

# Homework 1

## PSTAT Winter 2023

Due date: Jan 31st, 2023 at 23:59 PT

1. The dataset *trees* contains measurements of *Girth* (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 data set "trees" is contained in the R package "datasets"
require(datasets)
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

- (a) (1pt) Briefly describe the data set *trees*, i.e., how many observations (rows) and how many variables (columns) are there in the data set? What are the variable names?
- (b) (2pts) Use the *pairs* function to construct a scatter plot matrix of the logarithms of Girth, Height and Volume.
- (c) (2pts) Use the *cor* function to determine the correlation matrix for the three (logged) variables.
- (d) (2pts) Are there missing values?
- (e) (2pts) Use the *lm* function in R to fit the multiple regression model:

$$\log(\text{Volume}_i) = \beta_0 + \beta_1 \log(\text{Girth}_i) + \beta_2 \log(\text{Height}_i) + \epsilon_i$$

and print out the summary of the model fit.

- (f) (3pts) Create the design matrix (i.e., the matrix of predictor variables),  $X$ , for the model in (e), and verify that the least squares coefficient estimates in the summary output are given by the least squares formula:  $\hat{\beta} = (X^T X)^{-1} X^T y$ .
  - (g) (3pts) Compute the predicted response values from the fitted regression model, the residuals, and an estimate of the error variance  $\text{Var}(\epsilon) = \sigma^2$ .
2. Consider the simple linear regression model:

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

**Part 1:**  $\beta_0 = 0$

- (a) (3pts) Assume  $\beta_0 = 0$ . What is the interpretation of this assumption? What is the implication on the regression line? What does the regression line plot look like?
- (b) (4pts) Derive the LS estimate of  $\beta_1$  when  $\beta_0 = 0$ .
- (c) (3pts) How can we introduce this assumption within the *lm* function?

**Part 2:**  $\beta_1 = 0$

- (d) (3pts) For the same model, assume  $\beta_1 = 0$ . What is the interpretation of this assumption? What is the implication on the regression line? What does the regression line plot look like?
  - (e) (4pts) Derive the LS estimate of  $\beta_0$  when  $\beta_1 = 0$ .
  - (f) (3pts) How can we introduce this assumption within the *lm* function?
3. Consider the simple linear regression model:

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

- (a) (10pts) Use the LS estimation general result  $\hat{\beta} = (X^T X)^{-1} X^T y$  to find the explicit estimates for  $\beta_0$  and  $\beta_1$ .
- (b) (5pts) Show that the LS estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are unbiased estimates for  $\beta_0$  and  $\beta_1$  respectively.