

APPLIED ECONOMETRICS I

ACADEMIC YEAR 2021–2021

ASSIGNMENT ONE

Deadline for sending in answers: Monday, 5 October 2020, 23h59

To be sent by email to Tony O'Connor (tony.oconnor@coleurope.eu).

Please name the file you attach in the following way:

YourSurname.YourFirstName_Applied_Econometrics_Assignment_1

Example: OConnor.Tony_Applied_Econometrics_Assignment_1

Please supply answers to all the exercises in the attached document. You may work on the assignment together but each one has to send in an individual copy.

Write your name clearly at the **top** of the first page. (Use a title similar to the one above but with name preceded by “submitted by:” added below the assignment number. Number your pages.

The attached data are in Stata format, the software which must be used for the assignment.

QUESTION ONE

Suppose that a researcher using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression (std. error in parenthesis):

$$\text{TestScore} = 520.4 - 5.82 \times \text{CS}, \quad R^2 = 0.08, \quad \hat{\sigma} = 11.5$$

(20.4) (2.21)

- a. A classroom has 22 students. What is the regression's prediction for that classroom's average test score?
- b. The sample average class size across the 100 classrooms is 21.4. What is the sample average of the test score across the 100 classrooms?
- c. Construct a 95% confidence interval for β_1 , the regression slope coefficient.
- d. What is the p -value for the two-sided test of the null hypothesis, $H_0 : \beta_1 = 0$. What do you conclude from this?
- e. Calculate the p -value for the two-sided test of the null hypothesis, $H_0 : \beta_1 = -5.6$. Without doing any additional calculations, determine whether -5.6 is contained in the 95% confidence interval for β_1 .

QUESTION TWO

A regression of average weekly earnings (*AWE*, measured in dollars) on age (measured in years) using a random sample of college-educated full-time workers aged 25 – 65 yields the following:

$$AWE = 696.7 + 9.6 \times Age, \quad R^2 = 0.023, \quad \hat{\sigma} = 624.1$$

- Explain what the coefficient values, 696.7 and 9.6, mean.
- The standard error of the regression $\hat{\sigma}$, is 624.1. What are the units of measurement for $\hat{\sigma}$? (Dollars? Years? Or is it unit-free?)
- The regression R^2 is 0.023. What are the units of measurement for the R^2 ? (Dollars? Years? Or is it unit-free?)
- What is the regression's predicted earnings for a 25-year old worker? A 45-year old worker?
- Will the regression give reliable predictions for a 99-year old worker? Why or why not?

QUESTION THREE

The data file **Growth** contains data on average growth rates from 1960 through 1995 for 65 countries along with variables that are potentially related to growth. A detailed description is given in **Growth_Description**. In this exercise, you will investigate the relationship between growth and trade.

- Construct a scatterplot of average annual growth rate (*Growth*) on the average trade share (*TradeShare*). Does there appear to be a relationship between the two variables?
- One country, Malta, has a trade share much larger than the other countries. Find Malta on the scatterplot. Does Malta look like an outlier?
- Using all observations, run a regression of *Growth* on *TradeShare* (use the option `robust`). What is the estimated slope? What is the estimated intercept? Use the regression to predict the growth rate for a country with a trade share of 0.5 and with a trade share equal to 1.0.
- Estimate the same regression excluding the data for Malta. (Use the same command as in c., but add `if country_name != "Malta"` at the end without a comma.) Answer the same questions as in c.
- Should Malta be included in or excluded from the analysis?

QUESTION FOUR

The data file **TeachingRatings** contains data on course evaluations, course characteristics and professor characteristics for 463 courses at the University of Texas at Austin. A detailed description is given in **TeachingRatings_Description**. One of the characteristics is an index of the professor's "beauty" as rated by a panel of six judges. In this exercise, you will investigate how course evaluations are related to the professor's beauty.

- a. Construct a scatterplot of average course evaluations (*Course_Eval*) on the professor's beauty (*Beauty*). Does there appear to be a relationship between the two variables?
- b. Run a regression of average course evaluations (*Course_Eval*) on the professor's beauty (*Beauty*). Use the option `robust`. What is the estimated intercept? What is the estimated slope? Explain why the estimated intercept is equal to the sample mean of *Course_Eval*.
- c. Professor A has an average value of *Beauty*, while Professor B's value of *Beauty* is one standard deviation above the average. Predict Professor B's and Professor A's course evaluations.
- d. Comment on the size of the regression's slope. Is the estimated effect of *Beauty* on *Course_Eval* large or small? Explain what you mean by "large" and "small".
- e. Does *Beauty* explain a large fraction of the variance in evaluations across courses? Explain.
- f. What is the 95% confidence interval for the effect of *Beauty* on *Course_Eval*.
- g. What is the *p*-value for the null hypothesis that the slope coefficient is equal to zero? What do you conclude from this?

QUESTION FIVE

The data file **CollegeDistance** contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. In this exercise, you will use the data to investigate the relationship between the number of completed years of education for young adults and the distance from each student's high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete more years of higher education. It is assumed that students go to a high school close to their home.) A detailed description of the data is given in **CollegeDistance_Description**.

- a. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*), where *Dist* is measured in tens of miles. Use the option `robust`. What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How does the average value of years of completed schooling change when colleges are built closer to where students go to high school?

- b. Bob's high school was 20 miles from the nearest college. Predict Bob's years of completed education using the estimated regression.
- c. An education advocacy group argues that, on average, a person's educational attainment would increase by approximately 0.15 year if distance to the nearest college is decreased by 20 miles. Is the advocacy group's claim consistent with the estimated regression? Explain.
- d. Does distance to college explain a large fraction of the variance in educational % attainment across individuals? Explain.
- e. What is the value of the standard error of the regression? What are the units of measurement for the standard error?
- f. Construct a 95% confidence interval for the slope coefficient. What is the p -value for the null hypothesis that the slope is different from zero? What do you conclude from this?
- g. Run the regression using data only on females and repeat f.. (The data on the gender of the person is contained in the variable, *female*. If *female* equals one, the observation concerns a female; if *female* equals zero, the observation concerns a male.)
- h. Run the regression using data only on males and repeat f..
- i. Note that the value of the estimated coefficient for *Dist* in g. is contained in the confidence interval for *Dist* in h.; and that the value of the estimated coefficient for *Dist* in h. is contained in the confidence interval for *Dist* in g.. Can you conclude from this that the two coefficients are not statistically significantly different from each other?

QUESTION SIX

The dataset **TeacherSalary&PupilSpending** contains data on average public teacher pay (annual salary in dollars) and spending on public schools (dollars) in 1985 for 50 U.S. states and the district of Columbia.

- a. Plot the data: a scatterplot of salary on spending.
- b. Estimate a regression of salary on spending using the whole sample. Use the option `robust`.
- c. Establish a 95% confidence interval for the coefficient on spending.
- d. Test the hypothesis that the coefficient on spending is equal to 3.
- e. Re-estimate the regression of salary on spending excluding observation number 50.
- f. Do you think observation number 50 "significantly" biases the regression line? In other words, do you think that the observation is an outlier? A formal test will be provided in Chapter 6.