Econometrics

TA Sessions

Sebastian Ellingsen
Universitat Pompeu Fabra
2016/12/12 (updated: 2020-06-11)

Session 1

Excercise 1

Excercise 1

Background

In the late 1980s the Tennessee state legislature funded a four-year experiment to evaluate the effect of small class sizes on learning (Student Teacher Achievement Ratio or STAR).

The experiment compared three different class arrangements for children in kindergarten through third grade:

- A regular-size class (size 22-25 pupils) with a single teacher – The control group.
- A small class (size 13-17 pupils) with a single teacher.
- A regular-size class with a single teacher and a teacher's aide.

Participating schools were picked at random from the universe of public schools in Tennessee.

Each participating school had at least one class of each type.

Within schools both pupils and teachers were randomly assigned to one of the three types of classes. Each year the children were given standardized tests (the SAT) and these are the outcome measures

In this exercise you will analyze the test results for children in kindergarten by comparing the children in the small class with those in the regular class without aide.

Data

The variables included are:

- *sck* = whether kid in small class (the treatment 1, or control 0),
- *tscorek* = the test score of the child
- *boy* = whether kid a boy (1) or a girl (0),
- *freelunk* = whether kid gets a free lunch (proxy for being from poor household; 1),
- *totexpk* = years of teaching experience of teacher,
- *schidkn* = code for particular school.

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
TOTEXPK	5,749	9.31	5.77	0	5	9	13	27
SCHIDKN	5,749	39.83	22.96	1	20	39	60	80
TSCOREK	5,749	922.39	73.87	635	870	915	964	1,253
SCK	5,749	0.30	0.46	0	0	0	1	1
BOY	5,749	0.51	0.50	0	0	1	1	1
FREELUNK	5,749	0.48	0.50	0	0	0	1	1

- Look at the summary statistics to get a rough overview of the dataset.
- Spot errors.
- Get a sense of the variation, scale, etc.

Excercise 1

Empirical Model

Consider the bivariate regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

- Y_i denotes the tscorek.
- X_i an indicator variable equal to 1 if the kid was in a small class and 0 otherwise.
- u_i is the error term.

Question 1

Explain what the term u_i represents.

Excercise 1

Question 1

The term u_i represents all other factors that could have influenced the test score besides the size of the class.

For example,

- Factors such as teacher quality,
- Background of their students,
- Luck of the students on test day).

Taken together, u_i and X_i explain all the variation in Y_i.

Excercise 1

Question 1

Do the factors included in u_i change when we change the size of the class?

For example,

- is it the case that the school puts better students in the large classes?
- Assigns different quality teachers?
- Changes the class composition, and so on..

If so, we say that X_i is endogenous and $\hat{\beta}_1$ should be interpreted with caution.

Question 2

Suppose $\beta_1 > 0$. Does a kid who attended a small class necessarily have a higher test score than a kid who went to regular class? Explain.

Excercise 1

Question 2

This is not necessarliy true, however it is true on average since,

$$\beta_1 = E[Y_i|X_i = 1] - E[Y_i|X_i = 0].$$

It is possible that a student that attend a regular class $X_i = 0$, can get a higher score than one that attend a small class $X_i = 1$.

This is the case if the effect of the other factors is bigger than the effect of the class size.

i.e $u_i \geq \beta_1$.

Question 3

Run a regression of tscorek on whether you were in a small class (we are going to ignore the other treatment (being in a small class with aide).

- Interpret the coefficients.
- What is the expected test score for a kid in a small class?
- And for a kid in a regular class?

Excercise 1

Question 3

We run the following regression,

```
m1 ← lm(data=star, formula = tscorek~sck)
```

Excercise 1

Question 3

	tscorek
SCK	13.826***
	(2.115)
CONSTANT	918.225***
	(1.161)
Observations	5,749
R-squared	0.007
Adjusted R-squared	0.007
Residual standard error	73.604 (df = 5747)
F statistic	42.714*** (df = 1; 5747)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 1

Question 3

The point estimate is 13.83.

Both the estimate of the intercept and the slope are estimated with high precision.

This is our estimate of the causal class size effect, or the expected testscore gain from attending a small class, with respect to attending a large class.

- β_0 represent the expected score from a student of a regular class size
- β_1 is the expected differences in score between a student from a small size class and one from a regular class size.

Question 4

Show that and that when X_i is binary: $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$.

Excercise 1

Question 4

Treatment

Statistic	N	Mean	St. Dev.	Min	Median	Max
sck	1,733	932.05	76.43	747	924	1,253

Control

Statistic	N	Mean	St. Dev.	Min	Median	Max
sck	4,016	918.23	72.35	635	911.5	1,253

Note that the effect equals the difference in means for the treatment and control groups. Why?

Question 5

Consider now the case where an explanatory variable X is the number of teaching experience of a given teacher (*totexpk*).

- Run the regression of *tscorek* on *totexpk*.
- Interpret the estimated value for the coefficient of this variable.

Excercise 1

Question 5

	tscorek
TOTEXPK	1.418***
	(0.168)
CONSTANT	909.194***
	(1.838)
Observations	5,749
R-squared	0.012
Adjusted R-squared	0.012
Residual standard error	73.422 (df = 5747)
F statistic	71.340*** (df = 1; 5747)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 1

Question 5

The interpretation of the estimated coefficient is the following:

If the number of years of experience of the teacher increases by one then the average value of the test score will increase by 1.42 points.

Does this confirm your prior?

Excercise 1

Question 5

Do we interpret this as being informative about what would happen if we increased the experience of teachers?

- Possible to think about unobserved variables included in the error term which could be correlated with the variable *totexpk* and which have an effect on the dependent variable.

For example,

- it could be the case that new schools are hiring less experienced teachers and at the same time they are applying more modern methods which have a positive influence on the results of the tests.
- If so, the coefficient 1.42 is not capturing the causal effect of the experience of the teacher on the score, i.e. this estimated coefficient is partly capturing the effect of some other variables apart from teacher's experience.

Excercise 2

Excercise 2

Background

Ih the exercise we seek to understand how the average price differs for apartment with various features.

The dataset Stata *habitatge_BCN.dta* contains information a sample of houses that were sold in Barcelona during 1998-2000.

Data

The variables included are:

preu = price of the flat in en euros,

superf = flat size in square metres,

dorm = number of rooms,

edat = This variable refers to the age of the flat/building and takes values from 1 (very recent) to 7 (very old),

calef = Ficticious variable that takes value 1 if the flat has heating and 0 if not.

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
ANY	1,267	1,998.79	0.81	1,998	1,998	1,999	1,999	2,000
DORM	1,267	2.94	0.96	0	2	3	4	12
CALEF	1,267	0.23	0.42	0	0	0	0	1
EDAT	1,267	5.32	1.43	1	5	5	6	7
SUPERF	1,267	83.82	31.51	26	64.5	79.0	96.8	338
PREU	1,267	106,209.20	62,417.51	13,936.94	68,082.47	91,758.80	125,419.20	700,287.60

Question 1

Regress preu on calef.

Interpret the estimated coefficient.

Excercise 2

Question 1

We estimate the following regression,

```
m1 ← lm(data=flats, formula = preu~calef)
```

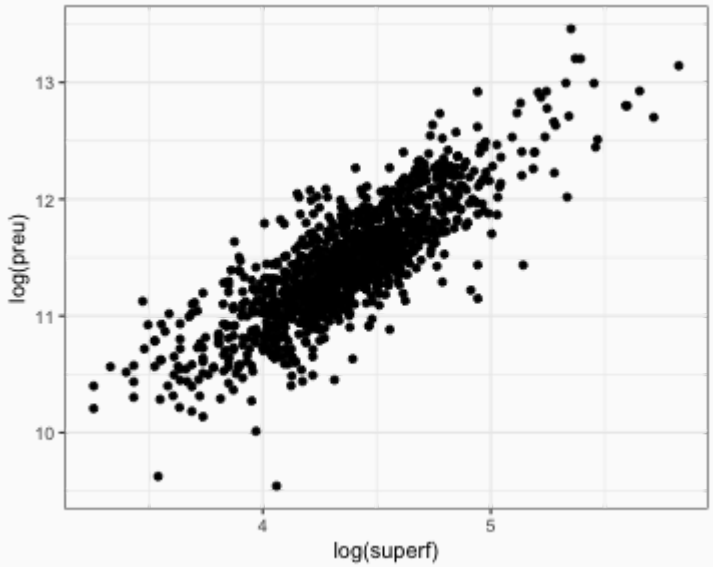
preu measures the price and *calef* is a binary variable.

How is the coefficient interpreted?

What assumptions should we make about the behavior of the error term?

Excercise 2

Question 1



Excercise 2

Question 1

	preu
CALEF	77,345.500***
	(3,552.846)
CONSTANT	88,383.720***
	(1,705.609)
Observations	1,267
R-squared	0.273
Adjusted R-squared	0.272
Residual standard error	53,257.630 (df = 1265)
F statistic	473.933*** (df = 1; 1265)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 2

Question 1

Those dwellings with heating have an expected price which is 77345.5 euros higher than the price of a dwelling without heating.

We reject null hypothesis of the coefficient of the calef variable being equal to zero.

This is because the t-statistic is greater than 1.96 in absolute value.

The p-value being 0.0000 is also evidence to reject the previous null hypothesis.

Excercise 2

Question 1

In the text they ask us to "Regress y on x"...

Comes from a geometric interpretation of the OLS estimator (can be thought of as the projection of y on the space spanned by X).

Think of it as dropping the y variable $on\ x$.

Question 2

What happens to the results of the previous regression if we rather express the price in thousands of euros?

What changes? Why?

Try to prove it.

Excercise 2

Question 2

```
flats_rescaled ← flats %>%
  mutate(preu=preu/1000)

m1 ← lm(data=flats_rescaled, formula = preu~calef)
```

Excercise 2

Question 2

	preu
CALEF	77.345***
	(3.553)
CONSTANT	88.384***
	(1.706)
Observations	1,267
R-squared	0.273
Adjusted R-squared	0.272
Residual standard error	53.258 (df = 1265)
F statistic	473.933*** (df = 1; 1265)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 2

Question 2

Both the estimates of the coefficient of calef and the intercept are exactly those obtained in question 1 divided by 1000.

The R^2 and the F statistic do not change, as it also happens with the t-statistics.

The latter is because both the coefficients and the corresponding standard errors are divided by 1000.

On the other hand, the residual sum of squares and, in general, all the sums of squares are divided by 1000.

Question 3

Estimate the model of the previous question using a variable (*calef0*) defined as 1 if the flat does not have heating and 0 if it has heating.

What changes and what does not change? Why?

