Homework 1

PSTAT Summer 2023

Due date: August 18th, 2023

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?

```
nrow(trees)
```

```
## [1] 31
```

```
ncol(trees)
```

```
## [1] 3
```

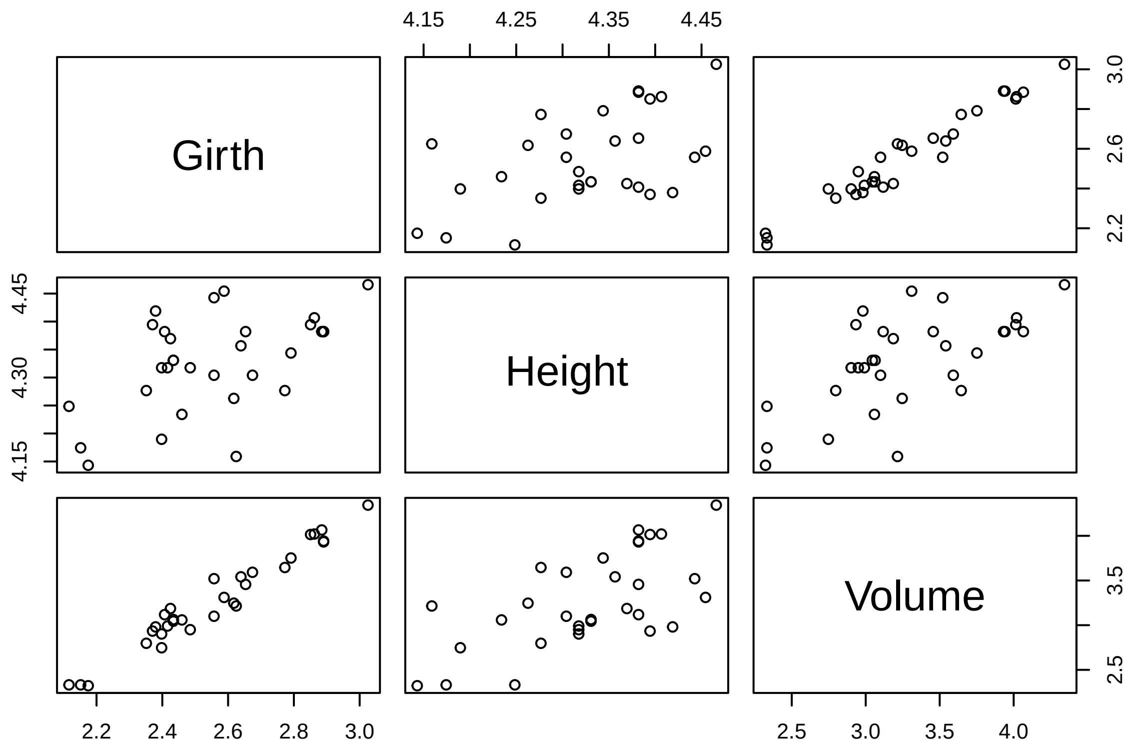
```
colnames(trees)
```

```
## [1] "Girth"  "Height" "Volume"
```

Data set trees has 31 observations and 3 variables. They name of the variable are Girth, Height and Volume.

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

```
pairs(log(trees))
```



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

```
cor(log(trees))
```

```
##            Girth    Height    Volume
## Girth  1.0000000 0.5301949 0.9766649
## Height 0.5301949 1.0000000 0.6486377
## Volume 0.9766649 0.6486377 1.0000000
```

(d) (2pts) Are there missing values?

```
any(is.na(trees))
```

```
## [1] FALSE
```

No, there are no missing values. (e) (2pts) Use the lm function in R to fit the multiple regression model:

$$log(Volume_i) = \beta_0 + \beta_1 log(Girth_i) + \beta_2 log(Height_i) + \epsilon_i$$

and print out the summary of the model fit.

```
md <- lm(log(Volume) ~ log(Girth) + log(Height), data = trees)
summary(md)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Girth) + log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
## log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
## log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

(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: $\hat{\beta} = (X^T X)^{-1} X^T y$.

```r
x = model.matrix(md)

beta_hat = solve(t(x) %*% x) %*% t(x) %*% log(trees$Volume)
beta_hat
```

```
##                  [,1]
## (Intercept) -6.631617
## log(Girth)   1.982650
## log(Height)  1.117123
```

(g) (3pts) Compute the predicted response valuvalues from the fitted regression model, the residuals, and an estimate of the error variance $Var(\epsilon) = \sigma^2$.

