ECONOMETRICS
ACADEMIC YEAR 2020–2021
ASSIGNMENT FIVE

Deadline for sending in answers: Monday, 09 November, 23.59 hours.

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_5

Example: OConnor.Tony_Applied_Econometrics_Assignment_5

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

Use the data in `wage2.dta` for this exercise and the option `robust`.

a) Estimate the model

$$\log(wage)_i = \beta_0 + \beta_1 black_i + \beta_2 married_i + \beta_3 educ_i + \beta_4 exper_i + \beta_5 tenure_i + \beta_6 south_i + \beta_7 urban_i + u_i.$$

Holding other factors fixed, what is the approximate difference in monthly salary between blacks and non-blacks? Is this difference statistically significant?

b) Add the variables $exper_i^2$ and $tenure_i^2$ to the equation and test their join significance.

c) Extend the original model in (a) to allow the return to education ($educ_i$) to depend also on race and test whether the return to education does depend on race.

d) Re-estimate the original model, but now allow the intercept for the wage equation to differ across the four groups of people: married and black, married and non-black, single and black, and single and non-black.

e) What is the estimated wage differential between the four groups?

f) Repeat (d) allowing both the intercept and the return to education ($educ_i$) to differ across the four groups of people.

g) Graph the relationship between (the log of) estimated wages in (f) and the level of education? Use the –`margins`– and –`marginsplot`– for this exercise. What do you conclude concerning the relative evolution of wages between the four categories as a function of wages?

Question Two

Exercises (a) through (g) are based on the following scenario: 700 income-earning individuals from a district were randomly selected and asked whether they were employed by the government ($Gov_i = 1$) or failed their entrance test ($Gov_i = 0$); data were also collected on their gender ($Male_i = 1$) if male and = 0 if female) and their years of schooling (*Schooling*$_i$, in years). The following table summarizes several estimated models.

Dependent Variable: *Gov*

	Probit (1)	Logit (2)	Linear Probability (3)	Probit (4)	Logit (5)	Linear Probability (6)	Probit (7)
Schooling	0.272 (0.029)	0.551 (0.062)	0.035 (0.003)				0.548 (0.091)
Male				−0.242 (0.125)	−0.455 (0.234)	−0.050 (0.025)	4.352 (1.291)
Male × Schooling							−.344 (0.096)
Constant	−4.107 (0.358)	−8.146 (0.800)	−0.172 (0.027)	−1.027 (0.098)	−1.717 (0.179)	0.152 (0.021)	−7.702 (1.238)

a) Using the results in column (1):
   - (i) Does the probability of working in the government depend on *Schooling*? Explain.
   - (ii) Matthew has 16 years of schooling. What is the probability that he will pass the test?
   - (iii) Christopher never went to college (12 years of schooling). What is the probability that he will get a job with the government?
   - (iv) The sample included values of *Schooling* between 0 and 18 years, and only five people in the sample had more than 15 years of schooling. Jed has completed a PhD and has been a student for 24 years. What is the model's prediction for the probability that Jed will get a job with the government? Do you think that this prediction is reliable? Why or why not?

b) Answer (i) through (iii) in (a) using the results in column (2).

c) Sketch the predicted probabilities from the probit and logit in columns (1) and (2) for values of *Schooling* between 0 and 18. Are the probit and logit models similar?

d) Answer (i) through (iii) in (a) using the results in column (3).

e) Sketch the predicted probabilities from the probit and linear probability in columns (1) and (3) as a function of $Schooling_i$ for values of *Schooling* between 0 and 18. Do you think that the linear probability is appropriate here? Why or why not?

f) Using the results in columns (4) through (6):
   (i) Compute the estimated probability of passing the test for men and for women.
   (ii) Are the models in (4) through (6) different? Why or why not?

g) Using the results in column (7):
   (i) Akira is a man with 10 years of schooling. What is the probability that the government will employ him?
   (ii) Jane is a woman with 12 years of schooling. What is the probability that the government will employ her?
   (iii) Does the effect of the years of schooling on employment in the government depend on gender? Explain.

QUESTION THREE

Use the data in `wage2.dta` for this exercise to estimate return to education for men and the option robust.

a) Run an OLS and IV regressions to estimate return to education for men, using number of siblings (*sibs*) as an instrument for *educ*. Compare the results of the two regressions and explain why OLS and IV give different estimation results. The regression which you need to estimate looks as follows. [Always report `robust` standard errors.]

$$\log(wage)_i = \alpha_0 + \alpha_1 educ_i + u_i$$

b) The R^2 computed in the IV regression from part (a) is negative. Please discuss the properties of R^2 in the context of IV estimation and whether it can be negative.

c) The variable *brthord* is birth order (*brthord* is one for a first-born child, two for a second-born child, and so on). Explain why *educ* and *brthord* might be negatively correlated. Regress *educ* on *brthord* to determine whether there is a statistically significant negative correlation.

