EC282: Homework Assignments

Onur Altındağ

Last update: 2020-03-23

Contents

Chapter 1

General Rules and Principles

1.1 Installing R and RStudio

Here are the instructions for installing R and RStudio on your Windows or Mac desktop. Skip the third part and do not install "SDSFoundations Package".

1.2 Basic rules and best practices

All files should exist in a local folder that syncs to a cloud-storage service. No file you ever work on should be at risk of being lost if your computer ceases to function or be in your possesion. NEVER place any file on "downloads" or "desktop" folders.

Get a free cloud-storage service with a desktop application that syncs to a cloud-storage service. I like the Dropbox desktop app but feel free to choose any other service. You don't need a lot of space so free version of any desktop cloud app would work. Under the Dropbox folder, create a designated folder for this course such as EC282.

All subfolders under EC282 and files in them should have unique and descriptive construction: DON'T use spaces in file or folder names. Here is an example of a folder structure that might work for a student in this class:

```
EC282


  Course_docs
        SyllabusEC282.pdf
        LectureNotes.pdf
  Assignments
      Assignment1
           dataset1name.Rda
           Lastname_Firstname_Assignment1_EC282.R
      Assignment2
           dataset2name.Rda
           Lastname_Firstname_Assignment1_EC282.R
      Assignment3
           dataset3name.Rda
           Lastname_Firstname_Assignment3_EC282.R
      Assignment4
           dataset4name.Rda
           Lastname_Firstname_Assignment4_EC282.R
           ...
  Exams
```

```
      Midterm1
            Midterm1Review.pdf
            Midterm1Review_myanswers.docx
|       |   ...
```

1.3 Header

At the beginning of any R script, you should have a standard header that you use across all scripts that clears the workspace, loads/installs packages as necessary, sets the working directory, etc. Here is an example that you can copy paste to the header of any script that you use:

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'MASS', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to your birthday for replication
set.seed(06081998)
################################################################################
```

When coding, use relative references to files. Typically, any script will begin looking for files in the working directory. At any time you can type `getwd()` on your Rstudio console to see the current working directory. The header above automatically sets the working directory to the folder that the R script is included. For example, if you are working on `Lastname_Firstname_Assignment1_EC282.R` script and need to load file `dataset1name.Rda` into an object, then you would simply run:

```
load(dataset1name.Rda)
```

However, if you were working in the same .R file, and needed to access `dataset2name.Rda`, you would need to point the program to a directory outside the current working directory – so, you go up one level, over one folder, and look there:

```
load(../Assignment2/dataset2name.Rda)
```

When learning R, the most important skill that you need to acquire is to be able to **google** your problem. There is probably not a single R question that you have yet has not been answered on Stack Overflow.

1.4 How to submit the homework assignment

You can use a snipping tool to copy and paste the relevant output and figures from RStudio console to a word file, type the answers, save everything and upload it on BlackBoard/Assignments. If you want to have a more elegant

looking homework output the `stargazer` package is very powerful in transforming your analysis into publishable formats.

Chapter 2

Homework Assigment I

Deadline: Feb 16, 2020, 11 PM

Source: Stock and Watson, 3rd Updated Edition. Exercise 3.1

Data description: You can find the data description here.

Question I

a. Compute the sample mean for average hourly earnings (`ahe`) in 1992 and 2008. Construct a 95% confidence internal for the population means of `ahe` in 1992 and 2008 and the change between 1992 and 2008.

b. In 2008, the values of the Consumer Price Index (CPI) was 215.2. In 1992, the value of the CPI was 140.3. Repeat (a) but use AHE measured in real 2008 dollars ($2008); that is, adjust the 1992 data for the price inflation that occured between 1992 and 2008.

c. If you were interested in the change in workers' purchasing power from 1992 to 2008, would you use the results from (a) or (b)? Explain.

d. Use the 2008 data to construct a 95% confidence interval for the mean of `ahe` for high school graduates. Construct a 95% confidence interval for the mean of `ahe` for workers with a college degree. Construct a 95% confidence interval for the difference between the two means.

e. Repeat (d) using the 1992 data expressed in $2008.

f. Did real (inflation-adjusted) wages of high school graduates increased from 1992 to 2008? Explain. Did real wages of college graduates increase? Did the gap between earnings of college and highschool graduates increase? Explain, using appropriate estimates, confidence intervals, and test statistics.

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment1` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your enviroment. Conduct the analysis below the header.

```
###############################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'MASS', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
```

```r
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication
set.seed(06061983)
################################################################################
#get the data url
df1.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/cps92_08.xlsx'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf)
head(df1)

#CONDUCT THE ANALYSIS BELOW
```

Chapter 3

Homework Assignment II

Deadline: March 8, 2020, 11 PM

Source: Stock and Watson, 3rd Updated Edition. Exercises 4.1 and 4.2

Data description: You can find the data descriptions for Question I here and for Question II here.

Question I

a. Run a regression of average hourly earnings (`ahe`) on `age`. Report the estimated intercept and the slope. Interpret each of the estimated coefficients.

b. Bob is a 26-year-old worker. Predict Bob's earnings using the estimated regression. Alexis is a 30-year-old worker. Predict Alex's earnings using the estimated regression.

c. Does age explain a large fraction of the variation in earnings across individuals? Explain.

Question II

a. Construct a scatterplot of average course evaluations – `course_eval` on the professor's `beauty`. Interpret the relationship based on your graph.

b. Run a regression of average course evaluations `course_eval` on the professor's beauty `beauty`. Report the estimated intercept and the slope. Report the sample means of `beauty` and `course_eval`. Explain why the estimated intercept is equal to the sample mean of `course_eval`.

c. Predict a course evaluation for a professor whose `beauty` is one standard deviation above the average. Compare this to the estimated intercept in (b) and explain the difference.

d. Interpret the size of the slope coefficient in (b). Is the estimated "effect" of beauty on course evaluations large or small?

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment2` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data sets `df1` and `df2` in your enviroment. Conduct the analysis below the header.

```
#############################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'MASS', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))
```

```r
# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication
set.seed(06061983)
################################################################################
#get the data urls
df1.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/cps08.xlsx'
df2.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/TeachingRatings.
#download the data
GET(df1.url, write_disk(tdf1 <- tempfile(fileext = ".xlsx")))
GET(df2.url, write_disk(tdf2 <- tempfile(fileext = ".xls")))
#check if it worked
df1 <- read_excel(tdf1)
df2 <- read_excel(tdf2)
head(df1)
head(df2)

#CONDUCT THE ANALYSIS BELOW
```

Chapter 4

Homework Assignment III

Deadline: March 22, 2020

Source: Stock and Watson, 3rd Updated Edition. Exercises 5.1 and 5.2

Data description: You can find the data descriptions for Question I here and for Question II here.

Question I

a. Run a regression of average hourly earnings `ahe` on `age` and report the regression output.

b. Is the estimated coefficient significant? That is, can you reject the null hypothesis $H_0 : \beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

c. Construct a 95% confidence interval for the slope coefficient.

d. Repeat (a) and (b) only using the data for college graduates.

Question II

a. Run a regression of `course_eval` on `beauty` and report the regression output.

b. Is the estimated coefficient significant? That is, can you reject the null hypothesis $H_0 : \beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment3` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data sets `df1` and `df2` in your enviroment. Conduct the analysis below the header.

```
###############################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'MASS', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory
```

```r
#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication
set.seed(0122124)
################################################################################
#get the data urls
df1.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/cps08.xlsx'
df2.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/TeachingRatings.
#download the data
GET(df1.url, write_disk(tdf1 <- tempfile(fileext = ".xlsx")))
GET(df2.url, write_disk(tdf2 <- tempfile(fileext = ".xls")))
#check if it worked
df1 <- read_excel(tdf1)
df2 <- read_excel(tdf2)
head(df1)
head(df2)

#CONDUCT THE ANALYSIS BELOW
```

Chapter 5

Homework Assignment IV

Deadline: April 12, 2020, 11 PM

Source: Stock and Watson, 3rd Updated Edition. Exercises 6.1 and 6.3.

Data description: You can find the data descriptions for Question I here and for Question II here.

Question I

a. Run a regression of `course_eval` on `beauty`. On a second regression, add the following control variables: `intro`, `onecredit`, `female`, `minority` and `nnenglish`. Report both regression outputs. Compare the estimated "effect" of `beauty` in the first to the second regression? Does the first estimated slope change substantially after adding the control variables to the model? What does that indicate?

b. Predict the outcome for a black male professor with average beauty and is a native English speaker. He teaches a three-credit upper-divisiocourse.

Question II

a. Drop the observation for Malta from the analysis data set.

b. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series `growth`, `tradeshare`, `yearsschool`, `oil`, `rev_coups`, `assasinations`, and `rgdp60`.

c. Run a regression of `growth` on `tradeshare`, `yearsschool`, `rev_coups`, `assasinations`, and `rgdp60`. Report the regression results in a table and interpret the coefficient on `rev_coups`.

d. Use the regression results to predict the average annual growth rate for a country that has average values for all regressors.

e. Include `oil` to regression (c) and interpret any **major** changes in regression (c).

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment4` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data sets `df1` and `df2` in your enviroment. Conduct the analysis below the header.

