R Lab I

Zeren Li

9/2/2019

Roadmap

- R Markdown
- Seeing theory
- Exploring CEO salary dataset
- Problem set

R Markdown

- This is an R Markdown (http://rmarkdown.rstudio.com) Notebook.
- Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents.
- R Markdown mainly consists of three parts: YAML header, texts, and `r` code chunck.
- R Markdown usually begins with a YAML header (optional) surrounded by `---`s, the header specifies meta information.
- You can write your texts with features like using header `#`, *italic*, **bold** , etc.
- When you run code within R Markdown, the results show below the chunk of code.
- You can set chunk global options that apply to every chunk in your file. This is done by calling `knitr::opts_chunk$set` in this code chunk. However, these global defaults can be overwritten in individual chunk headers.
- To understand more chunck options like `echo = TRUE`, `message = FALSE`, and `warning = FALSE`, check *RMarkdown tips and tricks*.
- Insert a new chunk: click the *Insert Chunk* button & using *Cmd+Option+I*.
- Execute chunk: click the *Run* button within the chunk or using *Cmd+Shift+Enter*.
- Click the **Knit** button to generate a document that includes both contents as well as the output of any embedded R code chunks within the document.

Seeing Theory

"**Seeing Theory** is a project designed and created by Daniel Kunin with support from Brown University's Royce Fellowship Program. The goal of the project is to make statistics more accessible to a wider range of students through interactive visualizations."

Check this: https://seeing-theory.brown.edu/basic-probability/index.html

Importing dataset

Here are various ways of importing data:

```r
# load packages
library(readr) # read csv,txt files
library(tidyverse) # powerful set of data science packages
library(haven) # read stata files
library(stargazer) # regression table
# set working directory (set your own directory)
setwd("./")

# read RData (R)
load("UNpop.RData")

# read csv
UNpop <- read_csv("./UNpop.csv") # readr package
```

Read CEO data

```r
# read dta (Stata)
ceo = read_dta("./CEOSAL2.DTA") # read CEO dataset using haven
```