d) Use *brthord* as an IV for *educ* in the above equation. Report and interpret the results. Compare the results to the OLS and IV (using number of siblings (*sibs*) as an instrument for *educ*) estimations.

e) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

$$\log(wage)_i = \alpha_0 + \alpha_1 educ_i + \alpha_2 sibs_i + u_i$$

Suppose that we want to use *brthord* as an IV for *educ*, assuming that *sibs* is exogenous. The reduced form for *educ* is

$$educ_i = \pi_0 + \pi_1 sibs_i + \pi_2 brthord_i + v_i$$

Estimated the reduced form. Test and discuss whether *brthord* passes the weak instruments threshold.

f) Estimate the equation from part (e) using *brthord* as an IV for *educ* (and *sibs* as its own IV). Comment on the standard errors for $\hat{\alpha}_{educ}$ and $\hat{\alpha}_{sibs}$.

g) Using the fitted values from part (e), $\widehat{educ}$, compute the correlation between $\widehat{educ}$, and *sibs*. Use this result to explain your findings from part (f).

QUESTION FOUR

Use the US state-level data on murder rates and executions in murder.dta for the following exercise.

Variable name	Description
state	state identifier
year	87, 90, or 93
mrdrte	murders per 100,000 population
exec	total executions, past 3 years
unem	annual unemployment rate
d90	=1 if year == 90
d93	=1 if year == 93
cmrdrte	$mrdrte_t - mrdrte_{t-1}$
cexec	$exec_t - exec_{t-1}$
cunem	$unem_t - unem_{t-1}$
cexec_1	$cexec_{t-1}$
cunem_1	$cunem_{t-1}$

a) Consider the following regression model

$$mrdrte_{it} = \mu + \beta_1 exec_{it} + \beta_2 unem_{it} + \beta_3 \alpha_t + u_i + e_{it}$$

where *mrdrte* are murders per 100,000 population, μ denotes overall mean, *exec* are total executions over the past 3 years, *unem* is annual unemployment rate, α_t denotes time fixed effects and u_i is the unobserved state effect; *i* denotes states and *t* denotes years. If past executions of convicted murderers have a deterrent effect, what should be the sign of β_1? What sign do you think β_2 should have? Explain.

b) Using just the years 1990 and 1993, estimate the equation from part (i) by *pooled* OLS. For the moment, ignore the serial correlation and heteroscedasticity problem in the composite error term. Do you find any evidence for a deterrent effect? Explain.

c) Using 1990 and 1993, estimate the equation from part (i) by *fixed effects*. Is there evidence of a deterrent effect? Is this a large effect? Discuss the size of the estimated effect considering the distribution of variables *mrdrte* and *exec*.

d) Compute the heteroskedasticity-robust standard errors for the estimation in part (iii). Interpret the estimated deterrent effect.

e) Since you only have two years of data, estimate the model by *first differencing* (FD). What does the theory suggest for the FE and FD estimators when $T = 2$? Do your results confirm this? [Hint: Use Stata `delta(#)` option when setting the data to panel data in order to specify how many time periods there are in between observations.]

f) Estimate the equation from part (i) by *fixed effects*, dropping Texas (TX) from the analysis. Compute the usual and heteroskedasticity-robust standard errors. What do you find with respect to the estimated deterrent effect? How do you explain this result?

g) Use all three years of data and estimate the model from part (i) by *fixed effects*. Include Texas (TX) in the analysis. Discuss the size and statistical significance of the deterrent effect compared with only using 1990 and 1993.

QUESTION FIVE

The file `traffic2.dta` contains 108 monthly observations on automobile accidents, traffic laws, and some other variables for California from January 1981 through December 1989. Use this data set and the option `robust` to answer the following questions.

a) During what month and year did California's seat belt law take effect? When did the highway speed limit increase to 65 miles per hour?

b) Regress the variable log(*totacc*) on a linear time trend and 11 monthly dummy variables, using January as the base month. Interpret the coefficient estimate on the time trend. Would you say there is seasonality in total accidents?

c) Add to the regression from part (b) the variables *wkends*, *unem*, *spdlaw*, and *belt-law*. Discuss the coefficient on the unemployment variable. Does its sign and magnitude make sense to you?

d) In the regression from part (c), interpret the coefficients on *spdlaw* and *beltlaw*. Are the estimated effects what you expected? Explain.

e) The variable *prcfat* is the percentage of accidents resulting in at least one fatality. Note that this variable is a percentage, not a proportion. What is the average of *prcfat* over this period? Does the magnitude seem about right?

f) Run the regression in part (c) but use *prcfat* as the dependent variable in place of log(*totacc*). Discuss the estimated effects and significance of the speed and seat belt law variables.