```
#########################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'MASS', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))
```

```r
# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication
set.seed(06061983)
################################################################################
#get the data urls
df1.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/TeachingRatings.
df2.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/Growth.xls'

#download the data
GET(df1.url, write_disk(tdf1 <- tempfile(fileext = ".xls")))
GET(df2.url, write_disk(tdf2 <- tempfile(fileext = ".xls")))

#check if it worked
df1 <- read_excel(tdf1)
df2 <- read_excel(tdf2)
head(df1)
head(df2)

#CONDUCT THE ANALYSIS BELOW
```

Chapter 6

Homework Assignment V

Deadline: April 26, 2020, 11 PM

Source: Stock and Watson, 3rd Updated Edition. Exercise 8.1

Data description: You can find the data descriptions for Question I here.

QUESTION I

a. Run a regression of average hourly earnings (`ahe`) on `age`,`female`, and `bachelor` and report the output. If `age` increases from 25 to 26, how are earnings expected to change? If `age` increases from 33 to 34, how are earnings expected to change?

b. Run a regression of the **logarithm** of average hourly earnings (`ln_ahe`) on `age`,`female`, and `bachelor` and report the output. If `age` increases from 25 to 26, how are earnings expected to change? If `age` increases from 33 to 34, how are earnings expected to change?

c. Run a regression of the **logarithm** of average hourly earnings (`ln_ahe`) on the **logarithm** of `ln_age`,`female`, and `bachelor` and report the output. If `age` increases from 25 to 26, how are earnings expected to change? If `age` increases from 33 to 34, how are earnings expected to change?

d. Run a regression of the **logarithm** of average hourly earnings (`ln_ahe`) on `age`, square of age (`age_sq`), `female`, and `bachelor` and report the output. If `age` increases from 25 to 26, how are earnings expected to change? If `age` increases from 33 to 34, how are earnings expected to change?

e. Comparing the regression results from (a),(b),(c),(d), choose one of the empirical models based on economic theory. Briefly explain why you choose the preferred model.

f. Run a regression of the **logarithm** of average hourly earnings (`ln_ahe`) on `age`, square of age (`age_sq`), `female`, and `bachelor` and the interaction term `female` $\times$ `bachelor`. Report the output. Consider the following individuals:

- Alexis: 30-year-old female with a bachelor's degree.
- Jane: 30-year-old female with a high school degree.
- Bob: 30-year-old male with a bachelor degree.
- Jim: 30-year-old male with a high school degree.

Using the regression results, predict `ln_ahe` and `ahe` for each individual. Calculate the college premium (predicted difference in log wage) for females and college premium for males. Is the college premium differ for female vs. males? Explain.

g. Is the effect of `age` on earnings different for man than women? Using the variables `ln_ahe`, `female`, `bachelor`, and `age`. Specify and estimate a regression that you can use to answer this question. Report the regression and explain your results.

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment5` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data sets `df1` and `df2` in your enviroment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'MASS', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication
set.seed(06061983)
################################################################################
#get the data urls
df1.url <- 'https://wps.pearsoned.com/wps/media/objects/11422/11696965/empirical/empex_tb/cps08.xlsx'
#download the data
GET(df1.url, write_disk(tdf1 <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf1)
head(df1)


#CONDUCT THE ANALYSIS BELOW
```

Chapter 7

Lecture notes, updates, and corrections:

7.1 Week1

Define a binary random variable Y for flipping a fair coin. Calculate the expected value and the variance. Flip the the coin for $\times 10$ and $\times 10000$, calculate the sample mean and sample variance.

```
#Bernouilli trails
coin <- c("H","T")
#flip a coin
sample(coin,1)
```

```
## [1] "H"
```

```
#flip a coin 10 times
sample(coin,10, replace=TRUE)
```

```
##  [1] "H" "T" "T" "T" "H" "T" "H" "T" "H" "H"
```

```
#create the random variable Y

out1 <- sample(coin,10, replace=TRUE)
Y1 <- as.data.frame(out1)
Y1$val <- 0
Y1$val[Y1$out1=="T"] <- 1

out2 <- sample(coin,10000, replace=TRUE)
Y2 <- as.data.frame(out2)
Y2$val <- 0
Y2$val[Y2$out2=="T"] <- 1

mean(Y1$val)
```

```
## [1] 0.5
```

```
var(Y1$val)
```

```
## [1] 0.2777778
```

```r
mean(Y2$val)
```

```
## [1] 0.4963
```

```r
var(Y2$val)
```

```
## [1] 0.2500113
```

7.2 Week2

If $Y \sim N(1, 4)$, standardize Y, interpret and calculate $Pr \leq 2$ Find the upper and lower tail range that yield 95% probability within the interval of these two values. Calculate $Pr(1 \leq Y \leq 2)$.

```r
pnorm(2,mean=1,sd=2, lower.tail = TRUE)
```

```
## [1] 0.6914625
```

```r
pnorm(2,mean=1,sd=2, lower.tail = TRUE) - pnorm(1,mean=1,sd=2, lower.tail = TRUE)
```

```
## [1] 0.1914625
```

```r
#the tails
qnorm(0.025, mean=1, sd=2, lower.tail = TRUE)
```

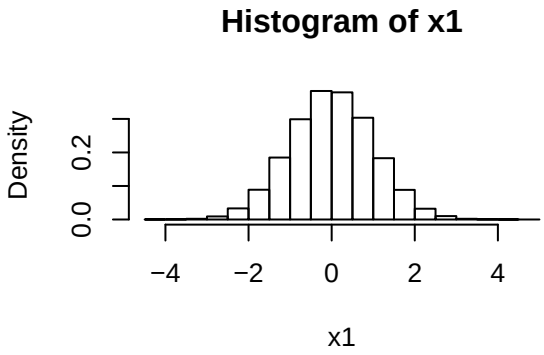
```
## [1] -2.919928
```

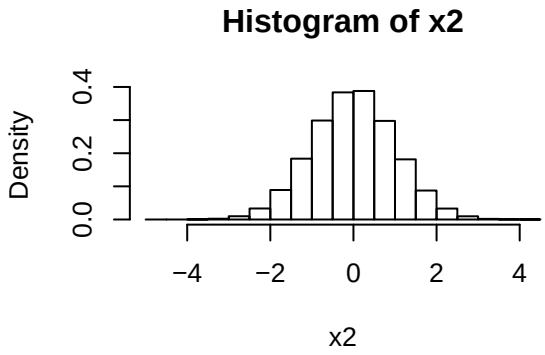
```r
qnorm(0.975, mean=1, sd=2, lower.tail = TRUE)
```

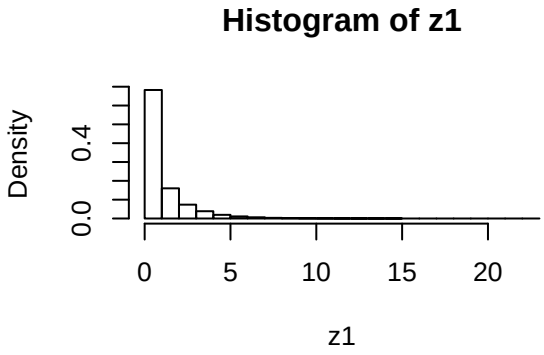
```
## [1] 4.919928
```

Generate normally distributed random variables and variable with chi-squared

```r
#x1 is a normally distributed RV
x1 <- rnorm(100000,0,1)
#x2 is a normally distributed RV
x2 <- rnorm(100000,0,1)
#z1 is the square of x1
z1 <- x1^2
#z2 is the sum of squared xs
z2 <- x1^2 + x2^2

par(mfrow=c(2,2))
hist(x1, prob=TRUE)
hist(x2, prob=TRUE)
hist(z1, prob=TRUE)
hist(z2, prob=TRUE)
```

