

EEP/IAS 118 - Introductory Applied Econometrics Problem Set 2

Problem Set 2, Answer Key, Spring 2023, Villas-Boas

Due in Gradescope – see deadline due time in Gradescope – Feb 23, 2023

Submit materials as **one pdf** on [Gradescope](#). After uploading the pdf to Gradescope, please **assign appropriate pages to each question**. Furthermore, please double check that, **on the pdf**, (i) codes are runnable and fully visible and (ii) associated outputs are fully visible. **Points will be deducted** if pages are not assigned to pages and codes/outputs are cut-off and not fully visible.

Chiara Luna Pilato Moncada

Exercise 1 (in R). Relationship between CEO Salary and Covariates

Guidelines:

- This exercise should be completed using R.
- When you want an output to show, you need to explicitly call the object. For example, if you want to show the mean airline fares, you can't just type `MeanSalary <- mean(data$salary)`, because that will just save the output to MeanSalary. Instead, you need to then type `MeanSalary` on its own, so it displays the output. Answers that do not display the desired output will be graded as incorrect.
- To write comments in your script or in code cells (text that will appear but will not be read as R commands), type a `#` at the beginning of the line. Use these as notes to keep track of which question you are trying to answer, the purpose of each command, etc.

```
In [1]: # a commented out line of code - nothing happens if you run this cell
```

```
In [2]: library(tidyverse)
library(haven)
```

```

— Attaching packages — tidyverse 1.
3.0 —

✓ ggplot2 3.4.0    ✓ purrr 1.0.1
✓ tibble 3.1.8     ✓ dplyr 1.1.0
✓ tidyr 1.3.0      ✓ stringr 1.5.0
✓ readr 2.1.3      ✓ forcats 1.0.0

— Conflicts — tidyverse_conflicts
() —
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()     masks stats::lag()

```

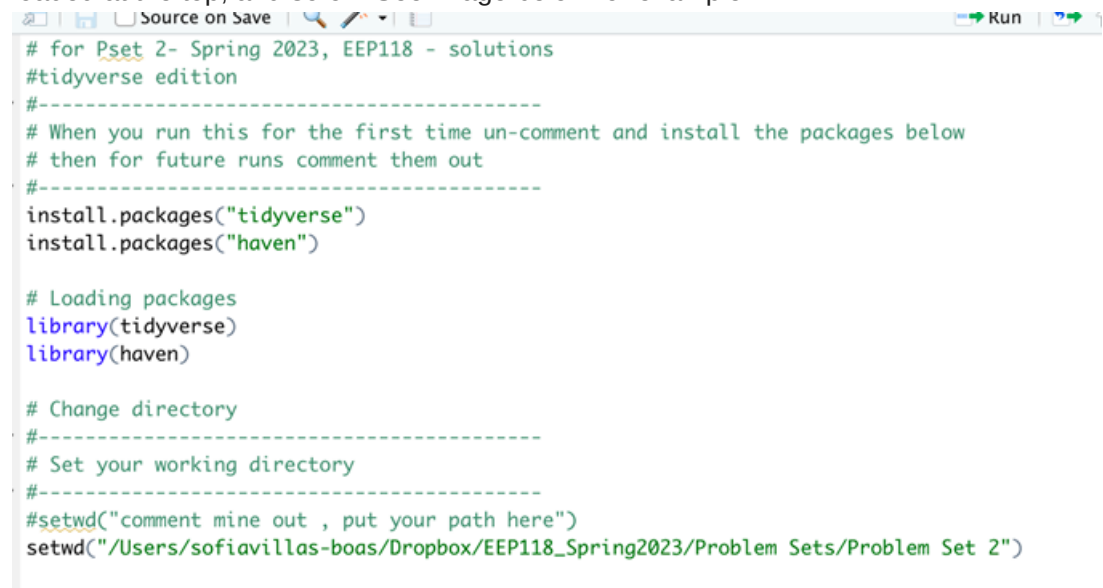
To get started:

