Applied Econometrics Assignment

Jeroen Kerkhof

25 November 2022

1 Practicalities

Assignment for the Econometrics class. Due on Friday 30 December 2022, 23.59h. Assignments can be made in groups of maximum 2 people. Discussions between groups regarding general content and programming techniques is allowed (actually encouraged). However, copy-pasting of material (code or report) is NOT allowed and will be reported. You are expected to hand-in a report with discussion of the questions and results, tables and figures. No code should be present in the report. This should be provided separatedly in a **working** .py (or .r) file. I need to be able to run that file if the data set is in the same folder as the .py file. DO NOT LINK TO folders like `'C:\Dropbox\blablabla\econometrics\assignment\data'`

For the report, you are expected to clearly translate your technical findings into plain English. Just reporting tables with estimatess and graphs is NOT enough. If you are unsure, what is expected, you might want to watch the plain_english.mp4 (again) and put yourself in the role of the analyst (though you might want to leave out the part about me being a golden retriever).

2 Cheating in the Sumo world

1. Comment on the scenes regarding cheating in sumo.mp4. Use your knowledge of econometrics, statistics and probability in order to explain what is described. Use terms such as randomness, random sample, statistical significance, selection bias, etc.

3 Simulation study

You will investigate the distributions of the OLS estimator and the corresponding t-values, and the model test.

$$\boldsymbol{y}_i = \beta_0 + \beta_1 \boldsymbol{x}_{1i} + \beta_2 \boldsymbol{x}_{2i} + \boldsymbol{\varepsilon}_i, \quad \text{with } \mathbb{E}[\boldsymbol{\varepsilon}_i] = 0 \text{ and } V[\boldsymbol{\varepsilon}_i] = \sigma^2 \tag{1}$$

Table 1: Simulation combinations

$n = 100, \boldsymbol{\varepsilon}_i \sim N\left(0, \sigma^2\right)$	$n = 100, \boldsymbol{\varepsilon}_i \sim \text{Laplace}(0, \sigma/\sqrt{2})$
$n = 10, \boldsymbol{\varepsilon}_i \sim N\left(0, \sigma^2\right)$	$n = 10, \boldsymbol{\varepsilon}_i \sim \text{Laplace}(0, \sigma/\sqrt{2})$

In order to have different experiments for each group, all groups need to use a different seed. Select as the seed the product of your birthdays (dd/mm) when you concatenate the day and month identifiers.

E.g. 01/01 and 31/12 becomes 101 * 3112 = 314312.

```
 # first birthday
bd_1 = 3112
# second birthday
bd_2 = 3112

group_seed = bd_1 * bd_2

# seed the random number generator
rng = np.random.default_rng(group_seed)
```

We are interested in the distribution of $\widehat{\beta}$, $\boldsymbol{t}$ and $\boldsymbol{F}$ (the model test).

Consider the 4 situations in Table 1.

1. Generate the variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for all observations one time. Assume that $\boldsymbol{x}_1 \sim N(5, 1)$ and $\boldsymbol{x}_2 \sim N(0, 2)$ and $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0.5$. Create a matrix X that includes both the realized data x_1 and x_2 in addition to a constant term. Note: You will only create one copy of this matrix. Using this matrix you will simulate multiple vectors for the dependent variable.

2. Set the true values of $\beta = (10, 0.5, 2.2)$ and $\sigma = 0.25$. Generate error terms for each observation for each simulation. This means generate a matrix $\boldsymbol{\varepsilon}$ with n rows (number of observations) and S columns (number of simulations). Hence, generate

$$\mathrm{E} = \begin{bmatrix} \varepsilon_1^{(1)} & \cdots & \varepsilon_1^{(S)} \\ \vdots & \ddots & \vdots \\ \varepsilon_n^{(1)} & \cdots & \varepsilon_n^{(S)} \end{bmatrix} \tag{2}$$

   for each of the combinations in Table 1. Use $2^{16} (= 65, 536)$ simulations. Use rng.normal for the normal random number generation and rng.laplace for the laplace random number generation.

3. Using the linear model

$$\boldsymbol{y}_i = \beta_0 + \beta_1 \boldsymbol{x}_{1i} + \beta_2 \boldsymbol{x}_{2i} + \boldsymbol{\varepsilon}_i$$

generate the dependent variable $y_1, ..., y_n$ for each simulation.

$$Y = \begin{bmatrix} y_1^{(1)} & \cdots & y_1^{(S)} \\ \vdots & \ddots & \vdots \\ y_n^{(1)} & \cdots & y_n^{(S)} \end{bmatrix} \tag{3}$$

4. Create a plot of the error terms for both distributions. Do they look alike?

5. For each simulation calculate $\widehat{\beta}$, the OLS estimate, the $t-$tests (vs the true values) and the model test (always vs zero values for the parameters). You should have S OLS estimates, S $t-$tests for each explanatory variable (including) the constant and S values for the model test.