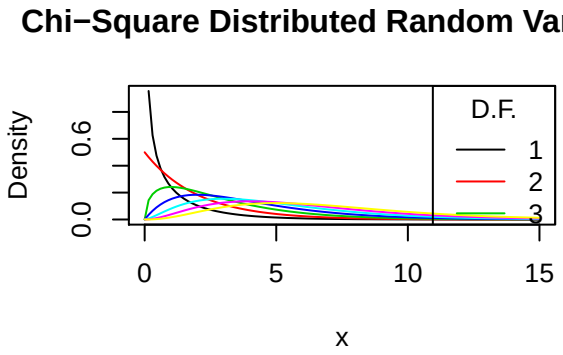


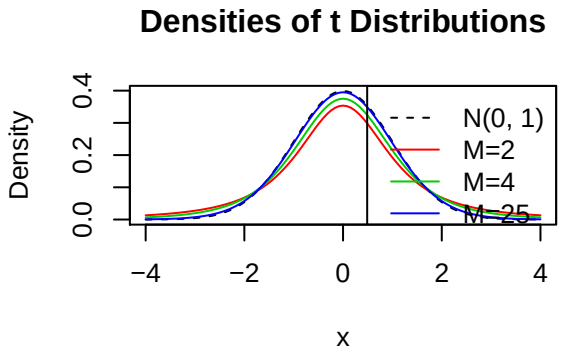


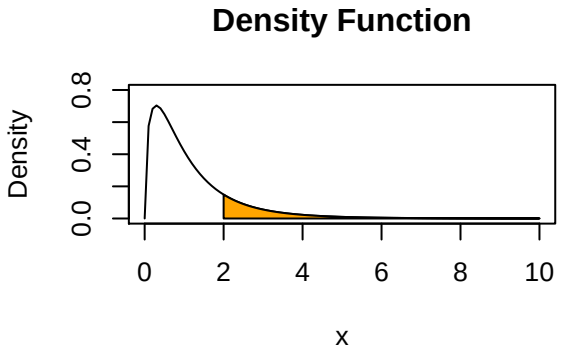



```r
# plot the density for M=1
curve(dchisq(x, df = 1),
      xlim = c(0, 15),
      xlab = "x",
      ylab = "Density",
      main = "Chi-Square Distributed Random Variables")

# add densities for M=2,...,7 to the plot using a 'for()' loop
for (M in 2:7) {
  curve(dchisq(x, df = M),
        xlim = c(0, 15),
        add = T,
        col = M)
}

# add a legend
legend("topright",
       as.character(1:7),
       col = 1:7 ,
       lty = 1,
       title = "D.F.")


# plot the standard normal density
curve(dnorm(x),
      xlim = c(-4, 4),
      xlab = "x",
      lty = 2,
```

```
      ylab = "Density",
      main = "Densities of t Distributions")

# plot the t density for M=2
curve(dt(x, df = 2),
      xlim = c(-4, 4),
      col = 2,
      add = T)

# plot the t density for M=4
curve(dt(x, df = 4),
      xlim = c(-4, 4),
      col = 3,
      add = T)

# plot the t density for M=25
curve(dt(x, df = 25),
      xlim = c(-4, 4),
      col = 4,
      add = T)

# add a legend
legend("topright",
       c("N(0, 1)", "M=2", "M=4", "M=25"),
       col = 1:4,
       lty = c(2, 1, 1, 1))


# define coordinate vectors for vertices of the polygon
x <- c(2, seq(2, 10, 0.01), 10)
y <- c(0, df(seq(2, 10, 0.01), 3, 14), 0)

# draw density of F_{3, 14}
curve(df(x ,3 ,14),
      ylim = c(0, 0.8),
      xlim = c(0, 10),
      ylab = "Density",
      main = "Density Function")

# draw the polygon
polygon(x, y, col = "orange")
```





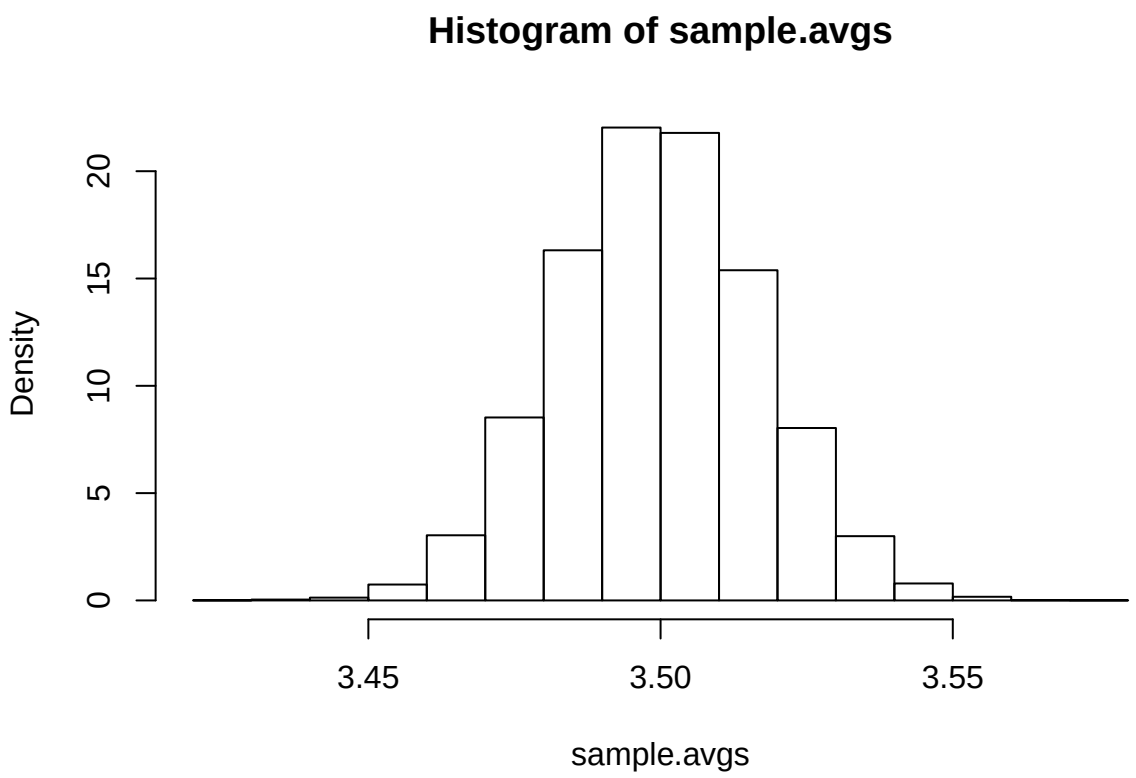



```r
# Vector of outcomes
S <- 2:12

# Vector of probabilities
PS <- c(1:6, 5:1) / 36

# divide the plotting area into one row with two columns
par(mfrow = c(1, 2))

# plot the distribution of S
barplot(PS,
        ylim = c(0, 0.2),
        xlab = "S",
        ylab = "Probability",
        col = "white",
        space = 0,
        main = "Sum of Two Dice Rolls")

# plot the distribution of D
probability <- rep(1/6, 6)
names(probability) <- 1:6

barplot(probability,
        ylim = c(0, 0.2),
        xlab = "D",
        col = "white",
        space = 0,
        main = "Outcome of a Single Dice Roll")
```

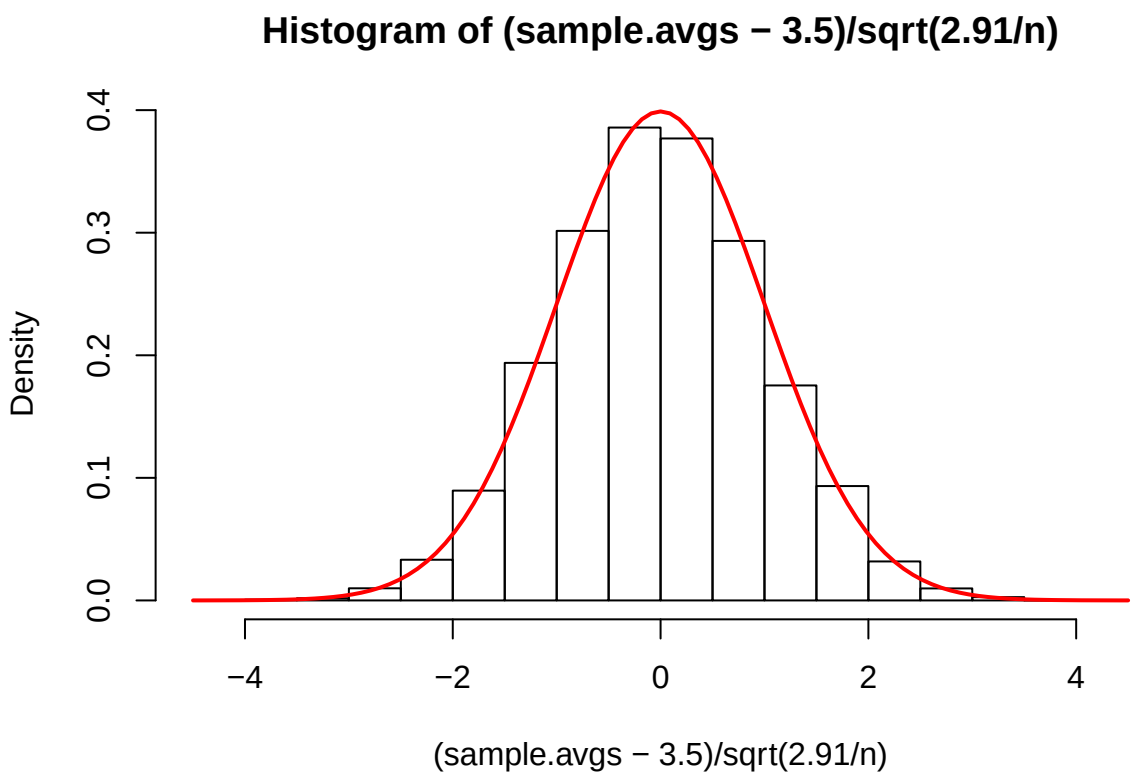

```
# set sample size and number of samples
n <- 10000
reps <- 20000

sample(1:6, 10, replace=TRUE)
```

```
##  [1] 2 1 4 2 1 3 1 5 6 2
```

```
# perform random sampling
samples <- replicate(reps,sample(1:6,n,replace=TRUE)) # 10 x 10000 sample matrix

# compute sample means
sample.avgs <- colMeans(samples)
hist(sample.avgs, freq=FALSE)
```




```
hist((sample.avgs-3.5)/sqrt(2.91/n), freq=FALSE)


curve(dnorm(x, sd = 1),
      col = "red",
      lwd = "2",
      add = T)
```



