Niclas Knecht (niclas.knecht@u-bordeaux.fr)
Econometrics
Spring 2025
Due: 27.03.2025, 10h45

Tutorial 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?

   Solution: It has been fairly well established that socioeconomic status affects student performance. The error term u contains, among other things, family income, which has a positive effect on GPA and is also very likely to be correlated with PC ownership.

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?

   Solution: Families with higher incomes can afford to buy computers for their children. Therefore, family income certainly satisfies the second requirement for an instrumental variable: it is correlated with the endogenous explanatory variable. But as we suggested in 1, *faminc* has a positive affect on GPA, so the first requirement for a good IV fails for *faminc*. If we had *faminc* we would include it as an explanatory variable in the equation; if it is the only important omitted variable correlated with PC, we could then estimate the expanded equation by OLS.

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*.

   Solution: This is a natural experiment that affects whether or not some students own computers. Some students who buy computers when given the grant would not have without the grant. (Students who did not receive the grants might still own computers.) Define a dummy variable, *grant*, equal to one if the student received a grant, and zero otherwise. Then, if *grant* was randomly assigned, it is uncorrelated with u. In particular, it is uncorrelated with family income and other socioeconomic factors in u. Further, *grant* should be correlated with PC: the probability of owning a PC should be significantly higher for student receiving grants. Incidentally, if the university gave grant priority to low-income students, grant would be negatively correlated with u, and IV would be inconsistent.

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.

   Solution: Disagree. Most time series processes are correlated over time, and many of them strongly correlated. This means they cannot be independent across observations, which simply represent different time periods. Even series that do appear to be roughly uncorrelated – such as stock returns – do not appear to be independently distributed.

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

   Solution: Agree. We do not need the homoskedasticity and no serial correlation assumptions.

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

   Solution: Disagree. Trending variables are used all the time as dependent variables in a regression model. We do need to be careful in interpreting the results because we may simply find a spurious association between y_t and trending explanatory variables. Including a trend in the regression is a good idea with trending dependent or independent variables.

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 $\beta_1 + \beta_3\, exper$.
Solution: Holding *exper* (and the elements in u) fixed, we have

$$\Delta log(wage) = \beta_1 \Delta educ + \beta_3 (\Delta educ) exper = (\beta_1 + \beta_3 exper)\Delta educ,$$

or

$$\frac{\Delta log(wage)}{\Delta educ} = (\beta_1 + \beta_3 exper).$$

This is the approximate proportionate change in *wage* given one more year of education.

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.
Solution: $H_0 : \beta_3 = 0$. If we think that education and experience interact positively – so that people with more experience are more productive when given another year of education – then $\beta_3 > 0$ is the appropriate 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.

Solution:

$$\begin{aligned}\mathbb{E}[\hat{\mu}] &= \mathbb{E}[\sum a_i y_i] \\ &= \sum a_i\, \mathbb{E}[y_i] \\ &= \sum a_i \mu\end{aligned}$$

Therefore, the restriction is $\sum a_i = 1$.

b) Derive the expression for the variance of the estimator $\hat{\mu}$.

Solution:

$$\begin{aligned} Var[\hat{\mu}] &= Var[\sum a_{ii}] \\ &= \sum Var[a_i y_i] && \text{by i.i.d.: no covariance} \\ &= \sum a_i^2 Var[y_i] \\ &= \sigma^2 \sum a_i^2 \end{aligned}$$

c) Derive the linear unbiased estimator that has the minimal variance in this class of estimators.

Solution: We need to find weights s.t. the variance is minimised under the constraint that $\sum a_i = 1$ (because of the unbiasedness). We can set up the Lagrangian:

$$\mathcal{L}(a_1, ..., a_n, \lambda) = \sigma^2 \sum a_i^2 + \lambda(1 - \sum a_i)$$

We then get the first FOC

$$\frac{\partial \mathcal{L}}{\partial a_i} : 2\sigma^2 a_i = \lambda$$

And we therefore find that $a_1 = a_2 = ... = a_n$, and through the second FOC

$$\frac{\partial \mathcal{L}}{\partial \lambda} : 1 = \sum a_i,$$

we find that $a_i = \frac{1}{n}$.