Excercise 2

Question 3

```
flats ← flats %>%
  mutate(calef_1=ifelse(calef==1,0,1))

m1 ← lm(data=flats, formula = preu~calef)
m2 ← lm(data=flats, formula = preu~calef_1)
```

	preu	
	(1)	(2)
CALEF	77,345.500***	
	(3,552.846)	
CALEF_1		-77,345.500***
		(3,552.846)
CONSTANT	88,383.720***	165,729.200***
	(1,705.609)	(3,116.667)
Observations	1,267	1,267
R-squared	0.273	0.273
Adjusted R-squared	0.272	0.272
Residual standard error (df = 1265)	53,257.630	53,257.630
F statistic (df = 1; 1265)	473.933***	473.933***
Notes:	***p < .01; **p < .05; *p < .1	

Excercise 2

Question 3

The models estimated in questions 2) and 3) are exactly the same. We are explaining the same “story” about how heating affects the price of a dwelling.

As you can see the only things which change are the estimated coefficients of the heating dummy and the intercept (and its standard error).

Now, the estimated coefficient of calef0 is indicating that not having heating decreases the expected price of a dwelling by 77.3 thousand euros compared to having heating.

Excercise 2

Question 3

Notice that the t-statistic of the coefficient of the heating variable is the same in absolute value in both models because the standard error is the same because the coefficient is measuring the same "difference" but the comparison between having heating or not is made in a different way in each case.

The model is the same in both cases because in each occasion you are choosing a different dummy among the two associated to each of the categories of heating in order to capture the effect of it (this is the "mechanical" application of how to model qualitative variables as explanatory variables).

calef0 is just a linear transformation of calef $[calef0 = 1 - calef]$ and if you make a linear transformation of any variable the model does not change in terms of the "story" you are explaining.

Of course, the estimated coefficients differ but the effect of the explanatory variable remains the same.

Question 4

Now use as a variable relative to dispose of heat or not the variable calef2 that takes the value 2 if the flat has heating and 1 if it does not. How do the estimated coefficients change?

Why?

Excercise 2

Question 4

```
flats_rescaled ← flats %>%
  mutate(relative=ifelse(calef==1,2,1))

m1 ← lm(data=flats_rescaled, formula = preu~relative)
```

Excercise 2

Question 4

	preu
RELATIVE	77,345.500***
	(3,552.846)
CONSTANT	11,038.220**
	(4,620.609)
Observations	1,267
R-squared	0.273
Adjusted R-squared	0.272
Residual standard error	53,257.630 (df = 1265)
F statistic	473.933*** (df = 1; 1265)
Notes:	***p < .01; **p < .05; *p < .1

Great job!

Questions? sebastian.ellingsen@upf[dot]edu

Regression Discontinuity Designs

Brief Intro Based on Calonico, Cattaneo and Titiunik (2015)

In the absence of randomized treatment assignment, some research designs identify the causal effects of non-experimental interventions under weak and transparent assumptions.

One particularly rigoruous such approach is the regression discontinuity design.

The key feature of the design is that the probability of a treatment changes abruptly at the known threshold.

Basic idea: units with scores barely below the cutoff can be used as counterfactuals for units with scores barely above it.

These slides will briefly cover the basics of this design and how to implement the analysis using R (see references below).

The potential outcomes framework

Each unit has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, corresponding to the outcomes that would be observed under the treatment or control conditions respectively.

T_i is the treatment dummy denoting 1 if i is assigned to the treatment condition and 0 otherwise.

Note, we never observe both potential outcomes for any one individual. We observe

$$Y_i = (1 - T_i) \times Y_i(0) + T_i \times Y_i(1).$$

The building blocks

- n units indexed by $i = 1, \ldots, n$.
- Each unit has a *running variable*, X_i.
- There is a known common cutoff for all the units, c.
- Units with $X_i \geq c$ are assigned to the treatment condition, while units with $X_i < c$ are assigned to the control condition.
- This assignment rule is defined as $T_i = \mathbf{1}(X_i \geq c)$.

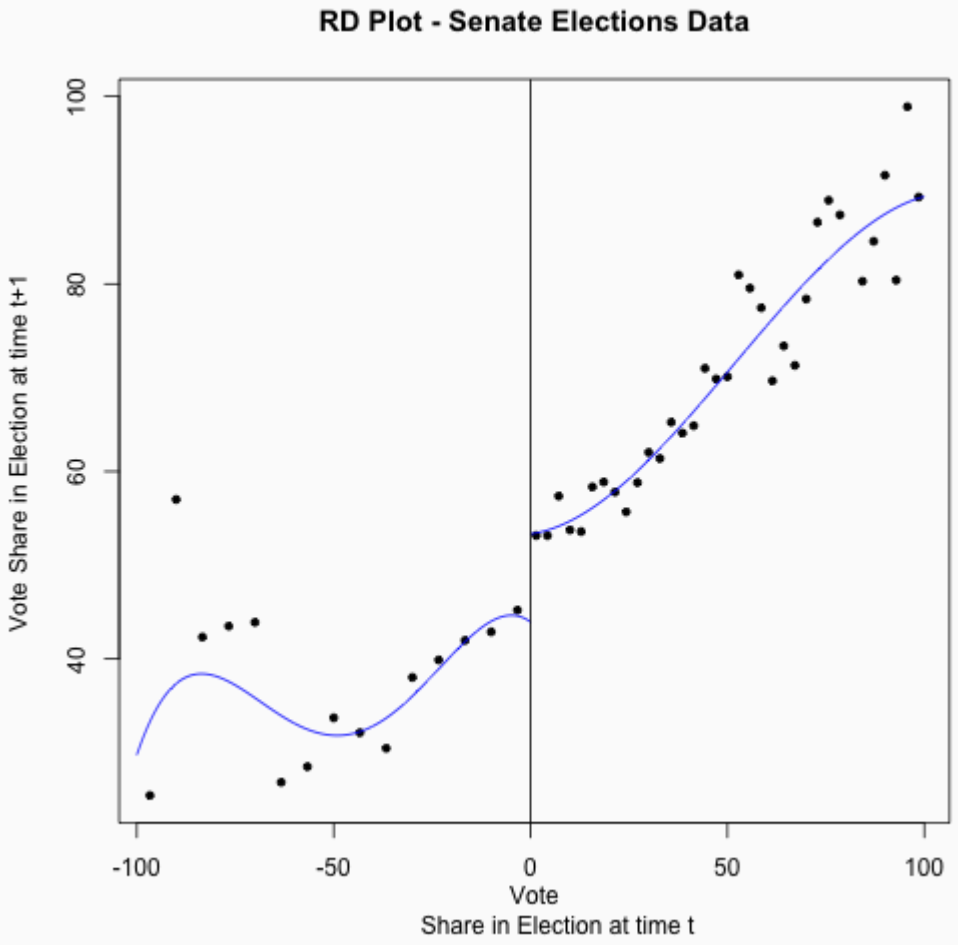
Key feature of the RD design: probability of treatment is a function of the score and it changes discontinuously at the cutoff.

Incumbency Advantage in House

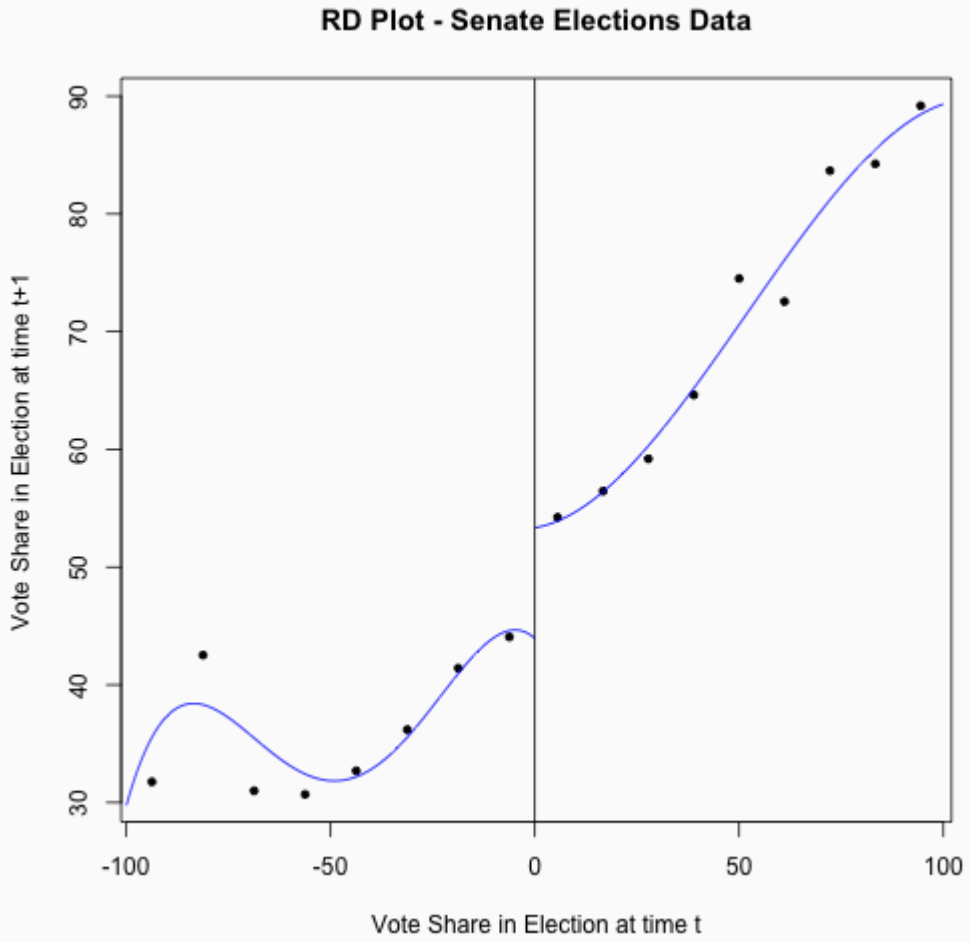
```
require(rdrobust)
data(rdrobust_RDsenate)

rdrobust_RDsenate %>%
  as.data.frame() %>%
  stargazer(type="html",
            style="io",
            font.size="tiny",
            header=FALSE,
            median=TRUE,
            star.char = c(""),
            digits = 2)
```

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
MARGIN	1,390	7.17	34.32	-100.00	-12.21	2.17	22.77	100.00
VOTE	1,297	52.67	18.12	0.00	42.67	50.55	61.35	100.00