- For those working with R studio on local installation, install the following two packages (as seen in header image below). For those working on Datahub, DO NOT install packages.
 - `install.packages("tidyverse")`
 - `install.packages("haven")`
- Then, for both local and Datahub users, load these packages.
 - `library(tidyverse)`
 - `library(haven)`
- This time you are opening a STATA-formatted ".dta" file, and so you will use the `read_dta()` command, provided by the haven package. Load data into R as.
 - `YourDataName <- read_dta("file_name")`

R Tips:

- See the Bootcamp Parts 1 and 2 for some basic. See also all the Lecture R script files, where we did most of what you are asked to do here: e.g., Lecture5.R, Lecture6.R, and Lecture7.R.
- The command `table()` lists all values a variable takes in the sample and the number of times it takes each value.
- To summarize data for a specified subset of the observations, you can use `filter()` to subset the data, and then either `summary()` for simple summary statistics or `summarise()` in tidyverse to generate more detailed summary statistics.
- If you are using R Studio, your header should be organized with notes such that purpose of the script file is clear, packages needed to run the script is installed and

loaded at the top, and so on. See image below for example:



```
# for Pset 2- Spring 2023, EEP118 - solutions
#tidyverse edition
#-----
# When you run this for the first time un-comment and install the packages below
# then for future runs comment them out
#-----
install.packages("tidyverse")
install.packages("haven")

# Loading packages
library(tidyverse)
library(haven)

# Change directory
#-----
# Set your working directory
#-----
#setwd("comment mine out , put your path here")
setwd("/Users/sofiavillas-boas/Dropbox/EEP118_Spring2023/Problem Sets/Problem Set 2")
```

Data Description:

The data set for this exercise comes from a 1990 sample of salaries for several firms' chief executive officers (CEO). In this problem set, we delve into that data set to provide insights into the CEO salaries and then relate the variation in salaries in the sample to firms and individual CEO characteristics.

The **pset2_2023data.dta** file includes the following variables, only which we will use in this pset:

Variable Name	Description and Units
salary	1990 compensation, 1000 USD
age	in years
college	= 1 if attended college
grad	= 1 if attended graduate school
comten	years with company
ceoten	years as ceo with company
sales	1990 firm sales, million USD
profits	1990 profits, million USD

Question 1: First, we would like you to become familiar with your data:

Note of Caution: The units for salary is 1,000 USD. This means, when your coefficient or summary stats on the (unit) variable salary reads 800, the correct answer or interpretation we are looking for is 800,000 USD.

(a) Read in the data: use `my_data <- read_dta("pset2_2023data.dta")`

```
In [73]: my_data <- read_dta("pset2_2023data.dta")
```

(b) How many individual CEO's are in the data set?

```
In [4]: nrow(my_data)
```

177

In this specific data set, there are 177 individual CEO's. Because this data is all according to different CEOs, we can get the number of CEOs by calculating the amount of rows in the data set.

(c) How many CEO's have salary greater than 1,000,000 dollars?

Hint: Using the tidyverse method. The command `filter()` trims dataframe to obs with salary more than 1,000,000 USD. Then, look at # rows after using `over1000 <- filter(my_data, salary > 1000)`

```
In [6]: over1000 <- filter(my_data, salary > 1000)
nrow(over1000)
```

57

In this case, as seen in the code, there are 57 CEOs that have a salary greater than 1,000,000 dollars.

(d) What is the average salary in the data set? (Reminder: Convert and indicate appropriate units.)

```
In [8]: average_salary <- my_data %>%
summarise(average_salary = mean(salary))
average_salary
```

A tibble: 1 × 1

average_salary

<dbl>

865.8644

```
In [16]: #To transform the number into the correct units we multiply the past
#value by 1000
865.8644*1000
```

865864.4

The average salary in the data set would be 865,864.4 dollars, as seen in the code

(e) What is the range for the variable salary in the data? (Reminder: Convert and indicate appropriate units.)

```
In [25]: range_of_variable_salary<- max(my_data$salary)-min(my_data$salary)
print(range_of_variable_salary)

[1] 5199
```

```
In [26]: #To transform the number into the correct units we multiply the past
#value by 1000
5199*1000

5199000
```

```
In [32]: #To make sure I got the right range of the values, I added
#this extra code of the range as a function of max and min,
#but this is an extra unnesseray step.
range_of_variable_salary<- range(my_data$salary)
print(range_of_variable_salary)

[1] 100 5299
```