View Data

```r
View(data) # View data
```

Explore CEO data

```r
class(ceo) # type of object
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```r
names(ceo) # variable names (column)
```

```
##  [1] "salary"   "age"      "college"  "grad"     "comten"   "ceoten"
##  [7] "sales"    "profits"  "mktval"   "lsalary"  "lsales"   "lmktval"
## [13] "comtensq" "ceotensq" "profmarg"
```

```r
nrow(ceo) # number of rows
```

```
## [1] 177
```

```r
ncol(ceo) # number of columns
```

```
## [1] 15
```

```r
summary(ceo)  # summarize the dataset
```

```
##      salary            age           college            grad
##  Min.   : 100.0   Min.   :33.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 471.0   1st Qu.:52.00   1st Qu.:1.0000   1st Qu.:0.0000
##  Median : 707.0   Median :57.00   Median :1.0000   Median :1.0000
##  Mean   : 865.9   Mean   :56.43   Mean   :0.9718   Mean   :0.5311
##  3rd Qu.:1119.0   3rd Qu.:62.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :5299.0   Max.   :86.00   Max.   :1.0000   Max.   :1.0000
##      comten         ceoten           sales          profits
```

```
##  Min.   : 2.0   Min.   : 0.000   Min.   :   29   Min.   :-463.0
##  1st Qu.:12.0   1st Qu.: 3.000   1st Qu.:  561   1st Qu.:  34.0
##  Median :23.0   Median : 6.000   Median : 1400   Median :  63.0
##  Mean   :22.5   Mean   : 7.955   Mean   : 3529   Mean   : 207.8
##  3rd Qu.:33.0   3rd Qu.:11.000   3rd Qu.: 3500   3rd Qu.: 208.0
##  Max.   :58.0   Max.   :37.000   Max.   :51300   Max.   :2700.0
##      mktval         lsalary          lsales          lmktval
##  Min.   :  387   Min.   :4.605   Min.   : 3.367   Min.   : 5.958
##  1st Qu.:  644   1st Qu.:6.155   1st Qu.: 6.330   1st Qu.: 6.468
##  Median : 1200   Median :6.561   Median : 7.244   Median : 7.090
##  Mean   : 3600   Mean   :6.583   Mean   : 7.231   Mean   : 7.399
##  3rd Qu.: 3500   3rd Qu.:7.020   3rd Qu.: 8.161   3rd Qu.: 8.161
##  Max.   :45400   Max.   :8.575   Max.   :10.845   Max.   :10.723
##     comtensq         ceotensq         profmarg
##  Min.   :   4.0   Min.   :   0.0   Min.   :-203.077
##  1st Qu.: 144.0   1st Qu.:   9.0   1st Qu.:   4.231
##  Median : 529.0   Median :  36.0   Median :   6.834
##  Mean   : 656.7   Mean   : 114.1   Mean   :   6.420
##  3rd Qu.:1089.0   3rd Qu.: 121.0   3rd Qu.:  10.947
##  Max.   :3364.0   Max.   :1369.0   Max.   :  47.458
```

```
summary(ceo$salary)  # summarize the variable
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.0   471.0   707.0   865.9  1119.0  5299.0
```

```
length(ceo) # length of a dataset means # of variables
```

```
## [1] 15
```

```
length(ceo$salary)  # length of a variable means # of obs
```

```
## [1] 177
```

```
head(ceo) # show the first 5 rows of the dataset
```

```
## # A tibble: 6 x 15
##   salary   age college  grad comten ceoten sales profits mktval lsalary
##    <dbl> <dbl>   <dbl> <dbl>  <dbl>  <dbl> <dbl>   <dbl>  <dbl>   <dbl>
## 1   1161    49       1     1      9      2  6200     966  23200    7.06
## 2    600    43       1     1     10     10   283      48   1100    6.40
## 3    379    51       1     1      9      3   169      40   1100    5.94
## 4    651    55       1     0     22     22  1100     -54   1000    6.48
## 5    497    44       1     1      8      6   351      28    387    6.21
## 6   1067    64       1     1      7      7 19000     614   3900    6.97
## # ... with 5 more variables: lsales <dbl>, lmktval <dbl>, comtensq <dbl>,
## #   ceotensq <dbl>, profmarg <dbl>
```

```
table(ceo$grad) # show the frequency of a categorical variable
```

```
##
##  0  1
## 83 94
```

```
ceo_grate <- ceo[ceo$grad == 1,] # fiter by condition(s)

ceo_over_1kk <- ceo[ceo$salary > 1000,] # fiter by conditionn(s)
```

```r
ceo_1to5 <- ceo[c(1:5), ] # filter by index

ceo_1 <- ceo[,c("salary", "profmarg")] # select by variable name

ceo_var1to5 <- ceo[,c(1:5)] # select by index

# rename variable
names(ceo_1)
```

```
## [1] "salary"   "profmarg"
```

```r
names(ceo_1)[2] <- "profit_margin"
names(ceo_1)
```

```
## [1] "salary"        "profit_margin"
```

```r
rm(ceo_1) # remove dataset
```

A bit beautiful summary statistics...

```r
stargazer(data.frame(ceo))  # summarize the variable
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Sep 05, 2019 - 10:57:16

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
salary	177	865.864	587.589	100	471	1,119	5,299
age	177	56.429	8.422	33	52	62	86
college	177	0.972	0.166	0	1	1	1
grad	177	0.531	0.500	0	0	1	1
comten	177	22.503	12.295	2	12	33	58
ceoten	177	7.955	7.151	0	3	11	37
sales	177	3,529.463	6,088.654	29	561	3,500	51,300
profits	177	207.831	404.454	−463	34	208	2,700
mktval	177	3,600.316	6,442.276	387	644	3,500	45,400
lsalary	177	6.583	0.606	4.605	6.155	7.020	8.575
lsales	177	7.231	1.432	3.367	6.330	8.161	10.845
lmktval	177	7.399	1.133	5.958	6.468	8.161	10.723
comtensq	177	656.684	577.123	4	144	1,089	3,364
ceotensq	177	114.124	212.566	0	9	121	1,369
profmarg	177	6.420	17.861	−203.077	4.231	10.947	47.458

Mean and Variance

population mean:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

```
m_salary <- sum(ceo$salary)/length(ceo$salary)
m_salary
```

```
## [1] 865.8644
```

```
mean(ceo$salary)
```

```
## [1] 865.8644
```

population variance:

$$\sigma^2 = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

```
sum( (ceo$salary - m_salary)^2 ) / (length(ceo$salary)-1)
```

```
## [1] 345261.2
```

```
var(ceo$salary) # R computes sample variance
```

```
## [1] 345261.2
```

Covariance & Correlation

population covariance:

$$Cov(X,Y) = E[(X - E(X)(Y - E(Y))]$$

sample covariance:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

We would like to look at the covariance and correlation between CEO's salary and firm performance measured by profit margins.

