STIJN MASSCHELEIN

JUST ENOUGH TO BE DAN-GEROUS

Contents

1
Preamble

The goal of the course is to give you enough information to start a quantitative research project in accounting and finance. The emphasis of the course is not so much on the actual accounting and finance focus of the thesis.[1] The aim for the course is threefold.

1. To teach general research skills such as doing a literature review, pitching a research idea, and understanding a theory.
2. To teach practical `R` programming skills
3. To teach how to perform common statistical procedures

This should give you just enough information to start a research project but also just enough to make mistakes. In the notes, I avoid most of the statistical theory and ignore the assumptions of a lot of the statistical methods we are going to use. First of all, the skills we will teach in this unit will translate better to jobs outside of academia. Second, there are excellent introductions to the theory available (Angrist and Pischke 2008; Cunningham 2018).[2] If you plan on becoming an academic in accounting and finance, these are must reads. Third, the literature in your chosen field will have a number of preferred statistical methods for a given research question. While it is a good idea to question current research practices, as a budding researcher you might not have yet developed the knowledge base to question the practices that have survived the peer review process and criticism from follow-up research.

The advantage of focusing on teaching just enough statistical knowledge to start a project is that I can spend more time and space on other ~~pet peeves of mine~~ topics. While most statistical books emphasise the role of theory in the analysis, there is often very little guidance on what that actually means in an actual research project. From the start, I will use an example with compensation data from S&P500 firms in the U.S to illustrate how to make sure that your data reflects your theory.[3]

[1] Everyone will have a different focus and it is difficult to give general advice that will be useful to everyone. I think these notes could be useful to people outside of accounting and finance who want to embark on their first research project with observational data

[2] See Scott Cunningham's website for the book and Jake Johnson's Github repository for the `R` datasets and the `R` code equivalent to the Stata code in the book

[3] I am not a specialist on the executive compensation literature and all the conclusions I draw should be taken with a grain of salt.

This is where the R statistical language comes in. To test whether a theory is correct, we can use graphs and the R language has excellent facilities to make graphics. Descriptive plots of the data can help to evaluate whether the data are measuring what we think they are measuring and whether the statistical method is actually appropriate. In addition, a lot of modern theories in accounting and finance require strong mathematical skills to really understand them. However, I find that simulations can often provide a better intuition for which variables are important in a theory and why. The R language again makes it very easy to program those simulations and I will use simulations to illustrate an advanced theory in the executive compensation literature.

Lastly, simulations can also help to better understand the assumptions and limitations of a statistical method. My aim with the focus on simulations is to give you the tools to evaluate your preferred method even if you do not have the mathematical background to understand all the assumptions of the method.[4]

[4] I do not want to imply that simulations can replace a strong mathematical understanding but they can get you half way there.

2
A Research Project: CEO Compensation

2.1 Introduction

CEO compensation is something that people in finance, accounting, economics, and outside of academia are interested in. The topic is probably the one with the most overlap between accounting and finance. In these notes, I will focus on two research questions:

1. How high can we expect the total compensation of a CEO to be based on some simple economic assumptions. Too high CEO compensation is sometimes seen as a signal of bad corporate goverance. To measure what 'too high' means, we first need to establish a baseline of what normal levels of compensation actually mean.
2. How should CEOs be incentivised: equity or options? And how schould we measure whether CEOs have appropriate incentives? Pay-for-performance is a big topic in accounting, finance, and economics and there is a long standing discussion of what the best performance measure is for a CEO.

Altough I am not a specialist in this literature, I am comforable with the type of economic theories that researchers test in this literature.[1] I am going to stress the role of theory in data analysis a lot. Some of you will have a topic that is at first sight less theory driven or rely more strongly on very specific knowledge about your setting. One of my goals is to convince you that even with these projects it is useful to think about the underlying story that you are testing.[2]

[1] The theories in these notes are all based on the overview paper by Edmans and Gabaix (2016). Finding a good overview paper is the best start to any research project.

[2] I have been mocked by multiple people in the department for asking one too many times: "I understand, but what is the story?"

2.2 Theory

The first theory on CEO compensation is mathematically easy to follow. We assume that the value of a firm depends (V) on three variables: how good the CEO is (T), how much labour is available to the firm (L), and how much capital the firm has available (K).[3]

[3] This is obviously a simplification and that is fine. All theories are simplifications and too much nuance in a theory can make a theory worthless (Healy 2017). We use (T)alent to describe how good the CEO is but you can also think of the CEO's ability or connections as part of T.

$$V = T^{\alpha_T}\left(\frac{K}{\alpha_K}\right)^{\alpha_K}\left(\frac{L}{\alpha_L}\right)^{\alpha_L} \tag{2.1}$$

$$\alpha_T + \alpha_K + \alpha_L = 1 \tag{2.2}$$

The condition (2.2) implies that there is nothing special about a specific size (V) of the firm. Bigger is not necessarily better, smaller is not necessarily better. This is a very traditional constant returns to scale assumption.

We also assume that the total compensation or wage (W) of the CEO is whatever value they create over the cost of labor (w_L) and the cost of capital (r).

$$W_T^* = \max_{K,L} V - w_L L - rK \tag{2.3}$$

With these equations in hand we can derive an empirical relation that we can test with data. The derivations are not very difficult, they are just a bit tedious.

First, let us find the optimal level of capital from the CEO's point of view. That is, the CEO will try to attract the amount of capital that will maximise their wage. Mathematically, that means that we need to find the level of K for which the first derivative of the CEO's wage to K equals 0.[4]

$$\begin{aligned}
0 = \frac{\partial W_T}{\partial K} &= T^{\alpha_T}\frac{\alpha_K}{\alpha_K}\left(\frac{K}{\alpha_K}\right)^{\alpha_K - 1}\left(\frac{L}{\alpha_L}\right)^{\alpha_L} - r \\
T^{\alpha_T}\left(\frac{K}{\alpha_K}\right)^{\alpha_K}\left(\frac{L}{\alpha_L}\right)^{\alpha_L}\frac{\alpha_K}{K} &= r \\
\frac{V\alpha_K}{K} &= r \\
\frac{V}{r} &= \frac{K}{\alpha_K}
\end{aligned}$$

[4] The $\frac{\alpha_K}{\alpha_K}$ is a bit weird and obviously it cancels out. It's the result of using the chain rule. The numerator is the result of taking the derivative of an exponential function. The denominator is the result of fact that we need to multiply the derivative of the exponential function with the derivative of $\frac{K}{\alpha_K}$. The whole reason why V was set-up this way was because it makes things cancel out; it's mathematically convenient.

Next, we can find the optimal level of labour that the CEO should attract.

$$\begin{aligned}
0 = \frac{\partial V}{\partial K} &= T^{\alpha_T}\left(\frac{K}{\alpha_K}\right)^{\alpha_K}\left(\frac{L}{\alpha_L}\right)^{\alpha_L - 1} - w_L \\
T^{\alpha_T}\left(\frac{K}{\alpha_K}\right)^{\alpha_K}\left(\frac{L}{\alpha_L}\right)^{\alpha_L}\frac{\alpha_L}{L} &= w_L \\
\frac{V}{w_L} &= \frac{L}{\alpha_L}
\end{aligned}$$

We can plug these results into equation (2.3) and we get

$$W_T^* = V - V\alpha_K - V\alpha_L = (1 - \alpha_K - \alpha_L)V = \alpha_T V \tag{2.4}$$

I like the basic intuition and deriviation of the model. The derivation is straightforward and (some of) the implicit assumptions are relatively easy to accept. The effect of the CEO depends on the size of the firm (V). When there is more capital and labour available a more talented CEO will have a bigger impact.[5] The model also predicts a clear quantitative relationship between firm size, V, and CEO compensation, W_T, i.e. that relationship should be linear. This is a nice result that we can test with data. In contrast we would not be able to test the relationship between CEO talent and compensation because talent is very difficult to measure. We can measure V but not T.

[5] If a CEO is a talented people manager, the advantage of this talent will be larger when there are more employees working in the company.

One of the assumptions in the model is that the CEO takes the ultimate decision and they have an incentive to maximise firm value because they keep the all the value after workers and investors have been paid. As it turns out when you assume competitive labour and financial markets, that assumptions does not really matter a lot for the predicted relationship between wage and firm size (Edmans and Gabaix 2016).[6] Nevertheless, This model is too simple to capture reality perfectly, but that is not the goal of the model and of this exercise. The idea is to see whether we can find a reasonable baseline for CEO compensation that we can test against the data. Here, we have established that there should be a positive relation between the company's size and the CEO's compensation and the relation is larger when the CEO is more talented.

[6] That is, you could set up the model so that investors take the final decision where they offer an incentive contract to the CEO who then chooses the number of workers that maximise the their compensation under the contract. The predictions for the model would be very similar to what we have here and we would have to deal with the contract as an extra complication. Good theory papers will explain when simplifications matter and when they do not.

2.3 *Empirical Test*

We will test the predictions of the model with data from S&P500 companies in the US. I made the data available for UWA students on LMS.[7] The data is downloaded from Compustat and Execucomp. A lot of you will use these are similar databases in your research project. I did not clean or check the data for this exercise. In your own project, you should show a better understanding of how the data are gathered and what they include than what I am displaying here.[8]

[7] I really want to find an open dataset where we can test the same model. In the Appendix of this chapter, you can find how you can download the same data straight from the WRDS database.

[8] The CEO compensation data is fairly complete. It includes changes in the value of equity and options. Market value also includes all outstanding financial instruments on the company.

The code below does multiple things in R. I do not expect you to understand or even replicate what I have done here. I include it as a reminder of what is possible with some coding. I hope it may serve as an inspiration further down the track when you are doing your own research project.

```
library(tidyverse)
library(cowplot)
# Read the data from a custom folder. If the data is not in this
# folder, change the path to the folder. Rename some of the
```

```
# variables to make them easier to work with.
folder = "data/"
us_comp = readRDS(paste0(folder, "us-compensation.RDS")) %>%
  rename(total_comp = tdc1)
us_value = readRDS(paste0(folder, "us-value.RDS")) %>%
  rename(year = fyear, market_value = mkvalt)
# Match the compensation data with the value data based on
# company key and year
us_comp_value = left_join(select(us_comp, gvkey, year, total_comp),
                          us_value, by = c("year", "gvkey"))
# Run the non-linear regression and save the results as an
# equation to put on the figure. This goes beyond what we are
# going to do in this module.
power_law_start = list(a= 5, b = .15)
power_law_comp =
  nls(total_comp/1000 ~ a * (market_value/1000)^b,
      data = us_comp_value, start = power_law_start)
eqn_comp = substitute(
  italic("compensation") == a * italic("MV") ^ b,
  list(a = as.numeric(format(coef(power_law_comp)[1],
                             digits = 2)),
       b = as.numeric(format(coef(power_law_comp)[2],
                             digits = 2))))
# create the plot
plot_comp_value =
# Add dataset and define x and y axis
  ggplot(us_comp_value, aes(x = market_value/1000,
                            y = total_comp/1000)) +
# Add the observations as points to the plot
  geom_point(alpha = .125) +
# Draw the results of the power_law regression on the plot
  stat_function(fun =
                function(x){
                  coef(power_law_comp)[1] *
                    x ^ coef(power_law_comp)[2]},
                colour = "blue") +
  ggtitle("CEO compensation (in million USD)") +
# Add label for the x-axis
  ylab("Compensation") +
  xlab("Company market value (in billion USD)")

print(plot_comp_value +
        annotate("text", y = 45, x = 500, parse = TRUE,
                 label = as.character(as.expression(eqn_comp))))
```

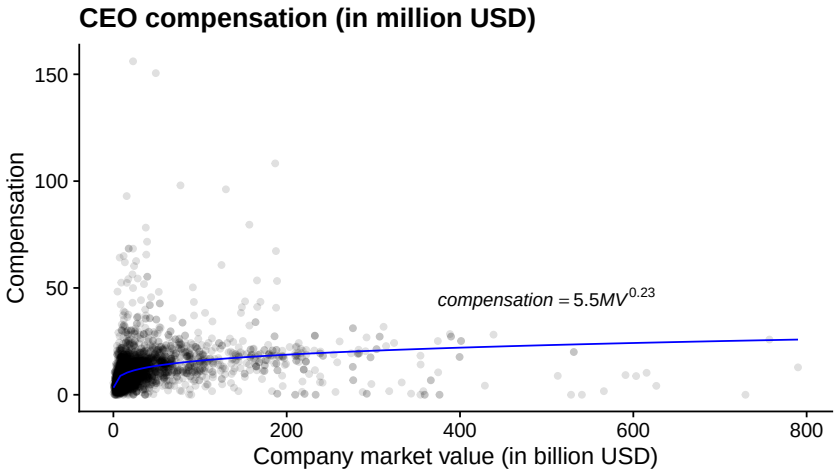


Figure 2.1: Relation between CEO compensation and market value for SP500 firms (2011-2018).

The qualitative relationship holds quite well. Larger companies have CEOs with higher compensation. However, the relationship is far from linear and looks more like a power function. Clearly there are other effects at play. In this sample, the power coefficient is 0.23.[9] Remember that in our setup the CEO can grow the firm at will by attracting more capital and more labour. That assumption is probably too strong. Later, we will see that there other theories that describe the relation between CEO compensation and firm size and they seem to describe the data better.

[9] Prior studies have found a coefficient more closely to 0.33 (Baker, Jensen, and Murphy 1988).

2.4 *Conclusion*

There are two things you should take away from this section. What we have just done here is a research project in a nutshell. First, we used theory to derive an empirical relation between two observable variables. Second, we tested the prediction against data. The evidence I present is not in the form of a complicated statistical analysis but in a graph. You will see in the remainder of this notes that I refer over and over to the connection between theory ('your story') and the data. The structure I used in this chapter (Introduction, Theory, Data Analysis, Conclusion) is also the structure of most research papers and your thesis.

An other important message is that I did not speak about a *causal* relation between market value and compensation. Our theory gave us a mathematical relation between the two variables and it is that relation we tried to find in the data. This is absolutely fine and I think a worthwile way to do research. However, this focus away from causal relation sometimes clashes with the popular statistical techniques and frame of mind that is presented in textbooks on causal inference

(Angrist and Pischke 2008; Cunningham 2018) and which permeates the literature in accounting and finance.[10]. I have seen students and established researchers alike get confused when their theory includes both equilibrium relations and a causal story. Throughout these notes I will try to explain the differences so that you can avoid the confusion when you start your own research project.

[10] This has been a source of frustration for me for awhile. I never could clearly articulate my unease. This podcast by physist Sean Caroll helped me crystalise my thoughts. He explains that from for a physicist on a fundamental level there are no causes. Fundamental particles follow equations and that is it. What we think of as causal relations, happens at a higher level of abstraction. If you have the time, give it a listen. The title of the podcast is appropriate: 'Mindscape'.

2.5 *Appendix: WRDS*

To get access to the Wharton Research Data Services, you first will have to obtain a username and password through your university's subscription. To get access to the databases, you have to install the `Rpostgres` package (Wickham, Ooms, and Müller 2019) and make a file where you can securely store your password. The WRDS website gives excellent guidelines on how to do that. You can then use the following code to establish a connection with the WRDS services.

```
library(RPostgres)
# use your own username for 'user'
wrds <- dbConnect(Postgres(),
                  host='wrds-pgdata.wharton.upenn.edu',
                  port=9737,
                  dbname='wrds',
                  user='stimas',
                  sslmode='require')
```

The following code choses the variables we are going to use from the ExecuComp and Compustat database, makes a connection with the database, gets the data and saves it as a `.RDS` file.

```
library(tidyverse)
sql1 = "select year, gvkey, cusip,
       exec_fullname, coname, ceoann,
       execid, bonus, salary,
       stock_awards_fv, stock_unvest_val,
       eip_unearn_num, eip_unearn_val,
       option_awards, option_awards_blk_value,
       option_awards_num, tdc1, tdc2,
       shrown_tot_pct,
       becameceo, joined_co, reason
       from EXECCOMP.ANNCOMP
       where year > 2010 and ceoann = 'CEO' and spcode = 'SP'"
res = dbSendQuery(wrds, sql1)
compensation = as_tibble(dbFetch(res)) %>%
```

```
## rename all variable to lowercase. It's a personal
## preference
  rename_all(tolower)
dbClearResult(res)
saveRDS(compensation,
        "data/us-compensation.RDS")

## unique keys in the compensation data
gvkeys = unique(compensation$gvkey)
cusips = unique(compensation$cusip)
cusips8 = cusips[str_length(cusips) == 8]
cusips7 = cusips[str_length(cusips) == 7]
cusips6 = cusips[str_length(cusips) == 6]

## Stock market value
sql2 = paste0("select fyear, gvkey, mkvalt, ni ",
              "from COMP.FUNDA ",
              "where fyear > 2010 ",
              "and gvkey in ('",
              paste(gvkeys, collapse = "','"),
              "')"
             )
res = dbSendQuery(wrds, sql2)
value = as_tibble(dbFetch(res)) %>%
  rename_all(tolower) %>%

## It turns out that there are different dataformats for
## some of the variables (but not market value) and there
## are multiple records per year-firm (one for each record).
  distinct()
dbClearResult(res)
saveRDS(value,
        "data/us-value.RDS")


## Unfortunately matching on CUSIP only works for 444
## companies with an 8 character CUSIP
sql3 = paste0("select begdat, cusip from crspa.dsfhdr ",
              "where substr(cusip, 1, 8) in ('",
              paste(cusips8, collapse = "','"),
              "')"
              )
res <- dbSendQuery(wrds, sql3)
company <- dbFetch(res)
```

```
dbClearResult(res)
saveRDS(company, "data/us-company.RDS")
```

3
Research Tools - Theory

In the previous chapter, I gave a small version of what a research project looks like. In this section, I go over some of the tools that I have been using in the past for my research. These are good starting points for your own research project.