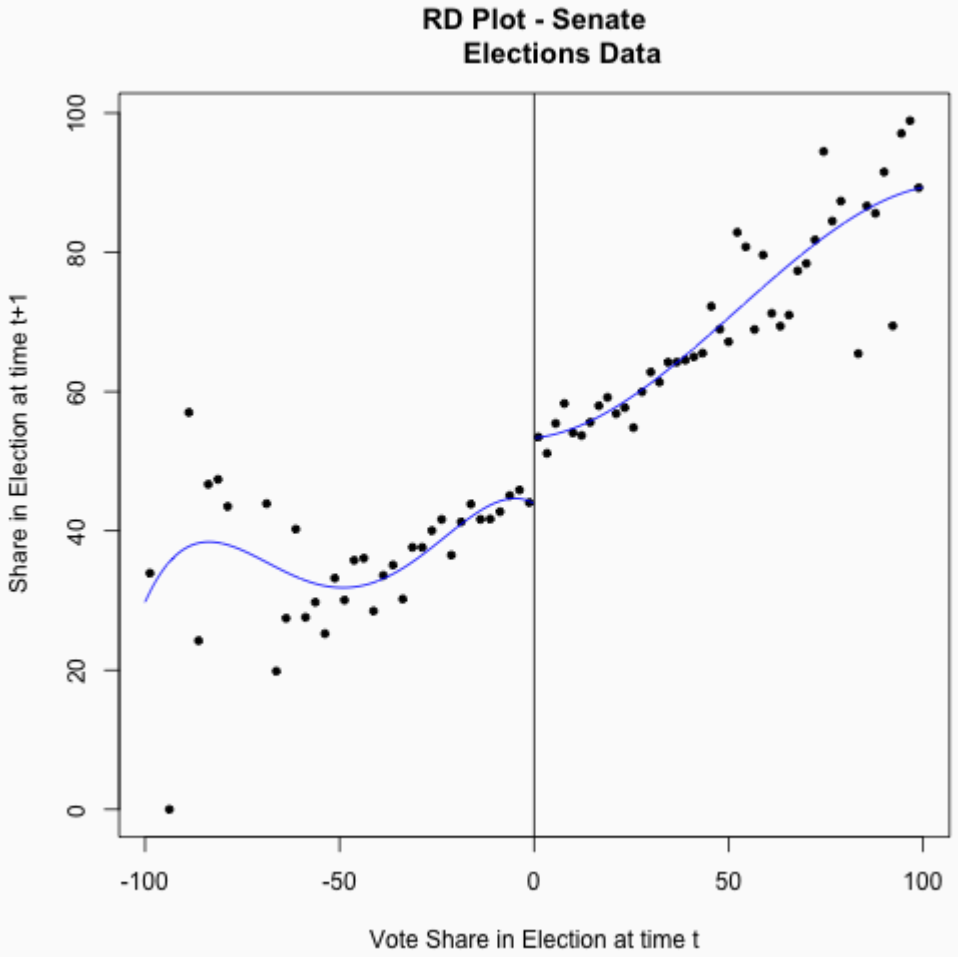
```
rdplot(y = rdrobust_RDsenate$vote, x = rdrobust_RDsenate$margin, title = "RD Plot - Se
       Share in Election at time t", y.label = "Vote Share in Election at time t+1")
```



```
rdplot(y = rdrobust_RDsenate$vote, x = rdrobust_RDsenate$margin, binselect = "es", tit
       x.label = "Vote Share in Election at time t",
       y.label = "Vote Share in Election at time t+1")
```



```
rdplot(y = rdrobust_RDsenate$vote, x = rdrobust_RDsenate$margin, binselect = "es", sca
       Elections Data", x.label = "Vote Share in Election at time t", y.label = "Vote
       Share in Election at time t+1")
```



```
summary(rdrobust(y = rdrobust_RDsenate$vote, x = rdrobust_RDsenate$margin))
```

```
## Call: rdrobust
##
## Number of Obs.                 1297
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 595         702
## Eff. Number of Obs.            359         322
## Order est. (p)                   1           1
## Order bias  (p)                  2           2
## BW est. (h)                 17.708      17.708
## BW bias (b)                 27.984      27.984
## rho (h/b)                    0.633       0.633
##
## ============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ============================================================================
##   Conventional     7.416     1.460     5.078     0.000     [4.554 , 10.278]
##         Robust         -         -     4.310     0.000     [4.094 , 10.925]
## ============================================================================
```

```
summary(rdrobust(y = rdrobust_RDsenate$vote, x = rdrobust_RDsenate$margin, all=TRUE))
```

```
## Call: rdrobust
##
## Number of Obs.                 1297
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  595         702
## Eff. Number of Obs.             359         322
## Order est. (p)                    1           1
## Order bias  (p)                   2           2
## BW est. (h)                  17.708      17.708
## BW bias (b)                  27.984      27.984
## rho (h/b)                     0.633       0.633
##
## ============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ============================================================================
##   Conventional     7.416     1.460     5.078     0.000     [4.554 , 10.278]
## Bias-Corrected     7.510     1.460     5.143     0.000     [4.648 , 10.372]
##         Robust     7.510     1.743     4.310     0.000     [4.094 , 10.925]
## ============================================================================
```

```
summary(rdbwselect(y = rdrobust_RDsenate$vote, x = rdrobust_RDsenate$margin, all = TRU
```

```
## Call: rdbwselect
##
## Number of Obs.                 1297
## BW type                         All
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 595         702
## Order est. (p)                   1           1
## Order bias  (q)                  2           2
##
## ===================================================
##                   BW est. (h)    BW bias (b)
##             Left of c Right of c  Left of c Right of c
## ===================================================
##      mserd    17.708     17.708     27.984     27.984
##     msetwo    16.154     18.009     27.096     29.205
##     msesum    18.326     18.326     31.280     31.280
##   msecomb1    17.708     17.708     27.984     27.984
##   msecomb2    17.708     18.009     27.984     29.205
##      cerrd    12.374     12.374     27.984     27.984
##     certwo    11.288     12.585     27.096     29.205
##     cersum    12.806     12.806     31.280     31.280
##   cercomb1    12.374     12.374     27.984     27.984
```

```
rdrobust(y = rdrobust_RDsenate$vote,x = rdrobust_RDsenate$margin,kernel = "uniform")
```

```
## Call: rdrobust
##
## Number of Obs.                 1297
## BW type                       mserd
## Kernel                      Uniform
## VCE method                       NN
##
## Number of Obs.                  595         702
## Eff. Number of Obs.             262         228
## Order est. (p)                    1           1
## Order bias  (p)                   2           2
## BW est. (h)                  11.157      11.157
## BW bias (b)                  22.670      22.670
## rho (h/b)                     0.492       0.492
```

```
rdrobust(y = rdrobust_RDsenate$vote,x = rdrobust_RDsenate$margin,p = 2,q = 4)
```

```
## Call: rdrobust
##
## Number of Obs.                 1297
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  595         702
## Eff. Number of Obs.             304         263
## Order est. (p)                    2           2
## Order bias  (p)                   4           4
## BW est. (h)                  13.660      13.660
## BW bias (b)                  39.927      39.927
## rho (h/b)                     0.342       0.342
```

```
rdrobust(y = rdrobust_RDsenate$vote,x = rdrobust_RDsenate$margin,h = 15,rho = 0.8)
```

```
## Call: rdrobust
##
## Number of Obs.                 1297
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  595         702
## Eff. Number of Obs.             319         288
## Order est. (p)                    1           1
## Order bias  (p)                   2           2
## BW est. (h)                  15.000      15.000
## BW bias (b)                  18.750      18.750
## rho (h/b)                     0.800       0.800
```

For Details

This application is closely based on the content and exposition in the following publications:

- Calonico, Cattaneo and Titiunik (2015): rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs, R Journal 7(1): 38-51.
- Cattaneo, Idrobo and Titiunik (2018): A Practical Introduction to Regression Discontinuity Designs: Volume I. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.
- Cattaneo, Idrobo and Titiunik (2018): A Practical Introduction to Regression Discontinuity Designs: Volume II. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

class: inverse, center, middle

Excercise 1

Excercise 1

Background

In the late 1980s the Tennessee state legislature funded a four-year experiment to evaluate the effect of small class sizes on learning (Student Teacher Achievement Ratio or STAR).

The experiment compared three different class arrangements for children in kindergarten through third grade:

- A regular-size class (size 22-25 pupils) with a single teacher – The control group.
- A small class (size 13-17 pupils) with a single teacher.
- A regular-size class with a single teacher and a teacher's aide.

Participating schools were picked at random from the universe of public schools in Tennessee.

Each participating school had at least one class of each type.

Within schools both pupils and teachers were randomly assigned to one of the three types of classes. Each year the children were given standardized tests (the SAT) and these are the outcome measures

In this exercise you will analyze the test results for children in kindergarten by comparing the children in the small class with those in the regular class without aide.

Data

The variables included are:

- *sck* = whether kid in small class (the treatment 1, or control 0),
- *tscorek* = the test score of the child
- *boy* = whether kid a boy (1) or a girl (0),
- *freelunk* = whether kid gets a free lunch (proxy for being from poor household; 1),
- *totexpk* = years of teaching experience of teacher,
- *schidkn* = code for particular school.

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
TOTEXPK	5,749	9.31	5.77	0	5	9	13	27
SCHIDKN	5,749	39.83	22.96	1	20	39	60	80
TSCOREK	5,749	922.39	73.87	635	870	915	964	1,253
SCK	5,749	0.30	0.46	0	0	0	1	1
BOY	5,749	0.51	0.50	0	0	1	1	1
FREELUNK	5,749	0.48	0.50	0	0	0	1	1

- Look at the summary statistics to get a rough overview of the dataset.
- Spot errors.
- Get a sense of the variation, scale, etc.

Excercise 1

Empirical Model

Consider the bivariate regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

- Y_i denotes the tscorek.
- X_i an indicator variable equal to 1 if the kid was in a small class and 0 otherwise.
- u_i is the error term.

Question 1

1. Assess whether the first LS assumption, $E[\epsilon_i|X_i] = 0$ is satisfied in this application.

Excercise 1

Question 1

In a randomized controlled experiment as the one examined in this exercise, subjects are randomly assigned to the treatment group $(X = 1)$ or to the control group $(X = 0)$.

Random assignment makes X and u independent, which in turn implies that the conditional mean of u given X is equal to the unconditional mean of u. Since b_0 captures any constant effect that could affect the test score then the unconditional mean of u should be zero.

So the first OLS assumption is likely to hold here.

Question 2

Comment on whether the second (independence) and third (finite fourth moments) assumptions of LS are satisfied.

Excercise 1

Question 2

The second assumption is a statement about how the sample is drawn.

If the observations are drawn by simple random sampling from a single large population, then the second assumption holds.

In this exercise it is said schools were sampled randomly and that pupils and teachers were randomly assigned to one of the three types of classes within schools.

