Niclas Knecht (niclas.knecht@u-bordeaux.fr) Econometrics Spring 2025 Due 03.04.2025 11h30

# Problem Set 4

# Problem 1

Consider a linear model to explain monthly beer consumption:

 $beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u,$ 

with  $\mathbb{E}[u|inc, price, educ, female] = 0$  and  $Var[u|inc, price, educ, female] = \sigma^2 inc^2$ . Write the transformed equation that has a homoskedastic error term.

#### Problem 2 (Stata)

Use the data in td4\_houseprices.dta to estimate the model

$$price = \beta_0 + \beta_1 \, sqrft + \beta_2 \, bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

- a) Write out the results in equation form.
- b) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?
- c) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in b).
- d) What percentage of the variation in price is explained by square footage and number of bedrooms?
- e) The first house in the sample has sqrft = 2,438 and bdrms = 4. What is the predicted selling price for this house from the OLS regression?
- f) The actual selling price of the first house in the sample was USD 300,000 (so price = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

## Problem 3 (Stata)

The file **td4\_ceo.dta** contains data on 177 chief executive officers, which can be used to examine the effects of firm performance on CEO salary.

a) Estimate a model relating annual salary to firm sales and market values. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.

 $log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 log(mktval) + u.$ 

- b) Add profits to the model from a). Why can this variable not be included in logarithmitic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?
- c) Add the variable *ceoten* (years as CEO with this company) to the model from b). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?
- d) Find the sample correlation coefficient between the variables log(mktval) and *profits*. Are these variables highly correlated? What does this say about the OLS estimators?

## Problem 4 (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 5 (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 6 (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  $\beta_2$  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*.

#### Problem 7 (Stata)

The data in td4\_fertility.dta includes, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

- 1. Estimate this model by OLS and interpret the estimates. In particular, holding age fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?
- 2. Frsthalf is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that frsthalf is uncorrelated with the error term from 1, show that frsthalf is a reasonable IV candidate for educ. (Hint: You need to do a regression.)
- 3. Estimate the model from 1 by using *frsthalf* as an IV for *educ*. Compare the estimated effect of education with the OLS estimate from 1.
- 4. Add the binary variables *electric*, *tv*, and *bicycle* to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.