```
cov(ceo$salary,ceo$profmarg) # covariance
```

```
## [1] -303.6705
```

```
m_profmarg = sum(ceo$profmarg)/length(ceo$profmarg)
sum((ceo$salary - m_salary) * (ceo$profmarg - m_profmarg ))/(length(ceo$profmarg) - 1)
```

```
## [1] -303.6705
```

$$Corr(X,Y) = \frac{E[(X - E(X)E(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

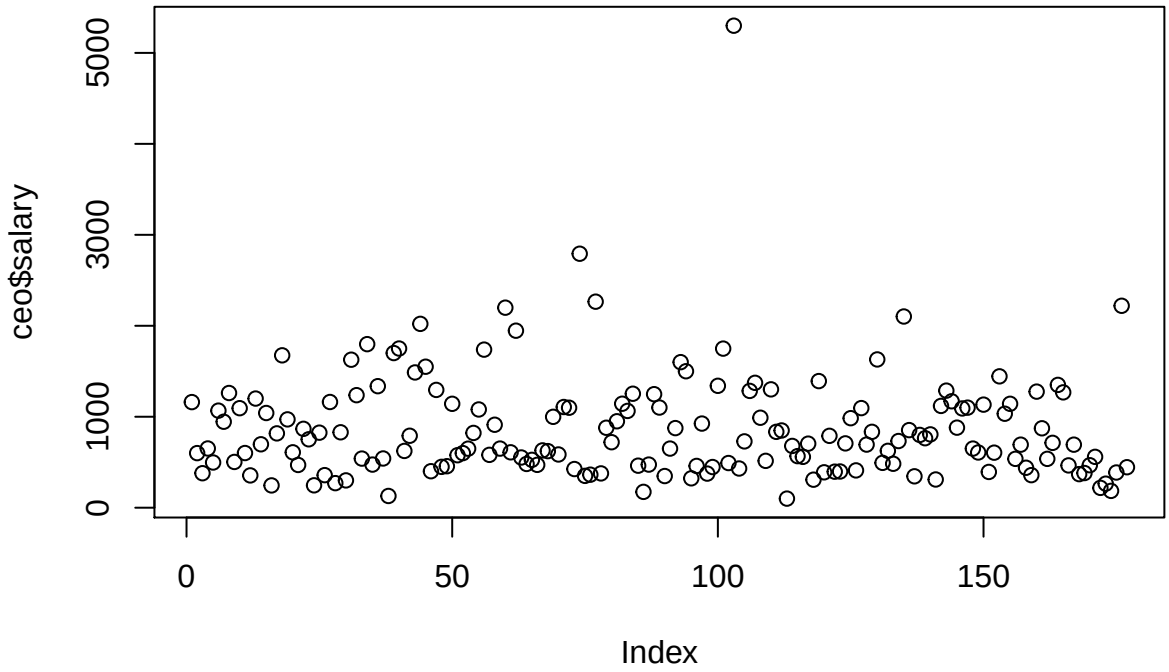
```
cor(ceo$salary,ceo$profmarg) # correlation
```