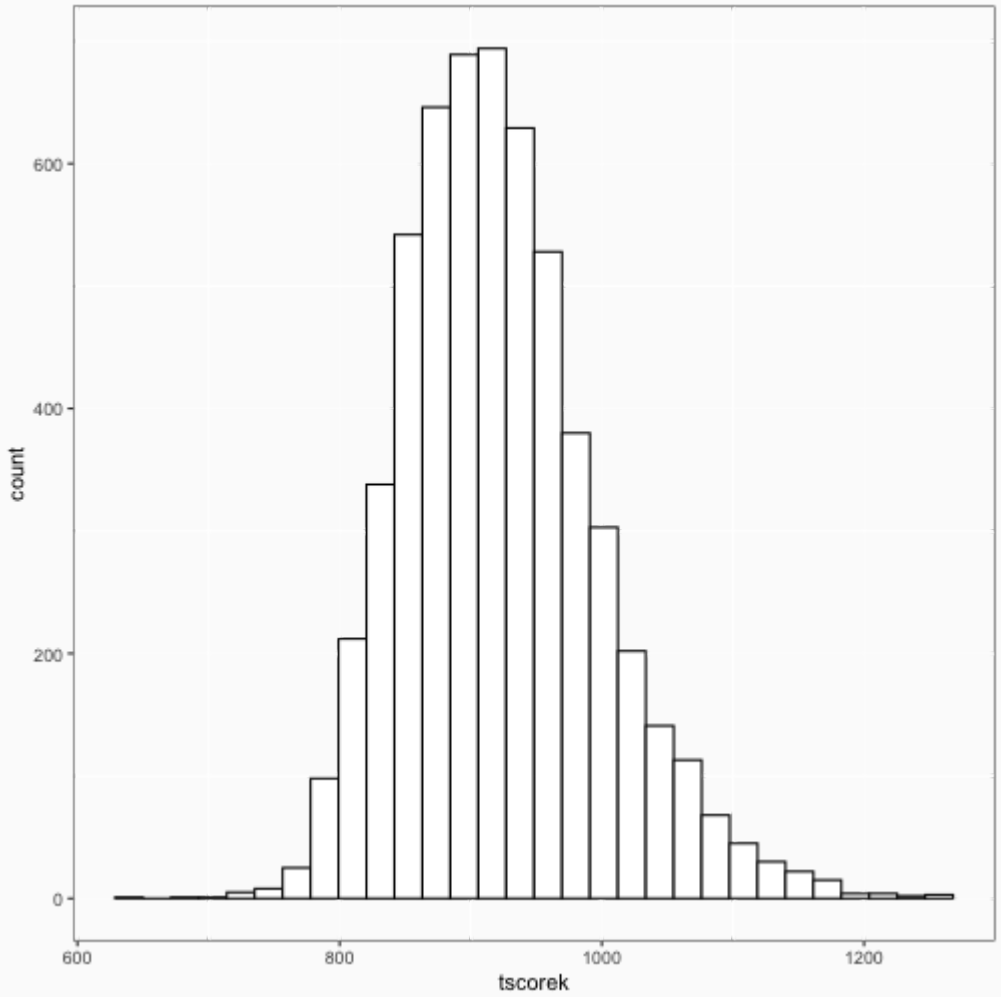
This means that students were NOT randomly sampled from the population.

The worry here is that observations might not be independent within schools although they likely are across schools.

The second OLS assumption is thus only partially valid.

Excercise 1

Question 2



Excercise 1

Question 2

The third assumption states that large outliers are unlikely (otherwise they can make OLS regression results misleading and the LLN and CLT do not apply).

The best that a student can do on standardized test is to get all the questions right and the worst he/she can do is to have all the questions wrong.

Because Y is bounded it must have finite fourth (all, in fact) moments.

On the other hand, if class size is bounded by the physical capacity or by legal/municipal restrictions, then X will not likely have outliers and we can assume it has a finite fourth moment.

Question 3

Run a regression of $tscorek$ on whether you were in a small class (we are going to ignore the other treatment – being in a regular class with aide).

- Interpret the coefficients.
- Contrast the significance of this estimated coefficient of variable sck.
- Generate a 95 percent confidence interval.

Excercise 1

Question 3

The following commands run linear regressions in R and Stata.

R:

```
m1 ← lm(data=star, formula = tscorek~sck)
```

Stata:

```
reg tscorek sck
```

Excercise 1

Question 3

	tscorek
SCK	13.826***
	(2.115)
CONSTANT	918.225***
	(1.161)
Observations	5,749
R-squared	0.007
Adjusted R-squared	0.007
Residual standard error	73.604 (df = 5747)
F statistic	42.714*** (df = 1; 5747)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 1

Question 3

The point estimate is 13.82.

This is our estimate of the causal class size effect, or the expected testscore gain from attending a small class, compared to attending a large class.

Question 4

Consider now the case where an explanatory variable X is the number of teaching experience of a given teacher ($totexpk$).

- Run the regression of tscorek on $totexpk$.
- Contrast the null hypothesis that the coefficient is zero.

Excercise 1

Question 4

	tscorek
TOTEXPK	1.418***
	(0.168)
CONSTANT	909.194***
	(1.838)
Observations	5,749
R-squared	0.012
Adjusted R-squared	0.012
Residual standard error	73.422 (df = 5747)
F statistic	71.340*** (df = 1; 5747)
Notes:	***p < .01; **p < .05; *p < .1

Question 4

Comment if the first OLS assumption, $E[\epsilon_i|X_i] = 0$, is satisfied in the previous subsection.

Verify that R^2 coincides with the square of the correlation coefficient between $tscorek$ and $totexpk$.

Excercise 1

Question 4

In this case it is possible to think about unobserved variables included in the error term which could be correlated with the variable *totexpk* and which have an effect on the dependent variable.

Even though professors and students were assigned randomly to students.

It could be the case for instance that some particular type of schools (for instance, new schools or liberal schools) are hiring less experienced teachers and at the same time they are applying more modern methods which have a positive influence on the results of the tests.

In this case, the coefficient 1.42 will not capturing the causal effect of the experience of the teacher on the score,

This estimated coefficient is partly capturing the effect of some other variables apart from teacher's experience.

Excercise 1

Question 4

	tscorek
TOTEXPK	1.418***
	(0.168)
CONSTANT	909.194***
	(1.838)
Observations	5,749
R-squared	0.012
Adjusted R-squared	0.012
Residual standard error	73.422 (df = 5747)
F statistic	71.340*** (df = 1; 5747)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 1

Question 4

R:

```
cor(star$tscorek,star$totexpk)
```

[1] 0.1107304

Stata:

```
corr tscorek totexpk
```

We see the correlation coefficient is the square root of R^2.

Excercise 2

Excercise 2

Background

Ih the exercise we seek to understand how the average price differs for apartment with various features.

The dataset Stata *habitatge_BCN.dta* contains information a sample of houses that were sold in Barcelona during 1998-2000.

Data

The variables included are:

preu = price of the flat in en euros,

superf = flat size in square metres,

dorm = number of rooms,

edat = This variable refers to the age of the flat/building and takes values from 1 (very recent) to 7 (very old),

calef = Ficticious variable that takes value 1 if the flat has heating and 0 if not.

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
ANY	1,267	1,998.79	0.81	1,998	1,998	1,999	1,999	2,000
DORM	1,267	2.94	0.96	0	2	3	4	12
CALEF	1,267	0.23	0.42	0	0	0	0	1
EDAT	1,267	5.32	1.43	1	5	5	6	7
SUPERF	1,267	83.82	31.51	26	64.5	79.0	96.8	338
PREU	1,267	106,209.20	62,417.51	13,936.94	68,082.47	91,758.80	125,419.20	700,287.60
CALEF_1	1,267	0.77	0.42	0	1	1	1	1

Question 1

Regress $preu$ on $calef$.

- Interpret the estimated coefficient.
- Test the null hypothesis that the true coefficient is zero.

Excercise 2

Question 1

R:

```
m1 ← lm(data=flats, formula = preu~calef)
```

Stata:

```
reg preu calef
```

Excercise 2

Question 1

	preu
CALEF	77,345.500***
	(3,552.846)
CONSTANT	88,383.720***
	(1,705.609)
Observations	1,267
R-squared	0.273
Adjusted R-squared	0.272
Residual standard error	53,257.630 (df = 1265)
F statistic	473.933*** (df = 1; 1265)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 2

Question 1

Those dwellings with heating have an expected price which is 77345.5 euros higher than the price of a dwelling without heating.

We reject null hypothesis of the coefficient of the calef variable being equal to zero. This is because the t-statistic is greater than 1.96 in absolute value. The p-value being 0.0000 is also evidence to reject the previous null hypothesis.

Question 2

Generate $calef2$ taking value 2 if the flat has heating and value 1 if not.

- How do the estimated coefficients change and why?
- Does anything else change?

Excercise 2

Question 2

R:

```
flats_rescaled ← flats %>%
  mutate(relative=ifelse(calef==1,2,1))

m1 ← lm(data=flats_rescaled, formula = preu~relative)
```

Stata:

```
gen calef2 = calef + 1
reg preu calef2
```

Excercise 2

Question 2

	preu
RELATIVE	77,345.500***
	(3,552.846)
CONSTANT	11,038.220**
	(4,620.609)
Observations	1,267
R-squared	0.273
Adjusted R-squared	0.272
Residual standard error	53,257.630 (df = 1265)
F statistic	473.933*** (df = 1; 1265)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 2

Question 2

This case can be understood as what happens if you use the codes of the categories associated to the heating variable which appear in the original version of the data set.

Notice that $calef2 = calef + 1$, so it is a linear transformation of calef and this does not affect the estimation of the model, i.e. the expected values of $preu000$ will not change it does not matter whether you use calef or $calef2$.

$$E(preus000|calef2) = 11.04 + 77.35 * calef2 = 11.04 + 77.35 * (calef + 1) = 88.38 +$$

Question 3

Run the regression on the $price$ (in thousands of euros) on the flat size $superf$.

- Contrast the null hypothesis that the true coefficient is zero. Contrast the null hypothesis that the true coefficient is 1.5.

Excercise 2

Question 3

Can be done in the following way in R and Stata:

R:

```
flats_rescaled ← flats %>%
  mutate(preu=preu/1000)

m1 ← lm(data=flats_rescaled, formula = preu~calef)
```

Stata:

```
gen preu000 = preu / 1000
reg preu000 superf
```

Excercise 2

Question 3

	preu
SUPERF	1.641***
	(0.031)
CONSTANT	-31.353***
	(2.791)
Observations	1,267
R-squared	0.687
Adjusted R-squared	0.686
Residual standard error	34.953 (df = 1265)
F statistic	2,772.231*** (df = 1; 1265)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 2

Question 3

R:

```
ttest ← function(reg, coefnum, val){
  co ← coef(summary(reg))
  tstat ← (co[coefnum,1]-val)/co[coefnum,2]
  2 * pt(abs(tstat), reg$df.residual, lower.tail = FALSE)
}
ttest(m1, 2,1.5)
```

Stata:

```
test _b[superf]==1.5
```

Excercise 2

Question 3

H_0: $\beta_1 = 0$:

[1] 4.535523e-321

H_0: $\beta_1 = 1.5$:

[1] 6.421389e-06

H_0: $\beta_1 = 1.6$:

[1] 0.1860488

Excercise 2

Question 3

$P - value = 0.0000$

We reject the null hypothesis that the superf coefficient is 0

$P - value = 0.00000$

In this case we can reject the null hypothesis that the coefficient is 1.5

Alternatively: Construct the confidence interval

Since 1.5 is outside the confidence interval at 95 confidence, we can reject the null hypothesis that the coefficient is different from 1.5.