The range for the variable salary in the data of the data set would be 5,199,000 dollars, with a max and min of [100,5299] (after adjusting for units we get the range of 5,199,000 dollars)

(f) What is the range for tenure of CEO's in the data

Hint: `summarise(my_data, "Range CEO Tenure (Yrs)" = max(ceoten) - min(ceoten))`

```
In [33]: summarise(my_data, "Range CEO Tenure (Yrs)" = max(ceoten) - min(ceoten))

A tibble: 1 × 1
  Range CEO Tenure (Yrs)
      <dbl>
1                37
```

In this data set, the range for the tenure of the CEO's would be 37 years.

(g) Construct a variable TenureBeforeCEO equal to tenure at the company minus the tenure as CEO.

You can create a new variable: `my_data <- mutate(my_data, TenureBeforeCEO = comten-ceoten)`

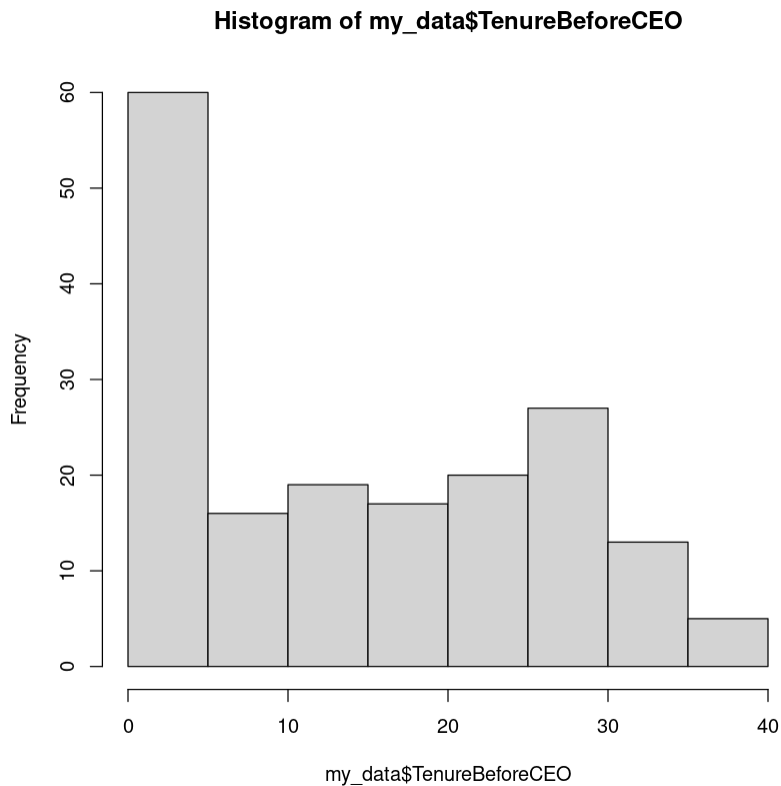
Note: you do not need to interpret this variable, this is just to practice making new variables in R.

```
In [34]: my_data <- mutate(my_data, TenureBeforeCEO = comten-ceoten)
```

(h) Plot a histogram of this constructed variable.

Hint: `hist(my_data$TenureBeforeCEO)`

```
In [35]: hist(my_data$TenureBeforeCEO)
```



(i) What is the range of tenure for CEOs whose salaries are over 1,000,000 USD?

```
In [36]: summarise(over1000, "Range CEO Tenure (Yrs)" = max(ceoten) - min(ceoten))
```

A tibble: 1 × 1

Range CEO Tenure (Yrs)

<dbl>

36

In this data set, the range of tenure of CEOs with salaries over 1,000,000 USD is 36 years.

