Use **td4_gpa.dta** for this exercise. The data set is for 366 student athletes from a large university for fall and spring semesters. Because you have two terms of data for each student, an unobserved effects model is appropriate. The primary question of interest is this: Do athletes perform more poorly in school during the semester their sport is in season?

a) Use pooled OLS to estimate a model with term GPA (*trmgpa*) as the dependent variable. The explanatory variables are *spring*, *sat*, *hsperc*, *female*, *black*, *white*, *frstsem*, *tothrs*, *crsgpa*, and *season*. Interpret the coefficient on *season*. Is it statistically significant?

b) Most of the athletes who play their sport only in the fall are football players. Suppose the ability levels of football players differ systematically from those of other athletes. If ability is not adequately captured by SAT score and high school percentile, explain why the pooled OLS estimators will be biased.

c) Now use the data differenced across the two terms. Which variables drop out? Now test for an in-season effect.

d) Can you think of one or more potentially important, time-varying variables that have been omitted from the analysis?

Problem 3 (Stata)

Use the data **td4_card.dta** for this exercise. Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education. He used a dummy variable for whether someone grew up near a four-year college (*nearc4*) as an instrumental variable for education. In a $log(wage)$ equation, he included other standard controls: experience (*exper*), a black dummy variable (*black*), dummy variables for living in an SMSA (*smsa*) and living in the south (*south*), and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966 (*smsa66*).

a) Estimate the $log(wage)$ equation using OLS. Interpret results.

b) In order for *nearc4* to be a valid instrument, it must be uncorrelated with the error term in the wage equation — we assume this — and it must be partially correlated with *educ*. To check the latter requirement, regress *educ* on *nearc4* and all of the exogenous variables appearing in the equation as in Card (1995) (that is, we estimate the reduced form for *educ*.)

c) Estimate the $log(wage)$ equation using *nearc4* as an IV for *educ* as in Card (1995). Compare results with OLS estimates from a).

d) The difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals from a). Use these to test whether *educ* is exogenous; that is, determine if the difference between OLS and IV is statistically significant.

e) In order for IV to be consistent, the IV for *educ* (*nearc4*) must be uncorrelated with u. Could *nearc4* be correlated with things in the error term, such as unobserved ability? Explain.

f) For a subsample of the men in the data set, an IQ score is available. Regress *IQ* on *nearc4* to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?

g) Now regress *IQ* on *nearc4*, *smsa66*, and the 1966 regional dummy variables *reg662*,..., *reg669*. Are *IQ* and *nearc4* related after the geographic dummy variables have been partialled out? Reconcile this with your findings from c).

h) From c) and d), what do you conclude about the importance of controlling for *smsa66* and the 1966 regional dummies in the $log(wage)$ equation?

Problem 4

Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u,$$

where PC is a binary variable indicating PC ownership.

1. Why might PC ownership be correlated with u?
2. Explain why PC is likely to be related to parents' annual income. Does this mean parental income is a good IV for PC? Why or why not?
3. Suppose that, four years ago, the university gave grants to buy computers to roughly one-half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for PC.

Problem 5

Decide if you agree or disagree with each of the following statements and give a brief explanation of your decision:

a) Like cross-sectional observations, we can assume that most time series observations are independently distributed.

b) The OLS estimator in a time series regression is unbiased under the first three Gauss-Markov assumptions.

c) A trending variable cannot be used as the dependent variable in multiple regression analysis.

Problem 6 (Stata)

Use the data in **td4_sleep.dta** from Biddle and Hamermesh (1990) to study whether there is a trade-off between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

a) Report your results in equation form along with the number of observations and R^2. What does the intercept in this equation mean?

b) If *totwrk* increases by 2 hours, by how much is *sleep* estimated to fall? Do you find this to be a large effect?

Problem 7 (Stata)

A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth rate that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognises this is

$$bwght = \beta_0 + \beta_1 \, cigs + \beta_2 \, faminc + u.$$

a) What is the most likely sign for β_2 and why?

b) Do you think *cigs* and *faminc* are likely to be correlated? Explain why the correlation might be positive or negative.

c) Now estimate the equation with and without faminc, using the data in **td4_cigs.dta**. Report the results in equation form, including the sample size and R-squared. Discuss your results, focusing on whether adding *faminc* substantially changes the estimated effect of *cigs* on *bwght*.