Excercise 2

Question 3

$P - value = 0.1860$.

In this case we cannot reject the null hypothesis that the coefficient is 1.6.

Alternatively: Construct the confidence interval.

Since 1.6 is inside the confidence interval at 95 confidence, we cannot reject the null hypothesis that the coefficient is different from 1.6.

Question 4

Do you consider that the causal effect of having heating on the price is captured by the estimated coefficient of question 1 of this exercise?

And how about the causal effect of superf given the estimated coefficient in question 3?

Excercise 2

Question 4

No. There are many reasons to suspect for the OLS Assumption #1 not to hold in this case.

For instance, more expensive houses might be built in neighborhoods that need more heating because they are close to the mountain, whereas cheaper houses are located in less demanded areas that need less heating

A similar argument can be exposed in the case of superf. In this case, people might build larger houses were land is cheaper due to more commuting time to work.

class: inverse, center, middle

Question 1

In this exercise you will analyze the test results for children in kindergarten by comparing the children in the small class with those in the regular class without aide.

Excercise 1

Question 1

In a randomized controlled experiment as the one examined in this exercise, subjects are randomly assigned to the treatment group $X = 1$ or to the control group $X = 0$.

Random assignment makes X and u independent, which in turn implies that the conditional mean of u given X is equal to the unconditional mean of u. Since b_0 captures any constant effect that could affect the test score then the unconditional mean of u should be zero.

So the first OLS assumption is likely to hold here.

Excercise 1

Question 1

Suppose that the linear population regression model that explains the test scores of students Y includes the following explanatory variables: student-teacher ratio X_1; school characteristics X_2; socio-economic condition of the family X_3; student ability X_4 and gender X_5.

Excercise 1

Question 1

Now, in the STAR experiment, since children were randomly assigned to small vs large class sizes, the indicator sck should be independent of any other variable.

Independence implies $cov(x_1, x_k) = 0$ for $k = 2, 3, 4, 5$.

Nonetheless, even though probably variables x_k will be correlated with Y and were omitted in the short regression, these variables x_k do not lead to omitted variable bias.

Excercise 1

Question 1

It is a different case with observational data from California, as there was no random assignment from students to classes. In this case, we should worry about omitted variable bias, i.e. variables that correlate with x_1 and at the same time determine Y.

For example, poor families might bring their kids to poorer schools that have higher STR. Therefore, not controlling for these variables might bias the estimation of β_1.

Question 2

Use *STAR_small.dta* and explain testscorek using as explanatory variables, apart from the constant, the student-teacher ratio *sck*, gender *boy* and if the student is eligible for free lunch *freelunk*.

Compare the estimated coefficient of sck in this model with the one obtained when running *testscorek* on *sck*.

- Does it change substantially? Why?
- Is this claim supported by the correlation coefficients matrix of these different variables? [Note: the Stata command corr allows you to obtain this information]

Exercise 1

Question 2

R:

```
m1 ← lm(data=star, formula = tscorek~sck)
m2 ← lm(data=star, formula = tscorek~sck + sck + boy + freelunk)

m1 ← coeftest(m1, vcov = vcovHC(m1, type="HC1"))
m2 ← coeftest(m2, vcov = vcovHC(m1, type="HC1"))
```

Stata:

```
reg tscorek sck, r

reg tscorek sck boy freelunk, r
```

Excercise 1

Question 2

	(1)	(2)
SCK	13.826***	13.276***
	(2.162)	(2.081)
BOY		-14.300***
		(1.862)
FREELUNK		-40.046***
		(1.856)
CONSTANT	918.225***	945.075***
	(1.142)	(1.770)
Notes:	***p < .01; **p < .05; *p < .1	

Excercise 1

Question 2

Calculate the correlation matrix:

R:

```
cor(star, method = c("pearson"))
```

Stata:

```
corr tscorek sck boy freelunk
```

```
mcor←round(cor(star),2)

lower.tri(mcor, diag = FALSE)
upper←mcor
upper[upper.tri(mcor)]←""
upper←as.data.frame(upper)
```

Question 3

Interpret the estimated coefficients of the different explanatory variables and analyze their significance.

Should we eliminate the free lunch program to improve the grades?

Excercise 1

Question 3

In the second model, the coefficient on sck indicates that the expected value of the test is 13.28 points higher for kids who went to small class size. Furthermore, boys have on average 14.30 points less than girls.

Literally taken, $freelunk$ indicates that kids participating in help programmes in the food hall have on average 40.05 points less.

But from here we cannot infer that elimination those “food subsidy” programmes would be a good idea to increase test scores. $Freelunk$ is just a proxy for family income.

Excercise 1

Question 3

All three coefficients are statistically significant.

Regarding the variable boy: it is conceptually different than $freelunk$ as you cannot change your gender, in the same way that you can change policies on free lunch.

Even if it is not causal, perhaps you are just interested if there are gender differences.

In a way, the gender coefficient is probably more informative than the free lunch coefficient.

Question 4

Contrast the null hypothesis (in the model with just sck) that going to small class leads to 15 points more on the test score comparted to large class, if all the other features are the same.

Excercise 1

Question 4

R:

```
ttest(m1, 2, 15)
```

Stata:

```
reg tscorek sck

test _b[sck]==15
```

We can not reject the hypothesis with a reasonable level of confidence.

[1] 0.578834

Question 5

Estimate the same model using only the first 100 observations , the observations from 101 to 200, and then the first 1000 observations.

How do the estimated coefficients and the standard errors compare with the full sample results? How we explain this?

Excercise 1

Question 5

R:

```
star_1 ← star[1:100,]
star_2 ← star[100:199,]

m1 ← lm(data=star_1, formula = tscorek~sck)
m1 ← coeftest(m1, vcov = vcovHC(m1, type="HC1"))
```

Stata:

```
reg tscorek sck boy freelunk if _n⩽100, r

reg tscorek sck boy freelunk if _n>100 & _n⩽200, r
```

Excercise 1

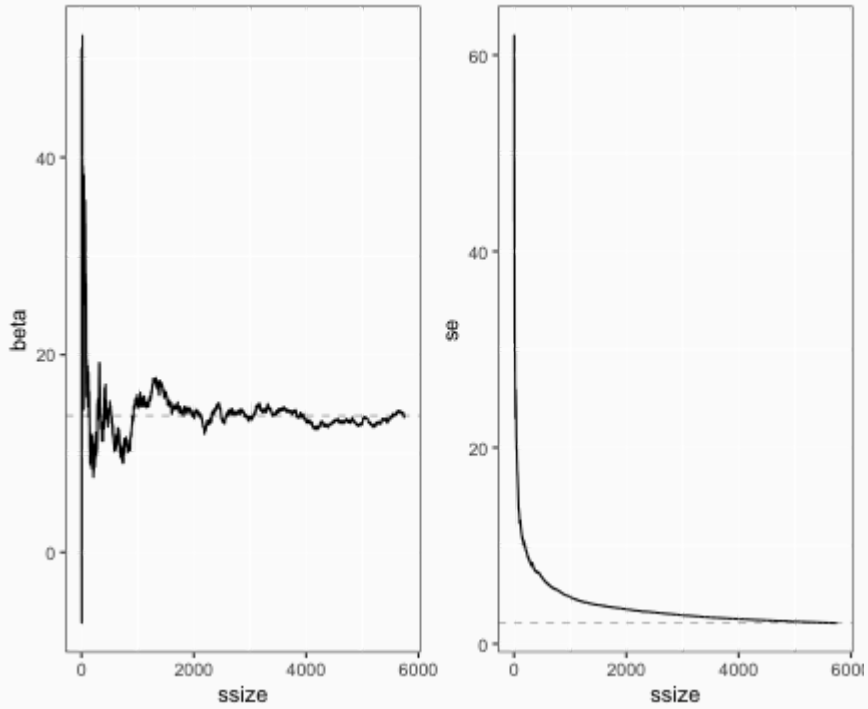
Question 5

	(1)	(2)	(3)
SCK	16.692	2.370	15.297***
	(13.042)	(16.039)	(4.981)
CONSTANT	922.733***	926.284***	915.061***
	(7.308)	(8.288)	(2.572)
Notes:	***p < .01; **p < .05; *p < .1		

Excercise 1

Question 5

```
library(broom)
library(ggpubr)

beta ← c()
se ← c()
for (i in 1:nrow(star)){
  beta[i] ← (lm(data=star[1:i,], formula = tscorek~sck) %>% tidy())[2,2]
  se[i] ← (lm(data=star[1:i,], formula = tscorek~sck) %>% tidy())[2,3]

}

data ← cbind(ssize=1:nrow(star), beta, se) %>%  as.data.frame() %>% filter(!is.na(bet

data ← as.data.frame(lapply(data, unlist))

p1 ← ggplot(data=data, aes(x=ssize, y=beta))+geom_line()+  geom_hline(yintercept=13.8
             color="black",
             linetype="dashed",
             size=0.3,
             alpha = 0.5)+theme(panel.background = element_rect(fill = "transparent",


p2 ← ggplot(data=data, aes(x=ssize, y=se))+geom_line()+  geom_hline(yintercept=2.216,202
```


beta
40
20
0
0
2000
4000
6000
ssize
se
60
40
20
0
0
2000
4000
6000
ssize

Excercise 1

Question 5

Precision has two elements:

(i) the larger the sample size, the smaller the changes in the estimated coefficient.

(ii) the larger the sample size, the smaller the variance of the estimated beta.

Excercise 1

Question 5

If we estimate a model with 100 observations, as in the first two cases, we have a substantial reduction in sample size. This leads to a loss in precision which has the following consequences:

(i) Standard errors of the estimated coefficients are larger and t- statistics are lower, apart from larger confidence intervals

(ii) The estimated coefficients for a given variable differ substantially across different samples of 100 observations.

Question 6

Estimate an explanatory model of the test score using boy as the only explanatory variable.

Then, another model with sck as the only explanatory variable. Then, another model with both variables.

For all these regressions, only use the first 100 observations and without the robust option. Comment how the coefficients of determination compare (R^2) and the corrected coefficients of determination (adjusted R^2).

Excercise 1

Question 6

	tscorek		
	(1)	(2)	(3)
BOY	6.233		9.323
	(12.429)		(12.549)
SCK		16.692	18.323
		(12.566)	(12.785)
CONSTANT	926.106***	922.733***	917.140***
	(9.048)	(7.947)	(10.961)
Observations	100	100	100
R-squared	0.003	0.018	0.023
Adjusted R-squared	-0.008	0.008	0.003
Residual standard error	62.033 (df = 98)	61.561 (df = 98)	61.702 (df = 97)

Excercise 1

Question 6

The R^2 always increases when we add a new additional variable. The RSS always goes down.