3.1 Literature Search

In the CEO compensation project in the previous chapter, we derived the theoretical prediction from a model in the literature. This is typical for the research process. You will build on prior theoretical and empirical research to build an argument for your predictions. To do that you will have to search for literature on your topic. There are a number of sources you can use to access the relevant academic literature. However, in most cases (ssrn is the exception), you will have to be on the university's network if you want to actually read the full paper.

- Google Scholar is probably the most comprehensive repository. This search engine work very similar to regular Google search. There are some additional tricks you can use `author:lastname-firstname` will help you to narrow down papers from a specific author. `intitle:keyword` let's you search for keywords in the title of papers. You can also narrow down your search based on year of publication. The advanced search features hidden in the left side bar give you additional options such as searching for certain journals. If you are on the university network, Google Scholar will tell you for every paper whether it is accessible or not.

- Onesearch is the university search engine. It's the best way to figure out whether there is an easily accessible version of the paper even when you are not on the university's network.

- Webofknowledge and EBSCOhost are two publisher driven search engines. They work pretty well. Each with their own quirks.
- SSRN (Social Science Research Network) and NBER (National Bureau of Economic Research both provide access to their own not-yet-peer-reviewed paper repositories. Here you go to find cutting edge research.

My favourite way to start a research project now is to find one or two excellent overview or review papers. A (systematic) review paper provides a state of a research field and identifies interesting new research questions. I find that a good review paper gives a good list of papers you can build on and they often already compare the most important papers in a field. The trick is to be not too picky. You will not find a review for your exact research problem but it is unlikely that you will not find a partly relevant overview paper.[1] You can search for review papers by adding `intitle:review` or `intitle:overview` to your Google Scholar search.

To find other papers relevant to your topic, you can build on the review paper by (1) looking up the papers referred to in the review paper and (2) search for papers that cite the review paper. You can do the latter via Google Scholar and Webofknowledge.

To find good reviews, I believe you should start your search in the better journals.[2]. There are even some journals that are dedicated to these literature reviews for instance *Journal of Economic Literature* and *Journal of Accounting Literature*.[3] I am not aware of a similar journal in finance but I will happily add it if you let me know.

[1] When you start your literature search, you do not want to start too narrow. You are not going to find an overview paper about "CEO compensation in Australian mining companies after the GFC". However, you can start with an overview paper about CEO compensation. Like the one I found: "Executive Compensation: A Modern Primer" by Edmans and Gabaix (2016) in *Journal of Economic Literature*.

[2] These links form a good starting point for Accounting and Finance. Unfortunately, you have to register your email address first.

[3] I also like *Psychological Bulletin* for theories in psychology.

3.2 *Writing (Under Construction)*

There are a couple of writing mistakes that are very common for first time researchers.[4] Below I highlight some of the problems I see quite often. Before I tackle those issuess it is useful to make a distinction between the different levels in a document. I will give you a quick overview so that you understand my other advice.

[4] There is a lot of good writing advice available. Dan Simons' collection is very pragmatic and is close to my sensibilities.

3.2.1 Different levels in a research paper or thesis.

By now, you probably know that there is a general structure in an empirical research paper: (1) Introduction, (2) Theory, (3) Method, (4) Results, (5) Conclusion. This is the highest structural level in a thesis.

The second level is within each of these sections. For instance, your introduction will consist of multiple paragraphs. The third level is at the level of the paragraph which consists of multiple sentences. The fourth level is within those sentence which contain multiple

words. The most common comment on writing is possibly that there are grammatical mistakes or typos in your writing. This is an issue at the lowest level, i.e. the level of the words. This is also the least interesting level. Most of my feedback will focus on the other issues.

You want to keep two rules in mind at each of the levels. (1) You should aim to keep each unit (section, paragraph, sentences) coherent.[5] (2) You should aim to give the reader the information that they need in the order that they need it. When you start using a new concept or theory, you should give the reader an introduction to that concept or theory, unless you can assume they already know the concept or theory.

[5] For instance, a paragraph should only contain one main message.

3.2.2 *Explain important concepts and theories.*

Introduce the key concepts and theories in your introduction. If you write a paper about market efficiency, you should give the reader at least an idea of what you mean by market efficiency. This should come at the start of your paper otherwise the reader will not know whether the literature review is relevant or not. An introduction of a concept should include a simplified description of what it is not only what it's effects are. If you look at the introduction of new legislation, tell the reader what the legislation requires of people or businesses. If you introduce a theory, explain it's important assumptions.

If your thesis focuses on a practical concept, you should still attempt to define your concept in theoretical terms. This definition will help you to make the connection with the relevant theories and with the mathematical specification in your statistical model.

3.2.3 *Theory before literature review.*

To be able to *do* a good research project, you should read a lot of literature first and then come up with your research question. However, when you *write* your research project up, you should explain your theory first and then present the literature.[6] The theory will provide structure of your literature review: which studies should be discussed together, which ones should be contrasted with each other, and which ones are relevant for your setting. If you do not start with a theoretical motivation, you run the risk that your literature review comes across (or is) just a series of empirical results without any connection.

[6] This one is a personal opinion and you should probably check with your supervisor.

You recognise these literature reviews by the following conclusion: some papers find a positive effect, some papers find a negative effect, we are testing it again in a different sample. This is not necessarily a good approach because no matter what the result of your study is, the next researcher can just do the same thing: some papers find

a positive effect, some papers find a negative effect, we are testing it again in a different sample. Nobody learns anything from these literature reviews and papers.

Sometimes I see students split up the literature review and the hypothesis section which artificially forces you to either postpone your theoretical arguments or to repeat them almost verbatim. You should integrate hypothesis and literature review. Don't split them artificially.

3.2.4 Delete unnecessary words.

One common mistake is to write too many words. This is an issue at the level of the sentence. The trick is to reread your sentences and see whether the meaning of a sentence changes if you drop certain words. You do not want to use more words than necessary because it will only make it harder for the reader to understand your main message.

3.2.5 Avoid vague or hedging words as much as possible.

When we explain a theoretical argument, we all want to express that we are not certain that this effect will be true. You maybe have the tendency to write " X may increase Y" to reflect that uncertainty. However, there are better ways to do this. You can write "According to theory, X increases Y" or "Under this assumption, X increases Y". Avoid words like "can", "may", "might" and try to be more precise under which circumstances your prediction or argument holds. This will help you to assess whether your theory fits your setting. Being precise will also help you to avoid the trap of writing a noncommital literature review

4
Research tools - Data Analysis

4.1 Data analysis tools

Our main tool for data analysis is the R statistical software (R Core Team 2019). R is free software specifically aimed at statistical analysis and graphical representation. It is a fully developed programming language that allows to extend it with new packages with new capabilities. We are going to heavily rely on some of those packages.[1]

We are going to interact with the R software through Rstudio.[2] Rstudio is an integrated development environment for R. It allows you to easily write, test, and run R code, and integrate the results with explanations. One advantage of R and Rstudio is that you can install them on as many computers as you want. You do not need a special license. This means that you do not have to be affiliated with the university to use the sofware.

In Rstudio, you write code and ask the R software to execute that code. You can save the different steps in your analysis in an R script which is just a plain text file with the `.R`-extenstion. The advantage of plain text file is that they are easy to share across different operating systems. The code in your script allows you and anyone else to run the same analysis at a later point in time which is very important to check for mistakes and for teaching.

We will make use of a special type of plain text files: Rmarkdown (`.Rmd`) files.[3]. RMarkdown is an R extension of the markdown format. The idea of markdown is to write plain text documents which can later be exported to other formats such as html, pdf, or word. The advantage of plain text scripts and markdown. The beauty of RMarkdown is that it lets you combine both R scripts and markdown into one document. This allows you to integrate your analysis and your description of your analysis in one document. When you go back to your analysis after a couple of weeks, you will be happy that you have more than the raw code to look at.

Finally, I am going to introduce you to a specific dialect of the R

[1] You can download R from this CRAN server

[2] You can download from the Rstudio website

[3] These notes are fully written in Rmarkdown files

language which is informally known as the tidyverse. If you want to be a good R programmer this might not be the best entry point into R. However, I assume that you want to quickly pick up some tools to facilitate your research project and for that purpose I believe the tidyverse will be excellent. I will introduce the most important bits and pieces throughout my lectures however I don't have the time to go into detail. Now and then, you will have to experiment on your own to get your code working. Some excellent resources are the free online books R for datascience and Data Visualization: A practical introduction. You can also buy reasonably priced physical copies if you your are interested in developing your R skills further.

For larger projects, you should use a project in its own separate folder.[4]. Once you need multiple scripts, you want to make sure that you can reliably point to the results or functions in different scripts. Projects help you manage all the parts of a larger piece of work. You can start a

[4] See the Project Chapter of Hadley Wickham's book

You start a new project by clicking `File > New Project ...` You can start a new file by clicking `File > New File > R Script`.

4.2 The R console

In Rstudio, you can use the R console directly. I mainly use the R console to test the code I have written in my scripts and I would advice you to do the same thing. If you want to understand how R works. You can experiment by typing the following lines in the console.[5]

[5] It is tempting to copy and paste the code but retyping the code is a surprising powerful way of learning to code. I would advise you to try to type as much of the code as possible when you are trying to figure out how R works.

```
x <- 1
x
```

```
## [1] 1
```

```
x <- x + 1
x
```

```
## [1] 2
```

The little code above already shows you two things.

1. You can assign (and overwrite) a numerical value to an object `x`
2. The right-hand side is assigned to the left-hand side. You can't just switch them around.

We will see further that we can assign almost anything to an object `x`. When `x` is a vector or a data set with multiple elements, this will allow us to perform the same function on each element of `x`. That

is the basic advantage of programming. You can tell your computer how to do a thing and than the computer can do it over and over for you.

4.3 R packages

```
install.packages("tidyverse")
library("tidyverse")
```

R packages are additions to R that give R extra functionality. One of the selling points of R is that it has a good way of integrating this extra functionality and a lot of people are highly motivated to add this extra functionality. We will heavily rely on one meta-packages. `tidyverse` is a package that helps with data transformations and working with tabular data in a tidy fashion. One of the packages in the `tidyverse` is `ggplot2` which provides tools to make pretty plots. The code above shows you how to install packages and use them. Normally, you will only have to install a package once. However, when you want to use a package in your script or code, you will have to load the package with the `library()` function. You only have to do load a package once at the start of your R session or at the start of your script.[6]

`dplyr` is another part of the `tidyverse` we will use extensively. With 5 verbs[7] and one pipe operator to glue them together helps you to explore the data.`filter()` let's you filter out a subset of the observations. `select()` selects a subset of the variables. `mutate()` changes variables and creates new ones. `group_by()` and `summarise()` group subsets of observations and summarise them. `%>%` is the pipe operator and joins together the verbs to create compound statements where you for instance filter a subset of the observations, create a new variable, create groups, and summarise the groups based on the average for the new variable.

[6] Installing packages is another reason to use the R console.

[7] i.e. functions to *do* something with a dataset.

4.4 Finally some useful code!

Let's run some code. First, we have to tell R to use the packages in the `tidyverse`.

```
library(tidyverse)
```

Next, we have a look at the data I used to draw the plot of the compensation of CEOs of S&P500 companies. First, we have to assign the file with the data to an object, `us_comp` in this case.[8] The glimpse function gives an overview of the variables in the data, the type of the variables, and a couple of examples of the values for that variable.

[8] The data is available on LMS or through the code in the Appendix 2.5

```
us_comp <- readRDS("data/us-compensation.RDS")
glimpse(us_comp)
```

```
## Observations: 3,458
## Variables: 22
## $ year                   <dbl> 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2...
## $ gvkey                  <chr> "001045", "001045", "001045", "001045", "00...
## $ cusip                  <chr> "02376R10", "02376R10", "02376R10", "02376R...
## $ exec_fullname          <chr> "Gerard J. Arpey", "Thomas W. Horton", "Tho...
## $ coname                 <chr> "AMERICAN AIRLINES GROUP INC", "AMERICAN AI...
## $ ceoann                 <chr> "CEO", "CEO", "CEO", "CEO", "CEO", "CEO", "...
## $ execid                 <chr> "14591", "26059", "26059", "46191", "46191"...
## $ bonus                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ salary                 <dbl> 613.842, 618.135, 591.911, 687.884, 231.538...
## $ stock_awards_fv        <dbl> 4020.380, 0.000, 4199.761, 7000.000, 10330....
## $ stock_unvest_val       <dbl> 0.000, 309.656, 0.000, 31258.138, 14653.946...
## $ eip_unearn_num         <dbl> 0.000, 151.900, 0.000, 410.111, 112.143, 14...
## $ eip_unearn_val         <dbl> 0.000, 121.520, 0.000, 21994.253, 4749.256,...
## $ option_awards          <dbl> 798.600, 0.000, 0.000, 0.000, 0.000, 0.000,...
## $ option_awards_blk_value <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ option_awards_num      <dbl> 220.000, 0.000, 0.000, 0.000, 0.000, 0.000,...
## $ tdc1                   <dbl> 5526.835, 656.226, 19092.978, 12301.976, 11...
## $ tdc2                   <dbl> 4728.235, 656.226, 21123.739, 18975.528, 40...
## $ shrown_tot_pct         <dbl> NA, NA, 0.061, 0.240, 0.224, 0.306, 0.404, ...
## $ becameceo              <date> 2003-04-25, 2011-11-28, 2011-11-28, 2013-1...
## $ joined_co              <date> NA, NA, NA, NA, NA, NA, NA, 2002-12-01, 20...
## $ reason                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

We can also look at a subset of the observations, i.e. we filter the observations from the company that has the key `001045`.

```
filter(us_comp, gvkey == "001045")
```

```
## # A tibble: 7 x 22
##    year gvkey cusip exec_fullname coname ceoann execid bonus salary
##   <dbl> <chr> <chr> <chr>         <chr>  <chr>  <chr>  <dbl>  <dbl>
## 1  2011 0010~ 0237~ Gerard J. Ar~ AMERI~ CEO    14591      0   614.
## 2  2012 0010~ 0237~ Thomas W. Ho~ AMERI~ CEO    26059      0   618.
## 3  2013 0010~ 0237~ Thomas W. Ho~ AMERI~ CEO    26059      0   592.
## 4  2014 0010~ 0237~ William Doug~ AMERI~ CEO    46191      0   688.
## 5  2015 0010~ 0237~ William Doug~ AMERI~ CEO    46191      0   232.
## 6  2016 0010~ 0237~ William Doug~ AMERI~ CEO    46191      0     0
## 7  2017 0010~ 0237~ William Doug~ AMERI~ CEO    46191      0     0
## # ... with 13 more variables: stock_awards_fv <dbl>, stock_unvest_val <dbl>,
## #   eip_unearn_num <dbl>, eip_unearn_val <dbl>, option_awards <dbl>,
## #   option_awards_blk_value <dbl>, option_awards_num <dbl>, tdc1 <dbl>,
```

```
## #   tdc2 <dbl>, shrown_tot_pct <dbl>, becameceo <date>, joined_co <date>,
## #   reason <chr>
```

You can see that the company, American Airlines, has had three CEOs since 2011. I used the `gvkey` to select a company and not its name. You will often use an id or a database key especially if you are working with multiple datasets and you want to link observations from one dataset to observations in the other dataset.

We can also filter data based on other variables. For instance, the below filters the observations from 2014 where the CEO had a salary over $1,000,000.