(j) What is the mean? What is the median? (of tenure for CEOs whose salaries are over 1 mil USD)

Hint 1: `summary(over1000$ceoten)`

Hint 2: Using tidyverse method,

```
summarise(over1000,
```

```
"Mean Tenure Years for Salaries > 1000" = mean(ceoten),
```

```
"Median Tenure Years for Salaries > 1000" = median(ceoten))
```

```
In [37]: summary(over1000$ceoten)
```

```
summarise(over1000,
  "Mean Tenure Years for Salaries > 1000" = mean(ceoten),
  "Median Tenure Years for Salaries > 1000" = median(ceoten))
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 3.000 6.000 8.211 11.000 37.000
```

A tibble: 1 × 2

Mean Tenure Years for Salaries > 1000	Median Tenure Years for Salaries > 1000
<dbl>	<dbl>
8.210526	6

In this data set, the mean of tenure for CEOs whose salaries are over 1 million USD is 8.211 years. Additionally, the median of tenure for CEOs whose salaries are over 1 million USD is 6 years.

(k) Filter out the college graduate CEOs into a separate data R object. Compute the mean salary separately for CEO's with and without graduate school attendance.

Hint: Use `group_by()` and `summarise()` in the tidyverse package. See Section 2 slides or notes for introduction to `group_by` function.

```
In [45]: grad_ceos <- my_data %>%
  filter(college == 1, grad == 1)
salary_summaries <- my_data %>% group_by(grad) %>%
  summarise(mean_salary = mean(salary))
```

salary_summaries

A tibble: 2 × 2

grad	mean_salary
<dbl>	<dbl>
0	867.7349
1	864.2128

```
In [47]: #To adjust to correct units of salary we multiply by 1000
```

```
967.7349*1000
864.2128*1000
```

967734.9

864212.8

(l) Do CEOs with graduate attendance have higher average salaries than CEOs without a graduate attendance?

In this case, the CEOs with graduate attendance (expressed as the 1 in the code) have a mean salary of 864,212.8 USD, while CEOs without graduate attendance (expressed as the

0 in the code) have a mean salary of 967,734.9 USD

We see how the CEOs with graduate attendance do not have higher average salaries as those CEOs with no graduate attendance.

Question 2

Consider the following model, where x_i = tenure in years of individual i as CEO

$$\text{Salary}_i = \beta_1 + \beta_2 x_i + u_i \quad i = \text{individuals } 1, 2, \dots, N \text{ in the data} \quad (\text{eq.1})$$

(a) Estimate the model in R with the `lm()` command. Interpret your β_1 and β_2 coefficients, remembering the triplet S(ign), S(ignificance), S(ize), though you don't need to comment on significance in this problem-set. Make sure your uploaded pdf includes the R output.

```
In [48]: model <- lm(salary ~ ceoten, data = my_data)
summary(model)
model

Call:
lm(formula = salary ~ ceoten, data = my_data)

Residuals:
    Min       1Q   Median       3Q      Max
-977.8 -345.9 -169.4  263.8 4373.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   772.426     65.676   11.761  <2e-16 ***
ceoten         11.746      6.148    1.911   0.0577 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 583.2 on 175 degrees of freedom
Multiple R-squared:  0.02043,    Adjusted R-squared:  0.01484
F-statistic: 3.651 on 1 and 175 DF,  p-value: 0.05769
Call:
lm(formula = salary ~ ceoten, data = my_data)

Coefficients:
(Intercept)          ceoten
    772.43             11.75
```

B_1 in this case is the estimated intercept of 772.426. B_2 is 11.746, meaning that the expected salary will increase by \$11,746 for each year you serve as a CEO, when holding all other variables constant. In this case, all signs are positive

(b) How well does the number of years of tenure as CEO predict the salary?

```
In [49]: 0.02043*100
```


2.043

Our R^2 value in this case is 0.02043, meaning about 2.043% of our model is explained away, which is a very low percentage. Our model can't really predict salary greatly, and our model can only "explain away" 2.043% of the variance.

The residual standard error is 583.2, which is far away from our mean salary.

We also know that our p-value in this case is 0.05769, and given the notion that for a p-value to be highly significant, it needs to be smaller than 0.05, in this case we are above. This means that our results are not that highly statistically significant.