The adjusted R^2 can be negative, like in the first model, if the significance of the coefficients is low.

When we go from the first to the third model, the adjusted R^2 increases, even though the added variable is not statistically significant. This is because the t-statistic is still above 1 and the improvement in RSS compensates the loss in the degree of freedom.

On the other hand, going from the second to the third model, the adjusted R^2 goes down Since the t-statistic of the variable boy is low.

Session 4

Excercise 1

Question 1

Use habitatge_BCN.dta and estimate a model to explain the price of flats using the number of rooms as the only explanatory variable. Then, add the superf variable.

How do the results compare in the two models? Why does the sign of the variable on the number of rooms change? Interpret the estimated coefficients of the second model.

Regress the number of rooms on superf and define the residuals as resid. Then, run the regression of price of the flat on resid.

How do the results compare to the ones of the estimated model that includes the number of rooms and superf as explanatory variables?

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
ANY	1,267	1,998.785	0.812	1,998	1,998	1,999	2,000
DORM	1,267	2.941	0.962	0	2	4	12
CALEF	1,267	0.230	0.421	0	0	0	1
EDAT	1,267	5.322	1.427	1	5	6	7
SUPERF	1,267	83.816	31.514	26	64.5	96.8	338
PREU	1,267	106,209.200	62,417.510	13,936.940	68,082.470	125,419.200	700,287.600
CALEF_1	1,267	0.770	0.421	0	1	1	1

	preu	
	(1)	(2)
DORM	32,382.420***	-9,396.245***
	(1,581.757)	(1,392.470)
SUPERF		1,839.882***
		(42.487)
CONSTANT	10,978.820**	-20,370.070***
	(4,893.799)	(3,189.779)
Observations	1,267	1,267
R-squared	0.249	0.698
Adjusted R-squared	0.248	0.697
Residual standard error	54,117.380 (df = 1265)	34,353.290 (df = 1264)
F statistic	419.121*** (df = 1; 1265)	1,457.680*** (df = 2; 1264)
Notes:	***p < .01; **p < .05; *p < .1	

Excercise 1

Question 1

It is worth noting that the variable changes sign when we include the surface area.

In column (1) dorm proxies for the size of the apartment.

Conditional on the surface area, more rooms is negative because there's less living space.

Excercise 1

Question 1

Example of the Frisch-Waugh theorem. Can obtain the coefficient from a mulitple regression by a step-wise procedure.

Consider the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + u_i,$$

One can obtain the estimate of β_1 by running two "small" regressions instead of the large one.

Will get the same point estimate, but different standard errors.

Excercise 1

Question 1

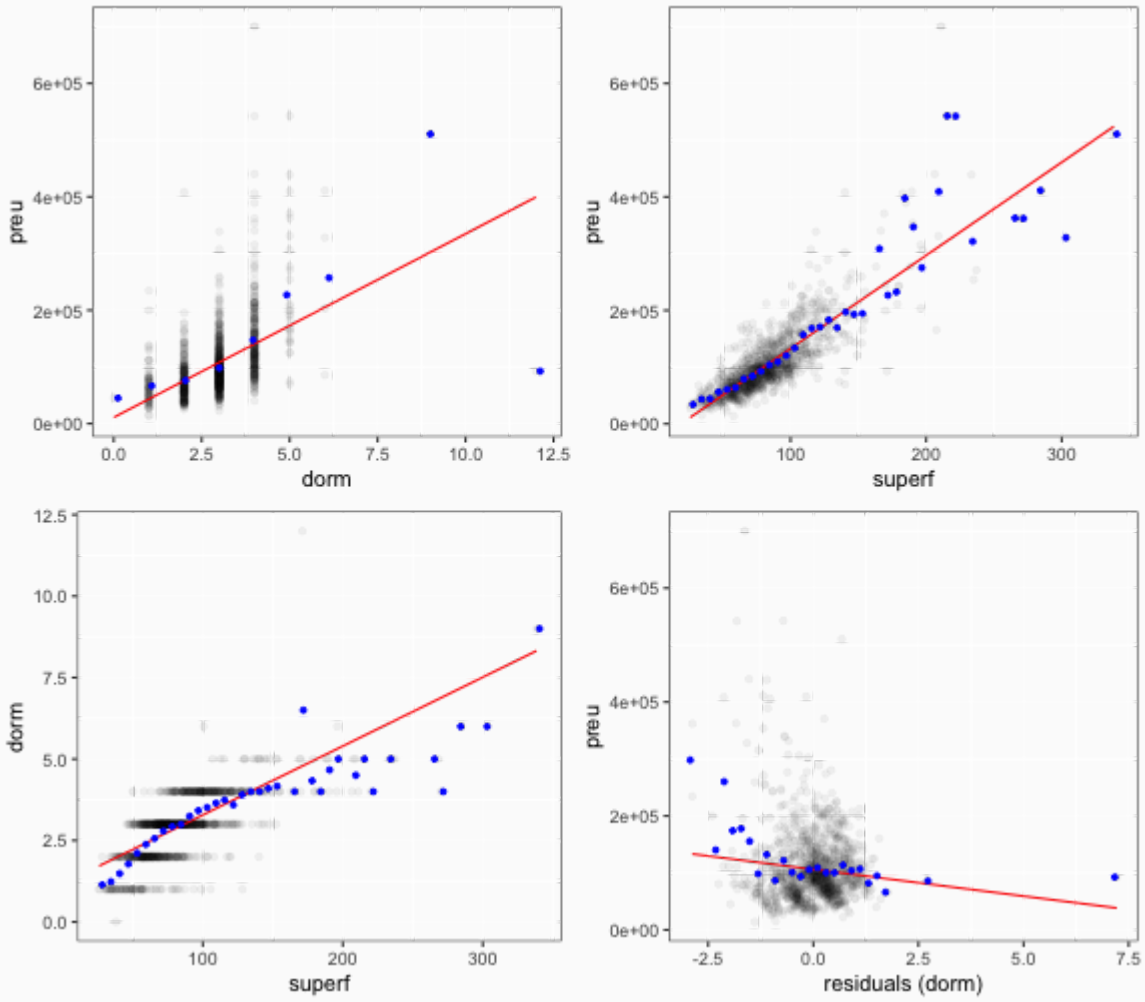
R:

```
m1 ← lm(data=flats, formula = dorm~superf)
flats$residuals ← m1[["residuals"]]
m2 ← lm(data=flats, formula = preu~residuals)
```

Stata:

```
reg dorm superf
predict resid, r
reg preu resid
```

	dorm	preu	
	(1)	(2)	(3)
DORM			-9,396.245***
			(1,392.470)
SUPERF	0.021***		1,839.882***
	(0.001)		(42.487)
RESIDUALS		-9,396.245***	
		(2,517.194)	
CONSTANT	1.169***	106,209.200***	-20,370.070***
	(0.055)	(1,744.661)	(3,189.779)
Observations	1,267	1,267	1,267
R-squared	0.480	0.011	0.698
Adjusted R-squared	0.480	0.010	0.697
Residual standard error	0.694 (df = 1265)	62,101.090 (df = 1265)	34,353.290 (df = 1264)


pre u
6e+05
4e+05
2e+05
0e+00
0.0
2.5
5.0
7.5
10.0
12.5
dorm
pre u
6e+05
4e+05
2e+05
0e+00
100
200
300
superf
dorm
12.5
10.0
7.5
5.0
2.5
0.0
100
200
300
superf
pre u
6e+05
4e+05
2e+05
0e+00
-2.5
0.0
2.5
5.0
7.5
residuals (dorm)

Excercise 2

Background

In the file *fútbol_9495.dta* there is information about 469 football matches of the 1a división in Spain corresponding to the seasons $1994-95$ and $1995-96$.

Background

The file contains the following variables:

Number of sold tickets (in thousands) (entradas)

Price of the cheapest ticket (€) (precio)

GDP per capita of the province (in thousands of €) (rentapc)

Number of players of the home team who has played with the national team (int_cas)

Number of players of the away team who has played with the national team (int_vis)

Number of wins in the last 3 games by the home team (nvic3_cas)

Background

Number of wins in the last 3 games by the away team (nvic3_vis)

Season (starting year) (temp)

Climatology (1=Heat without rain, 2=Cold without rain, 3=Rain) (clima)

Budget of the home team (millions €) (pres_cas)

Budget of the away team (millions €) (pres_vis)

Number of tickets on sale (thousands) (capac)

Visiting team (1=Barcelona, 2=Real Madrid, 3=Other) (eq_vis)

Rivalry game (1=Yes, 2=No) (rival)

Background

Estimate the following demand equation (using robust unless it is explicitly said otherwise):

entradas = f(precio, rentapc, intcas, intvis, nvic3cas, nvic3vis, temp, rival, prescas, presvis) + u

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
PRECIO	469	20.618	6.944	4.079	17.147	24.472	48.944
RENDAPC	469	14.091	2.742	9.457	12.976	16.290	18.247
INT_CAS	469	14.243	4.260	6	11	16	24
INT_VIS	469	11.778	4.925	2	8	16	24
NVIC3_CAS	469	0.977	0.825	0	0	2	3
NVIC3_VIS	469	1.058	0.893	0	0	2	3
TEMP	469	94.514	0.500	94	94	95	95
CLIM	469	1.505	0.633	1	1	2	3
PRES_CAS	469	25.445	21.168	4.883	13.959	27.214	71.895
PRES_VIS	469	18.337	17.866	4.692	7.842	19.547	71.895
CAPAC	469	21.596	12.990	4.095	10.377	35.017	46.476
EQ_VIS	469	2.859	0.464	1	3	3	3

Question 1

Interpret the estimated coefficients and comment their individual meaning.

Contrast the joint hypothesis of the coefficients of the variables that refer to the recent trajectory of the two teams (number of wins in the last 3 games). (i.e. a joint hypothesis of 2 variables)

Excercise 2

Question 1

R:

```
m1 ← lm(data=futbol, formula = entradas~precio+rendapc+int_cas+int_vis+
          pres_cas+pres_vis+ nvic3_cas+nvic3_vis+d_rival+temp)
```

Stata:

```
gen d_rival=rival==1
reg entradas precio rendapc int_cas int_vis pres_cas pres_vis
nvic3_cas nvic3_vis d_rival temp, r
```

Question 1

	entradas
PRECIO	-0.078*
	(0.041)

Excercise 2

Question 1

Less attendance when the price increases.

Rival games games have more people attending on average.

Contrast that the coefficients of the variables regarding the home team are equal to the ones referring to the away team.

Repeat (c) but when the errors do not present any heteroskedasticity problem. Then, make a contrast without using the test command, that is, making use of the formula that compares the sum of squared residuals of the restricted and unrestricted model.

Contrast if the effect of the quality of the home and away teams, measured by the number of victories, is the same. Use the F-statistic and the t-statistic.