```r
predicted = predict(md)

residual = residuals(md)

error_var = (sum(residual^2)) / (length(residual)-3)


list(predicted = predicted, residual =  residual, error_var = error_var)
```

```
## $predicted
##        1        2        3        4        5        6        7        8
## 2.310270 2.297879 2.308547 2.807900 2.976888 3.022580 2.802931 2.945736
##        9       10       11       12       13       14       15       16
## 3.035777 2.981461 3.057130 3.031349 3.031349 2.974906 3.118250 3.246641
##       17       18       19       20       21       22       23       24
```

```
## 3.401459 3.475068 3.319702 3.218167 3.467691 3.524097 3.478455 3.643019
##       25       26       27       28       29       30       31
## 3.754853 3.929478 3.965974 3.983197 3.994242 3.994242 4.355446
##
## $residual
##            1            2            3            4            5            6
##  0.021874049  0.034264461  0.013841066 -0.010618992 -0.043031233 -0.041961116
##            7            8            9           10           11           12
## -0.055659877 -0.044314840  0.082173329  0.009258910  0.129222704  0.013172999
##           13           14           15           16           17           18
##  0.032041483  0.083801431 -0.168561198 -0.146548628  0.119002049 -0.164525292
##           19           20           21           22           23           24
## -0.073210648 -0.003299352  0.073268586 -0.067780336  0.113362744  0.002430731
##           25           26           27           28           29           30
## -0.002998263  0.085102043  0.054006059  0.082405145 -0.052660573 -0.062416748
##           31
## -0.011640695
##
## $error_var
## [1] 0.006623692
```

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?

beta_0=0 means the y intercept of this line is 0. When x=0, the value of y should be 0, so this line will pass through the origin (0,0). This regression line plot should be a straight line that starts at (0,0) and the slope of this line according to the beat_1.

(b) (4pts) Derive the LS estimate of β_1 when $\beta_0 = 0$.

$SSR = \sum_{i=1}^{n}(y_i - \hat{y})^2$
$SSR = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
$\hat{\beta}_0 = 0$
$\frac{dSSR}{d\hat{\beta}_1} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_1 x_i)$
$\sum_{i=1}^{n} -2(y_i - \hat{\beta}_1 x_i) = 0$
$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$
$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{y_i x_i}{x_i^2}$

(c) (3pts) How can we introduce this assumption within the lm function?

lm(y~x-1)

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?

beta_1=0 means x and y do not have linear relationship. The implication of this regression line is that this line will be a horizontal line with y intercept at beat_0. This regression line plot should be a horizontal line that has y intercept according to the value of beta_0.

(e) (4pts) Derive the LS estimate of β_0 when $\beta_1 = 0$.

$SSR = \sum_{i=1}^{n}(y_i - \hat{y})^2$
$SSR = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
$\hat{\beta}_1 = 0$
$\frac{dSSR}{d\hat{\beta}_1} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0) \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0) = 0$
$\sum_{i=1}^{n}(y_i - \hat{\beta}_0) = 0$
$n\hat{\beta}_0 = n\bar{y}$
$\hat{\beta}_0 = \bar{y}$

(f) (3pts) How can we introduce this assumption within the lm function?

lm(y~0+x)

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 β_0 and β_1.

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \; y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$
$$X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$
$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$
$$X^T y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$
$$(X^T X)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$
$$\hat{\beta} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(b) (5pts) Show that the LS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates for β_0 and β_1 respectively.
For $\hat{\beta}_1$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
$Sxx = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})xi$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{Sxx} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i)(x_i-\bar{x})-\bar{y})}{Sxx}$
$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i)}{Sxx}$
$E(\hat{\beta}_1) = E(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i)}{Sxx}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})E((y_i))}{Sxx} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 xi)}{Sxx}$
$E(\hat{\beta}_1) = \beta_0\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{Sxx} + \beta_1\frac{\sum_{i=1}^{n}(x_i-\bar{x})xi}{Sxx}$
$E(\hat{\beta}_1) = 0 + \beta_1\frac{Sxx}{Sxx} = \beta_1$ Therefor, it is unbiased.

For $\hat{\beta}_0$
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
$E(\bar{y}) = E(\frac{1}{n}\sum_{i=1}^{n} y_i) = \beta_0 + \beta_1\bar{x}$
$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1\bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1)$ $E(\hat{\beta}_0) = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$ Therefor, it is unbiased.