Based on these notions, we could conclude that the number of years of tenure as CEO or ceoten, is not a very good model to predict the salary.

(c) What is the predicted salary for a CEO with an average tenure as CEO of 4 years?

```
In [51]: ceo_4year_salary= 772.43+11.75*4  
ceo_4year_salary
```

819.43

```
In [52]: 819.43*1000
```

819430

After multiplying this by 100 to get the correct units, we get that the average salary of a CEO with 4 years of tenure is \$819,430 USD

Question 3

Consider the following model for observations of CEO with **more than 0 years of tenure**, where x_i =tenure in years of individual i as CEO

$$\log(\text{Salary}_i) = \beta_3 + \beta_4 \log(x_i) + v_i \quad i = \text{individuals } 1, 2, \dots, N \text{ in the data}$$

(a) Estimate the model in R with the `lm()` command. Interpret your β_4 coefficient, remembering the triplet S(ign), S(ignificance), S(ize), though you don't need to comment you need to generate the logarithm of the variables of interest for this question's model. Use `log()`, namely the log of salary and log of tenure as CEO. Run the model only for those that have non-zero tenure as CEO, that is for a filtered subset of the data that `ceoten > 0`.

```
In [57]: new_data<-my_data[my_data$ceoten > 0, ]  
new_data$log_ceoten <- log(new_data$ceoten)  
  
lm1 <- lm(log(salary) ~ log_ceoten,  
data = new_data[new_data$ceoten > 0, ])  
  
summary(lm1)
```

```
Call:
lm(formula = log(salary) ~ log_ceoten, data = new_data[new_data$ceoten >
0, ])
```

Residuals:

Min	1Q	Median	3Q	Max
-2.15314	-0.38051	-0.01206	0.44648	1.89128

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.40895	0.09822	65.254	<2e-16 ***
log_ceoten	0.10723	0.05022	2.135	0.0342 *

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

Residual standard error: 0.6028 on 170 degrees of freedom

Multiple R-squared: 0.02612, Adjusted R-squared: 0.02039

F-statistic: 4.559 on 1 and 170 DF, p-value: 0.03418

In this case, our B_4 is 0.10723, meaning that as we incorporated this new beta, there was a positive relationship between the log salary and log ceoten of CEOs that have more than 0 years of tenure. For every 1% unit increase in our logtenure, there is a 0.10723% increase in the salaries.

Here we see a significant p-value, as its less than 0.05

(b) Using the results from estimating (eq. 2), how would you expect the salary to change if the years of CEO tenure increases by 25%?

In [59]: $0.25 * 0.10723$

$0.0268075 * 100$

0.0268075

2.68075

After the previous results, we could expect the salaries to change by 2.68% if the years of CEO tenure increase by 25%

Question 4

We will now explore the role of CEO and firm characteristics in explaining the salary where x is the CEO tenure years as before and $grad_i$ is an indicator equal to one if the CEO attended graduate school and equal to zero otherwise.

$$Salary_i = \beta_5 + \beta_2 x_i + \beta_3 grad_i + \gamma_i \quad i = \text{individuals } 1, 2, \dots, N \text{ in the data} \quad (eq.3)$$

(a) Estimate equation (eq.3). How did your estimate β_2 for x (that is, CEO tenure) change between equation (eq.1) and equation (eq.3)?

```
In [60]: eq3 <- lm(salary ~ ceoten + grad, data = my_data)
summary(eq3)
```

Call:

```
lm(formula = salary ~ ceoten + grad, data = my_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-986.1	-348.7	-167.1	264.7	4380.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	764.261	83.975	9.101	<2e-16 ***
ceoten	11.846	6.198	1.911	0.0576 .
grad	13.879	88.559	0.157	0.8756

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

Residual standard error: 584.8 on 174 degrees of freedom

Multiple R-squared: 0.02057, Adjusted R-squared: 0.009315

F-statistic: 1.827 on 2 and 174 DF, p-value: 0.1639

Our B₂ estimate between equation 1 (where it was 11.75) and equation 3 (where we have 11.846), we see the B₂ increased in its estimate.