- Handout answer Q2 (h-l)

```
# Generate the data for coin experiment
n <- 100
n1 <- 61
n2 <- 39

#create the data of 100 obs with 61=1 and 39=0
df1 <- as.data.frame(c(rep(1,n1),rep(0,n2)))
names(df1) <- "Y"
#scramble so it looks random

df1$Y <- sample(df1$Y)
View(df1)

df1$Y.bar <- mean(df1$Y)
df1$sq.dev <- (df1$Y - df1$Y.bar)^2

sample.mean <- mean(df1$Y)
sample.var <- sum(df1$sq.dev)/(n-1)
sample.std <- sqrt(sample.var)
sample.se <- sample.std/sqrt(n)

mean(df1$Y)
```

```
## [1] 0.61
```

```r
var(df1$Y)
```

```
## [1] 0.240303
```

```r
#describe(df1$Y)

#t-statistic
t.statistic <- -((sample.mean-0.5)/sample.se)
print(t.statistic)
```

```
## [1] -2.243949
```

```r
pt(t.statistic, n-1,lower.tail = TRUE)*2
```

```
## [1] 0.02706281
```

```r
qt(0.025, n-1, lower.tail = TRUE)
```

```
## [1] -1.984217
```

```r
#much easier way to do it
t.test(df1$Y,mu=0.5)
```

```
##
##  One Sample t-test
##
## data:  df1$Y
## t = 2.2439, df = 99, p-value = 0.02706
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.5127323 0.7072677
## sample estimates:
## mean of x
##      0.61
```

```r
lb95<- sample.mean + qt(0.025, n-1, lower.tail = TRUE)*sample.se
ub95<- sample.mean - qt(0.025, n-1, lower.tail = TRUE)*sample.se

print(c(lb95,ub95))
```

```
## [1] 0.5127323 0.7072677
```

7.3 Week3

```r
df1 <- read.csv("https://www.dropbox.com/s/c9aj0kjftso7qzi/birth19682002.csv?dl=1") %>%
  filter(bwei<9000 & is.na(bwei)==FALSE)

df2 <- df1[df1$yearb==1968 & df1$race3=='black', ]
df3 <- df2[df1$yearb==1968 & df1$race3=='white', ]
```

```
black.mean.1968 <- mean(df2$bwei, na.rm = TRUE)
black.var.1968 <- var(df2$bwei, na.rm = TRUE)
white.mean.1968 <- mean(df3$bwei, na.rm = TRUE)
white.var.1968 <- var(df3$bwei, na.rm = TRUE)



race.gap.1968 <-  white.mean.1968 - black.mean.1968
race.gap.1968.se <- sqrt((black.var.1968/447) + (white.var.1968/2462))
race.gap.1968.ub <- race.gap.1968 + (1.96*race.gap.1968.se)
race.gap.1968.lb <- race.gap.1968 - (1.96*race.gap.1968.se)

res <- c(black.mean.1968,white.mean.1968,race.gap.1968, race.gap.1968.lb, race.gap.1968.ub  )
print(round(res,0))
```

```
## [1] 3090 3231  141   81  201
```

```
t1 <- t.test(df1$bwei, df2$bwei, na.rm = TRUE)
t1$estimate
```

```
## mean of x mean of y
##  3324.193  3089.745
```

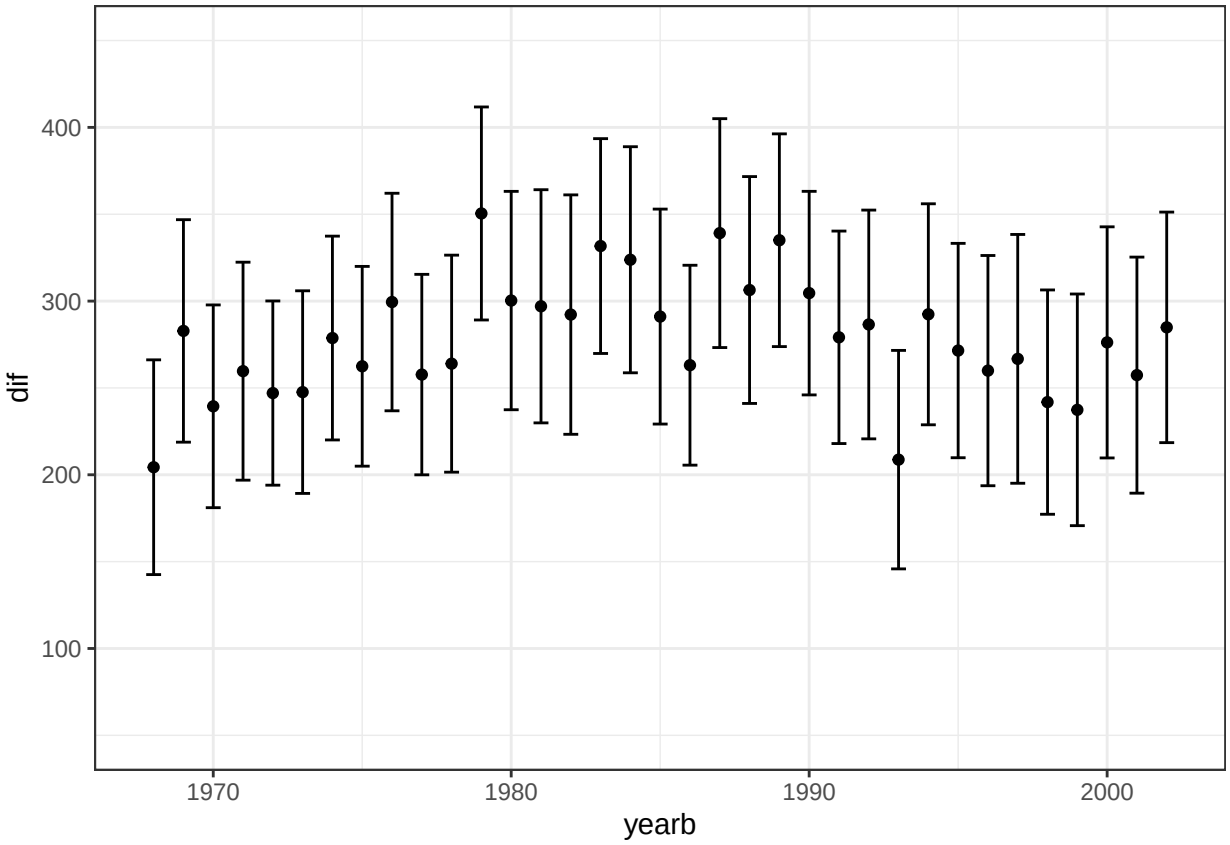
```
t1$stderr
```

```
## [1] 29.44446
```

```
t1$conf.int
```

```
## [1] 176.5824 292.3142
## attr(,"conf.level")
## [1] 0.95
```

```
df4 <- df1 %>%
  filter(race3!='oth') %>%
  dplyr::select(yearb,bwei,race3) %>%
  group_by(race3,yearb) %>%
  summarise_each(funs(mean,n(),sd,se=sd(.)/sqrt(n())))

df5 <- cbind(df4[df4$race3=='black',], df4[df4$race3=='white',] ) %>%
  mutate(dif=mean1-mean) %>%
  mutate(dif.se=sqrt((sd1^2/n1 + sd^2/n))) %>%
  ggplot(aes(x=yearb,y=dif)) +
  geom_point() +
  geom_errorbar(aes(ymin=dif-1.96*dif.se, ymax=dif+1.96*dif.se), width=0.5) +
  theme_bw() +
  ylim(c(50,450))

df5
```



7.4 Week4

- See the data called Anscombe's Quartet below for four pairs of x and y: $(x1, y1), (x2, y2), (x3, y3), (x4, y4)$

```
df1 <- anscombe
print(df1)
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

- Interpret the estimated sample covariance and correlation coefficients below.

```
#summarize the variables in the data set
df1.summary <- df1 %>%
  summarise_all(list(mean,sd))
```

```
df1.s1 <- as.data.frame(round(df1.summary,2))
print(df1.s1)
```

```
##   x1_fn1 x2_fn1 x3_fn1 x4_fn1 y1_fn1 y2_fn1 y3_fn1 y4_fn1 x1_fn2 x2_fn2 x3_fn2
## 1      9      9      9      9    7.5    7.5    7.5    7.5   3.32   3.32   3.32
##   x4_fn2 y1_fn2 y2_fn2 y3_fn2 y4_fn2
## 1   3.32   2.03   2.03   2.03   2.03
```