```
## [1] -0.02893538
```

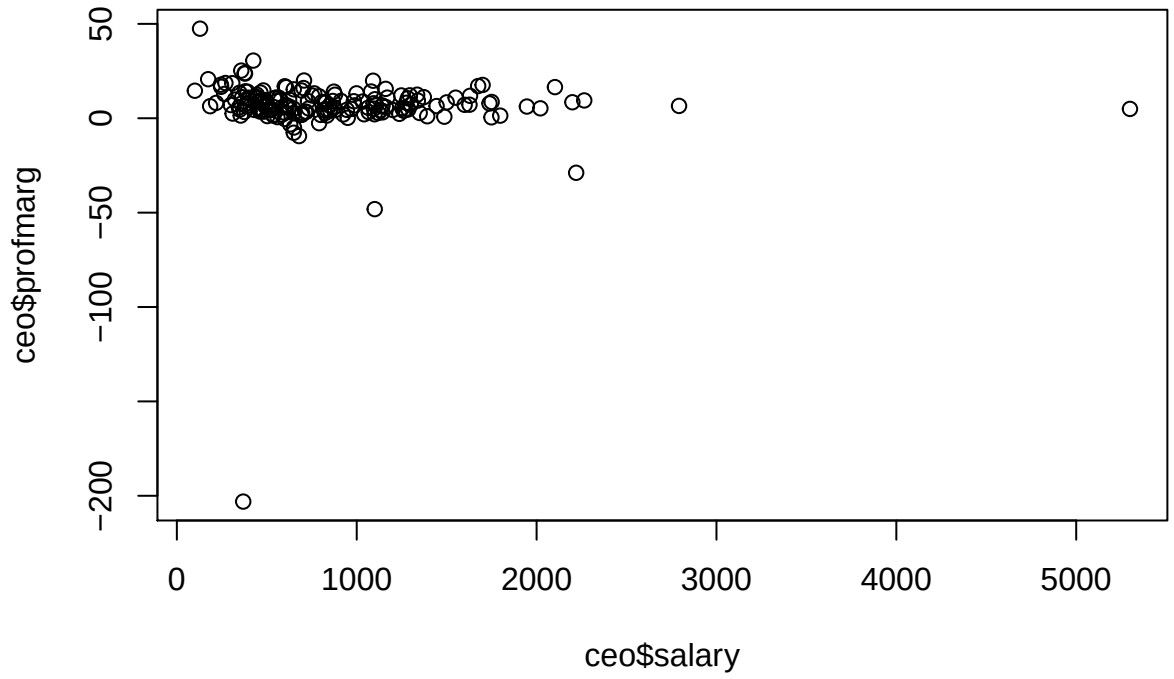
```
# How to compute manually?
# Show it in the problem set, it should be the same as the result from cor()
```

R graph

```
plot(ceo$salary) # one-way scatterplot
```

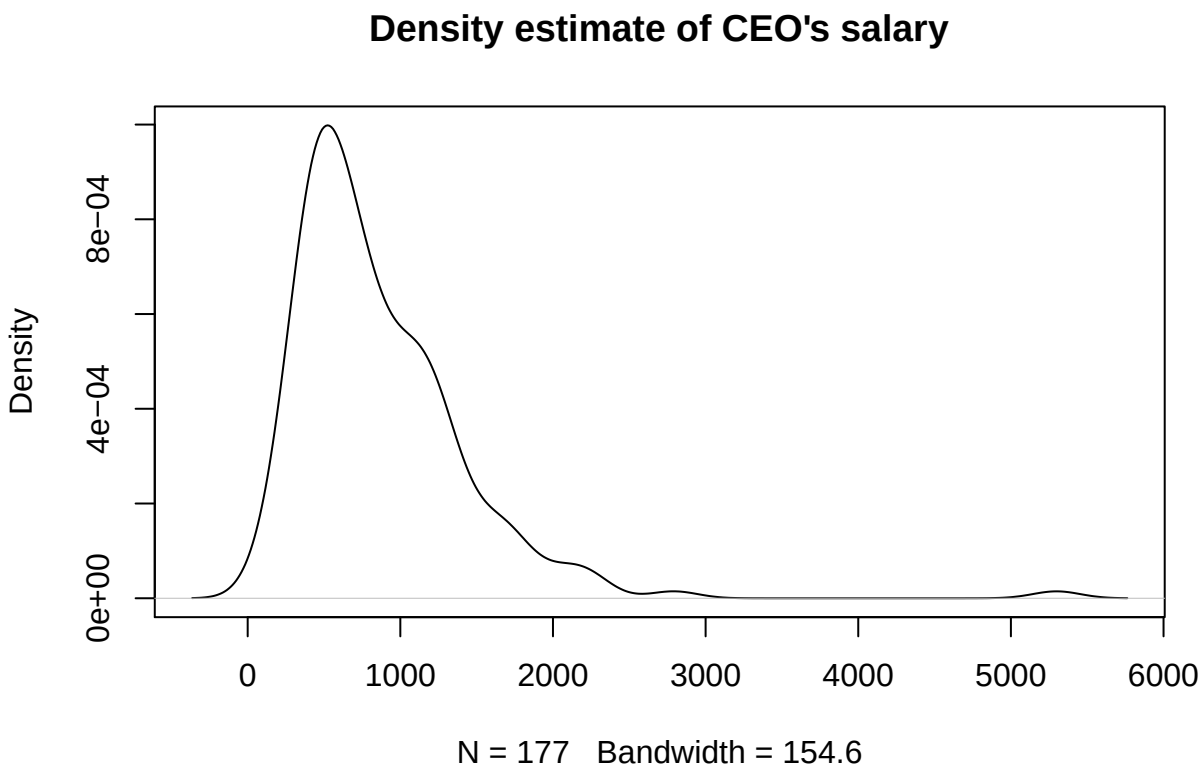