```
In [69]: 11.84-11.75
1-0.08999999999999999
```

0.08999999999999999

0.91

(b) Without performing any calculations, what information does this give you about the correlation between the CEO tenure years and whether a CEO attended graduate school? (Explain your reasoning in no more than 4 sentences. Hint: OVB)

Without calculations, this tells me that the correlation between CEO tenure years and if a CEO attended grad school is that there is not really a strong correlation between these two: tenure years and grad school attendance, when its about predicting salary.

We see in our initial equation, we used less variables in the model to determine the salary, but in our third equation we also incorporated the grad school attendance variable; so by doing this we reduced our omitted Variable Bias, as more variables always gets us away from bias a gets us more accurate results

(c) Compute the correlation between the years of CEO tenure and the profits of the firm of the CEO in question. If you include the variable profits in the model in equation (eq.3) and that slope coefficient for profits estimate is 0.58, what do you think will happen to the estimated coefficient on the CEO tenure years when compared to the estimate you got in equation (eq. 3)? Explain your reasoning using OVB briefly here again.

```
In [70]: cor(my_data$ceoten, my_data$profits)
cor(my_data$grad, my_data$profits)
eq3_w_profits <- lm(salary ~ ceoten + grad + profits, data = my_data)
summary(eq3_w_profits)
```

-0.0216067502859885

0.0978255293723162

Call:

lm(formula = salary ~ ceoten + grad + profits, data = my_data)

Residuals:

	Min	1Q	Median	3Q	Max
	-1079.9	-303.7	-107.9	220.5	4406.3

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	664.5934	79.0221	8.410	1.47e-14	***
ceoten	12.2294	5.6918	2.149	0.0331	*
grad	-31.4741	81.7015	-0.385	0.7005	
profits	0.5808	0.1006	5.774	3.51e-08	***

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

Residual standard error: 537.1 on 173 degrees of freedom

Multiple R-squared: 0.1788, Adjusted R-squared: 0.1646

F-statistic: 12.56 on 3 and 173 DF, p-value: 1.813e-07

We do in fact see the estimated slope to be 0.58 in the data, meaning that when holding everything else equal, there is a positive relationship between company profits and CEO salaries. In this case by introducing more variables into the model we can analyze to see if we have any changes in the values due to omitted variables (which were not there in equation 3). Also important to note that the grad estimated variable in this case is negative.

(d) Predict the expected salary for a CEO with no graduate education and with 20 years of tenure as CEO using your estimates from equation (eq.3).

```
In [72]: 764.261+11.846*20+13.879*0
1001.181*1000
```

1001.181

1001181

In this case, the predicted salary for a CEO with no graduate education, and 20 years of tenure as CEO (using estimates from eq.3) is \$1,001,181 USD.

Exercise 2 (Intuition Only; No R or Calculation Involved)

Policy makers are interested in understanding important determinants of number of electric cars sold in US cities and run the following model:

$$\text{Per Capita Number of Electric Cars Sold}_j = \beta_1 + \beta_2 GP_j + \beta_3 ER_j + u_j$$

where GP_j corresponds to gasoline retail average price in a US city j and is assumed to be related to the Number of Electric vehicles sold in city j (PerCapitaNumber of Electric Cars Sold), and also related to whether city j has environmental regulation for gasoline or not (indicator $ER=1$, or $=0$). Finally, u_j is the disturbance term related to the electric vehicle per capita sales in city j .

(a) What do you expect the sign of β_3 to be in equation (eq.4)? Why?

The sign of ER , the environmental regulation being present or not I would expect to be positive. That would definitely affect if electric cars are sold in US Cities where there is a lot of environmental regulation or not. If we control or regulate gasoline (for gas using cars), that would help with the selling of electrical cars, so it would only increase sales of Electric car sales.

(b) What would probably happen to β_2 if you omit $ER=EnvRegulation$ from the estimation, assuming that cities with environmental regulation can only sell gasoline that is more costly? Explain (very briefly) why.

In there, if you omit the ER variable from the estimation, you will have a biased model. In this case there's a possibility of over estimating the impact of gas prices on electric car sales if we can only look at the gasoline that is sold more costly.