```
#calculate sample covariance and correlatio between pairs of
#(x1,y1), (x2,y2), (x3,y3), and (x4,y4)


df1.s2 <- round(c(cov(df1$x1,df1$y1), cov(df1$x2,df1$y2), cov(df1$x3,df1$y3), cov(df1$x4,df1$y4)),2)
df1.s3 <- round(c(cor(df1$x1,df1$y1), cor(df1$x2,df1$y2), cor(df1$x3,df1$y3), cor(df1$x4,df1$y4)),2)

print(df1.s2)
```

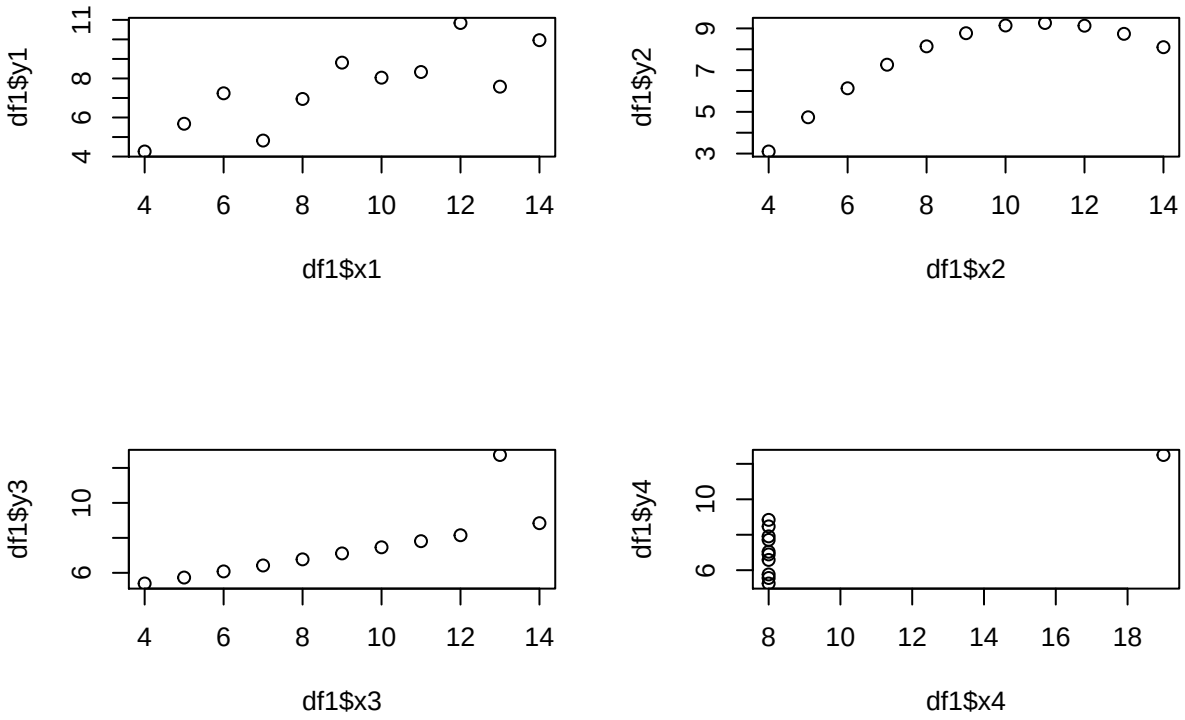
```
## [1] 5.5 5.5 5.5 5.5
```

```
print(df1.s3)
```

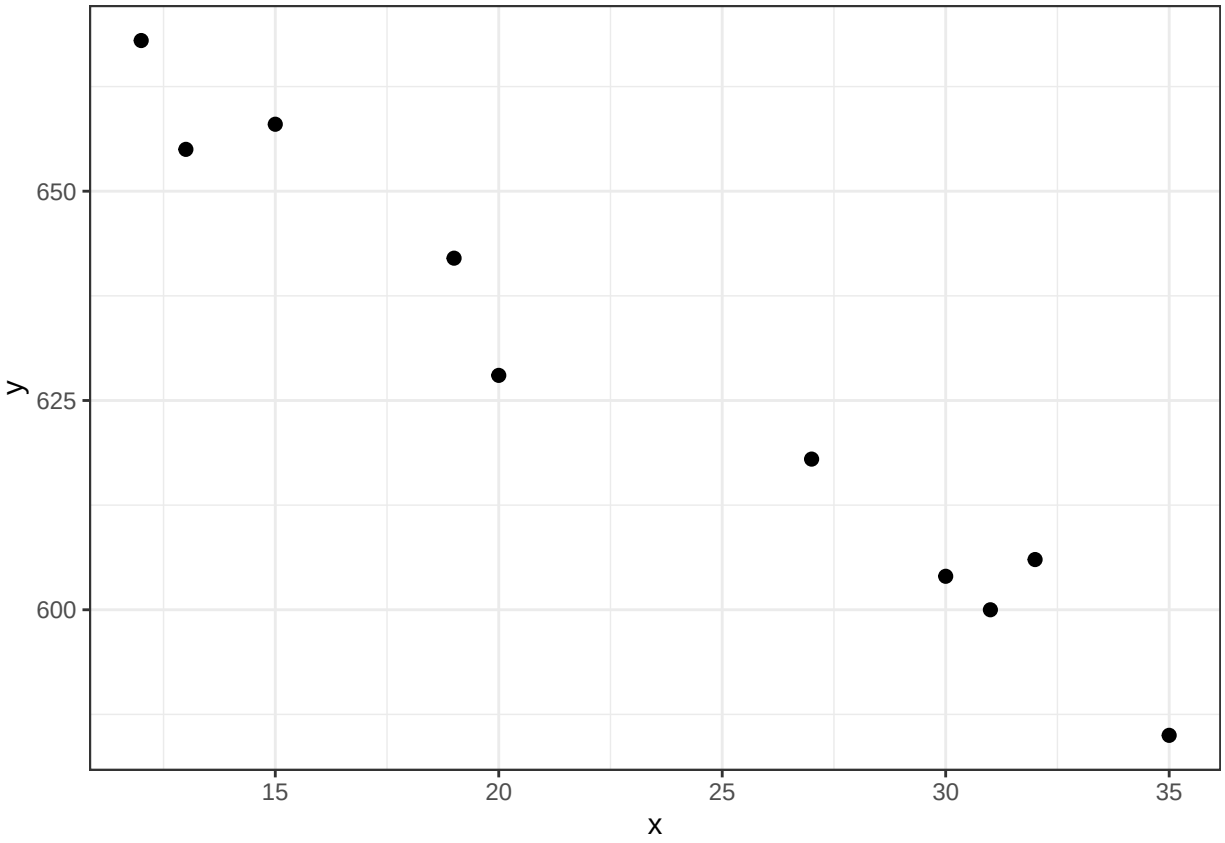
```
## [1] 0.82 0.82 0.82 0.82
```

- Below is the scatter plot for each pair of x and y, how do you reconcile the difference in graphs and the correlation coefficients shown above?

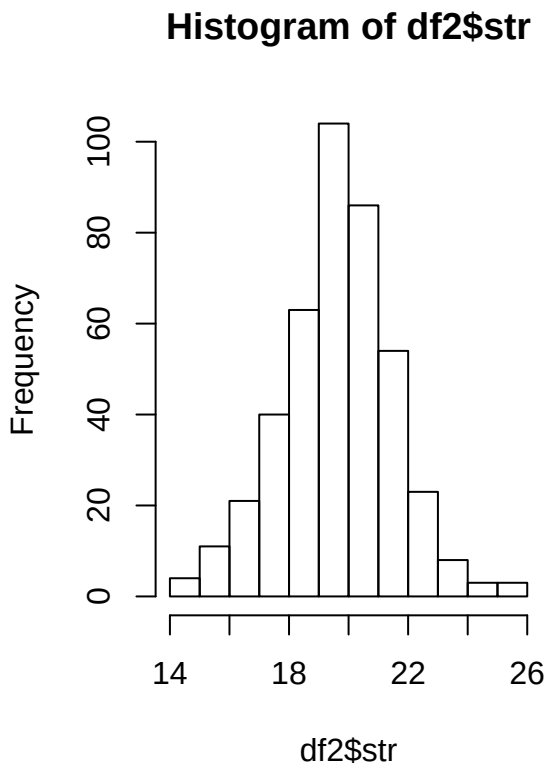
```
par(mfrow=c(2,2))
plot(df1$x1,df1$y1)
plot(df1$x2,df1$y2)
plot(df1$x3,df1$y3)
plot(df1$x4,df1$y4)
```

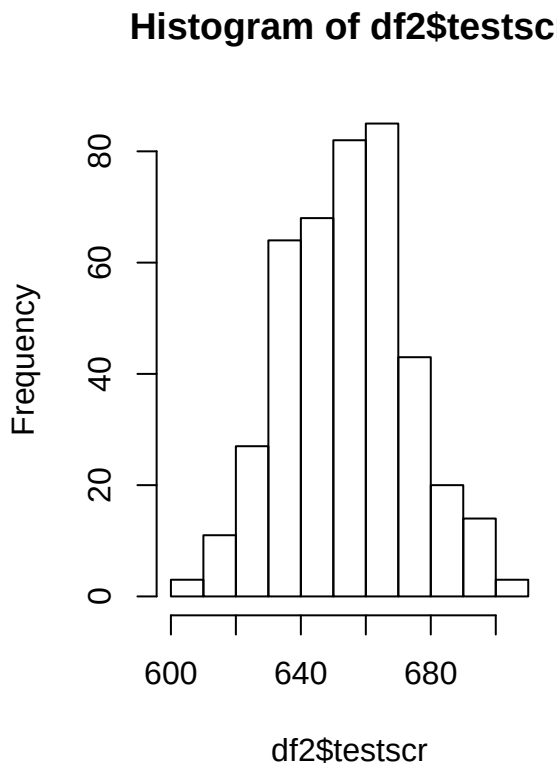



```
x <- 11:35
x <- sample(x,size=10,replace=FALSE)
y <- round(690 - 3.1*x + rnorm(10,mean=10,sd=5))

ggplot() +
geom_point(aes(x,y), size=2) +
geom_smooth(method = "lm", se = FALSE) +
theme_bw()
```




```
df2 <- foreign::read.dta('http://fmwww.bc.edu/ec-p/data/stockwatson/caschool.dta')

par(mfrow=c(1,2))
hist(df2$str)
hist(df2$testscr)
```





7.5 Week5

Suppose that a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the following OLS regression:

```
#set the model parameters
set.seed(01012000)
n <- 100
b0_hat <- 520.4
b1_hat <- -5.82
u_hat <-  rnorm(n, mean = 0, sd = 100)
CS <- sample(10:30,n,TRUE )
TestScore <-  b0_hat + b1_hat*CS + u_hat

df1 <- as.data.frame(cbind(TestScore,CS))
mo1 <- lm(TestScore~CS)
summary(mo1)
```

```
##
## Call:
## lm(formula = TestScore ~ CS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -258.058  -68.062   -1.417   69.704  297.534
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  523.756     38.011  13.779  < 2e-16 ***
## CS            -6.479      1.741  -3.721 0.000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.8 on 98 degrees of freedom
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.1149
## F-statistic: 13.85 on 1 and 98 DF,  p-value: 0.0003311
```

```
summary(CS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   15.75   21.50   21.00   26.00   30.00
```

```
confint.lm(mo1,level=0.99)
```

```
##                 0.5 %    99.5 %
## (Intercept) 423.90261 623.60884
## CS          -11.05307  -1.90493
```

```
df1$small.CS <- 0
df1$small.CS[df1$CS<20] <- 1

head(df1)
```

```
##   TestScore CS small.CS
## 1  378.6047 30        0
## 2  468.6586 21        0
## 3  331.0196 29        0
## 4  292.1520 23        0
## 5  388.1037 30        0
## 6  479.7165 20        0
```

```
mo2 <- lm(df1$TestScore~df1$small.CS)
summary(mo2)
```

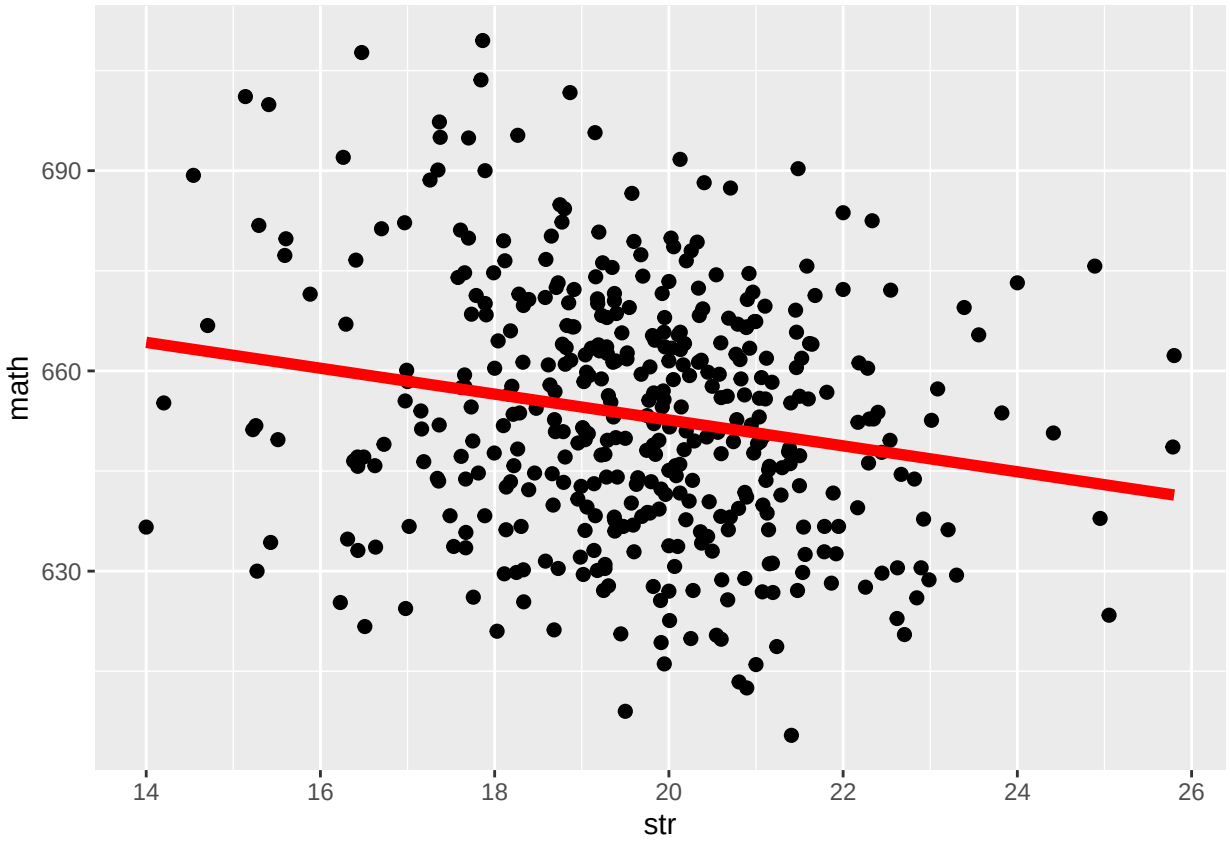
```
##
## Call:
## lm(formula = df1$TestScore ~ df1$small.CS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -298.588  -76.795    4.799   73.665  309.736
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    369.92      14.04  26.355   <2e-16 ***
## df1$small.CS    44.45      22.19   2.003   0.0479 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108.7 on 98 degrees of freedom
## Multiple R-squared:  0.03933,    Adjusted R-squared:  0.02953
## F-statistic: 4.012 on 1 and 98 DF,  p-value: 0.04793
```

7.6 Week 6

```
data("CASchools")

View(CASchools)

CASchools$str <- CASchools$students/CASchools$teachers

mo1 <- lm(CASchools$math~CASchools$str)
summary(mo1)
```

```
##
## Call:
## lm(formula = CASchools$math ~ CASchools$str)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.615 -13.374  -0.828  12.728  52.711
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   691.4174     9.3825  73.692  < 2e-16 ***
## CASchools$str  -1.9386     0.4755  -4.077 5.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.41 on 418 degrees of freedom
## Multiple R-squared:  0.03824,    Adjusted R-squared:  0.03594
## F-statistic: 16.62 on 1 and 418 DF,  p-value: 5.467e-05
```

```
CASchools$math.hat <- predict.lm(mo1)

g1 <- ggplot(data=CASchools) +
  geom_point(aes(x=str,y=math),   size=2) +
  geom_line(aes(x=str,y=math.hat),   size=2, colour='red')

CASchools$small <- 0
CASchools$small[CASchools$str>=20] <- 1

plot(g1)
```

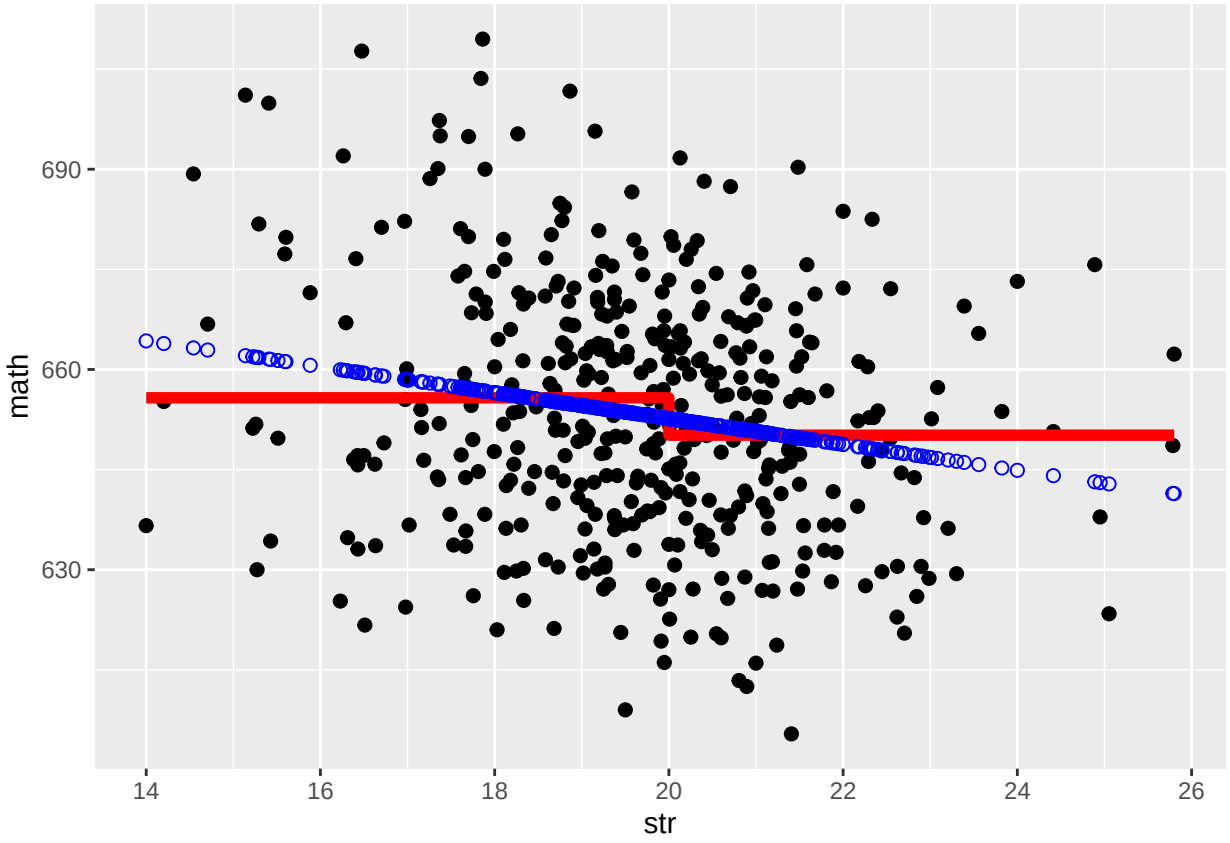



```
mo2 <- lm(CASchools$math~CASchools$small)
summary(mo2)
```

```
##
## Call:
## lm(formula = CASchools$math ~ CASchools$small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.756 -13.456  -0.706  12.769  53.744
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      655.756      1.201  545.93  < 2e-16 ***
## CASchools$small   -5.599      1.830   -3.06  0.00235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.57 on 418 degrees of freedom
## Multiple R-squared:  0.02191,    Adjusted R-squared:  0.01957
## F-statistic: 9.364 on 1 and 418 DF,  p-value: 0.002355
```

