Homework 1

PSTAT 126, Winter 2021

Due date: Janurary 30, 2021 at 23:59 PT

Note: Please show all the procedures of your analysis, and prepare the homework solution using RMarkdown. All code should be well documented. A RMarkdown homework template is available on GauchoSpace. Homework should be submitted on GauchoSpace.

You should write up your homework solution on your own. In particular, do not share your homework RMarkdown file with other student.

Q1. The dataset *cov.data* (constructed below) contains number of *confirmed* COVID-19 cases (in thousand people), and *population* (in thousand people) of 9 counties in California (as of 1/18/2021).

```
cov.data <- data.frame(row.names = c("San Bernardino County",
                                     "Riverside County",
                                     "Orange County",
                                     "San Diego County",
                                     "Santa Clara County",
                                     "Kern County",
                                     "Sacramento County",
                                     "Fresno County",
                                     "Alameda County"))
cov.data$population <- c(2149, 2411, 3168, 3316, 1927, 887, 1525, 985, 1657)
cov.data$confirmed <- c(251, 239, 223, 212, 93, 85, 79, 81, 67)
```

Our primary goal is to predict the number of confirmed COVID-19 cases of a county based on the county population.

(1a). In this problem, what is the predictor variable, and what is the response variable?

(1b). Calculate (without using *lm* function) $\hat{\beta}_0$ and $\hat{\beta}_1$, i.e., the least squares coefficient estimates of β_0 and β_1 in this simple linear regression model.

(1c). What is the interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$ in this dataset?

(1d). Write out the simple linear regression model in this problem.

(1e). What is the predicted number of confirmed COVID-19 cases of a county with 1,000,000 population?

Q2. The dataset *trees* contains measurements of *Girth* (actually, tree diameter) in inches, *Height* in feet, and *Volume* of timber (in cubic feet) of a sample of 31 felled black cherry trees. The following commands can be used to read the data into R. The dataset can be accessed under the name *trees*.

```
# the dataset "trees" is contained in the R package "datasets"
require(datasets)
trees
```

(2a). Briefly describe the dataset *trees*, i.e., how many observations (rows) and how many variables (columns) are there in the dataset? What are the variable names?

(2b). Use the *pairs* function to construct a scatterplot matrix of the *logarithms* of Girth, Height and Volume.

(2c). Use the *cor* function to determine the correlation matrix for the three (logged) variables.

(2d). Are there missing values?

(2e). Use the *lm* function in R to fit the multiple regression model,

$$\log \text{Volume}_i = \beta_0 + \beta_1 \log \text{Girth}_i + \beta_2 \log \text{Height}_i + \varepsilon_i$$

and print out the summary of the model fit.

(2f). Create the design matrix (i.e., the matrix of predictor variables), $\mathbf{X}$, for the model in (2e), and verify that the least squares coefficient estimates in the summary output are given by the least squares formula: $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

(2g). Compute the predicted response values from the fitted regression model, the residuals, and an estimate of the error variance $\text{Var}(\varepsilon) = \sigma^2$.

Q3. Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Show that minimizing the sum of squared residuals lead to the following least squares coefficient estimates:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.