```
filter(us_comp, salary > 1000, year == 2014)
```

```
## # A tibble: 271 x 22
##     year gvkey cusip exec_fullname coname ceoann execid bonus salary
##    <dbl> <chr> <chr> <chr>         <chr>  <chr>  <chr>  <dbl>  <dbl>
##  1  2014 0010~ 7234~ Donald E. Br~ PINNA~ CEO    05835      0   1240
##  2  2014 0010~ 2824~ Miles D. Whi~ ABBOT~ CEO    14300      0   1973.
##  3  2014 0011~ 7903~ Rory P. Read  ADVAN~ CEO    42390      0   1000.
##  4  2014 0013~ 4385~ David M. Cote HONEY~ CEO    20931   5500   1866.
##  5  2014 0013~ 4280~ John B. Hess  HESS ~ CEO    02132      0   1500
##  6  2014 0014~ 2553~ Nicholas K. ~ AMERI~ CEO    40778      0   1241.
##  7  2014 0014~ 2581~ Kenneth I. C~ AMERI~ CEO    02157   4500   2000
##  8  2014 0014~ 1055~ Daniel Paul ~ AFLAC~ CEO    00013      0   1441.
##  9  2014 0014~ 2687~ Robert Herma~ AMERI~ CEO    20972      0   1385.
## 10  2014 0015~ 3110~ Frank S. Her~ AMETE~ CEO    04252    520.  1182.
## # ... with 261 more rows, and 13 more variables: stock_awards_fv <dbl>,
## #   stock_unvest_val <dbl>, eip_unearn_num <dbl>, eip_unearn_val <dbl>,
## #   option_awards <dbl>, option_awards_blk_value <dbl>,
## #   option_awards_num <dbl>, tdc1 <dbl>, tdc2 <dbl>, shrown_tot_pct <dbl>,
## #   becameceo <date>, joined_co <date>, reason <chr>
```

You can also select some variables if you are only interested in those. If a variable has an undescriptive name, you can use select to rename the variable. For instance, `tdc1` is the total compensation of a the CEO. A more descriptive name will help you to remember what the variable actually means.

```
select(us_comp, year, coname, bonus, salary, total = tdc1)
```

```
## # A tibble: 3,458 x 5
##     year coname                     bonus salary  total
##    <dbl> <chr>                      <dbl>  <dbl>  <dbl>
##  1  2011 AMERICAN AIRLINES GROUP INC     0   614.  5527.
##  2  2012 AMERICAN AIRLINES GROUP INC     0   618.   656.
```

```
##  3  2013 AMERICAN AIRLINES GROUP INC     0   592. 19093.
##  4  2014 AMERICAN AIRLINES GROUP INC     0   688. 12302.
##  5  2015 AMERICAN AIRLINES GROUP INC     0   232. 11419.
##  6  2016 AMERICAN AIRLINES GROUP INC     0     0    141.
##  7  2017 AMERICAN AIRLINES GROUP INC     0     0  18115.
##  8  2011 PINNACLE WEST CAPITAL CORP      0  1091   6698.
##  9  2012 PINNACLE WEST CAPITAL CORP      0  1146  10123.
## 10  2013 PINNACLE WEST CAPITAL CORP      0  1203.  7124.
## # ... with 3,448 more rows
```

You can also create new variables with the `mutate` function. I created a variable that calculates what percentage of total compensation is the CEO's salary.

```
select(us_comp, year, coname, bonus, salary, total = tdc1) %>%
  mutate(salary_percentage = salary / total)
```

```
## # A tibble: 3,458 x 6
##     year coname                      bonus salary  total salary_percentage
##    <dbl> <chr>                       <dbl>  <dbl>  <dbl>             <dbl>
##  1  2011 AMERICAN AIRLINES GROUP INC     0   614.  5527.            0.111
##  2  2012 AMERICAN AIRLINES GROUP INC     0   618.   656.            0.942
##  3  2013 AMERICAN AIRLINES GROUP INC     0   592. 19093.            0.0310
##  4  2014 AMERICAN AIRLINES GROUP INC     0   688. 12302.            0.0559
##  5  2015 AMERICAN AIRLINES GROUP INC     0   232. 11419.            0.0203
##  6  2016 AMERICAN AIRLINES GROUP INC     0     0    141.            0
##  7  2017 AMERICAN AIRLINES GROUP INC     0     0  18115.            0
##  8  2011 PINNACLE WEST CAPITAL CORP      0  1091   6698.            0.163
##  9  2012 PINNACLE WEST CAPITAL CORP      0  1146  10123.            0.113
## 10  2013 PINNACLE WEST CAPITAL CORP      0  1203.  7124.            0.169
## # ... with 3,448 more rows
```

Remark how the `select` statement from before is chained together with the `mutate` statement through the pipe operator (`%>%`). This operator pipes the results from the first statemetn to the second statement. The second statement implicitly uses the result from the first statement as the data it is going to work with.

You can use pipe statement to make your own descriptive statistics table. I created a table with some statistics for each firm. I first group the observations based on the `gvkey` and then summarise each group with the same key by defining a number of new variables.

```
group_by(us_comp, gvkey) %>%
  summarise(N = n(), N_CEO = n_distinct(execid),
            average = mean(salary), sd = sd(salary),
            med = median(salary), minimum = min(salary),
```

```
            maximum = max(salary)) %>%
  ungroup()
```

```
## # A tibble: 500 x 8
##    gvkey      N N_CEO average    sd   med minimum maximum
##    <chr>  <int> <int>   <dbl> <dbl> <dbl>   <dbl>   <dbl>
##  1 001045     7     3    392. 306.   592.      0     688.
##  2 001075     7     1   1232.  93.3 1240    1091    1355
##  3 001078     7     1   1910.  27.6 1900    1900    1973.
##  4 001161     7     3    890. 155.   925.    566.   1000.
##  5 001209     7     2   1158. 112.  1200     905.   1200
##  6 001230     7     2    453.  44.5  437.    410.    535.
##  7 001300     7     2   1780. 166.  1800    1415.   1890
##  8 001327     7     2    738.  83.3  748.    635.    850
##  9 001380     7     1   1500    0   1500    1500    1500
## 10 001440     7     2   1233. 156.  1280.    903.   1375
## # ... with 490 more rows
```

4.5 Looking for help

Programming is hard work. You will make mistakes. You will get error messages. An important programming skill is to efficiently debug your code and find out what is going wrong.

- The most important technique is trial and error. Change one thing in your code and see what the output is. Do you get an error message? What does the error message tell you? If you are carefull and do not change everything at once, this should at least help you to find out which part of the code does not work. This is one reason why scripts are so important. Because a script contains your entire analyis, you can always go back and (let your computer) redo the analysis from the start.
- If you are not sure how you should use a certain function in R, you can read its help files in the Rstudio *Help* window. You can search for more information what a function is doing.
- R is a very popular language. If you have a problem it is likely that someone else had the same problem before you. If you Google the error message, there is a decent chance you will find a solution to your problem.
- A specific website where a lot of these questions are asked is StackOverflow. You can often directly search for your error on the website and find multiple solutions. You can improve your chances of finding a related answer to your problem by adding the `[r]`, `[tidyverse]`, `[dplyr]`, `[ggplot]` tags to your error message.

- You can also ask me directly for help on LMS. I set up a forum for questions and I promise to help you with any R questions you might have for the unit and for your thesis.

4.6 RMarkdown = Markdown + R

One of the big advantages of the R world is that you can easily combine explanations with your analysis in `.Rmd` files. The assignements can all be completed in Rmarkdown. You can find a lot of good resources on Rmarkdown on the Rstudio website.

In short, markdown is a simple markup language[9] that lets you include R code.

[9] This means that you can add little bits of code to your text that indicate how the text should look

4.6.1 Markdown

You can have titles in markdown.

```
# Title
## Subtitle
### Lower level titles.
```

You can emphasise some words.
Italic, *****bold*****
Add links to pages and include pictures.

```
[weblink](https://www.google.com)
![pictures](https://rstudio.com/wp-content/uploads/2015/10/r-packages.png)
```

You can write enumerations and lists.

```
1. Item 1
2. Item 2
3. Item 3

- one
- two
- three
```

You can also write tables. You will rarely have to use tables. Typically, you can directly create the tables from R without the need to type in the results of your analysis.

```
First Header  | Second Header
------------- | -------------
Content Cell  | Content Cell
Content Cell  | Content Cell
```

4.6.2 R-code in markdown: R chunks

The largest advantage of Rmarkdown files is that we can include pieces of R code.[^ Again, all these notes are written in Rmarkdown. Code chuncks go between three backticks and we tell Rmarkdown that the language we are using is `{r}`. The example below creates a code chuck where we create a random vector `x` with then elements where the elements are drawn from a random distribution with mean 4 and standard deviation 3.

````
```{r}
x <- rnorm(n = 10, mean = 4, sd = 3)
```
````

5
Theory: Maths and Simulations

5.1 Introduction

A good research project relies on strong theoretical foundation. Theories are a summary of prior research findings and present predictions you can test. In accounting and finance, a lot of theories are mathematical theories. Sometimes, research articles will give a good explanation of the arguments in a theory but sometimes you will have to put in some extra effort to understand the theory. In this section, I will give you two techniques that can help to understand the theory better. The first is making the theory simpler and less general and redo the derivations. The second is to simulate data based on the theory and to visualise the theory with plots. Computers are very good at doing calculations. Whenever possible, you should let computers do the work for you. Simulate and visualise is a technique, we will use a lot more in the rest of the notes.

5.2 New theory: CEO-firm matching.

Let us introduce a new theory how the size of the company is related to the compensation of the CEO. In Chapter 2 I presented a basic model where talented CEOs hire more people and attract more capital and thus grow the company. This theory ignored that companies and CEOs can choose to work with each other. In this section, we introduce a new theory about matching firms and CEOs (Edmans and Gabaix 2016; Tervio 2008).

The theory assumes that the increase in *V*alue of a firm from time 0 to time 1 is given by the following equation.

$$V_1(n) - V_0(n) = CV_0(n)T(m) \tag{5.1}$$

The increase in value depends on the *T*alent of the CEO, the inital *V*alue of the firm, and a scaling factor C.[1] n is the rank of the size

[1] The scaling factor is not crucial to the theory. If performance is measured in millions of dollars or in billions of dollars, the scaling factor will differ but that is not telling us anything about the economic mechanism.

of the firm and m is the rank of the talent of the CEO. $n = 1$ is the largest firm, $m = 1$ is the most talented CEO. The question the theory is trying to answer which CEO, m, is going to work for firm, n.

The theory assumes that firms will make a decision about which CEO to hire and CEOs will only accept to work for a firm if they cannot do better in another firm. The firms have to compensate a manager with a $w$age and will maximise the residual value of the firm, $V_1(n) - V_0(n) - w(m)$. The model assumes that the managers will need a compensation above w_0.

Tervio (2008) and Edmans and Gabaix (2016) show that when $m = n$, no firm or CEO can improve themselves by switching. This means that the most talented CEO works for the largest firm, the second most talented CEO works for the second largest firm, until we reach the least talented CEO and the smallest firm. The intuition is that CEOs have a larger impact in larger firms.[2] Therefor, most value is created when the largest firms are managed by the best CEOs. The difficulty is to determine how much each firm should pay the CEO. I first go over a simplified mathematical model that gets some of the intuition across, then I explain how you can simulate from the more complicated model

[2] If a CEO is good at managing people, the impact of the CEO will be better if they are managing more people.

5.3 *Three firms - three CEOs model*

Let us assume that there are only three CEOs and only three firms. In equilibrium, we want to make sure that the largest firm ($n = 1$) cannot do better than hiring the most talented CEO ($m = 1$). In other words, the performance of the most talented CEO after paying their compensation, should be higher than the residual performance of the other two CEOs.

$$\begin{aligned} CV_0(1)T(1) - w(1) &\geq CV_0(1)T(2) - w(2) \\ CV_0(1)T(1) - w(1) &\geq CV_0(1)T(3) - w(3) \end{aligned} \tag{5.2}$$

Next, the second largest firm has to be better off hiring the the second best CEO.

$$CV_0(2)T(2) - w(2) \geq CV_0(2)T(3) - w(3) \tag{5.3}$$

If we add the first condition of (5.2) to condition (5.3). We can rewrite everything and get the second condition of (5.3).

$$\begin{aligned} CV_0(1)T(1) - w(1) + CV_0(2)T(2) - w(2) &\geq CV_0(1)T(2) - w(2) + CV_0(2)T(3) - w(3) \\ CV_0(1)T(1) - w(1) &\geq C(V_0(1) - V_0(2))T(2) + C(V_0(2) - V_0(1))T(3) + CV_0(1)T(3) - w(3) \\ CV_0(1)T(1) - w(1) &\geq C(V_0(1) - V_0(2))(T(2) - T(3)) + CV_0(1)T(3) - w(3) \end{aligned}$$

Because $V_0(1) > V_0(2)$ and $T_0(2) > T_0(3)$, we can delete the first term on the right hand side. So, the two inequalities give us the third inequality from the previous slide for free.

Because firms will prefer to paying a lower compensation than more compensation, firms pay their CEO just enough so that smaller firms are not willing to pay the same amount of money to poach the CEO away. For each firm, we have to make sure that the advantage of having a more talented CEO is smaller than the extra wage of hiring the CEO.

$$\begin{aligned} CV_0(2)(T(1) - T(2)) &\geq w(1) - w(2) \\ CV_0(3)(T(2) - T(3)) &\geq w(2) - w(3) \end{aligned}$$

Because firm $n = 1$ and 2 will set the wage just high enough to make sure that a smaller firm does not poach their CEO, they will set the compensation $w(1)$ and $w(2)$ just high enough but not higher. We can simplify the resulting conditions than to equalities.

$$\begin{aligned} CV_0(2)(T(1) - T(2)) + w(2) &= w(1) \\ CV_0(3)(T(2) - T(3)) + w(3) &= w(2) \end{aligned} \tag{5.4}$$

The deriviations are a bit tedious and I do not necessarily want you to be able to do this entirely on your own. However, it shows that with some small calculations and with some economic intuition about what we want to calculate, we can again derive the relation between firm value and CEO compensation.

5.4 *N firms - N CEOs model*

If the literature points you to a mathematical model and you want to understand it better. Breaking it down to a simpler model with only 2 or 3 firms can be very illuminating. I hope the three firm model helped you to get some of the intuition behind the model without resorting to too complicated maths. It is relatively easy to see that we can write the inequalities in (5.4) in a more general form.

$$\begin{aligned} CV_0(n+1)(T(n) - T(n+1)) + w(n+1) &= w(n) \\ \forall n &= 1, .., N-1 \end{aligned} \tag{5.5}$$

The basic idea is that a larger firm will pay more for a CEO than a smaller firm. If the larger firm wants to make sure that a the more talented CEO works for them, they need to pay the CEO a high enough wage. The difference between the two wages for a talented CEO will be the surplus that the CEO would create in the smaller

firm. So the smaller firm will not be willing to poach the more talented CEO because the costs (higher wage) would outweight the benefit (higher surplus).

The original papers go further with the derivations (Edmans and Gabaix 2016; Tervio 2008). However, this is an algorithm we can give to R when we add some further assumptions. So instead of going over all the math, we are going to simulate wages and firm values based on the algorithm.

The original theoretical papers also need the extra assumptions. The goal of the simulation is to show that sometimes you do not need all the fancy maths when you can write a computer program to do the work for you. We will use similar assumptions as the original papers but implement them in an R program.

5.5 *Simulation*

First, let us load the `tidyverse` package.

```
library(tidyverse)
```

Next, we simulate data for `obs=500` observations. `size_rate` is a parameter that controls the size of firms. A value of 1 means that firms have constant returns to size, the same assumption as in Chapter 2. `talent_rate` is something similar for the Talent of the CEOs. A larger number for both rate parameters implies that differences between sizes or CEOs become larger at the top. `C` is the C constant in the model. `scale` is an additional parameter that helps me scale the size of the firms so that I get similar numbers as the real data.[3] `w0` is the base wage for the least talented CEO.

[3] This is not necessarily a fudge. The theory does not say anything about whether we should measure firm size in USD, in AUD or in CNY. So the scale we use is arbitrary.

