Regression 2: Implementation in R

Instructor: Yuta Toyama

Last updated: 2020-03-30

Section 1

Introduction

Acknowledgement

This note is based on “Introduction to Econometrics with R”.
https://www.econometrics-with-r.org/index.html

Preliminary: packages

- We use the following packages:
  - `AER` :
  - `dplyr` : data manipulation
  - `stargazer` : output of regression results

```
# Install package if you have not done so
# install.packages("AER")
# install.packages("dplyr")
# install.packages("stargazer")
# install.packages("lmtest")

# load packages
library("AER")
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest
```

Empirical setting: Data from California School

- Question: How does the student-teacher ratio affects test scores?
- We use data from California school, which is included in `AER` package.
  - See here for the details: https://www.rdocumentation.org/packages/AER/versions/1.2-6/topics/CASchools

```
# load the the data set in the workspace
data(CASchools)
```

- Use `class()` function to see `CASchools` is `data.frame` object.

```
class(CASchools)
```

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

- We take 2 steps for the analysis.
  - Step 1: Look at data (descriptive analysis)
  - Step 2: Run regression

Step 1: Descriptive analysis

- It is always important to grasp your data before running regression.
- `head()` function give you a first overview of the data.

```
head(CASchools)
```

```
## # A tibble: 6 x 14
##   district school county grades students teachers calworks
##   <chr>    <chr>  <fct>  <fct>     <dbl>    <dbl>    <dbl>
## 1 75119    Sunol~ Alame~ KK-08       195    10.9      0.510
## 2 61499    Manza~ Butte  KK-08       240    11.1     15.4
## 3 61549    Therm~ Butte  KK-08      1550    82.9     55.0
## 4 61457    Golde~ Butte  KK-08       243    14       36.5
## 5 61523    Paler~ Butte  KK-08      1335    71.5     33.1
## 6 62042    Burre~ Fresno KK-08       137     6.40    12.3
## # ... with 5 more variables: expenditure <dbl>, income <dbl
## #   read <dbl>, math <dbl>
```

- Alternatively, you can use `browse()` to see the entire dataset in browser

Create variables

- Create several variables that are needed for the analysis.
- We use `dplyr` for this purpose.

```
CASchools %>%
  mutate( STR = students / teachers ) %>%
  mutate( score = (read + math) / 2 ) -> CASchools
```

Descriptive statistics

- There are several ways to show descriptive statistics
- The standard one is to use `summary()` function

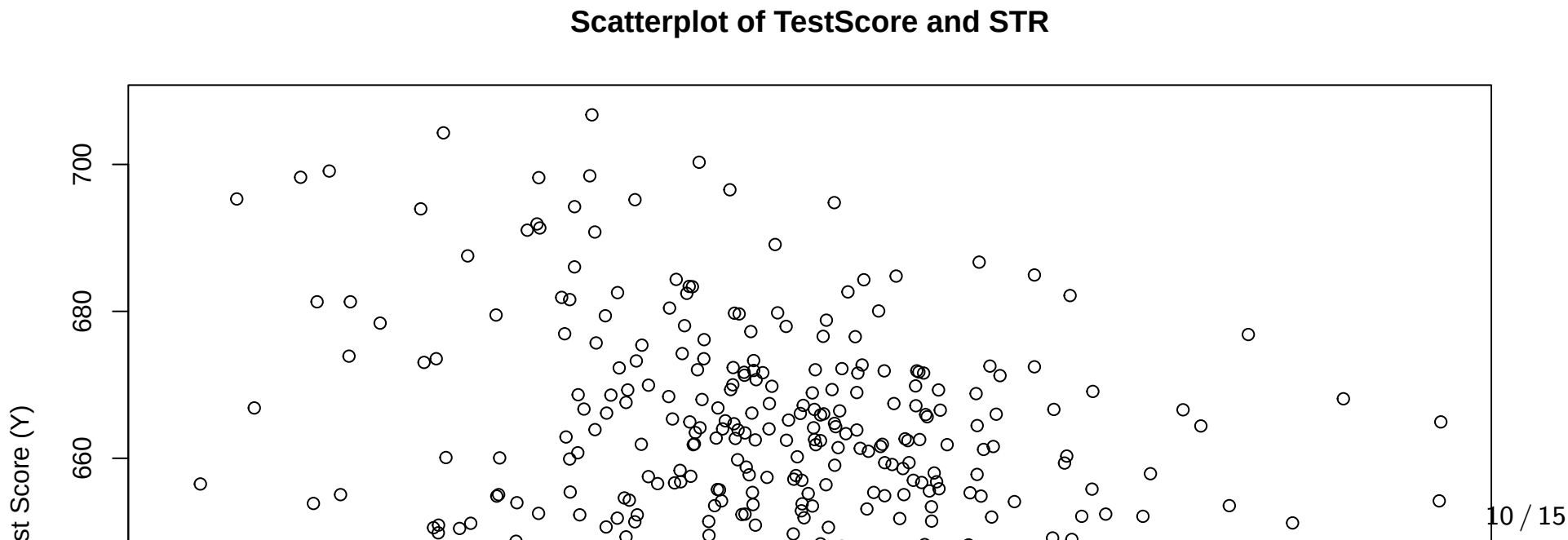
```
summary(CASchools)
```

```
##    district                    school                      county
##  Length:420         Length:420         Sonoma      : 29   KK
##  Class :character   Class :character   Kern        : 27   KK
##  Mode  :character   Mode  :character   Los Angeles: 27
##                                        Tulare      : 24
##                                        San Diego   : 21
##                                        Santa Clara: 20
##                                        (Other)     :272
##     students          teachers         calworks           l
##  Min.   :   81.0   Min.   :   4.85   Min.   : 0.000   Min.
##  1st Qu.:  379.0   1st Qu.:  19.66   1st Qu.: 4.395   1st Q
##  Median :  950.5   Median :  48.56   Median :10.520   Media
```