Excercise 2

Question 1

R:

```
m1 ← lm(data=futbol, formula = entradas~precio+rendapc+int_cas+int_vis+
           pres_cas+pres_vis+ nvic3_cas+nvic3_vis+d_rival+temp)
test ← linearHypothesis(m1,
                        c("nvic3_cas=0", "nvic3_vis=0"),
                        white.adjust = "hc1")
```

Stata:

```
reg entradas precio rendapc int_cas int_vis pres_cas pres_vis
nvic3_cas nvic3_vis d_rival temp, r
test nvic3_cas nvic3_vis
```

Excercise 2

Question 1

F - statistic: [1] 2.67491

P - value: [1] 0.06998967

Excercise 2

Question 1

R:

```
futbol ← futbol %>%
  mutate(d_rival=ifelse(rival==1,1,0))
m1 ← lm(data=futbol, formula = entradas~precio+rendapc+int_cas+int_vis+
          pres_cas+pres_vis+ nvic3_cas+nvic3_vis+d_rival+temp)
test ← linearHypothesis(m1,
                        c("int_cas=int_vis",
                          "pres_cas=pres_vis",
                          "nvic3_cas=nvic3_vis"),
                         white.adjust = "hc1")
```

Excercise 2

Question 1

Stata:

```
reg entradas precio rendapc int_cas int_vis pres_cas pres_vis
nvic3_cas nvic3_vis d_rival temp, r
test (int_cas=int_vis) (pres_cas=pres_vis) (nvic3_cas=nvic3_vis)
```

Excercise 2

Question 1

Without robust standard errors we obtain the following test statistic. The difference arises because the covariance matrix is estimated in a different manner.

F - statistic: [1] 8.546838

P - value: [1] 1.562304e-05

Excercise 2

Question 1

F - statistic: [1] 9.524635

P - value: [1] 4.102253e-06

Excercise 2

Question 1

Can estimate it manually under the assumption of homoskedasticity. The information needed can be obtained from the regression table:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n-k-1)},$$

where SSR_R is from the restricted model, SSR_{UR} is from the unrestricted model, q is the number of restrictions and $n - k$ is the number of degrees of freedom.

Stata:

```
display ((12000.1666-11295.4605)/3)/(11295.4605/(469-(10+1)))
```

Excercise 2

Question 1

F - statistic: [1] 2.789825

P - value: [1] 0.09554782

We reject the hypothesis at the 10 percent-level, but not at 5 percent.

Excercise 2

Question 1

Alternatively we can calculate the t-statistic manually.

Stata:

```
. matrix coef=e(b)
. matrix varcov=e(V)
.
. display (coef[1,7]-coef[1,8])/sqrt(varcov[7,7]+varcov[8,8]-2*varcov[8,7])
1.6702768
```

Can also define a variable given by the difference between the two and include it in the regression. Then the value can be read off directly.

Excercise 2

Question 1

What significant changes happen when we eliminate the variables corresponding to the budgets of the two teams? How do you explain it? [Note: Use the info on the correlations between explanatory variables]

Excercise 2

Question 1

	entradas	
	(1)	(2)
PRECIO	-0.078*	-0.013
	(0.041)	(0.042)
RENDAPC	0.504***	0.745***
	(0.105)	(0.102)
INT_CAS	0.383***	0.720***
	(0.087)	(0.072)
INT_VIS	-0.018	0.224***
	(0.070)	(0.052)
PRES_CAS	0.102***	

Excercise 2

Question 1

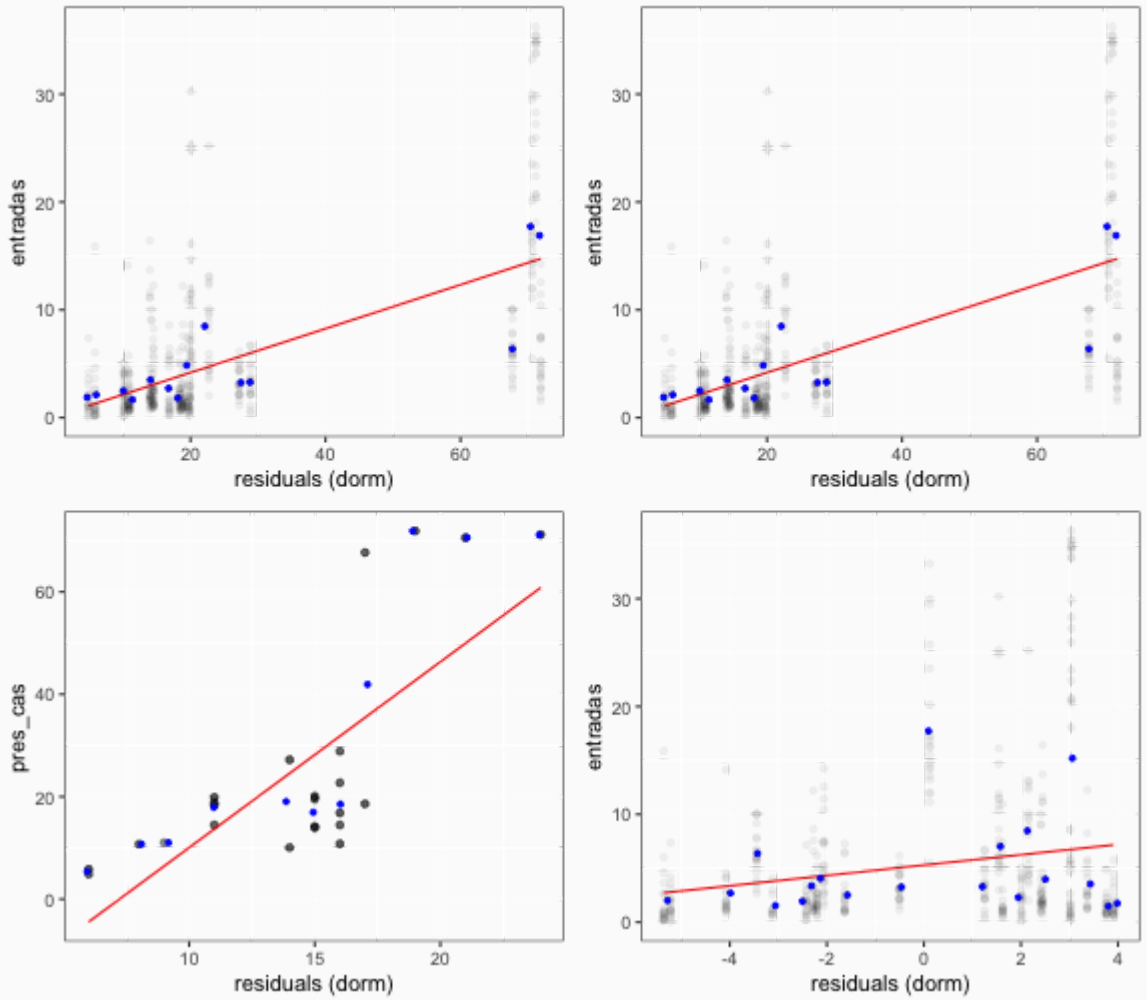
Dropping the budget variables biases upwards the coefficients of the variables corresponding to the number of international players.

This is because the budget variables proxy for a similar variable (the quality of teams), as is reflected in the correlation coefficient of both variables (see coefficients in red) and the omitted variables (the budgets) positively affect the number of tickets sold.

Highlight the negative coefficient of int_vis in the complete specification, even if not statistically different from zero.

Excercise 2

Question 1



Excercise 2

Question 1

Once we remove the correlation between the income of the club and the number of international players, there is no relationship between the attendance and the number of international players.

The relationship seems to be completely driven by the relationship with the clubs income.

Useful for writing term papers when you want to plot the relationship from a multivariate regresson.

Always very important to plot the data. Are there outliers? Which ones are the influential observations?

Session 5

Excercise 1

Background

In the file pwt90Year2014.dta you will find information of the Penn World Tables http://www.rug.nl/ggdc/productivity/pwt/ for the year 2014 and 182 countries.

Background

The file contains, among others, the following variables:

- *rgdpna*: Real GDP at constant prices of 2011 (in millions of USD of 2011)
- *rkna*: Capital stock at constant prices of 2011 (in millions of USD of 2011)
- *rtfpna*: Total factor productivity at constant prices of 2011
- *pop*: Population (in millions)
- *emp*: Employed population (in millions)
- *avh*: Average hours worked per year by active people
- *hc*: Human capital index, based on years of schooling and returns to education

Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
rgdpna	182	569,321.700	1,921,326.000	100.275	20,023.860	355,564.600	17,150,538.000
rkna	180	2,045,551.000	6,931,829.000	1,428.255	55,357.240	1,400,028.000	67,590,072.000
rtfpna	116	1.004	0.073	0.659	0.968	1.034	1.283
pop	182	39.312	143.978	0.005	2.084	27.109	1,369.436
emp	169	18.883	74.343	0.044	1.202	12.280	798.368
avh	68	1,864.113	260.592	1,371.101	1,684.616	2,050.126	2,510.406
hc	144	2.595	0.690	1.193	2.018	3.156	3.734

Question 1

Estimate a regression by OLS where the dependent variable is the real GDP (rgdpna) and the explanatory variable is the capital stock (rkna).

Based on your intuition and knowledge of economics, which relevant variable(s) are we omitting in this equation?

Excercise 1

Question 1

Know from macroeconomics that the relationship is often modelled in the following manner,

$$Y = A \times F(K, L).$$

Therefore we are omitting L and A.

Excercise 1

Question 1

R:

```
m ← lm(data=pwt, formula = rgdpna~rkna)
m ← coeftest(m, vcov = vcovHC(m, type="HC1"))
```

Stata:

```
reg rgdpna rkna,r
```

Excercise 1

Question 1

RKNA	0.275***
	(0.018)
CONSTANT	13,493.660
	(24,123.910)
Notes:	***p < .01; **p < .05; *p < .1

Question 2

Estimate a second regression by OLS where the dependent variable is the real GDP (rgdpna) and the explanatory variables are the capital stock (rkna), the population (pop) and the total factor productivity (rtfpna).

Interpret the coefficients and test the hypothesis that the coefficients of pop and rkna are jointly zero. Which are your conclusions?