```
obs <- 500
size_rate <- 1; talent_rate <- 2/3;
C <- 10; scale <- 600; w0 <- 0;
n <- c(1:obs)
size <- scale * n ^ (-size_rate)
talent <- - 1/talent_rate * n ^ (talent_rate)
```

`n` is an R vector of length `obs` (i.e. 500) with values from 1 to `obs`. So it is the rank in size and talent for each firm and CEO. You can see what `n` look like by just typing `n` and enter in the R console.

Size and talent follow an exponential distribution which has some theoretical motivation given in Tervio (2008) and Edmans and Gabaix (2016). If that interests you, please go have a look but it goes beyond what we need today. What we do is give a value for the size of each firm from 1 to 500 and for the talent of each CEO from 1 to 500.

We can also calculate the wage of each CEO-firm combination from equation (5.5). We start by creating a wage vector with 500 `NA` values.[4] At the last (`obs = 500`) position of the vector, we set the wage equal to `w0` for the least talented CEO and the smallest firm. Than *for* each firm (we go from `i = 499` to `i = 1`), we set the wage as the wage of the smaller firm (`i + 1`) and subtract the surplus our CEO would generate in the smaller firm over what their CEO is now generating.[5]

[4] The `rep` function creates a vector of `obs` *rep*itions of `NA`

[5] Again `scale` and `C` just scale some of the values so that they are easier to interpret. These parameters are less important for the economic intuition.

```
wage <- rep(NA, obs)
wage[obs] <- w0
for (i in (obs - 1):1){
  wage[i] <- wage[i + 1] + 1/scale * C * size[i + 1] *
    (talent[i] - talent[i + 1])
}
```

Now we can put all our variables in a dataset. The `tidyverse` calls datasets `tibbles` and they are the main object that `tidyverse` functions work on.

```
simulated_data <- tibble(
  n = n,
  size = size,
  talent = talent,
  wage = wage
)
```

5.6 *Visualisations*

With the data we simulated, we can visualise the theory and see whether our theory matches our intuition. Visulising a theory is one way to understand the assumptions and to check whether it matches the data. Even if you do not have your data yet, it will give you an idea of what the data should look like.

Figure 5.1 shows the relation between the simulated CEO wage and simulated firm size. It's not perfect but the plot does follow a similar non-linear pattern as what we found in Chapter 2 (see Figure 5.2).[6]

[6] `qplot` is a funtion to make *q*uick plots and it is part of the `ggplot` package which is part of the `tidyverse`.

```
qplot(data = simulated_data, y = wage, x = size)
```

```
print(plot_comp_value)
```

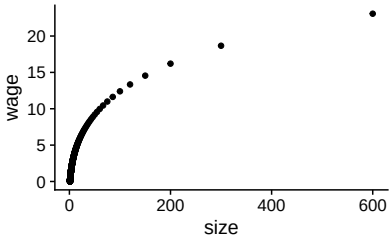
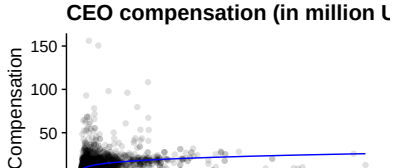


Figure 5.1: Relation between simulated CEO wage and firm size.

To better understand the assumptions in the theory, we can also plot how talent and size change as a function of the rank of respectively the CEO and the firm. I glossed over the details before but that



does not mean that we cannot check whether those functions make sense.

Figure 5.3 shows us which assumptions were necessary for the theory to work. We see that the difference in talent at the top of the distribution (`n = 1`) is not that large, the difference in firm sizes is much more pronounced and is driving the difference in wages according to this theory.

```
p_talent <- qplot(data = simulated_data, y = talent, x = n)
p_size <- qplot(data = simulated_data, y = size, x = n)
cowplot::plot_grid(p_talent, p_size, ncol = 1, labels = "AUTO")
```

In the code, I use a function from the `cowplot` package to plot different plots (`p_talent` and `p_size`) in 1 column. The automatic labels will add the A and B labels for the plots.

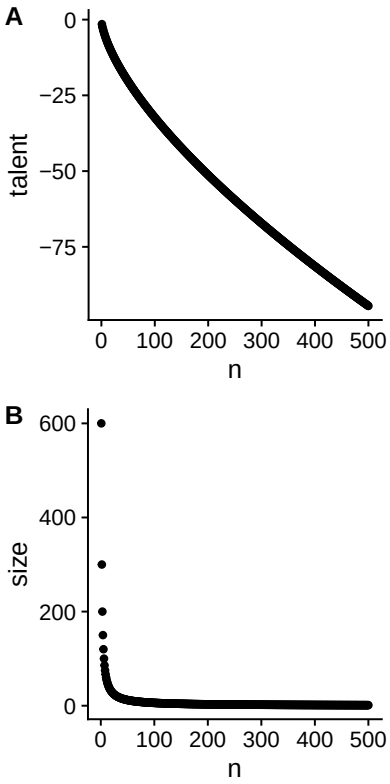


Figure 5.3: Relation between talent and rank of CEO (A) and between size and rank of the firm (B).

5.7 *Functions in R*

One of the most valuable aspects of `R` is that you can write new functions. Functions allow you to create your own verbs to apply to objects. In the previous section, we simulated data with a number of parameters in the theory set at a fixed value. If we want to compare the sensitivity of the theory to changes in the parameters, we want to simulate new datasets with different parameter values. A function to simulate data is what we need.

Functions are created with the `function` function. In between brackets, you define the parameter you want to use in your functions and their default values. In between the curly braces `{}` you tell `R` what it should do with those parameters. Ideally, you should not rely on any parameter or data that is not defined in your function. `R` has some liberal defaults which might give you unexpected results if you do that. You can rely on external functions like I do to create the `tibble`.

I do nothing in the function that I have not done before. The only addition is that at the end, the function `returns` the simulated data. You can then use the function to create a new simulated dataset.[7]

```
create_fake_data <- function(obs = 500, size_rate = 1,
                             talent_rate = 1,
                             w0 = .001, C = 1){
  scale = 600
  n = 1:obs
  size = scale * n ^ (-size_rate)
  talent =  -1/talent_rate * n ^ talent_rate
  wage = rep(NA, obs)
```

[7] I call the function `create_fake_data` because I want to emphasise that there is nothing special about simulated data. Some people prefer more dignified names. If you prefer that, you can use any other name for the function. Just make sure that it is clear what your function is doing.

```
  wage[obs] = w0
  for (i in (obs - 1):1){
    wage[i] =  wage[i + 1] + 1/scale * C * size[i + 1] *
      (talent[i] - talent[i + 1])
  }
  fake_data = dplyr::tibble(n = n, size = size, talent = talent,
                            wage = wage)
  return(fake_data)
}
fake_data <- create_fake_data(talent_rate = 2/3, C = 0.01)
```

We can create new datasets where the talent_rate is increased to 1 (high talent) and the effect of CEOs on the surplus is increased to .015 instead of .10. I choose .10 in this simulation instead of 10 so that all values can be interpreted in billion USD to mimic the real data. Again, that is a rather arbitrary scaling issue.

```
data1 <- create_fake_data(obs = 500, talent_rate = 2/3, C = .01) %>%
  mutate(talent_rate = "low talent", C = "low effect")
data2 <- create_fake_data(obs = 500, talent_rate = 1, C = .01) %>%
  mutate(talent_rate = "high talent", C = "low effect")
data3 <- create_fake_data(obs = 500, talent_rate = 2/3, C = .015) %>%
  mutate(talent_rate = "low talent", C = "high effect")
data4 <- create_fake_data(obs = 500, talent_rate = 1, C = .015) %>%
  mutate(talent_rate = "high talent", C = "high effect")
data_exp <- bind_rows(data1, data2, data3, data4)
data_exp
```

```
## # A tibble: 2,000 x 6
##        n  size talent   wage talent_rate C
##    <int> <dbl>  <dbl>  <dbl> <chr>       <chr>
##  1     1 600    -1.5  0.0241 low talent  low effect
##  2     2 300    -2.38 0.0197 low talent  low effect
##  3     3 200    -3.12 0.0172 low talent  low effect
##  4     4 150    -3.78 0.0156 low talent  low effect
##  5     5 120    -4.39 0.0143 low talent  low effect
##  6     6 100    -4.95 0.0134 low talent  low effect
##  7     7  85.7  -5.49 0.0126 low talent  low effect
##  8     8  75    -6.00 0.0120 low talent  low effect
##  9     9  66.7  -6.49 0.0114 low talent  low effect
## 10    10  60    -6.96 0.0110 low talent  low effect
## # ... with 1,990 more rows
```

The `bind_rows` function allows you to combine the four datasets in one big dataset `data_exp`. This will help us to plot the data in Figure 5.4 with the more complicated but also more flexible `ggplot`

function. In the function, we first define the data we want to use. Within the `aes()` specification, we define the variables that are going to be plotted. We tell ggplot that we want the data to be plotted as the following *geom*etric elements a point and a line.[8] I make the line colour grey. The `facet_grid` creates different subplots depending on whether an observation has a different value for the `talent_rate` variable and the `C` variable.

[8] Obviously, you do not need both. This is just to show you what you can do.

```
plot_exp <- ggplot(data_exp, aes(x = size, y = wage)) +
  geom_point() +
  geom_line(colour = "gray") +
  facet_grid(talent_rate ~ C)
print(plot_exp)
```

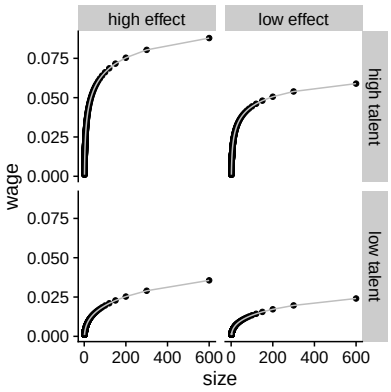


Figure 5.4: (ref:sim-combined)

It looks like our theory is much more sensitive to changes in the talent rate than in changes to the scaling effect.

5.8 *Why should you simulate data?*

In these notes, I will come back to the idea of simulation over and over. For three big reasons:

1. Simulating data from a theory and visualising the theory helps you sharpen your intuition for your theory and for which values are reasonable and which ones are not. If you know upfront which values are not reasonable that will help you to interpret your findings in your statistical analysis. If you estimate parameters in your model that are too big or too small, it might be an indication that something went wrong with your analysis.

2. When you simulate data you can simulate variables that you cannot observe (e.g. CEO talent). Sometimes these variables need to be included in your statistical analysis to get unbiased parameter estimates. With simulated datasets, you can run the analysis with and without the unobservable variables to investigate the impact of including and excluding the variable.

3. If there are different statistical tests you can use to investigate your research question, you do not want to test the different tests on your real data. If you pick the statistical test that gives you the "right" answer, you are likely to delude yourself. The right way to compare different statistical tests is to see whether they can estimate parameters in simulated data where you know what the right value of the parameter is.

In short, the advantage of being able to generete simulated data is that it sharpens your understanding of your theory and of your statistical test.

6
Linear regression in R

Most of the statistical tests in accounting and finance are some variation on a linear regression. In this section, I am assuming that you vaguely know what a linear regression is. I am not going into the mathematical details but I am focusing on the how to run these tests in R.

6.1 Introduction

There are a number of different ways how we can represent a linear regression. The first expression shows how the outcome vector $\mathbf{y}$ is a linear combination of a number of observable vectors $\mathbf{x_i}$ and an unobserved vector $\mathbf{e}$. The second line is an abreviated notation where we group all the observable vectors in a matrix X. The typical assumption is that the unobserved vector follows a normal distribution with a standard deviation σ. The last line explicitly states that $\mathbf{y}$ has a normally distributed random component.

$$\mathbf{y} = a + b_1\mathbf{x_1} + ... + b_n\mathbf{x_n} + \mathbf{e} \tag{6.1}$$

$$\mathbf{y} = a + \mathbf{b}X + \mathbf{e} \tag{6.2}$$

$$\mathbf{y} \sim \mathcal{N}(a + \mathbf{b}X, \sigma) \tag{6.3}$$

To esimate the coefficients a, b_1 and b_2 in R, we have to run the following piece of code.

```
reg <- lm(y ~ x1 + x2, data = my_data_set)
summary(reg)
```

6.2 Linear regression with non-linear theories

Linear regressions can be used with a non-linear theory with some transformations. The matching theory of CEO compensation implies

that when there are an infinite amount of firms, the relation between the CEO compensation (W) and the firm value (V_0) is given by equation (6.4).

$$W = aV_0^b \implies ln(W) = ln(a) + b \times ln(V_0) \tag{6.4}$$

We can see how well the logarithmic transformation works on the simulated data by transformating the scale of Figure 5.4

```
plot_exp +
  scale_x_continuous(trans = "log",
                     breaks = scales::pretty_breaks(n = 2))  +
  scale_y_continuous(trans = "log",
                     breaks = scales::pretty_breaks(n = 2))
```

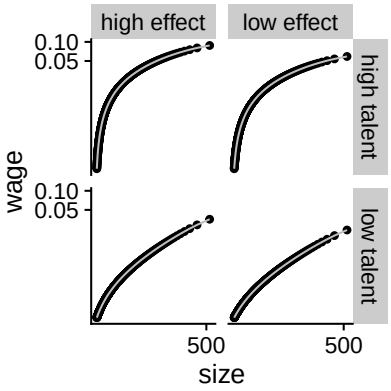


Figure 6.1: The logarithmic transformation of simulated data

One reason why it does not work perfectly is that the reservation wage w_0 might be set too low in the simulation. You can see that the line is approximately linear for the largest firms. The real models assume that we have an infinite number of firms and we are only interested in the largest 500 firms, that is the most right part of the graph. This explanation also illustrates how simulating from your theory can help you to understand the theory better. One of the homework exercises asks you to check whether the transformation works better once you simulate more firms.

Let us now see how well the log transformation works on the real S&P500 compensation data. First, we load the data in.

```
us_comp <- readRDS("data/us-compensation.RDS") %>%
  rename(total_comp = tdc1)
us_value <- readRDS("data/us-value.RDS") %>%
  rename(year = fyear, market_value = mkvalt)
summary(us_value$market_value)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##     117.3    9142.3   16006.8   37036.0   33581.9 1073390.5       682
```

We have a bunch of observations with missing values (`NA`) for the market value of the firm. I will not further investigate this but you should definitely do this in your own project.

We combine the market value data with the CEO compensation data with the `left_join` function. `left_join` is a function that joins two datasets together based on key variables that identify which observations from one dataset should be matched with which observations from the other dataset. The join functions in the `tidyverse` can be a lifesaver if you are working with multiple datasets that need to be merged.

```
us_comp_value <- left_join(select(us_comp, gvkey, year, total_comp),
                           us_value, by = c("year", "gvkey"))
glimpse(us_comp_value)

## Observations: 4,564
## Variables: 5
## $ gvkey        <chr> "001045", "001045", "001045", "001045", "001045", "001...
## $ year         <dbl> 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2011, 2011, ...
## $ total_comp   <dbl> 5526.835, 656.226, 19092.978, 12301.976, 11418.547, 14...
## $ market_value <dbl> 117.3438, 266.5571, 6591.9923, 37405.5843, 26452.7417,...
## $ ni           <dbl> -1979.000, -1876.000, -1834.000, 2882.000, 7610.000, 2...
```

Because there are some CEOs with 0$ compensation, I use the $log(x+1)$ transformation. Another option would have been to exclude these obervations.

```
reg <- lm(log(total_comp + 1) ~ log(market_value + 1),
          data = us_comp_value)
# summary(reg)
print(summary(reg), digits = 1L)

##
## Call:
## lm(formula = log(total_comp + 1) ~ log(market_value + 1), data = us_comp_value)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -9.9   -0.2    0.1    0.4    2.9
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.77       0.14      47   <2e-16 ***
## log(market_value + 1)     0.23       0.01      16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 3946 degrees of freedom
##   (616 observations deleted due to missingness)
## Multiple R-squared:  0.06,   Adjusted R-squared:  0.06
## F-statistic: 2e+02 on 1 and 3946 DF,  p-value: <2e-16
```

The `summary` gives way to much information but I am just restricting the number of significant digits because that is my biggest concern. You can just use `summary(reg)`. We can also plot the log-log relation between firm value and CEO compensation in figure.