Scatter plot

- Let's see how test score and student-teacher-ratio is correlated.

```
plot(score ~ STR,
     data = CASchools,
     main = "Scatterplot of TestScore and STR",
     xlab = "STR (X)",
     ylab = "Test Score (Y)")
```



Step 2: Run regression

Simple linear regression

- We use `lm()` function to run linear regression
- First, consider the simple linear regression

$$score_i = \beta_0 + \beta_1 size_i + \epsilon_i$$

  where $size_i$ is the class size (student-teacher-ratio).
  - From now on we call student-teacher-ratio (STR) class size.
- To run this regression, we use `lm`

```
# First, we rename the variable `STR`
CASchools %>%
  dplyr::rename( size = STR) -> CASchools

# Run regression and save results in the varaiable `model1_sum
model1_summary <- lm( score ~ size, data = CASchools)

# See the results
summary(model1_summary)
```

Correction of Robust standard error

- We use `vcovHC()` function, a partof the package `sandwich`, to obtain the robust standard errors.
  - The package `sandwich` is automatically loaded if you load `AER` package.

```
# compute heteroskedasticity-robust standard errors
vcov <- vcovHC(model1_summary, type = "HC1")

# get standard error: the square root of the diagonal element
robust_se <- sqrt(diag(vcov))
robust_se
```

```
## (Intercept)        size
##  10.3643617   0.5194893
```

- Notice that robust standard errors are larger than the one we obtained from `lm`!
- How to combine the robust standard errors with the original summary? Use `coeftest()` from the package `lmtest`

Report by Stargazer

- stargazer is useful to show the regression result.

```
# load
library(stargazer)

# Create output by stargazer
stargazer::stargazer(model1_summary, type ="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                score
## -----------------------------------------------
## size                          -2.280***
##                               (0.480)
##
## Constant                      608.933***
```

Full results

Taken from https://www.econometrics-with-r.org/7-6-analysis-of-the-test-score-data-set.html

```
# load the stargazer library

# estimate different model specifications
spec1 <- lm(score ~ size, data = CASchools)
spec2 <- lm(score ~ size + english, data = CASchools)
spec3 <- lm(score ~ size + english + lunch, data = CASchools)
spec4 <- lm(score ~ size + english + calworks, data = CASchool
spec5 <- lm(score ~ size + english + lunch + calworks, data =

# gather robust standard errors in a listh
rob_se <- list(sqrt(diag(vcovHC(spec1, type = "HC1"))),
               sqrt(diag(vcovHC(spec2, type = "HC1"))),
               sqrt(diag(vcovHC(spec3, type = "HC1"))),
               sqrt(diag(vcovHC(spec4, type = "HC1"))),
               sqrt(diag(vcovHC(spec5, type = "HC1"))))
```