



# **KAN-CEAPV2505U**

## **Econometric Analysis of Firm Data**

July 30, 2025

Copenhagen Business School  
Department of Economics  
Ralf A. Wilke  
Mette Franck  
2025/2026, Autumn semester

### **Problem Set 1: First Steps with STATA/R and Revision of OLS Estimation**

#### **Part 1: First steps with STATA or R**

1. Use the data in wage1.dta for this exercise.
  - (a) Find the average education level in the sample. What are the lowest and highest years of education?
  - (b) Find the average hourly wage in the sample. Does it seem high or low?

- (c) The wage data are reported in 1976 dollars. The Consumer Price Index (CPI) was 56.9 in 1976 and 184.0 in 2003. Use the CPI values to find the average hourly wage in 2003 dollars. Now does the average hourly wage seem reasonable?
  - (d) How many women are in the sample? How many men?
2. The data in *meap01.dta* are for the state of Michigan in the year 2001. Use the data to answer the following questions.
- (a) Find the largest and smallest values of *math4*. Does the range make sense?
  - (b) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?
  - (c) How many schools have math pass rates of exactly 50 percent?
  - (d) Compare the average pass rates for the math and reading scores. Which test is harder to pass?
  - (e) Find the correlation between *math4* and *read4*. What do you conclude?
  - (f) The variable *exppp* is expenditure per pupil. Find the average of *exppp* along with its standard deviation. Would you say that there is wide variation in per pupil spending?

## Part 2: OLS: Estimation

3. The file *CEOSAL2.dta* contains data on 177 chief executive officers and 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 value. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.
  - (b) Add *profits* to the model from part (a). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?
  - (c) Add the variable *ceoten* to the model in part (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?
4. Use BWGHT2.dta to answer this question. 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 weight 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.
- (a) Estimate the model

$$\log(bwght) = \beta_0 + \beta_1cigs + \beta_2npvis + u,$$

and report the results in the usual form, including the sample size and R-squared. Are the signs of the slope coefficients what you expected? Explain.

- (b) If  $npvis$  increases by one sample standard deviation, what is the estimated effect on birth weight?
- (c) Now run the simple regression of  $\log(bwght)$  on  $cigs$ , and compare the slope coefficient with the estimate obtained in part (a). Is the estimated effect of cigarette smoking much different than in part (a).
- (d) Find the correlation between  $cigs$  and  $npvis$  and use it to explain the similarity of the simple and multiple regression estimates of  $\beta_1$ .
- (e) Add the variables  $mage$ ,  $meduc$ ,  $fage$ , and  $feduc$  to the regression from part (a). Is birth weight [more precisely  $\log(bwght)$ ] an easy variable to explain?

These problems have been taken from the textbook "Introductory Econometrics" by J.Wooldridge, 7th edition, 2020.