```
plot_check <-
  ggplot(data = us_comp_value,
         aes(y = total_comp, x = market_value)) +
  geom_point(alpha = .25) +
  scale_x_continuous(trans = "log1p",
                breaks = scales::log_breaks(n = 5, base = 10),
                labels = function(x)format(x/1000, digits = 2)) +
  scale_y_continuous(trans = "log1p",
                labels = function(x)format(x/1000, digits = 2)) +
  ylab("compensation ($ million)") +
  xlab("market value ($ billion)")
print(plot_check)
```

```
## Warning: Removed 616 rows containing missing values (geom_point).
```

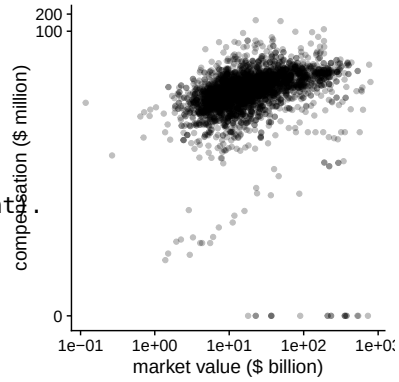


Figure 6.2: CEO compensation and firm value on the log-log scale

On the log-log scale, we see that there are some CEOs with very low compensation compared to the bulk of the observations. In the next figure, we will just ignore those.

```
plot_check2 <- plot_check +
  scale_x_continuous(trans = "log1p", limits = c(500, NA),
                breaks = scales::log_breaks(n = 5, base = 10),
                labels = function(x)format(x/1000, digits = 2)) +
  scale_y_continuous(trans = "log1p", limits = c(500, NA),
                breaks = scales::log_breaks(n = 5, base = 10),
                labels = function(x)format(x/1000, digits = 2))
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

```
print(plot_check2)
```

```
## Warning: Removed 665 rows containing missing values (geom_point).
```

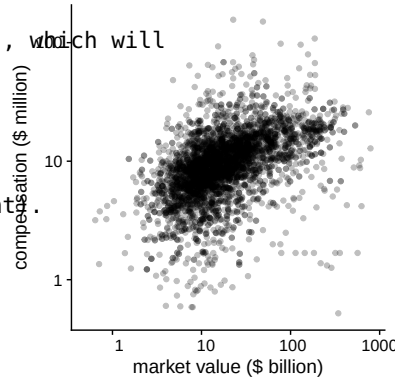


Figure 6.3: CEO compensation and firm value on the log-log scale

Although there still a lot of variation visible, the relation in Figure 6.3 looks reasonably linear. The goal of plotting the data is to get an idea whether the data reflect the explicit and implicit assumptions in your theory and your statistical model. One of those assumptions is that the relation between compensation and value is linear on the log-log scale. Plotting the data also showed us that there are some CEOs with very low or 0 compensation. In a real, study you would want to investigate why this is the case. For instance, you might have an error in the database.

7
Measurement and Theory

```
library(kableExtra)

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows

if (knitr::is_latex_output()){
  options(knitr.table.format = "latex")
} else {
  options(knitr.table.format = "html")
}
remake_diagrams <- FALSE
```

7.1 Introduction

Chapter 5 explained how different functions can describe the relation between total CEO compensation and firm size (or firm value). While the total amount of CEO compensation is a hot-button issue, another discussion is whether CEO compensation reflects pay-for-performance. That is, do CEOs get more compensation when the firm performance better and vice versa.

In 1990, Jensen and Murphy (1990) reported that for every $1000 loss (gain) in firm value, a CEO only loses (gains) $3.25 and took that as evidence that CEO compensation was not aligned with shareholder value.[1] In this chapter, I will use advances in the literature to show that their conclusion depends on how pay-performance sensitivity is measured. The main take-away from this chapter should be that the way you measure your variables is guided by the assumptions in your theory.

[1] This study is 30 years old and the authors would argue that CEO compensation is currently better aligned with firm performance.

7.2 *Assumptions About CEOs*

There are two assumptions we can make about CEOS. The first is what the impact is of a CEO on the firm and the second is about the CEO's preferences. Let us tackle these assumptions in order..[2]

The first assumption is about whether the actions of a CEO scale with the size of the firm. That is, do the actions of a CEO have the same impact on firm value in a small firm and in a large firm? Some actions such as using the corporate jet or sponsoring a sports team have the same impact independent of the firm size. If the company pays for a corporate jet, they pay the same amount of money whether they are large or small. Other actions such as implementing a better strategy, attracting better employees, or developing better relations with investors will impact the entire firm and thus will have a larger monetary impact in large firms.[3]

The second assumption is how CEOs value their leisure time. They can see leisure time as independent of the compensation. For instance, they could think of the opportunity cost of being at work. They could value their time off work as independent of their income. A day with family is as important independent of your compensation. However, the other assumption is that leisure time is more valuable if you have high compensation. For instance, if you earn a lot of money, you can do a lot more on your holidays.[4] Another, possibility is that CEOs are already rich and they could just invest their own fortune and earn a lot of money that way.[5]

[2] This chapter is entirely inspired by Edmans and Gabaix (2016)

[3] In a way, the question is whether shareholders should care more about the ability of CEOs to directly destroy value or trust in the CEOs ability to create value. Although, in a larger firm it might also be easier to fraudulently embezzle more money.

[4] Elon Musk and Richard Branson seem to have a lot of fun when they are not working.

[5] Think of hedge fund managers who decide to close their fund and run a family fund to just invest their own money and that of some associates and family members.

7.3 *Optimal CEO incentives*

When I say optimal CEO incentives I mean optimal *from an optimal contracting* between shareholders and CEO point of view. This could still be suboptimal from a social perspective. It also means that we take into account that the CEO might be risk averse and we might need to pay them more money to overcome the risk aversion.

The two assumptions have in impact on how a firm should incentivise the CEO to maximise firm value. If the actions of a CEO are independent of firm size, then the right performance measure for CEOs is the increase in the monetary value of the firm. If the actions of the CEO have a bigger impact in large firms,[6] the right performance measure is the increase in firm value compared to the size of the firm.

If the CEOs' preferences for not working are independent of their compensation, firms can reward the CEOs independent of how wealthy they already are. That is, the better they perform the more money they get. In contrast, if the CEOs' preferences depend on

[6] It's easier to increase firm value by $1 Million for Apple than for a start-up company.

how wealthy they are, they expect more money if they are already wealthy. The reward of these CEOs should be proportional to their wealth.

As a result, the literature has defined pay-performance sensitivity in three different ways.

1. **$ Pay - $ Firm value:** This is the Jensen and Murphy (1990) measure. For every change in dollar firm value, the CEOs compensation changes with a *fixed* dollar amount. The consequence of these assumptions is that you would expect CEOs to hold the same equity stake in small and large firms.

2. **$ Pay - % Firm value:** For every % increase in firm value, there is a *fixed* dollar amount paid to the CEO. You an think of this measure of pay-performance sensitivity as a bonus system where the CEO gets a bonus for every increase in the companies stock market return.

3. **% Wealth - % Firm value:** For every increase % increase in the value of the firm, the CEO should receive a fixed % increase in their wealth. This is the measure that is advocated for in Edmans and Gabaix (2016) and I also belief that the evidence is most in favour of this definition and thus the assumptions that a CEO's actions and preferences depend on firm size and on CEO wealth respectively.

Mathematically, we can represent each definition as follows:

1. $\frac{\partial W}{\partial V} = \frac{\partial W}{\partial r}\frac{1}{V_0}$
2. $\frac{\partial W}{\partial r}$
3. $\frac{\partial W}{\partial r}\frac{1}{W_0}$

Where V is the value of the firm at a certain point, W is the CEO's wealth, and r is the return of the stock compared to the starting value V_0. Because derivatives are not empirically observable, we have to look at changes in wealth and firm value over two different time periods. The measures that follow from the mathematical definitions are as follows where we use the fact that $\Delta W = W - W_0$ and $ln(W) - ln(W_0) = ln(\frac{W}{W_0})$

1. $\frac{\Delta Wealth}{\Delta Firm\ Value}$
2. $\frac{\Delta Wealth}{\Delta ln(Firm\ Value)}$
3. $\frac{\Delta ln(Wealth)}{\Delta ln(Firm\ Value)}$

To reiterate, we have three different measures of pay-performance sensitivity, i.e. how well CEO compensation reflects the performance.

For each measure, we have assumed that the firm will use the optimal incentives given our assumption about the CEO's actions and their preferences. The point of this exercise is that different assumptions determine how you should measure the variable you want to analyse. In the next section, we are going to investigate how these three measures behave in the S&P500 sample.

7.4 What does the data say?

Let us first reload the data. We need a measure for the wealth of the CEO. I am going to assume that the most important component of the CEO's wealth is their shares in the company. This is obviously not perfect but it is not a bad guess for S&P500 CEOs either.

```
library(tidyverse)
us_comp <- readRDS("data/us-compensation.RDS") %>%
  rename(total_comp = tdc1, shares = shrown_tot_pct)
us_value <- readRDS("data/us-value.RDS") %>%
  rename(year = fyear, market_value = mkvalt)
us_comp_limit <- select(us_comp, gvkey, execid, year, shares, total_comp)
us_comp_value <- left_join(us_comp_limit, us_value,
                           by = c("year", "gvkey")) %>%
# Delete the observations with missing values.
  filter(!is.na(market_value) & !(is.na(shares))) %>%
  mutate(wealth = shares * market_value / 100)

max(us_comp_value$market_value)/1e3

## [1] 790.0501

max(us_comp_value$wealth)/1e3

## [1] 92.24485
```

To get an idea wheather I did not make any mistakes I let R print the maximum market value and the maximum CEO wealth in the dataset. We can compare this to some known values to check whether we made any mistakes in putting the dataset together. Apple's market value at the time of writing hovers around $800 billion and Jeff Bezos[7] is worth around $100 billion. So the proxies are not a terrible first attempt.[8]

Above I explained that the first measure ($-$) assumes that the percentage share of stock the CEO owns will be independent of the size of the company. To test this hypothesis, Figure 7.1 plots the percentage share the CEO owns against the size of the company.

[7] before the divorce

[8] Especially once you consider that some papers throw away the maxima for some variables because they are considered outliers.

```
shares_size <- ggplot(data = us_comp_value,
                      aes(x = market_value/1000, y = shares)) +
  geom_point(alpha = .20) +
  ylab("CEO Ownership") +
  xlab("Firm Market Value (in Billions)")
plot(shares_size)
```

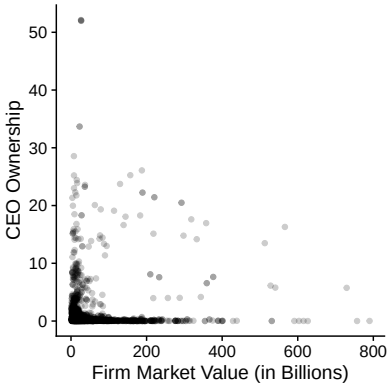


Figure 7.1: Relation between CEO ownership and market value for SP500 firms (2011-2018).

The figure is not that instructive so let's plot the data on the log-log scale in Figure 7.2. The bulk of the data now shows a downward trend. This is not consistent with the hypothesis of the $-$ measure of pay-performance sensitivity where we would expect the link between firm size and CEO ownership to be constant.

```
shares_size_log <- shares_size +
  scale_x_continuous(trans = "log",
                     labels = function(x) prettyNum(x, digits = 2),
                     breaks = scales::log_breaks(n = 5, base = 10)) +
  scale_y_continuous(trans = "log",
                     labels = function(x) prettyNum(x, digits = 2),
                     breaks = scales::log_breaks(n = 5, base = 10))
plot(shares_size_log)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

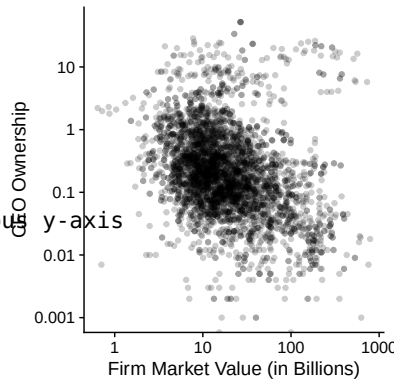


Figure 7.2: Relation between CEO ownership and market value for SP500 firms (2011-2018).

Nevertheless, there seems to be a cloud of points at the top of Figure 7.2 where CEOs have a large stake in the company and this cloud looks largely independent of firm size. I suspect that these CEOs are founders who have a larger ownership share for other reasons than incentives. We can test this idea. The code below filter all the observations with a market value above $10 Billion and ownership stake above 1%. I then join those observations with the CEO's name and the firm's name from the original data, group the observations by CEO and company, and count the number of years they are in the new data set. The `knitr::kable` function formats the table for pdf and html output.

```
filter(us_comp_value, market_value > 1e5, shares > 1) %>%
  select(gvkey, year) %>%
  left_join(select(us_comp, exec_fullname, coname, year, gvkey),
            by = c("year", "gvkey")) %>%
  group_by(gvkey, exec_fullname, coname) %>%
  summarise(nr_years = n()) %>%
  knitr::kable(booktabs = TRUE, caption = 'Outliers')
```

My hunch looks correct.[9] This whole process of plotting the data and checking assumptions on a subset of the data might seem a long

[9] Also, I think the `tidyverse` functions are amazing because they make it so easy to quickly look at the data and summarise the data.

Table 7.1: Outliers

gvkey	exec_fullname	coname	nr_years
002176	Warren E. Buffett	BERKSHIRE HATHAWAY	6
012141	Steven A. Ballmer	MICROSOFT CORP	4
012142	Lawrence J. Ellison	ORACLE CORP	3
064768	Jeffrey P. Bezos	AMAZON.COM INC	6
160329	Lawrence Edward Page	ALPHABET INC	11
170617	Mark Elliot Zuckerberg	FACEBOOK INC	5

detour. You could ask yourself why we are not immediatelly calculating the three measures that we are actually interested in. Checking your assumptions and checking the validity of your dataset for your research question is an important part of the research process.[10] And as you can see you do not have to run a statistical test to test your assumptions.

[10] Remember that we derived the three performance-pay sensitivity measures based on the optimal contract between shareholders and owners. The firm needs a contract to govern the conflict of interest between the shareholders and the CEO. If the founder is the CEO, this conflict of interest is different. You could ask yourself whehter you should keep founder-CEOs in the dataset because you do not necessarily expect their contract to follow the optimal contract.

7.5 *Three pay-performance sensitivities*

Now that we understand our dataset, we can calculate the changes in CEO wealth, total compensation, CEO compensation and market value. The last `filter` is a bit of hack. There are a few wealth change that are infinite at the log value. That can happen when a CEO ends up with 0 shares and no compensation. There are only a handful of these observations and I am ignoring them. Ideally, you would have a look whether anything funky is going on.

```
us_comp_value <- group_by(us_comp_value, gvkey, execid) %>%
  arrange(year) %>%
  mutate(prev_market_value = lag(market_value),
         prev_wealth = lag(wealth),
         prev_comp = lag(total_comp)) %>%
  ungroup() %>%
  mutate(change_market_value = market_value - prev_market_value,
         change_wealth = wealth - prev_wealth,
         change_comp = total_comp - prev_comp,
         change_log_value = log(market_value) - log(prev_market_value),
         change_log_wealth = log(wealth) - log(prev_wealth),
         change_log_comp = log(total_comp) - log(prev_comp)) %>%
  filter(!is.infinite(change_log_wealth),
         !is.infinite(change_log_comp)) %>%
  arrange(gvkey)
```

With these variables, we can replicate the basic approach of some key papers on executive compensation. The first regression takes the

Jensen and Murphy (1990) approach and investigates the relation between a change in market value and a change in CEO compensation (Measure 1 in 7.3). The second and third regression follow the approach of Hall and Liebman (1998) where we look the relation between relative changes in market value and relative changes in total CEO compensation.[11] Finally, the fourth regression follows the Core, Guay, and Thomas (2005) and looks at the relation between changes in relative market value and relative CEO wealth.[12]

[11] The difference between the two specification is that the first uses market returns, just like the original paper, which is not exactly theoretically correct but a very good approximation of the theory.

[12] In all cases, the authors of the original papers do a much better job than I do in measuring the variables they are interested in. Do not take the results too seriously although they are a good approximation of the results in the original papers.