```
CASchools$math.hat2 <- predict.lm(mo2)

g2 <- ggplot(data=CASchools) +
  geom_point(aes(x=str,y=math),   size=2) +
  geom_line(aes(x=str,y=math.hat2),   size=2, colour='red') +
  geom_point(aes(x=str,y=math.hat),   shape=1, size=2, colour='blue')
```

```
plot(g2)
```

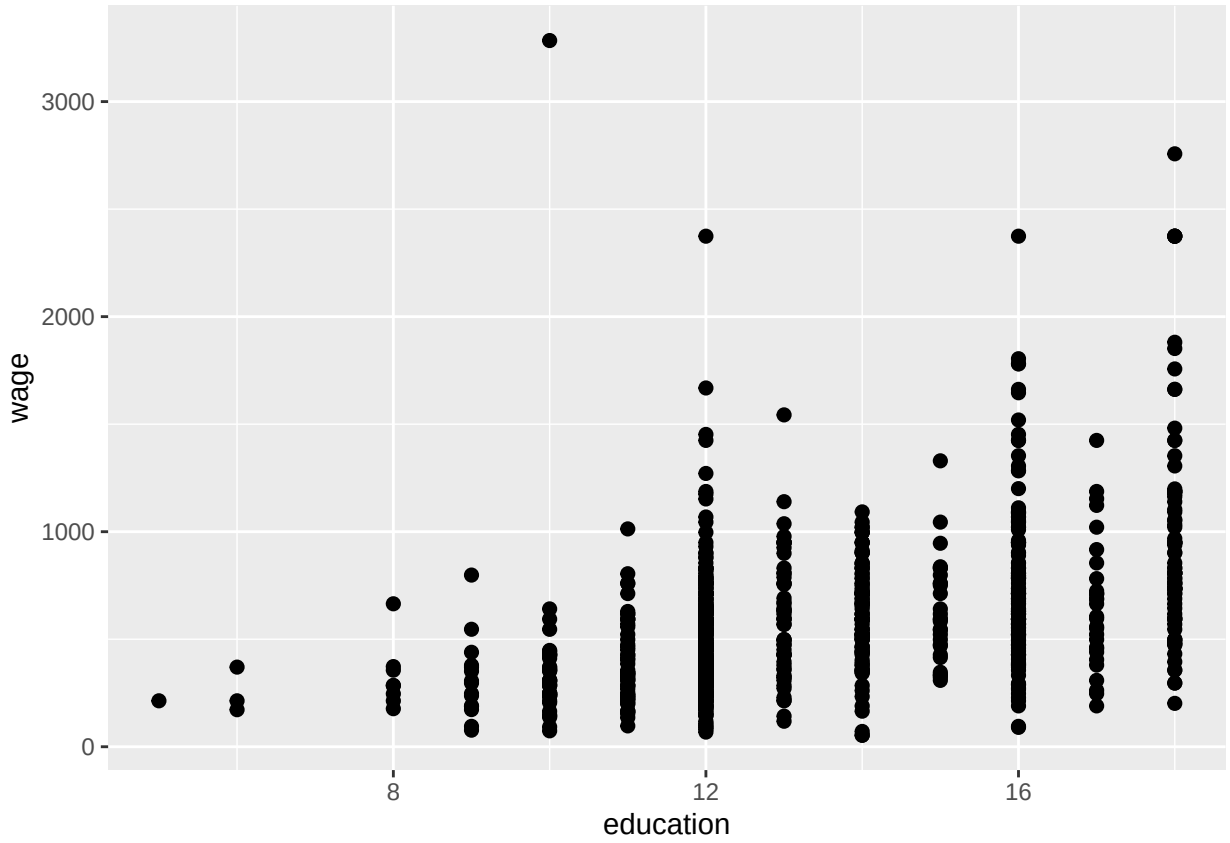



```
table(CASchools$math.hat2)
```

```
##
## 650.156355546983 655.755646135019
##              181              239
```

7.7 Week 7

```
data("CPS1988")
df1 <- CPS1988 %>%
  filter(experience==10)

g2 <- ggplot(data=df1) +
  geom_point(aes(x=education,y=wage),   size=2)

plot(g2)
```




```
mo1 <- lm(data=df1, formula=wage ~ education)

summary(mo1)
```

```
##
## Call:
## lm(formula = wage ~ education, data = df1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -689.3 -186.9  -44.4  133.4 2920.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -295.803     58.635  -5.045 5.49e-07 ***
## education     65.941      4.303  15.326  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317 on 901 degrees of freedom
## Multiple R-squared:  0.2068, Adjusted R-squared:  0.2059
## F-statistic: 234.9 on 1 and 901 DF,  p-value: < 2.2e-16
```

```
#homoskedastic
coeftest(mo1)
```

```
##
```

```
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -295.8033    58.6348 -5.0448 5.489e-07 ***
## education     65.9409     4.3026 15.3257 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

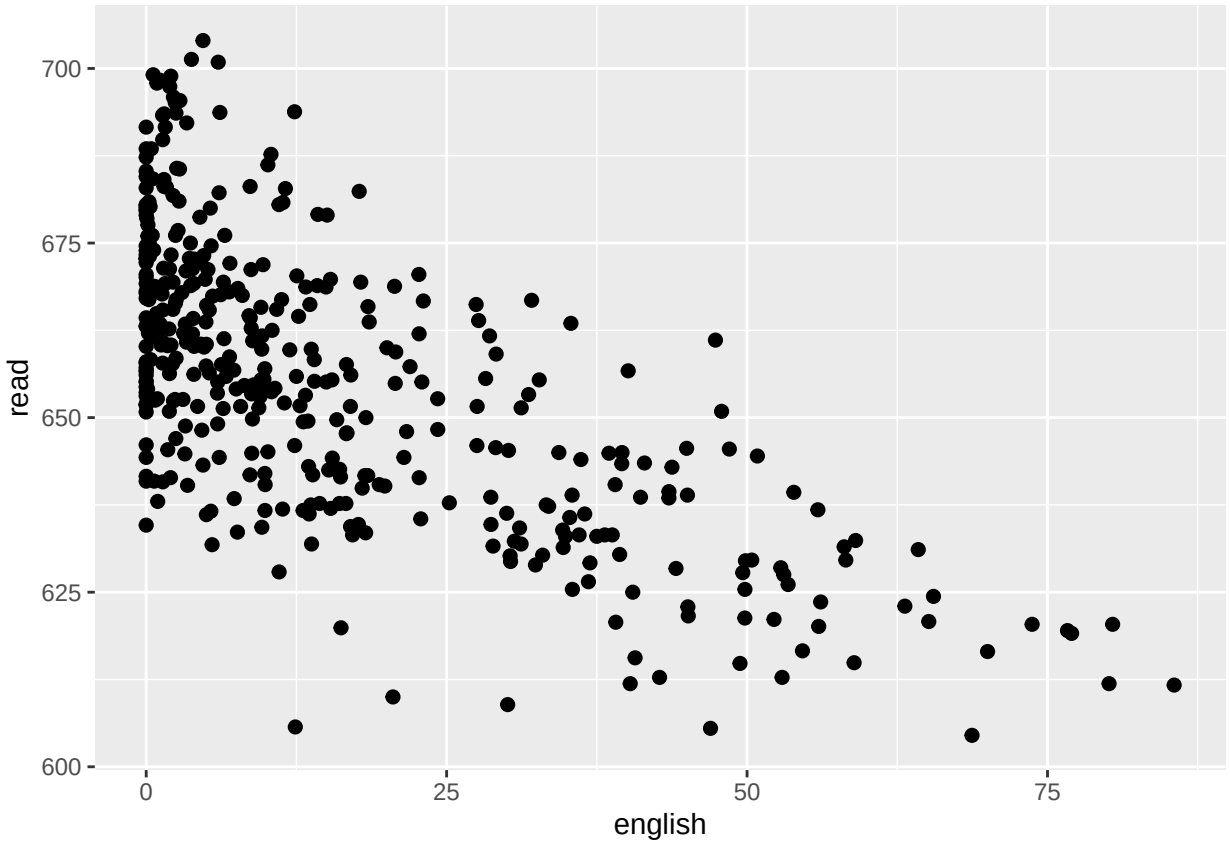
```
#Breusch-Pagan test
bptest(mo1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mo1
## BP = 11.6, df = 1, p-value = 0.0006596
```

