STATS 191: Homework Assignment 1

Pratheepa Jeganathan

09/26/2019

You may discuss homework problems with other students, but you have to prepare the written assignments yourself.

Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.

Grading scheme: $\{0, 1, 2\}$ **points per question, total of 50.**

Due date: 11:59 PM October 4, 2019 (Friday evening).

Question 1

This question is from the textbook: **CH** page 23, Exercises 1.2.

Give an example in any area of interest to you where regression analysis can be used as a data analytic tool to answer some questions of interest.

1. What is the question of interest.
2. Indentify the response and the predictor variables.
3. Classify each of the variables as either quantitative or qualitative.
4. Which type of regression (see **CH** page 18, Table 1.15) can be used to analyze the data?
5. Give a possible form of the model and identify its parameters.

Question 2

On Groundhog Day, February 2, a famous groundhog in Punxsutawney, PA is used to predict whether a winter will be long or not based on whether or not he sees his shadow. Jonathan Taylor collected data on whether Phil saw his shadow or not from here . He stored some of this data in this table .

Although Phil is on the East Coast, Jonathan wondered if the information says anything about whether or not we will experience a rainy winter out here in California. For this, Jonathan found rainfall data .

I saved it in a table. Here is how this was extracted using `R` packages.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
url = "http://cdec.water.ca.gov/cgi-progs/precip1/8STATIONHIST"
webpage = readLines(url)[476:593]
ind_of_rows_with_space = which(webpage == webpage[2])
webpage = webpage[-ind_of_rows_with_space]
# remove whitespace in each line
webpage = lapply(as.list(webpage), function(x){trimws(x)}) %>% unlist()
library(stringr)
temp = lapply(as.list(webpage), function(x){strsplit(x, " ", fixed = TRUE)}) %>% unlist()
temp = temp[-which(temp == temp[2])]
df = matrix(temp, ncol = 14, byrow = TRUE)
col.names.df = df[1, ]
df = df[-1, ] %>% data.frame()
colnames(df) = col.names.df
df = apply(df, 2, function(x){x  = as.numeric(x)}) %>% data.frame()
write.csv(df, file = "rainfall.csv", row.names = FALSE)
```

1. Using the `ggplot2` package, make a boxplot of the mean monthly rainfall (total annual rainfall divided by 12 months) rainfall in Northern California comparing the years Phil sees his shadow versus the years he does not. [Read Phil's data from this link. Download and read rainfall data from Canvas@Stanford. [Use `dplyr::filter` to select years of rainfall data that Phil sees his shadow or not. Use `dplyr::left_join` to merge Phil's shadow data and rainfall data.]
2. Construct a 93% confidence interval for the difference between the mean monthly rainfall (total annual rainfall divided by 12 months) in years Phil sees his shadow and years he does not.
3. Interpret the procedure used to construct in part 2. What do we really know about confidence intervals?
4. At level, $\alpha = 0.05$ would you reject the null hypothesis that the average rainfall in Northern California during the month of February was the same in years Phil sees his shadow versus years he does not?
5. What assumptions are you making in forming your confidence interval and in your hypothesis test?

Question 3

In Question 2, part 4 above, you are asked to carry out a hypothesis test. In part 5, you are asked to justify your confidence interval and hypothesis test. Both are typically based on a T statistic of some form.

1. Write functions in `R` to generate new data sets for the two different groups of years, calling them `shadow` and `noshadow`. The functions should be such that you can specify the average rainfall within the two years separately, as well as the variability of the rainfall within those years (for example, you might use `rnorm` with different mean and variance parameters).