```
jensen_murphy90 <- lm(change_comp ~ change_market_value,
                      data = us_comp_value)
hall_liebman98 <- lm(change_log_comp ~
                     I(change_market_value/prev_market_value),
                     data = us_comp_value)
hall_liebman98_alt <- lm(change_log_comp ~ change_log_value,
                         data = us_comp_value)
core_guay_thomas05 <- lm(change_log_wealth ~ change_log_value,
                         data = us_comp_value)

coef(jensen_murphy90)[2]

## change_market_value
##          0.03032071

coef(hall_liebman98)[2]

## I(change_market_value/prev_market_value)
##                                 0.239141

coef(hall_liebman98_alt)[2]

## change_log_value
##        0.3096497

coef(core_guay_thomas05)[2]

## change_log_value
##        0.8112861
```

Jensen and Murphy found that a CEO gets $3.5 for every $1000 change in market value. Our estimate is $0.03 for every $1000 change in market value. The difference is first of all an indication that this simple analysis is not the best estimate but it probably also means that the $-$ incentives are not the best measure. Recall from Figure 7.2 that CEO ownership decreases with firm size. Since the study was published, companies have become larger which pushes the $-$ measure lower.[13]

[13] Jensen and Murphy (1990) looked at more than the direct incentives. They also estimated the potential loss of getting fired. Moreover, total compensation does not take into account wealth. Other studies have found that the $-$ incentives measure of pay-performance sensitivity are smaller for larger firms.

The original Hall and Liebman (1998) measure shows that CEOs see a 0.24% increase in pay for every 1% increase in firm value, the alternative measure shows a 0.31% pay-performance sensitivity. Finally, the Core, Guay, and Thomas (2005) measure shows a 0.81% wealth-performance sensitivity.

7.6 Comparison plot

We can plot the results as well to get a better intuition of the what this results mean. The code is long and boring but not so difficult to understand. First, make a new dataset with all the change in pay and change in performance variables. Next, just copy the new dataset three times. Create a new variable `analysis` for each type of analysis. Create a new variable `incentive` and `performance` for each type of analysis.

```
data_comparison = select(us_comp_value, change_comp,
                         change_market_value,
                         change_log_comp, change_log_value,
                         change_log_wealth) %>%
  filter(complete.cases(.))

n = nrow(data_comparison)

data_comparison = bind_rows(data_comparison, data_comparison,
                            data_comparison) %>%
  mutate(analysis = c(rep("jensen_murphy90", n),
                      rep("hall_liebman98_alt", n),
                      rep("core_guay_thomas05", n))) %>%
  mutate(
    incentive = recode(analysis,
                       "jensen_murphy90" = change_comp/1000,
                       "hall_liebman98_alt" = change_log_comp,
                       "core_guay_thomas05" = change_log_wealth),
    performance = recode(analysis,
                         "jensen_murphy90" = change_market_value/1000,
                         "hall_liebman98_alt" = change_log_value,
                         "core_guay_thomas05" = change_log_value)) %>%
  select(analysis, performance, incentive)
```

Now, we can plot the the incentive to performance relation for the data in Figure 7.3. Despite all all the caveats about how unpolished this analysis is, the figure shows that neglecting CEO ownership is going to lead you to underestimate the pay-performance sensitivity for CEOs.

```
comparison_plot = ggplot(data_comparison,
                         aes(x = performance, y = incentive)) +
  geom_point(alpha = .1) +
  facet_wrap( ~ analysis, ncol = 1, scales = "free") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 5),
                     labels = function(x) prettyNum(x, dig = 2)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 5),
                     labels = function(x) prettyNum(x, dig = 2))
print(comparison_plot)
```

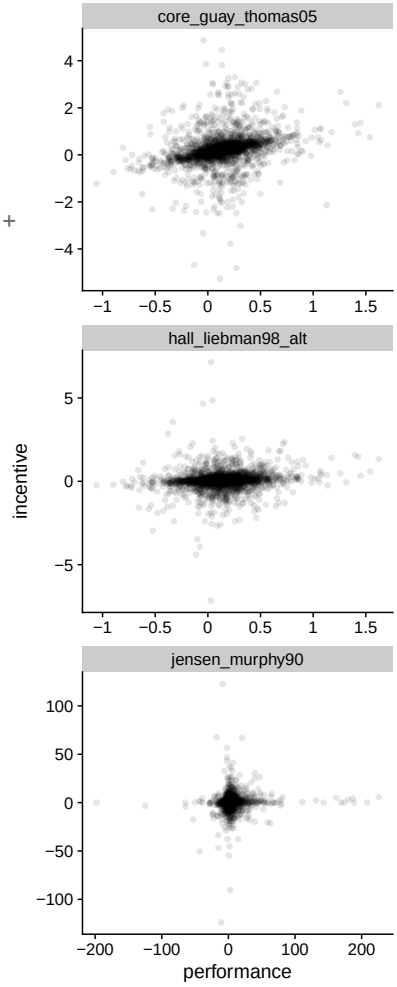


Figure 7.3: Comparison of three measures of pay-performance sensitivity for S&P500 Companies (2011-2018)

7.7 *Libby boxes*

An empirical study can often be represented in the form of a diagram as in Figure 7.4. These diagrams are often referred to as the Predictive Validity Framework or, in the accounting literature, Libby Boxes (Libby, Bloomfield, and Nelson 2002). The top of the diagram (1) represents the research question. We want to ask whether there is a relation between two theoretical concepts. In our case, we want to know whether there is a relation between CEO incentive pay and firm performance. Chapter 5 dealt with how to use theory to predict a relation between two concepts. In reality, we cannot directly observe pay and performance, so we have to come up with a measure of these concepts (3 and 4). This Chapter has dealt with the issue of how we can find a good measure for our two concepts. The downwards arrows represent the connection between the theoretical concepts and the empirical measures. At first, you might have thought that it is obvious how to measure CEO performance and firm performance. The pay-performance sensitivity example should make you realise that you need to think careful about what you measure and how your measure connects to your theory. Most statistics textbooks focus on the link between the two measures (2), ´change_log_wealth´ and ´change_log_performance´ in our case. This is the next step in this Chapter. Later in the notes, I wil look at the issue of control variables.

```
library(DiagrammeR)
libby_boxes <- grViz("
    digraph libby_boxes{
    graph [fontsize = 18, layout = 'dot']

    node [shape = box, group = 'cause_group']
    cause [label = 'Performance']
    m_cause [label = 'Change_Log_Value']
    node [shape = box, group = 'effect_group']
```

```
    effect [label = 'Pay']
    m_effect [label = 'Change_Log_Wealth']
    controls

    cause -> effect [label = '(1)']
    m_cause -> m_effect [label = '(2)']
    cause -> m_cause [label = ' (3) ']
    effect -> m_effect [label = ' (4) ']
    m_effect -> controls [dir=back, label = '(5)']

    subgraph{
    rank = same; cause; effect;
    }
    subgraph{
    rank = same; m_cause; m_effect; controls;
    }
  }
")
trelliscope::widgetThumbnail(libby_boxes,
                             "_bookdown_files/libby_boxes.pdf")
trelliscope::widgetThumbnail(libby_boxes,
                             "_bookdown_files/libby_boxes.png")

knitr::include_graphics("_bookdown_files/libby_boxes.png",
                        auto_pdf = TRUE, dpi = 300)
```

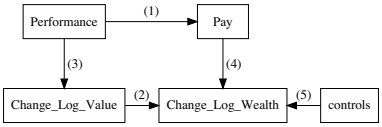


Figure 7.4: Predictive Validity Framework or Libby Boxes for pay-performance sensitivity

7.8 *Are incentives independent of size?*

Section 7.4 showed that there was a negative relation between CEO ownership and the size of the company. The relation is pretty strong and we did not look at it in more detail. In a real study with more subtle effects, you want to get some sense whether the relation is real or whether it can be explained by random variation.

I will illustrate this issue with the new pay-performance sensitivity. The optimal performance contract should not depend on the size of the firm only on the performance of the firm. Figure 7.5 shows the relation between pay-performance sensitivity and the size of the firm. The figure does not show any strong relation between firm size and pay-performance sensitivity.

```
hypothesis = ggplot(us_comp_value,
                    aes(y = change_log_wealth / change_log_value,
                        x = market_value/1000)) +
  geom_point(alpha = .2) +
```

```
  scale_x_continuous(
    trans = "log",
    breaks = scales::log_breaks(n = 5, base = 10),
    labels = function(x) prettyNum(x, dig = 2)) +
  scale_y_continuous(
    limits = c(-50, 50)) +
  xlab("market value") +
  ylab("sensitivity")
print(hypothesis)
```

```
## Warning: Removed 1599 rows containing missing values (geom_point).
```

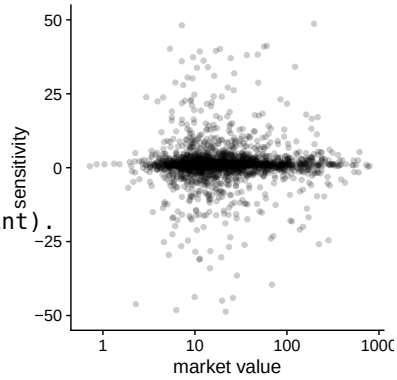


Figure 7.5: Relation between pay-performance sensitivity and market value

There a number of different ways to test whether this relation is real or whether it can be explained by random fluctuations. I first present two simulation based approaches and then present an approach based on formulas. Statistical software typically implements the formulas for you so you do not have to remember them.

The starting point for each of these tests is that if there is no relation, the true effect that we are interested in[14] equals 0. Because of random variation, the number in the statistical analysis[15] might be slightly off. Each of the three approaches tries to quantify that random variation around a the null effect.

[14] That is the top half of the Libby Boxes

[15] The bottom half of the Libby Boxes

7.8.1 Randomisation Test

The first way to quantify the random variation is to randomly assign one observation of the firm-year to another obervation of a CEO-year. That is we create a dataset where we link the sensitivity of a CEO's pay with the market value of another firm. The general idea of a randomisation test is that you break the relation that you want to test and see whether the value in the real dataset is extreme compared to the randomised dataset.

```
data_hypo <- mutate(us_comp_value,
                    sensitivity = change_log_wealth / change_log_value) %>%
  select(sensitivity, market_value) %>%
  filter(complete.cases(.))
true_cor <- cor(
  data_hypo$sensitivity, data_hypo$market_value)
rand_cor <- cor(
  data_hypo$sensitivity, sample(data_hypo$market_value))
print(prettyNum(c(true_cor, rand_cor), dig = 2))
```

```
## [1] "-0.036" "-0.014"
```

So, we find that the correlation in the real dataset is different from the correlation in the random dataset but it does not allow us to

draw any conclusions. We need to draw a lot of random datasets and compare that random dataset to the real correlation. That is what the next lines of code do. First, create a function that resamples the market value and calculates the correlation. Second, rerun that function 10,000 times. `rand_cor` is a vector with randomly created correlations.

```
simulate_cor = function(data){
  return(cor(data$sensitivity, sample(data$market_value)))}
rand_cor = replicate(1e4, simulate_cor(data_hypo))
```

Figure 7.6 visualises the test. We plot a histogram of the random correlations and compare them to the size of the of the real correlation. We want to test whether the value of the real correlation is extreme, so we compare the random distribution to correlations with an absolute value larger than the real correlation in orange.

```
hist_sim <- qplot(rand_cor, bins = 100,
                  col = abs(rand_cor) < abs(true_cor)) +
  xlab("Random Correlations") +
  scale_color_viridis(discrete = TRUE) +
  theme(legend.position = "none")
print(hist_sim)
```

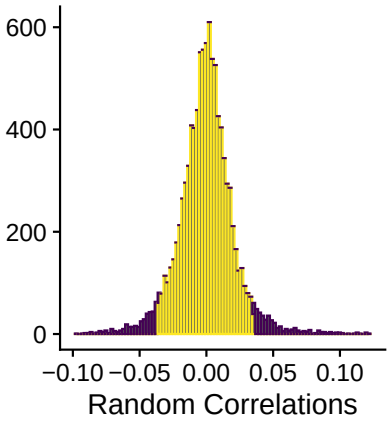


Figure 7.6: Random correlation compared to real correlation

```
pvalue_sim =  mean(ifelse(abs(rand_cor) > abs(true_cor), 1, 0))
print(prettyNum(pvalue_sim, digits = 2))
```

```
## [1] "0.082"
```

We can calculate how likely it is that a random correlation is larger (in absolute value) than the true correlation. We find that there is 0.082 probability of getting the true value with randomisation. That means that there is little evidence of a relation between pay-performance sensitivity and firm size.

7.8.2 Bootstrap

Another way to test whether there is a relation between two variables is the bootstrap (Efron and Hastie 2017). The idea of the bootstrap is that the sample of firms that we have is a good reprentation of a bigger population of (very large) firms.[16] In this procedure, we assume that the random variation comes from the sampling error compared to the full population. The bootstrap will resample the observations in the real dataset with replacement. We keep CEOs and firms together this time but in the newly created datasets some observations might occur more than once. If we create enough of

[16] You could think of for instance the same firms in 2019 and 2020 or 2008 and 2009.

these datasets, we will get a good representation of the correlations we get from repeatedly sampling from the larger population.

```
calc_corr <- function(d){
  n <- nrow(d)
  id_sample <- sample(1:n, size = n, replace = TRUE)
  sample <- d[id_sample, ]
  corr <- cor(sample$sensitivity, sample$market_value)
  return(corr)
}
boot_corr <- replicate(2000, calc_corr(data_hypo))
```

We visualise the sampling variation in the correlation in Figure 7.7. We can now compare this distribution to a 0 correlation. If the distribution is (almost) entirely on one side of a null effect, we would conclude that sampling variation around a nulll effect is not a good explanation for the effect that we found in the real sample. However, in this case, we find that a considerable amount of the distribution is on either side of the orange line. Again, we find no good reason to belief there is anything more than random variation around a null effect.

```
qplot(boot_corr, bins = 100) +
  geom_vline(aes(xintercept = 0), col = "darkorange") +
  xlab("Bootstrapped Correlation")
```

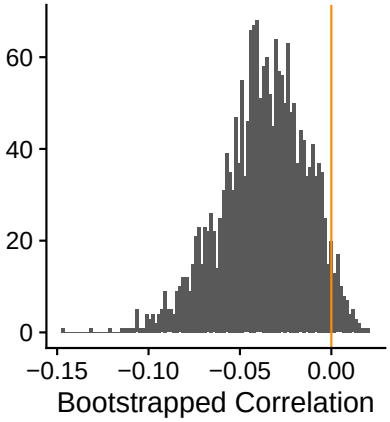


Figure 7.7: 2000 bootstrapped correlations

TODO: `boot` package

7.8.3 *P-values*

Randomisation and bootstrapping make different assumptions about where the random variation comes from and what it means that there is no effect. There is an active discussion in the econometrics literature about which assumptions are more appropriate under which conditions. However, most empirical researchers ignore these questions and trust in formula based p-values. The interpretation of p-values is closer to the bootstrap. We test whether the real correlation is extreme enough so that it is unlikely that it arises sampling variation around a null effect.

```
cor = cor.test(data_hypo$sensitivity, data_hypo$market_value)
pvalue_cor = cor$p.value
print(prettyNum(pvalue_cor, dig = 2))
```

```
## [1] "0.086"
```

We can let R do the calculation for us and we see that the p-value for the correlation between sensitivity and market value is 0.086. This is really close to the value we got from the randomisation test. Again, it looks like that the correlation is quite likely to emerge from sampling variation around a null hypothesis.

Finally, I would like to point out the relation between linear regression and a correlation. When there is only one independent variable in a regression model, the correlation test is equivalent to the statistics we get from a regression.

$$y \sim \mathcal{N}(\alpha + \beta x, \sigma)$$
$$\beta = \frac{sd(y)}{sd(x)} cor(x, y)$$

```
regr_sens = lm(sensitivity ~ I(market_value/1e6), data = data_hypo)
coefficients(summary(regr_sens)) %>%
  kable(dig = 2, booktabs = TRUE, caption = 'Regression Results')
```

Table 7.2: Regression Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.30	0.81	1.60	0.11
I(market_value/1e+06)	-17.98	10.46	-1.72	0.09

In summary, most empirical researchers use p-values to estimate how far away the estimated coefficient is from the null effect. To do that they use formulas that estimate the standard error of the coefficient. If the coefficient is large and the standard error is small, we expect the effect to be extreme compared to the null effect. To estimate how likely it is that the real coefficient is the result of random variation around the null, your software calculates a statistic, the coefficient divided by the standard error. This t-statistic is expected to follow a t-distribution and from this distribution we can calculate the p-value.

Most empirical papers will try to find small p-values. That is they try to show that there is a relation. I deliberately chose an example where there is no relation between the two variables of interest because and where that is also what we hoped to find. Our theory said that there should not be a relation between sensitivity and size and after looking at the data we have no reason to believe otherwise. This type of test is a natural fit for p-values and this is why I introduced them in this way.

There is a lot of discussion in the literatur about abuses and misinterpretations of p-values. My hope is that I put you on the right path

to interpreting p-values despite all the traps that have been laid for you in the literature.

8
Control Variables

Oh, someone's really smart
Oh, complete control, yeah that's a laugh
The Clash - Complete Control

Let's quickly revise the Libby boxes. In Chapter 5, we looked how we could better understand a mathematical theory. Simulations can help us to better understand the relation between firm size and compensation. This is link (1) in the Libby boxes. In Chapter 7, we focus on the links (3) and (4). The measures in the statistical analysis depend on the theoretical assumptions. In section 7.8, we focused on how we establish statistically a link between the two measures which is link (2) in the boxes.

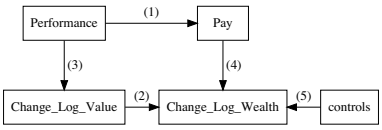


Figure 8.1: Predictive Validity Framework or Libby Boxes for pay-performance sensitivity

8.1 R housecleaning.

We are going to run a lot of regressions in this chapter. I load the `stargazer` package which helps to make tables of regressions.

