Niclas Knecht (niclas.knecht@u-bordeaux.fr) Econometrics Spring 2025 Due 20.03.2025 11h30

Problem Set 3

Problem 1 (Stata)

Use the data in td3_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 the data td3_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 3

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 onehalf 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 4

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 5 (Stata)

Use the dataset td3_gpa.dta on 4,137 college students and estimate the following equation by OLS:

 $colgpa = \beta_0 + \beta_1 hsperc + \beta_2 sat + u,$

where colgpa is the grade point average (GPA) after the fall semester measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (e.g. if a student is in the top-5% of their class, *hsperc* = 5), and *sat* is the combined maths and verbal score on the Student Achievement Test.

- a) Why does it make sense for the coefficient on *hsperc* to be negative?
- b) What is the predicted college GPA when hsperc = 20 and sat = 1050?
- c) Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but student A's SAT schore was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?
- d) Holding *hsperc* fixed, what difference in SAT scores leads to a predicted *colgpa* difference of 0.50, or one half of a grade point? Comment on your answer.

Problem 6 (Stata)

Consider a model where the return to education depends upon the amount of work experience:

 $log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$

- a) Show that the return to another year of education, holding *exper* fixed, is $\beta_1 + \beta_3 exper$.
- b) State the null hypothesis that the return to education does not depend on the level of *exper*. What do you think is the appropriate alternative? Use the data in td3_wage.dta to test the null hypothesis against your stated alternative.
- c) Let θ_1 denote the return to education, when exper = 10: $\theta_1 = \beta_1 + 10\beta_3$. Obtain $\hat{\theta}_1$ and a 95% confidence interval for θ_1 . (*Hint*: Write $\beta_1 = \theta_1 10\beta_3$ and plug into the equation; then, rearrange. This gives the regression for obtaining the confidence interval for θ_1 .)

Problem 7 (Stata)

Consider the data in td3_sleep.dta. The variable sleep is total minutes per week spent sleeping at work, totwrk is total weekly minutes spent working, educ and age are measured in years, and male is a gender dummy. Estimate the following equation with OLS:

 $sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 male + u.$

- a) All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?
- b) Is there a statistically significant trade-off between working and sleeping? What is the estimated trade-off?

Problem 8 (Stata)

Using data from td3_sat.dta consider the following equation:

 $sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 female + \beta_4 black + \beta_5 female \cdot black + u$

where the variable sat is the combined SAT score, hsize is the size of the student's high school graduating class (in hundreds), female is a gender dummy variable equal to one for women, and black is a race dummy variable equal to one for black people, and zero otherwise.

- a) Estimate the equation. Is there strong evidence that $hsize^2$ should be included in the model? From this equation, what is the optimal (for student's SAT scores) high school size?
- b) Holding *hsize* fixed, what is the estimated difference in SAT score between non-black women and non-black men? Is this estimated difference statistically significant?
- c) What is the estimated difference in SAT score between non-black men and black men? Test the null-hypothesis that there is no difference between their scores.
- d) What is the estimated difference in SAT score between black women and non-black women? What would you need to do to test whether the difference is statistically significant?

Problem 9

Let y_i be independently and identically distributed with mean μ and variance σ^2 , and consider linear estimators of the mean μ of the form $\hat{\mu} = \sum_{i=1}^n a_i y_i$.

- a) Derive the restriction on a_i needed to guarantee that the estimator $\hat{\mu}$ is unbiased.
- b) Derive the expression for the variance of the estimator $\hat{\mu}$.
- c) Derive the linear unbiased estimator that has the minimal variance in this class of estimators.