Applied Econometrics Assignment

Jeroen Kerkhof

December 2023

1 Practicalities

Assignment for the Econometrics class 2023-2024. Due on Thursday 28 December 2023, 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).

In order to use the code examples without issues or need for your own customization, you should use the folder structure in Figure 1. You can also find the recommended folder structure with examples on Canvas.

2 Selection bias

1. Provide several (minimum 3) real-life examples of selection bias that were not covered in class. Discuss the impact on econometric studies and make suggestions on how they could be handled.
2. Comment on the following fragment Musk on Rogan Hint: You can still become very rich even if you don't understand the basics of Statistics 101.

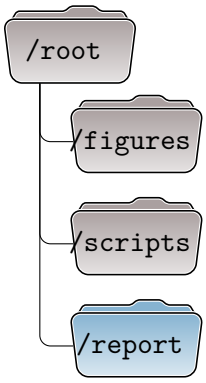

Figure 1: Recommended folder structure. Figures for graphs, scripts for Python / R scripts, report for the LaTeX report.

3 Simulation study: Heteroskedasticity

You will investigate the distributions of the OLS and GLS estimator in the presence of heteroskedasticity.

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

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. Heteroskedastic and homoskedastic models for 2 different sample sizes.

1. Generate the variable $\boldsymbol{x}_1$ for all observations one time. Assume that $\boldsymbol{x}_1 \sim N(20, 2)$. Create a matrix X that includes both the realized data x_1 in addition to a constant term.

Table 1: Simulation combinations

num obs	heteroskedastic	homoskedastic
$n = 1000,$	$\boldsymbol{\varepsilon}_i \sim N\left(0, \sigma^2 x_i\right)$	$\boldsymbol{\varepsilon}_i \sim N\left(0, \sigma^2 \bar{x}\right)$
$n = 100,$	$\boldsymbol{\varepsilon}_i \sim N\left(0, \sigma^2 x_i\right)$	$\boldsymbol{\varepsilon}_i \sim N\left(0, \sigma^2 \bar{x}\right)$

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 $\beta = (1, 2)$ and $\sigma = 0.3/\sqrt{20}$. 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.

3. Using the linear model

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

generate the dependent variable $y_1, ..., y_n$ for each simulation using both the homoskedastic and heteroskedatic model.

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

4. Create plots of the error terms for the heteroskedastic and the homoskedastic case. First, create a histogram of the average error for each observation for both the heteroskedastic and homoskedastic case. Second, create a scatter plot of the *squared* error terms. Discuss any differences.

5. What is the Ω matrix for the GLS estimator in the heteroskedastic case. For completeness, also calculate Ω for the homoskedastic case and point out the difference.

6. For each simulation calculate $\widehat{\beta}$, the OLS and GLS estimates (For the homoskedastic case OLS and GLS should be the same, so only 1 is required) and the accompanying $t-$tests (vs the true values, NOT 0). You should have S OLS and GLS estimates, S $t-$tests for each explanatory variable (including) the constant and S values for the model test.

7. Create histograms for each of the quantities you found in the previous question.
8. For the $t-$tests (assuming a level of 5%) check how often the tests are rejected compared to the desired 5%. Do they seem accurate in all cases?
9. For the heteroskedastic case, adjust the standard errors to White standard errors. Does this make the $t-$tests more precise?

4 Empirical investigation

In this part you are asked to perform an empirical analysis on the the connection between work and sleep. It considers the relation between work and sleep. The data is in sleep_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('sleep_data.csv')
num_obs = 600
# select 600 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 Work and sleep

The dependent variable is the amount of sleep per week, $sleep_i$. To start, we have the basic model

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \varepsilon_i \tag{4}$$

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. If people trade off work and sleep what the sign of β_1 be?
3. Report your results in equation form along with the number of observations and R^2. What does the constant in this equation mean?
4. If $totwrk$ increases by 2 hours, by how much is sleep estimated to fall? Do you find this to be a large effect?
5. Now add the variables, $educ$ and age to the model. Before estimating, consider the sign of β_2 ($educ$) and β_3 (age) and explain your reasoning.

6. Re-estimate the model including $educ$ and age. Report your results.

7. If someone works five more hours per week, by how many minutes is sleep predicted to fall? Is this a large tradeoff?

8. Discuss the sign and magnitude of the estimated coefficient on $educ$.

9. Would you say $totwrk$, $educ$, and age explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with totwrk?

10. Now add the variables age^2 and $yngkid$. Estimate the model again and report your results.

11. Now estimate the model from the previous question separately for men and women. Report and interpret the results.

12. Instead of estimating the model for men and women separately, estimate it once, but including a gender dummy variable ($male$). Report and interpret your results (compare to your previous estimations). Why is this model better than the ones above?

13. Discuss the statistical significance of each variable.

14. Report the model test and discuss if the model is of any use in explaining sleep.

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

16. 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!

Table 2: Description of the variables in the dataset

Variable	Description
age	in years
black	=1 if black
case	identifier
clerical	=1 if clerical worker
construc	=1 if construction worker
educ	years of schooling
earns74	total earnings, 1974
gdhlth	=1 if in good or excellent health
inlf	=1 if in labor force
leis1	sleep - totwrk
leis2	slpnaps - totwrk
leis3	rlxall - totwrk
smsa	= 1 if live in smsa
lhrwage	log hourly wage
lothinc	log othinc, unless othinc < 0
male	= 1 if male
marr	= 1 if married
prot	= 1 if Protestant
rlxall	slpnaps + personal activs
selfe	=1 if self employed
sleep	mins sleep at night, per week
slpnaps	mins sleep, including naps, per week
south	=1 if live in south
spsepay	spousal wage income
spwrk75	=1 if spouse works
totwrk	mins worked per week
union	=1 if belong to union
worknrm	mins work main job
workscnd	mins work second job
exper	age - educ - 6
yngkid	=1 if children < 3 present
yrsmarr	years married
hrwage	hourly wage
agesq	age^2