6. Create histograms for each of the quantities you found in the previous question.

7. For the $t-$tests plot the density of the appropriate $t-$ distribution in the graph. Similarly, for the model test plot the appropriate F distribution. Do they seem accurate in all cases?

8. Instead of the $t-$distributions, now plot the density of the normal distribution in the graph. Similary, for the model test, plot the appropriate χ^2 distribution. Do they seem accurate in all cases?

9. Explain the consequences of your results. Which of the four (sample size, distribution) combinations is problematic?

4 Empirical investigation

In this part you are asked to perform an empirical analysis on the return on education. It considers the wage as a function of a number of explanatory variables among which whether you go to junior college (2-years) or university (4-years). The data is in data.csv and the variable descriptions are in Table 2. You will use a subset of this data. However, everyone will use a different subset depending on your group seed.

```
# read the full data set
data_full = pd.read_csv('data.csv')
num_obs = 6000
# select 6000 observations randomly (the rng uses your seed)
observations = rng.choice(len(data_full), num_obs,
                          replace=False)
# select on the observations for your group
data = data_full.iloc[observations, :].copy()
```

4.1 Earnings from schooling

The dependent variable is $\log(\text{hourly wage}) = lwage$. Hence, we have

$$\boldsymbol{lwage_i} = \beta_0 + \beta_1 \boldsymbol{x}_{1i} + ... + \beta_2 \boldsymbol{x}_{Ki} + \boldsymbol{\varepsilon}_i \tag{4}$$

You are encouraged to go beyond the questions asked here, but please motivate what you are doing.

1. Investigate the descriptive statistics and note any peculiarities (if any).

2. Which of the non-linear models we saw in class is used here?

3. Use as explanatory variables, jc, $univ$ and $exper$ (in this order). Estimate the model using OLS. Which variables are economically significant, and which variables are statistically significant? Explain.

4. Now test whether β_1 (jc) is equal to β_2 $(univ)$. Explain which test you use, the distribution and whether or not you can reject the null hypothesis.

5. In the test above you tested for equality. What if you wanted to test if $\beta_2 > \beta_1$? This cannot be done in the usual manner. However, if you write $\theta = \beta_2 - \beta_1$ we have $\beta_1 = \beta_2 + \theta$. Plug this into your regression equation and collect terms. Now re-estimate the model using $totcoll = jc + univ$ instead of $univ$ using OLS. Compare your estimates to the results above. Can you do a one-sided test now? If so, can you reject the null-hypothesis?

6. The variable phsrank is the persons high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Add phsrank to your original model and report the OLS estimates in the usual form. Is phsrank statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?

7. Does adding phsrank to the original model substantively change the conclusions on the returns to two- and four-year colleges? Explain.

   Continue to use the model above for the next questions.

8. Test if the overall model makes sense. Write down the null hypothesis, the test statistic and its distribution (under the null) and the test results.

9. Compute the confidence and prediction intervals for all observations. That is, construct, confidence and prediction intervals around $\widehat{y}$. Provide a plot (y vs $\widehat{y}$ with the confidence and prediction intervals.)

10. There is the potential for heteroskedasticity. Perform a test for this and if needed adjust the standard errors to heteroskedasticity consistent [HC] standard errors. Write down the null hypothesis. Do you still find the same level of significance?

Table 2: Description of the variables in the dataset

Variable	Description
female	=1 if female
phsrank	high school rank; 100 = best
BA	=1 if Bachelor's degree
AA	=1 if Associate's degree
black	=1 if African-American
hispanic	=1 if Hispanic
id	ID Number
exper	total (actual) work experience
jc	total 2-year credits (junior college)
univ	total 4-year credits (university)
lwage	log hourly wage
stotal	total standardized test score
smcity	=1 if small city, 1972
medcity	=1 if med. city, 1972
submed	=1 if suburb med. city, 1972
lgcity	=1 if large city, 1972
sublg	=1 if suburb large city, 1972
vlgcity	=1 if very large city, 1972
subvlg	=1 if sub. very lge. city, 1972
ne	=1 if northeast
nc	=1 if north central
south	=1 if south
totcoll	jc + univ

11. Try to improve the model by adding additional variables in the dataset (or combinations of them). Explain your choices and results.

12. Before handing in, read the last line of Section 1 again and make sure that the code runs from a clean (freshly started) Python or R session. You can do this by restarting the kernel or (if you want to be absolutely sure) the whole of Python / R.

Best of luck!