Excercise 1

Question 2

R:

```
m2 ← lm(data=pwt, formula = rgdpna~rkna + pop + rtfpna)
m2 ← coeftest(m2, vcov = vcovHC(m2, type="HC1"))
```

Stata:

```
reg rgdpna rkna pop rtfpna,r
test rkna pop
corr rkna pop
```

Excercise 1

Question 2

	(1)	(2)
RKNA	0.275***	0.274***
	(0.018)	(0.033)
POP		77.188
		(1,252.219)
RTFPNA		-157,824.200
		(228,047.700)
CONSTANT	13,493.660	165,297.500
	(24,123.910)	(248,476.100)
Notes:	***p < .01; **p < .05; *p < .1	

Excercise 1

Question 2

Testing the hypothesis:

R:

```
m2 ← lm(data=pwt, formula = rgdpna~rkna + pop + rtfpna)
test ← linearHypothesis(m2,
                        c("pop=0", "rkna=0"),
                        white.adjust = "hc1")
```

Stata:

```
reg rgdpna rkna pop rtfpna,r
test rkna pop
corr rkna pop
```

Excercise 1

Question 2

Linear hypothesis test

Hypothesis: pop = 0 rkna = 0

Model 1: restricted model Model 2: rgdpna ~ rkna + pop + rtfpna

Note: Coefficient covariance matrix supplied.

Res.Df Df F Pr(>F)
1 114
2 112 2 125.18 < 2.2e-16 *

Signif. codes: 0 '***' 0.001 "** 0.01 '*' 0.05 '.' 0.1 ' ' 1

Question 3

Do a graphical scatter plot (in Stata, the command is: scatter y x) of real GDP (rgdpna) and the stock of capital (rkna) and another scatter plot of real GDP (rgdpna) and population (pop).

Are both relationships linear?

Excercise 1

Question 2

R:

```
p1 ← ggplot(data=pwt, aes(x=rkna, y =  rgdpna)) +
  stat_summary_bin(shape=1,fun.y='mean', alpha=0.7,
                   bins=800, size=1, geom='point')
p2 ← ggplot(data=pwt, aes(x=pop, y =  rgdpna)) +
  stat_summary_bin(shape=1,fun.y='mean', alpha=0.7,
                   bins=800, size=1, geom='point')
ggarrange(p1,p2)
```

Excercise 1

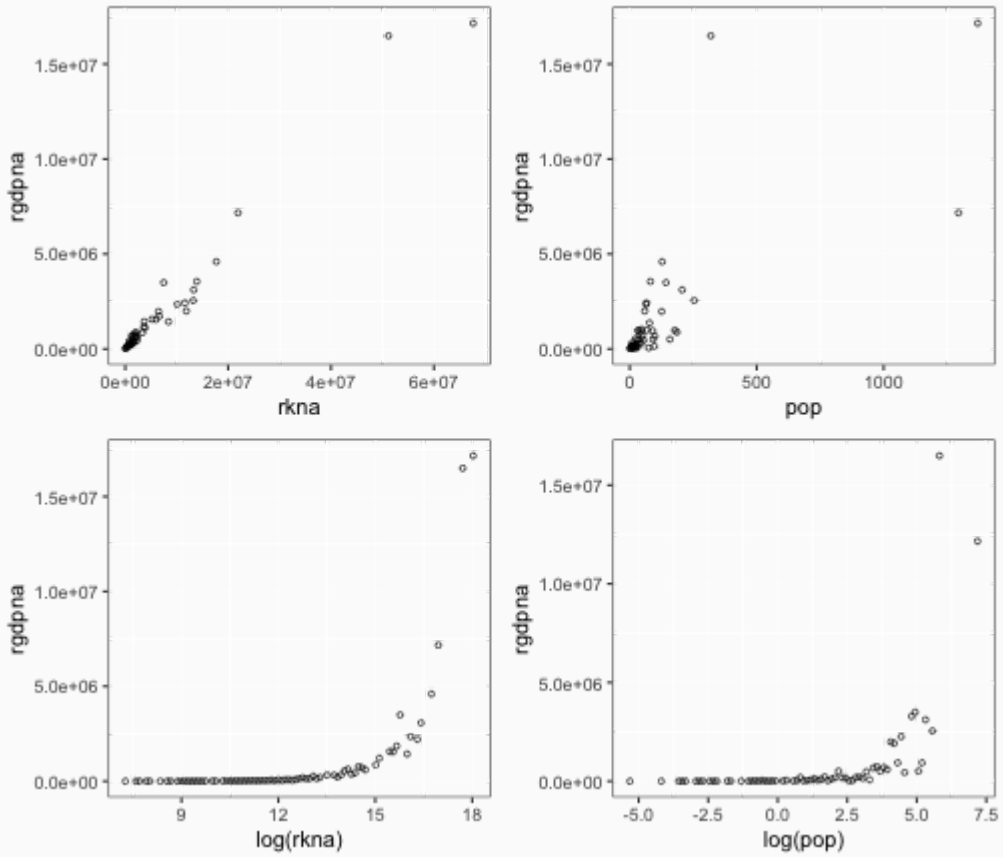
Question 2

Stata:

```
scatter rgdpna rkna
scatter rgdpna pop
```

Excercise 1

Question 2



Question 3

Create the logarithm of real GDP (rgdpna), the stock of capital (rkna) and the population (pop), and estimate by OLS a regression where the dependent variable is the logarithm of real GDP (rgdpna), and the explanatory variables are the logarithm of the stock of capital (rkna), the logarithm of the population (pop) and the total factor productivity (rtfpna). Interpret the estimated coefficients.

Does it make sense to interpret β_0 in this case?

Excercise 1

Question 3

R:

```
pwt_log ← pwt %>% mutate(lrgdpna=log(rgdpna),
                         lrkna=log(rkna),
                         lpop=log(pop))
```

Stata:

```
gen lrgdpna=ln(rgdpna)
gen lrkna=ln(rkna)
gen lpop=ln(pop)
reg lrgdpna lrkna lpop rtfpna,r
```

Excercise 1

Question 3

LRKNA	0.851^{***}
	(0.029)
LPOP	0.163^{***}
	(0.033)
RTFPNA	0.084
	(0.753)
CONSTANT	0.270
	(0.954)
Notes:	***p < .01; **p < .05; *p < .1

Excercise 1

Question 3

No sense to interpret _constant: it would be a country with 1 K, 1 L, and zero A.

Question 4

Using the results of the previous question, test the hypothesis of constant returns to scale (that is, that the sum of the two estimated coefficients is equal to 1).

Does the data verify this hypothesis? Rewrite the previous equations to test the same hypothesis but now based on a single estimated coefficient.

Excercise 1

Question 4

Consider the standard production function:

$$F = A \times K^{\alpha} \times L^{\beta}$$

The scale properties are summarized by the α coefficients:

$$A(\lambda K)^{\alpha} \times (\lambda L)^{\beta} = Y \times \lambda^{\alpha+\beta}$$

Therefore the scale properties deped on the value of $\alpha + \beta$:

CRTS if $\alpha + \beta = 1$, IRTS if $\alpha + \beta > 1$ and decreasing if $\alpha + \beta < 1$.

Excercise 1

Question 4

We want to test if the coefficients add up to one.

R:

```
m1 ← lm(data=pwt_log, formula = lrgdpna~lrkna+lpop+rtfpna)
test ← linearHypothesis(m1,
                        c("lrkna+lpop=1"),
                        white.adjust = "hc1")
```

Stata:

```
reg lrgdpna lrkna lpop rtfpna, r
test lrkna+lpop=1
```

Excercise 1

Question 4

Linear hypothesis test

Hypothesis: lrkna - lpop = 0

Model 1: restricted model Model 2: lrgdpna ~ lrkna + lpop + rtfpna

Note: Coefficient covariance matrix supplied.

Res.Df Df F Pr(>F)
1 113
2 112 1 134.18 < 2.2e-16 *

Signif. codes: 0 '***' 0.001 "** 0.01 '*' 0.05 '.' 0.1 ' ' 1

Excercise 1

Question 4

Cannot reject constant returns to scale.

Question 5

Estimate by OLS a regression where the dependent variable is the logarithm of real GDP (rgdpna) and the explanatory variables are the logarithm of the stock of capital (rkna) and the logarithm of the population (pop).

Then, estimate another regression where you use the logarithm of the employed population (emp) instead of the logarithm of the population (pop). Does the estimated coefficient change much?

Excercise 1

Question 5

R:

```
pwt_log ← pwt %>% mutate(lrgdpna=log(rgdpna),
                         lrkna=log(rkna),
                         lpop=log(pop),
                         lemp=log(emp))

m1 ← lm(data=pwt_log, formula = lrgdpna~lrkna+lpop)
m1 ← coeftest(m1, vcov = vcovHC(m1, type="HC1"))
m2 ← lm(data=pwt_log, formula = lrgdpna~lrkna +lemp)
m2 ← coeftest(m2, vcov = vcovHC(m2, type="HC1"))
```

Excercise 1

Question 5

Stata:

```
reg lrgdpna lrkna lpop,r
gen lemp=ln(emp)
reg lrgdpna lrkna lemp,r
corr lpop lemp
```

Excercise 1

Question 5

	Dependent variable:	
	(1)	(2)
lrkna	0.828***	0.815***
	(0.020)	(0.020)
lpop	0.226***	
	(0.023)	
lemp		0.215***
		(0.024)
Constant	0.505**	0.889***
	(0.221)	(0.243)

Excercise 1

Question 5

Not very much, because highly correlated and they are in logs so we are looking at percentage changes in the X.

Question 6

Estimate again the last equation of the previous question, but now also include four additional variables: avh, the square of avh, hc and the square of hc.

Test if the relation between real GDP and the average hours worked per year by active people is linear or quadratic. Then do the same for human capital.

From a statistical viewpoint, did it make sense to include these variables? But what if you only include hc (and its square)? Why do results change?

Excercise 1

Question 6

R:

```
pwt_log ← pwt %>% mutate(lrgdpna=log(rgdpna),
                         lrkna=log(rkna),
                         lpop=log(pop),
                         lemp=log(emp),
                         avh_sq=avh^2,
                         hc_sq=hc^2)

m1 ← lm(data=pwt_log, formula = lrgdpna~lrkna+lemp+avh+avh_sq+hc+hc_sq)
m1 ← coeftest(m1, vcov = vcovHC(m1, type="HC1"))
```

Excercise 1

Question 6

Stata:

```
gen avh_sq=avh*avh
gen hc_sq=hc*hc
reg lrgdpna lrkna lemp avh avh_sq hc hc_sq,r
```

	(1)	(2)	(3)
LRKNA	0.828***	0.815***	0.606***
	(0.020)	(0.020)	(0.049)
LPOP	0.226***		
	(0.023)		
LEMP		0.215***	0.416***
		(0.024)	(0.050)
AVH			0.0003
			(0.001)
AVH_SQ			-0.00000
			(0.00000)
HC			0.357
			(0.909)
HC_SQ			-0.003

Excercise 1

Question 6

Quadratic terms are not significant in the output of the regression

Probably not much sense to add the square terms, as they are not significant.

Results change, due to imperfect multicollinearity. They square term is significant and shows a concave function.

If you have hours and human capital and their squares is too many variables being highly correlated.

Excercise 1

Question 6

When variables are highly correlated we have a trade off between unbiasedness and precision. If we include both variables the estimator will be unbiased but will be to imprecise (that is, large standard error).

If we don't include the variable, then the included variable will capture the effect of both variables