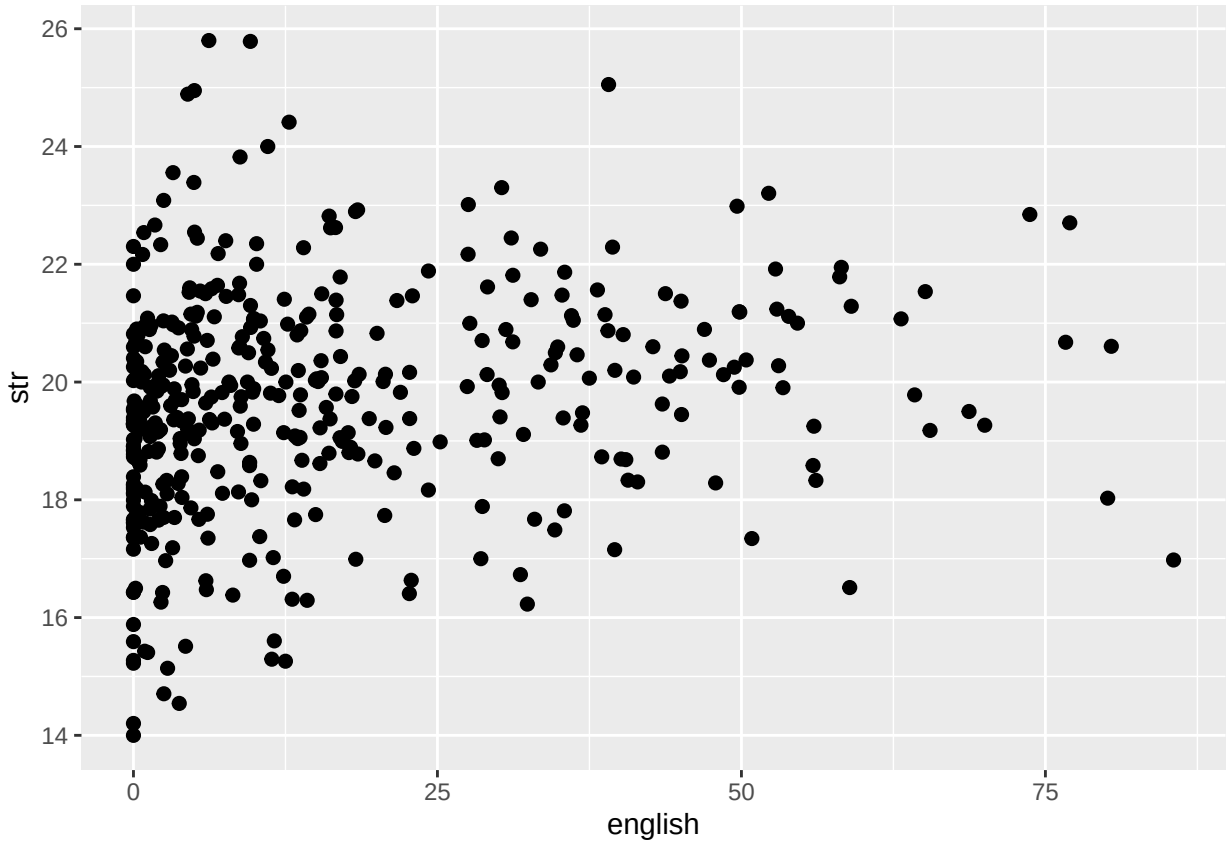
```
coeftest(mo1, vcov = vcovHC(mo1, "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -295.8033    68.8836 -4.2942 1.943e-05 ***
## education     65.9409     5.3622 12.2974 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data(CASchools)

CASchools$str <- CASchools$students/CASchools$teachers

g2 <- ggplot(data=CASchools) +
  geom_point(aes(x=english,y=read),   size=2)

plot(g2)
```




```
g2 <- ggplot(data=CASchools) +
  geom_point(aes(x=english,y=str),   size=2)

plot(g2)
```




```
CASchools$small <- 0
CASchools$small[CASchools$str<20] <- 1

mo1 <- lm(data=CASchools, formula=read~small)
summary(mo1)
```

```
##
## Call:
## lm(formula = read ~ small, data = CASchools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.237 -14.637   1.013  13.703  47.903
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  649.997      1.461 444.824  < 2e-16 ***
## small          8.740      1.937   4.512 8.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.66 on 418 degrees of freedom
## Multiple R-squared:  0.04644,    Adjusted R-squared:  0.04416
## F-statistic: 20.36 on 1 and 418 DF,  p-value: 8.368e-06
```

```
mo2 <- lm(data=CASchools[CASchools$english<1.9,], formula=read~small)
mo3 <- lm(data=CASchools[CASchools$english>1.9 & CASchools$english<8.8,], formula=read~small)
mo4 <- lm(data=CASchools[CASchools$english>8.8 & CASchools$english<23,], formula=read~small)
```

```
mo5 <- lm(data=CASchools[CASchools$english<23,], formula=read~small)

summary(mo2)
```

```
##
## Call:
## lm(formula = read ~ small, data = CASchools[CASchools$english <
##     1.9, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.042 -10.142  -0.196  11.108  31.458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 667.8963     2.8032 238.258   <2e-16 ***
## small        -0.2542     3.2634  -0.078    0.938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.57 on 101 degrees of freedom
## Multiple R-squared:  6.007e-05,  Adjusted R-squared:  -0.00984
## F-statistic: 0.006067 on 1 and 101 DF,  p-value: 0.9381
```

```
summary(mo3)
```

```
##
## Call:
## lm(formula = read ~ small, data = CASchools[CASchools$english >
##     1.9 & CASchools$english < 8.8, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.570 -10.768   0.031   8.832  36.830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  662.368      2.374 279.062   <2e-16 ***
## small          4.802      3.083   1.557    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.74 on 106 degrees of freedom
## Multiple R-squared:  0.02237,    Adjusted R-squared:  0.01315
## F-statistic: 2.426 on 1 and 106 DF,  p-value: 0.1223
```

```
summary(mo4)
```

```
##
## Call:
## lm(formula = read ~ small, data = CASchools[CASchools$english >
##     8.8 & CASchools$english < 23, ])
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -43.980 -12.098  -0.217  10.371  38.346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  649.680      2.211 293.774   <2e-16 ***
## small          5.774      3.069   1.881   0.0628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.64 on 102 degrees of freedom
## Multiple R-squared:  0.03353,    Adjusted R-squared:  0.02406
## F-statistic: 3.539 on 1 and 102 DF,  p-value: 0.06279
```

```
summary(mo5)
```

```
##
## Call:
## lm(formula = read ~ small, data = CASchools[CASchools$english <
##     23, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.659 -11.176   0.106  10.524  39.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  658.359      1.500 439.017   <2e-16 ***
## small          5.735      1.911   3.001   0.0029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.5 on 313 degrees of freedom
## Multiple R-squared:  0.02797,    Adjusted R-squared:  0.02487
## F-statistic: 9.008 on 1 and 313 DF,  p-value: 0.002905
```

7.8 Week 9

```
data("CASchools")

CASchools$str <- CASchools$students/CASchools$teachers

mo1 <- lm(data=CASchools, formula=read~str)
summary(mo1)
```

```
##
## Call:
## lm(formula = read ~ str, data = CASchools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.839 -14.479   1.121  14.495  44.370
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 706.4485     9.9410  71.064  < 2e-16 ***
## str          -2.6210     0.5038  -5.202 3.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.51 on 418 degrees of freedom
## Multiple R-squared:  0.06081,    Adjusted R-squared:  0.05856
## F-statistic: 27.06 on 1 and 418 DF,  p-value: 3.091e-07
```

```
mo2 <- lm(data=CASchools, formula=read~str+english)
summary(mo2)
```

```
##
## Call:
## lm(formula = read ~ str + english, data = CASchools)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49.46 -10.28  -0.32   9.64  38.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 691.87505    7.37033   93.87  < 2e-16 ***
## str          -1.28970    0.37818   -3.41 0.000712 ***
## english      -0.73403    0.03912  -18.76  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.38 on 417 degrees of freedom
## Multiple R-squared:  0.4907, Adjusted R-squared:  0.4883
## F-statistic: 200.9 on 2 and 417 DF,  p-value: < 2.2e-16
```

```
mo3 <- lm(data=CASchools, formula=english~str)
summary(mo3)
```

```
##
## Call:
## lm(formula = english ~ str, data = CASchools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.823 -13.006  -6.849   7.834  74.601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.8541     9.1626  -2.167  0.03081 *
## str           1.8137     0.4644   3.906  0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.98 on 418 degrees of freedom
## Multiple R-squared:  0.03521,    Adjusted R-squared:  0.0329
## F-statistic: 15.25 on 1 and 418 DF,  p-value: 0.0001095
```

```
RMSE <- sqrt(mean(mo2$residuals^2))
print(RMSE)
```

```
## [1] 14.33303
```

```
CASchools$noise <- sample(200, size = nrow(CASchools), replace = TRUE)

mo4 <- lm(data=CASchools, formula=read~str+english+noise)
summary(mo4)
```

```
##
## Call:
## lm(formula = read ~ str + english + noise, data = CASchools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.821 -10.489  -0.403   9.854  39.315
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 690.796982   7.501590  92.087  < 2e-16 ***
## str          -1.285530   0.378389  -3.397 0.000746 ***
## english      -0.735135   0.039169 -18.768  < 2e-16 ***
## noise         0.009862   0.012614   0.782 0.434766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 416 degrees of freedom
## Multiple R-squared:  0.4914, Adjusted R-squared:  0.4878
## F-statistic:   134 on 3 and 416 DF,  p-value: < 2.2e-16
```