```
plot(ceo$salary, ceo$profmarg) # two-way scatterplot
```

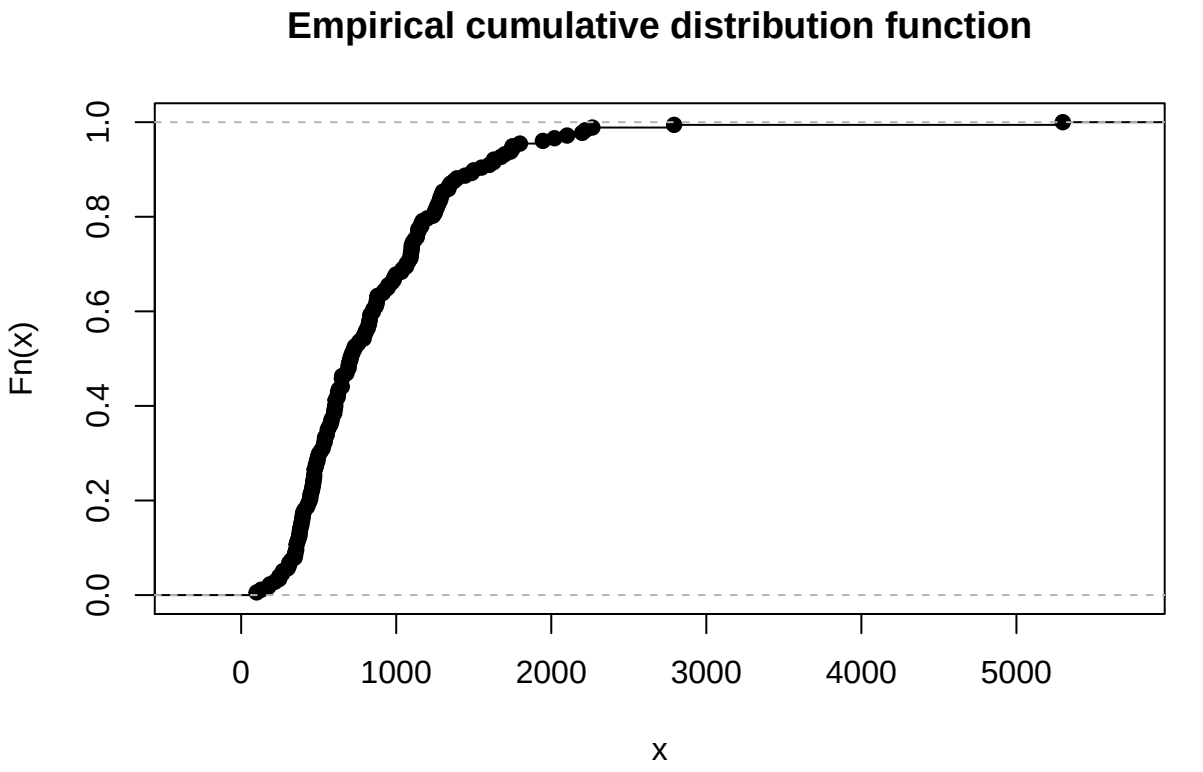



```r
hist(ceo$salary, main = "Histogram of CEO's salary") # histogram
```


```r
plot(density(ceo$salary), main = "Density estimate of CEO's salary") # pdf
```




```
plot(ecdf(ceo$salary),main = "Empirical cumulative distribution function") # cdf
```



Other resources

Installing RMarkdown: https://bookdown.org/yihui/rmarkdown/

Frequently asked questions: https://yihui.name/knitr/faq/

RMarkdown cheatsheet: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf

R Style: http://adv-r.had.co.nz/Style.html