2. Using your two functions above, simulate data under the null hypothesis that the data from `shadow` years is the same as the data from `noshadow` years, computing the T statistic each time. Plot a density of a sample of 5000 such T statistics, overlaying it with a "true" density that holds under the null hypothesis. Explain how these densities relate to the test you carried out in Question 1, part 4.
3. Again using the same two functions, simulate data under the null hypothesis that the average rainfall from `shadow` years is the same as the average rainfall from `noshadow` years, allowing for the possibility that the variability of the average is different among the two groups. The function `t.test` allows specifying `var.equal` to be true or false. Compare the density of the T statistics when the variability is not the same within the two groups. There are 4 possibilities (2 choices of variances for `shadow` and `noshadow`, and setting `var.equal` to be `TRUE` or `FALSE`). So we should see 4 histograms possibly comparing to the "true" density of part 2.

Question 4

The data set `walleye` in the package `alr4` (remember you may have to run `install.packages("alr4")`) of data measured on walleye fish in Wisconsin.

1. Create a boxplot (use `ggplot2`) of `length`, for `age` in `5:8`. Use `filter` in `dplyr` package to filter rows from the data frame.
2. Compute the sample mean, sample standard deviation `length` in the four age groups (5:8) (use `dplyr`).
3. Create a histogram (use `ggplot2`) of `length` within `age` of `5:8` putting the plots in a 2x2 grid in one file.
4. Compute a 95% confidence interval for the difference in `length` in years 5 and 7. What assumptions are you making?
5. At level $\alpha = 10\%$, test the null hypothesis that the average `length` in the group `age==5` is the same as the in the group `age==7`. What assumptions are you making? What can you conclude?
6. Repeat the test in 5. using the function `lm`.

Question 5

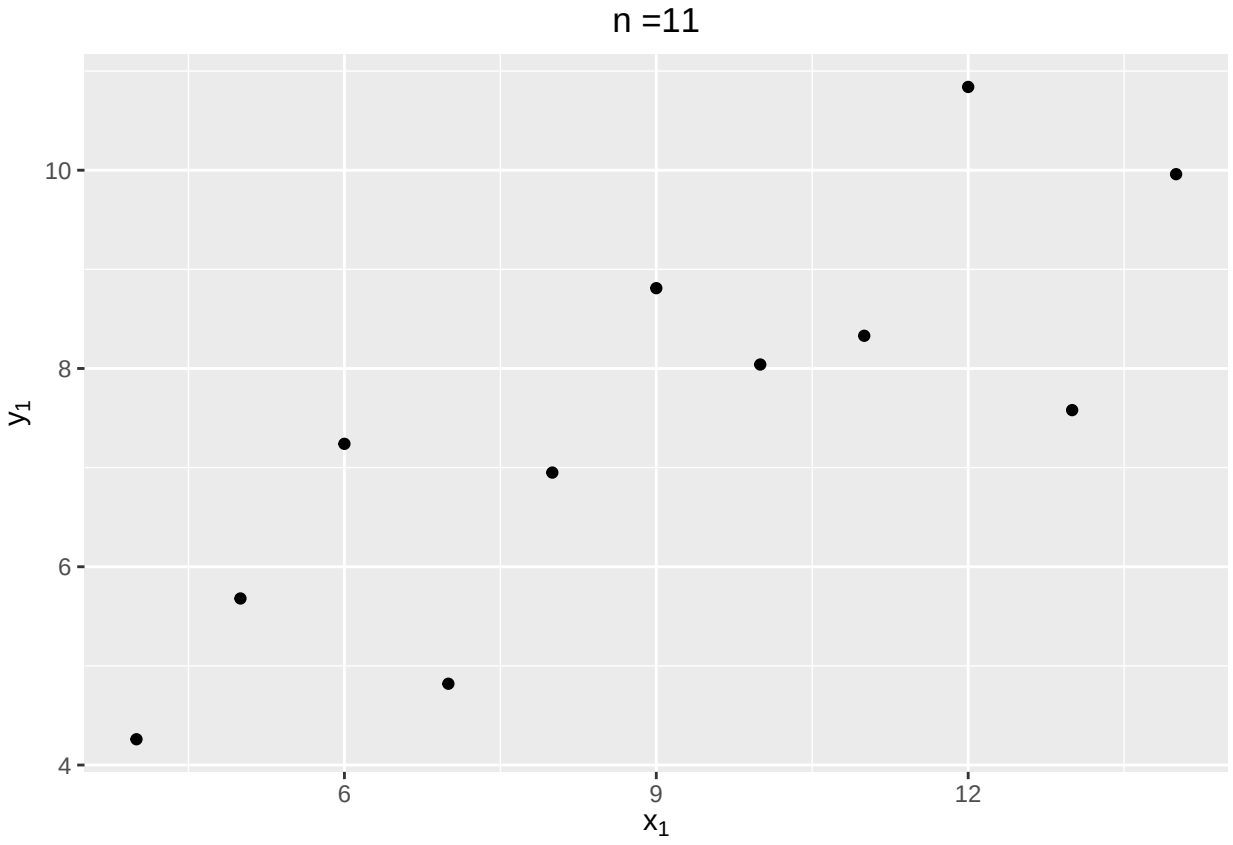
Following questions are based on the `anscombe` data in `R`.

1. Plot the 4 data sets (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) on a 2-by-2 grid of plots using `ggplot2` and `gridExtra` package.

   Add the number of the dataset to each plot as the main title on each plot.

   Add the axis-labels using `bquote` to each plot. For example,

```
library(ggplot2)
library(magrittr)
library(dplyr)
ggplot(data = anscombe) +
  geom_point(aes(x = x1, y = y1)) +
  xlab(bquote(x[1])) +
  ylab(bquote(y[1])) +
  ggtitle(paste0("n =", dim(anscombe %>% select(x1,y1))[1])) +
  theme(plot.title = element_text(hjust = 0.5))
```



2. Fit a regression model to the data sets:

   a. $y_1 \sim x_1$

   b. $y_2 \sim x_2$

   c. $y_3 \sim x_3$

   d. $y_4 \sim x_4$

using the command `lm`. Verify that all the fitted models have the exact same coefficients (up to numerical tolerance).

3. Using the command `cor`, compute the sample correlation for each data set.
4. Fit the same models in 3. but with the x and y reversed. Using the command `summary`, does anything about the results stay the same when you reverse x and y?
5. Compute the SSE, SST and R^2 value for each data set. Use the commands `mean, sum, predict` and / or `resid`. (Use the original models, i.e. $y_i \sim x_i$ so only 4 SSE values)
6. Using the `ggplot2` package, replot the data, adding the regression line to each plot. (Use the original models, i.e. $y_i \sim x_i$ so only 4 plots)

```
anscombe
```

```
##    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
```