```
suppressMessages(library(stargazer))
suppressMessages(library(tidyverse))
library(DiagrammeR)
tab_format <- getOption("knitr.table.format")
```

We also need to reread the data and calculate some extra variables. The most important feature is that we have a new dataset `us_firm` with two important variables. The first one is `ipo_employee` or whether the CEO was an employee when the company did their initial public offering. The second one is `ipo_ceo` indicating whether the CEO was already in their position when the company launched their initial public offering. From section 7.4, we know that CEOs who are present early in the life of the company might not have a compensation contract that follows the efficient incentive contract.[1]

[1] Not because they are abusing their power necessarily, but because they are intrinsically motivated to do well for the company and they are motivated to stay in control of the company which is easier if they have more shares than necessary from an incentive point of view.

```
us_comp <- readRDS("data/us-compensation.RDS") %>%
  rename(total_comp = tdc1, shares = shrown_tot_pct)
us_value <- readRDS("data/us-value.RDS") %>%
  rename(year = fyear, market_value = mkvalt)
us_firm <- readRDS("data/us-company.RDS")

combined <- left_join(us_comp, us_firm, by = "cusip") %>%
  left_join(us_value, by = c("year", "gvkey")) %>%
  mutate(ipo_ceo = I(becameceo < begdat),
         ipo_employee = I(joined_co < begdat),
         wealth = shares * market_value / 100) %>%
  mutate(ipo_employee = if_else(is.na(ipo_employee), ipo_ceo,
                                ipo_employee)) %>%
  mutate(ipo = ifelse(ipo_ceo, "ceo",
                      ifelse(ipo_employee, "employee", "none")))

combined_change <- group_by(combined, gvkey) %>%
  arrange(year) %>%
  mutate(
    delta_total = log(total_comp) - log(lag(total_comp)),
    delta_wealth = log(wealth) - log(lag(wealth)),
    delta_value = log(market_value) - log(lag(market_value))) %>%
  filter(!is.infinite(delta_total),
         !is.infinite(delta_value),
         !is.infinite(delta_wealth),
         year > 2011)
```

8.2 *Directed Acyclical Graphs*

In this chapter, I will use Directed Acyclical Graphs (or DAGs) to visualise and understand the theory that applies to the empirical data. DAGs follow from the work of Pearl (2009a) and Pearl (2009b). I personnaly like the introductions in Cunningham (2018) and Rohrer (2018). Figure 8.2 shows an example of a DAG for a paper that was presented in our seminar. The setting is typical for a finance and accounting paper.

The key variable `xbrl` indicates whether a company has been mandated to adopt the XBRL format to publish its financial statements. XBRL is a mark-up language like XML that puts structure on the financial statements and makes them more easily machine readable which should improve the information environment for investors. While the information environment can not be directly measured, it is possible to measure the informativeness of the stock prices in companies. XBRL adoption has been staggered over three years, where

larger firms were mandated to adopt in the first year, the medium ones in the second year, and lastly the smaller firms.

The paper is also interested in whether the adoption of `xbrl` has an impact on institutional owenship in the company. The difficulties with studying this relation are given by rest of the DAG. The size of the company also affects institutional ownership before XBRL adoption and studies have shown that institutional ownership affects the information environment.

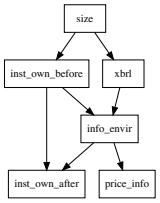


Figure 8.2: Example of a Directed Acyclical Graph

Figure 8.2 serves as an example of what a DAG for a single study can look like. For a typical study, the DAG can quickly become quite complex. One of the reasons is that a social science observational study always have to account for a large number of potential effects. The second reason is that DAGs do not allow for direct feedback loops[2] and we have to account for the feedback between institutional ownership and information environment by adding a time modifier (before and after) to the institutional ownership variables.

[2] That is what acyclical means

8.3 Controlling for measurement error

8.3.1 Directed Acyclical Graph (or DAG)

In the remainder of this chapter, we are interested in testing the relation between `x` and `y` and we have to make a decision whether we should control for `z`. The first reason to control for a variable is because it affects the outcome of interest. This is often the main reason given to control for a variable, it affects the outcome.

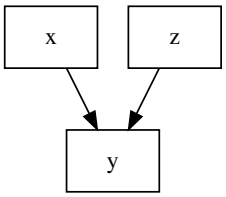


Figure 8.3: Measurement error

8.3.2 Simulation

We can illustrate the problem with a simulation. We are interested in the effect of `x` on `y` and there is another factor `z` with a large effect on `z`. We print the regression of `y` on `x` with and without `z` as a control variable with the `stargazer` package (Hlavac 2018).

```
set.seed(12345)
obs = 500
x <- rnorm(n = obs); z <- rnorm(n = obs)
y <- rnorm(n = obs, mean = 1 * x + 20 * z, sd = 1)
d <- tibble(y = y, x = x, z = z)
yx <- lm(y ~ x, data = d)
yxz <- lm(y ~ x + z, data = d)
stargazer(yx, yxz, type = tab_format,
          label = "measurement_error",
          omit = c("Constant"), digits = 2,
          intercept.bottom = FALSE,
```

```
          star.cutoffs = c(0.05, 0.01, 0.001),
          keep.stat = c("n", "rsq"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Mar 04, 2020 - 12:27:46

	Dependent variable:	
	y	
	(1)	(2)
x	0.97	1.07***
	(0.91)	(0.05)
z		19.99***
		(0.04)
Observations	500	500
R^2	0.002	1.00
Note:	*p<0.05; **p<0.01; ***p<0.001	

The estimate of the effect of x regressions is close to the true value, 1, but the standard error is much smaller when we include z as a control variable. The reason to control for z is to make sure that we have a precise estimate of the effect of interest, however the estimate itself is not going to change when z only has an effect on y.

8.3.3 Real Data

We can turn to the CEO data to illustrate the idea of controlling for other factors that might effect the variable of interest. We investigate the relationship between CEO ownership and market value of the company in Figure 8.4.

```
meas_data <-  select(combined, market_value, shares, ipo) %>%
              filter(complete.cases(.))
ownership <-
  ggplot(data = meas_data,
      aes(x = market_value/1000, y = shares + 0.001)) +
  geom_point() +
  scale_x_continuous(trans = "log",
    breaks = scales::log_breaks(n = 5, base = 10),
    labels = function(x) prettyNum(x, dig = 2)) +
  scale_y_continuous(trans = "log",
```

```
    breaks = scales::log_breaks(n = 5, base = 10),
    labels = function(x) prettyNum(x, dig = 2),
    limits = c(NA, 50)) +
  labs(y = "CEO ownership", x = "Market Value in $ Billion")
print(ownership)
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

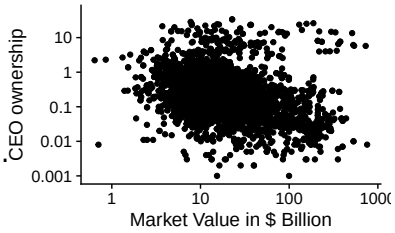


Figure 8.4: Relation between CEO ownership and market value for SP500 firms (2011-2018).

We hypothesised back in Chapter 7 that CEOs who are founders will have higher ownership for reasons that have nothing to do with the optimal incentives. We cannot directly measure whether CEOs are founders but we know wether they are already CEO (`ipo_ceo`) or employee (`ipo_employee`) at the time of the initial public offering of the company. We include those two variables as indicator variables in the regression because it is more likely that these CEOs are also the founders of the company[3].

[3] I will ignore the fact that we lose observations in calculating the new variables. That is not necessarily a good idea.

```
form <- log(shares + 0.001) ~ log(market_value/1000)
base <- lm(form, data = combined)
form_error <- update(form, . ~ . + ipo_ceo + ipo_employee)
error <- lm(form_error, data = combined)
stargazer(base, error, type = tab_format,
          label = "ownership",
          digits = 2, intercept.bottom = FALSE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          keep.stat = c("n", "rsq"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Mar 04, 2020 - 12:27:48

Unfortunately, controlling for `ipo_ceo` and `ipo_employee` does not really improve the precision of our estimates for the effect of market value on ownership. The standard error for the estimate stays about the same (i.e. 0.2).

8.3.4 Interpretation of Regression Outcome (or Economic Significance)

This is a bit of disgression of the main storay and you skip this section. Nevertheless, I would encourage you to go over the example as it gives a good example of how to interpret the numbers in a regression. Something, you will have to do for your thesis as well.

The regression we have estimated is

$$\log(S + 0.001) = \beta_0 + \beta_1 \log(MV) + \beta_2 \text{ipo\_ceo} + \beta_3 \text{ipo\_employee}$$

and based on the theory, we are mainly interested in the parameter β_1. However, something intriguing is going on with the coefficients

	Dependent variable:	
	log(shares + 0.001)	
	(1)	(2)
Constant	-0.23^{**}	-0.50^{***}
	(0.07)	(0.07)
log(market_value/1000)	-0.46^{***}	-0.43^{***}
	(0.02)	(0.02)
ipo_ceo		0.14
		(0.12)
ipo_employee		1.11^{***}
		(0.10)
Observations	3,842	3,373
R^2	0.09	0.19
Note:	*p<0.05; **p<0.01; ***p<0.001	

for the two control variable. The results seem to imply that whether the CEO was an employee at the time of the IPO is a better indicator for share ownership than whether the CEO was already CEO at the time of the IPO.

That conclusion is not correct. There are a number of ways you can see this. When the CEO was already CEO at the time of the IPO, they were also already an employee. That means that the `ipo_ceo` variable is dependent on the `ipo_employee` variable. If `ipo_employee = 0` then `ipo_ceo = 0` and if `ipo_ceo = 1` then `ipo_employee = 1`. You could also see this if you make a descriptive table with the percentage of observations with each theoretical combination of the `ipo_ceo` and `ipo_employee` variable.

```
combined %>% select(ipo_ceo, ipo_employee, year) %>%
  group_by(year) %>%
  summarise(
    ipo_both = mean(ipo_ceo * ipo_employee, na.rm = T),
    ipo_ceo_only = mean(ipo_ceo * (1 - ipo_employee),
                        na.rm = T),
    ipo_empl_only = mean((1 - ipo_ceo) * ipo_employee,
                         na.rm = T),
    ipo_none = mean((1 - ipo_ceo) * (1 - ipo_employee),
```

```
                    na.rm = T)
) %>% kableExtra::kable(digits = 2, booktabs = TRUE)
```

year	ipo_both	ipo_ceo_only	ipo_empl_only	ipo_none
2011	0.16	0	0.08	0.77
2012	0.17	0	0.08	0.75
2013	0.17	0	0.07	0.76
2014	0.15	0	0.07	0.78
2015	0.14	0	0.05	0.81
2016	0.13	0	0.05	0.82
2017	0.12	0	0.05	0.83
2018	0.10	0	0.05	0.85

The table again shows that it is impossble for a CEO to have been the CEO at the time of the IPO but not an employee[4]. This example shows the importance of using descriptive statistics to better understand the data. Even if you had not realised that there is a relation between `ipo_ceo` and `ipo_employee`, the table would have alerted you to it.

[4] `ipo_ceo_only` equals 0 in every year.

It also shows us that we cannot interpret the coefficients separately. The right way to interpret the coefficients is to interpret them together. However, there is a second problem with interpreting our regression outcomes. We modelled `log(shares + 0.001)` and so the effect are a bit harder to interpret. With some rearrangement of the regression function, we can make the interpreation easier.

$$\begin{aligned}\log(S + 0.001) &= \beta_0 + \beta_1 \log(MV) + \beta_2 \text{ipo\_ceo} + \beta_3 \text{ipo\_employee} \\ S + 0.001 &= e^{\beta_0} e^{\beta_1 \log(MV)} e^{\beta_2 \text{ipo\_ceo}} e^{\beta_3 \text{ipo\_employee}} \\ S + 0.001 &= e^{\beta_0} MV^{\beta_1} e^{\beta_2 \text{ipo\_ceo}} e^{\beta_3 \text{ipo\_employee}}\end{aligned}$$

This means that the relation between ownership and whether CEO is a (likely) founder is multiplicative. For a given firm size, the CEO owns e^{β_3} times more shares when they were already an employee at time of the IPO, and $e^{\beta_2+\beta_3}$ times more shares when they were already CEO (and thus also employee) at the time of the IPO[5]. We can easily calculate these values in R.

[5] This ignores the 0.001 adjustment but it does not really make a difference in this example.

```
beta2 <- error$coefficients["ipo_ceoTRUE"]
beta3 <- error$coefficients["ipo_employeeTRUE"]
exp(beta3)
```

```
## ipo_employeeTRUE
##         3.040505
```

```
exp(beta2 + beta3)
```

```
## ipo_ceoTRUE
##    3.494751
```

This means that the CEO who has already an employee at the time of IPO, has 3 times higher ownership and a CEO who was already CEO at the time of IPO has 3.5 times more shares than a CEO who was not in their company of similar size at the time of the IPO.

8.3.5 Better model of CEO founders

9
Assessment

9.1 Overview

There is no final or midsemester exam for this unit. You will have at least three assignments for this unit which count towards 70% of your total mark for the unit. The goal of the unit is that you are able to do the data analysis part of your thesis. It makes sense that this is what we are going to evaluate you on. All assignments will require you to analyse some real data.

The other 30% goes towards three activities that help you to prepare for your thesis.

1. The research pitch (10% - April 3)
2. The proposal (10% - June 21)
3. The proposal presentation (10% - Probably 23-24 July)

In this document, I want to set out some guidelines on the assessment for the thesis, the proposal, and the proposal presentation. Most important, the weight of the proposal and the presentation mark pale in comparison to the weight of the thesis mark for your final honours mark. Don't focus too much on the mark and the assessment criteria for the proposal and the presentation. Realise that both activities have a function: to make sure that you write an excellent thesis. The proposal forces you to work on your literature review and think about your empirical strategy. You will receive feedback on how you are progressing in terms of writing and defining your research project. The presentation requires you to present your research concisely to get feedback on your empirical strategy. This feedback should help you to write a better thesis and get a higher mark for your thesis. So, you should use the proposal and the presentation to get the most feedback as possible on how to improve your thesis.

9.2 *Proposal (5000 words)*

9.2.1 *Deadline*

Friday, June 21.

9.2.2 *Guidelines*

- Introduction and motivation (500-1000 words)
- Literature review and hypothesis development (2500-3500 words)
- Measurement and Methodology (750 - 1250 words)

The number of words are guidelines and do not have to be strictly followed. Nevertheless, I do ask that you limit yourself to 5000 words in total so that I can give quick feedback on all proposals quickly. My suggestion is to limit the literature review in the proposal to papers that are directly relevant to your research question. Your literature review in your final thesis should probably be longer. It can also be fine to have a shorter measurement and methodology section, if you are mainly following existing measures and existing statistical tests.

I will mark the proposals and provide feedback in order of submission date (earlier first) and length (shorter first).

9.2.3 *Assessment Criteria*

See the rubric on LMS in the "Assessment 1: Proposal" folder. It's the "Honours Marking Guide.xlsx" file or see the last section. Obviously, you will not be evaluated based on the presentatation and discussion of results. The other four criteria will apply to the proposal.

9.3 *Presentation (10 minutes + 20 minutes for feedback and questions)*

You have 10 minutes to present your intended research project. The academic staff will give you feedback. This means that your presentation should be focused on the elements where the audience can help you.

9.3.1 *Deadline*

Probably 23-24 July

9.3.2 *Structure*

In a 10 minute presentation, you should not have a slide with the outline of your presentation.

- Introduction.

  The introduction should make clear what your research question is, in which setting your are working, why your research question is interesting or relevant.

- Key papers.

  In your presentation, you do not need a long literature review. It's better to identify which key papers you are trying to extent, base your method on, base your theory on, ... It can be 1 paper, it can be 5 papers.

- Hypotheses or research question

  Explain the theory or arguments behind the predictions and expectations that you want to test against data.

- Data and method

  Describe your (expected) dataset and method. Focus on the elements that help you test your hypotheses or that or relatively new in the literature.

This structure allows the audience to figure out whether your proposed test and data is appropriate for your research question and whether there is other literature that could help you improve your empirical strategy.

9.3.3 Assessment

Most of the focus of the presentation is on the content but this mark will be moderated for presentation style.

9.3.4 Content

See the Honours Marking Guide.

9.3.5 Presentation style

Criterium	High Performance	Good Performance	Satisfactory	Unsatisfactory
Use of visual aids	The slides or other visual aids are consistent in style and, support the talk (through graphs, pictures, and tables) without containing the whole talk.	The visual aids or consistent in style. The aids are sometimes supportive of the talk, sometimes a mere summary of the talk. The visual aids contain some small typos or they are sometimes difficult to read.	The visual aids present an adequate summary of the talk. There is a lot of in-formation on the slides which is read out loud. The slides look sloppy and unpre-pared in some sections of the presentation.	The slides are incomplete, mostly sloppy and unprepared.
Presentation style	The presenter has good posture and makes eye contact with the audience.	The presenter makes eye contact with the audience most of the time.	The presenter tries to make eye contact with the audience.	The presenters have no eye contact with the audience and do not try. They merely look at their written preparation or the slides.

Criterium	High Performance	Good Performance	Satisfactory	Unsatisfactory
Speaking skills	The presenter speaks clearly with good pace and volume. They use normal speaking/conversational language, different from writing language.	The presenter speaks mostly clearly with adequate pacing and volume. The presenter uses normal conversational language. Sometimes they rely too much on jargon and abbreviations without explaining them.	The presenter are sometimes difficult to understand but pacing and volume are adequate. Overall, they rely too much on jargon and abbreviations without explaining them.	The presenter are hard to understand, talk too fast and/or to silent.

9.4 Thesis.

The thesis itself is evaluated based on the aforementioned honours marking guide by two academics (who are not your supervisor). We developed the assessment criteria to break down a good empirical research project in its separate elements. The criteria are a good guideline of what to focus on if you want a high mark and a good thesis. However, even if you nominally should score good on each separate element, the overall thesis might still be less than the sum of its parts which probably will be reflected in your mark. It is a good idea to not get too focused on the criteria if that distracts from conducting and writing up a good research project.

The length of your thesis will depend a lot on your topic. In a less well studied research area your literature review may well be shorter but it becomes more important to cover all the relevant literature. A simple and well established statistical model may well require less explanation but it becomes paramount to show why this simple model is sufficient for your research question.

9.4.1 Deadline

Monday 26 October 2020 at 5pm.

9.5 Pitching Document

9.6 Assignments and Homework (15%)

1. If you do homework two, you will receive at least 7.5/15 for my assignment.
2. The assignment is worth 15%

9.6.1 Homework 1

There is going to be some trial-and-error and debugging. That is fine. Carefully read the errors you get and use the resouces for help. Don't be afraid to ask me or each other for help.

1. Answer in RMarkdown format. `File > New File > R Markdown > ...` Give a name to your document and give your name as the author.

2. Search for a relevant review article on your topic and give the reference and a link to the article. Don't search to narrow!

3. Use R code chunks with backtick to

    a. Load the CEO compensation data from LMS in your file.
    b. Create a data set of the CEOs without a cash bonus in 2013
    c. Calculate the number of observations, and the average and median bonus per year. Use the examples in 4.4 to figure out how to do this.

4. Knit the report. Click on the `knit` button in Rstudio.

5. Upload the Rmarkdown and HTML version to LMS. You will have to make .zip file to be able to upload it.

9.6.2 Homework 2

Choose *three* out of four of the exercises below. If you upload the Rmarkdown and HTML file (as a .zip file) to LMS before the next seminar, you will get at least 7.5/15 for my assignment.

1. Check the conjecture that the log-log transformation makes the theoretical relation between wage-size almost linear for the largest firms.Linearity after log-log transformation

- Adapt the simulation with the `create_fake_data()` function for the same `C` and `talent_rate` values but with `obs = 10000`.
- For each simulated dataset keep only the 500 largest firms.
- You can use the variable `n` and the data manipulation functions we introduced last week
- Plot the log(wage)-log(size) relation for the four different parameters spaces
- What have you learned from this exercise?

2. One important assumption for the statistical tests of linear models is that the error terms are independently distributed. One could suspect that different years from the same firm have more similar error terms than the rest of the sample. In addition, a lot of tests in the accounting and finance literature use panel data (multiple firms, mutiple years) and assume that the relationship is more or less stable over time. Or we expect that the relationship is stable until a major change in legislation.

   - Plot the log(compensation)-log(size) relation for each year using the `facet_wrap()` function to assess whether the linear trend is visible in each year. Use the help resources if you want to know how to use `facet_wrap()`.
   - Does the linear relation hold from year to year?

3. An exploding time series. When you deal with time series, you will talk about the stationarity property of a time series. To illustrate this issue, let's simulate two time series. The simplest time series have a structure that depends on the previous obervation and some independent random noise. For simplicity let's use normally distributed noise. We can then write.

$$V_t = \beta V_{t-1} + \epsilon_t$$
$$V_t \sim \mathcal{N}(\beta V_{t-1}, \sigma)$$

   I want you to generate two time series `x` and `y` with 365 obervations each. For both time series, you can generate the noise as normally distribute with mean 0 and standard devation .5. For `x` $\beta = .9$ and for `y` $\beta = 1.1$. You can use the same random noise for both timeseries. Have a look at the `wage` simulation in Chapter 5. There are striking similarities with time series. Put both variables in a dataset (`tibble`) and plot the two time series over time in two separate plots. This means you will have to have a variable that designates the time (i.e. the day). You will need the `rnorm` function to simulation a normally distributed variable.

4. A lot of you will deal with the problem of which control variables to introduce in your statistical model. I have a lot of opinions about this. However for now, I just want to show you how simulations can help you decide whether you should include some variables. Let's assume you have the following theoretical model in mind. You have your variable of interest x (= CEO experience) and you believe that x leads to more of y (firm profit). You think that because x leads to z (optimal investment policy) and z leads to y. So we have x –> z –> y as your causal model. If this is what you think than I would argue you should not control for z when you regress x on y. Let's put some numbers on the causal model.

$$
\begin{aligned}
x &\sim \mathcal{N}(0, 1) \\
z &\sim \mathcal{N}(1 + x, 1) \\
y &\sim \mathcal{N}(1 + z, 1)
\end{aligned}
$$

   In this causal model, more of x leads ultimately to more of y. Your tests should reflect that. You will need the `rnorm` function to simulation a normally distributed variable.

   - Simulate 1000 values of x, y, z for this causal structure.
   - Run the linear regression with z as dependent variable and x as independent variable.
   - Run the linear regression with y as dependent variable and x as independent variable.
   - Run the linear regression with y as dependent variable and x and z as independent variable.
   - What have you learned from this exercise?

9.6.3 Assignment 1

The deadline for Assignment 1 is April 7

1. The CEO compensation data

variable name	Description
year	Fiscal year for this row
gvkey	a 6 digit character variable identifying the company
coname	This item represents the name of the company.
execid	Permanent Executive ID Number
page	The age of the named executive officer.

variable name	Description
joined_co	The date the named executive officer joined the company
becameceo	The date the individual became chief executive officer.
salary	The dollar value of the base salary (cash and non-cash) earned by the named executive officer during the fiscal year.
bonus	The dollar value of a bonus (cash and non-cash) earned by the named executive officer during the fiscal year.
stock_awards_fv	Grant Date Fair Value of Stock Awarded
eip_unearn_val	Equity Incentive Plan - Value of Unearned/Unvested Equity
option_awards_fv	Grant Data Fair Value of Options Granted
tdc1	Total compensation for the individual year, comprised of the following: Salary, Bonus, Other Annual, Total Value of Restricted Stock Granted, Total Value of Stock Options Granted (using Black-Scholes), Long-Term Incentive Payouts, and All Other Total.
shrown_tot_pct	Percentage of Total Shares Owned - As reported

2. The Market Value Data

variable name	Description
gvkey	A 6 digit character variable identifying the company
fyear	This item represents the fiscal year of the current fiscal year-end month. If the current fiscal year-end month falls in January through May, this item is the current calendar year minus 1 year. If the current fiscal year-end month falls in June through December, this item is the current calendar year.
mkvalt	Consolidated company-level market value is the sum of all issue-level market values, including trading and non-trading issues.

3. Yearly Data

variable name	Description
year	Year of row
sp500	Level of the SP500 at the beginning of the year
cpi	The Consumer Price Index in the US (inflation) at the beginning of the year

```
library(tidyverse)
year = c(2010: 2017)
sp500 = c(1123.58, 1282.62, 1300.58, 1480.40, 1822.36, 2028.18,
          1918.60, 2275.12)
cpi = c(100, 103.156841568622, 105.291504532867, 106.833848874866,
        108.566932118964, 108.695721960693, 110.06700893427,
        112.411557302308)
year_data = tibble(year = year, sp500 = sp500, cpi = cpi)
```

4. Helper function

To calculate the quantiles (or deciles) of a vector `x` you need the quantile function.

```
quantile(x, probs = c(0.25, .5, .75))
```

5. Assignment

Downoad and read the paper "Are CEOs Really Paid Like Bureaucrats" by Brian Hall and Jeffrey Liebman (1998) in the *Quarterly Journal of Economics*. Your assignment is simple. Use the data described above to replicate one (part of a) table in the paper with the more up-to-date data and create a figure to illustrate the same information as in the table. You do not have to replicate a whole table because some tables are more difficult and you might not have all the data to replicate every table. I will grade your work compared to the difficulty of the table. Half a difficult table or a difficult table with some small errors will get you a higher mark than a complete easy table. Make sure that I see all your code.

The plot needs to demonstrate the same information as in the table. However, a plot often accommodates more data than a table. For instance, a table can show the distribution of a variable by presenting the deciles. In a plot, you can show the full distribution with each observation as a point.

9.7 Honours Marking Guide

	Unsatisfactory (0 - 10)	Pass (10 - 14)	Distinction (14-17)	High Distinction (17 - 20)
Motivation and Significance	- The thesis is mainly motivated by data availability. - The thesis is mainly motivated as "others have investigated something similar"	The thesis identifies a gap in the existing literature without further explanation why the study is relevant.	The thesis identifies the main audience the results will speak to. It is clear that academics, practitioners, and/or policy makers care about the topic. It is not clear how the audience should use the information in the thesis.	The thesis clearly identifies the main audience the results will speak to. It is clear to an informed, non-specialist reader how the thesis will update the beliefs of academics, practitioners, and/or policy makers.

	Unsatisfactory (0 - 10)	Pass (10 - 14)	Distinction (14-17)	High Distinction (17 - 20)
Literature Review, Research Questions, and Expectations	The literature review is a list of article summaries and the connection with the research questions and expectations is lost The literature review is complete but lacks integration.	The literature review focuses only on empirical results not on the underlying theory. There is a connection between the literature review and the research question and expectations.	The literature review is a fair overview of the current literature. The literature review discusses theory and empirical results. The research question(s) and expections follow directly from the literature review.	The literature review is a fair overview of the current literature and integrates the different studies in a coherent 'story'. The research question(s) and expectations follow logically from the literature review and the underlying 'story'/theory.

	Unsatisfactory (0 - 10)	Pass (10 - 14)	Distinction (14-17)	High Distinction (17 - 20)
Description of Sample, Variables, and Method	Measurement of key variables is difficult to understand. - It is not clear which observations are in the sample and which are omitted - Descriptive statistics of key variables make no sense or they are inconsistent with prior studies	A well-informed researcher is able to replicate the key parts of the study and assess the quality of the empirical work. The descriptive statistics largely confirm the sample and measures are appropriate.	A well-informed researcher is able to replicate the study and assess the quality of the empirical work. There are some references back to the literature review. The descriptive statistics confirm the sample and measures are appropriate.	A well-informed researcher is able to replicate the study and assess the quality of the empirical work. It is clear for an informed reader why the sample, variables, and method are appropriate to investigate the resarch question. The descriptive statistics confirm the sample and measures are appropriate.

	Unsatisfactory (0 - 10)	Pass (10 - 14)	Distinction (14-17)	High Distinction (17 - 20)
Presentation and discussion of results	The thesis merely presents the results without much discussion or interpretation.	The discussion of the results refers back to the research question. It is clear whether the results support or reject the expectations	The discussion of the different results refers back to the research question. It is clear how the different results support or reject the expectations. The thesis reports the common robustness tests in the literature.	The discussion of the different results refers back to the research question and the underlying theory. It is clear how the different results support or reject the expectations. The thesis discusses and/or tests for the most important alternative explanations to the results.

	Unsatisfactory (0 - 10)	Pass (10 - 14)	Distinction (14-17)	High Distinction (17 - 20)
Writing and Style	There are many grammatical and spelling errors. The organisation of writing lacks logic and coherence. Sentences are difficult to understand and word choice is inappropriate. The thesis style is unclear.	There are minor grammatical and spelling errors. The organisation of writing lacks logic and coherence. Most sentences are easy to understand and word choice is appropriate. The thesis style is almost clear and easy to follow.	There are minor grammatical and spelling errors. The organisation of writing is almost logical and coherent. Most sentences are easy to understand and word choice is appropriate. Thesis style is always clear and easy to follow.	There are no grammatical errors. The organisation of writing is logical and coherent. Fluent sentences are easy to understand and word choice is appropriate. The thesis style is clear and easy to follow.

10
References

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Baker, George P, Michael C Jensen, and Kevin J Murphy. 1988. "Compensation and Incentives: Practice Vs. Theory." *The Journal of Finance* 43 (3). Wiley Online Library: 593–616.

Core, John E., Wayne R. Guay, and Randall S. Thomas. 2005. "Is U.S. CEO Compensation Broken?" *Journal of Applied Corporate Finance* 17 (4): 97–104. https://doi.org/10.1111/j.1745-6622.2005.00063.x.

Cunningham, Scott. 2018. *Causal Inference: The Mixtape*.

Edmans, Alex, and Xavier Gabaix. 2016. "Executive Compensation: A Modern Primer." *Journal of Economic Literature* 54 (4): 1232–87.

Efron, Bradley, and Trevor Hastie. 2017. *Computer Age Statistical Inference*.

Hall, Brian J., and Jeffrey B. Liebman. 1998. "Are CEOs Really Paid Like Bureaucrats?" *The Quarterly Journal of Economics* 113 (3): 653–91. https://doi.org/10.1162/003355398555702.

Healy, Kieran. 2017. "Fuck Nuance." *Sociological Theory* 35 (2). SAGE Publications Sage CA: Los Angeles, CA: 118–27.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. https://CRAN.R-project.org/package=stargazer.

Jensen, Michael C., and Kevin J. Murphy. 1990. "Performance Pay and Top-Management Incentives." *Journal of Political Economy* 98 (2): 225–64. https://doi.org/10.1086/261677.

Libby, Robert, Robert J Bloomfield, and M Nelson. 2002. "Experimental Research in Financial Accounting." *Accounting, Organizations and Society* 27 (8): 775–810. https://doi.org/10.1016/S0361-3682(01)00011-3.

Pearl, Judea. 2009a. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3 (0): 96–146. https://doi.org/10.1214/09-SS057.

———. 2009b. *Causality*. Cambridge University Press.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rohrer, Julia M. 2018. "Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data." *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42. https://doi.org/10.1177/2515245917745629.

Tervio, Marko. 2008. "The Difference That CEOs Make: An Assignment Model Approach." *American Economic Review* 98 (3): 642–68. https://doi.org/10.1257/aer.98.3.642.

Wickham, Hadley, Jeroen Ooms, and Kirill Müller. 2019. *RPostgres: 'Rcpp' Interface to 'Postgresql'*. https://CRAN.R-project.org/package=RPostgres.