Homework Assignment 1

Sam Fang (8114613)

01/30/2021

Q1: (a) The predictor variable is the population. The response variable is the number of confirmed COVID-19 cases.

(b) Since $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$,we can calculate that $\hat{\beta}_0 = 2.85395$ and $\hat{\beta}_1$= 0.07236.

```
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)
x <- cov.data$population
y <- cov.data$confirmed

betaone <- ((x[1] - mean(x)) * (y[1] - mean(y)) + (x[2] - mean(x)) * (y[2] - mean(y)) +
    (x[3] - mean(x)) * (y[3] - mean(y)) + (x[4] - mean(x)) * (y[4] - mean(y)) + (x[5] -
    mean(x)) * (y[5] - mean(y)) + (x[6] - mean(x)) * (y[6] - mean(y)) + (x[7] - mean(x)) *
    (y[7] - mean(y)) + (x[8] - mean(x)) * (y[8] - mean(y)) + (x[9] - mean(x)) * (y[9] -
    mean(y)))/((x[1] - mean(x))^2 + (x[2] - mean(x))^2 + (x[3] - mean(x))^2 + (x[4] -
    mean(x))^2 + (x[5] - mean(x))^2 + (x[6] - mean(x))^2 + (x[7] - mean(x))^2 + (x[8] -
    mean(x))^2 + (x[9] - mean(x))^2)

betazero <- mean(y) - betaone * mean(x)

betaone
```

```
## [1] 0.07236
```

```
betazero
```

```
## [1] 2.854
```

(c)$\hat{\beta}_1$ is the average change of people confirmed for every 1000 increase in population. $\hat{\beta}_0$ is the number of confirmed people when population is zero.

(d) Simple linear regression model: y = 2.854 + 0.07236*x

(e) When x = 1000000, y = 72364

```r
predicted <- betazero + 1e+06 * betaone
predicted
```

```
## [1] 72364
```

Q2

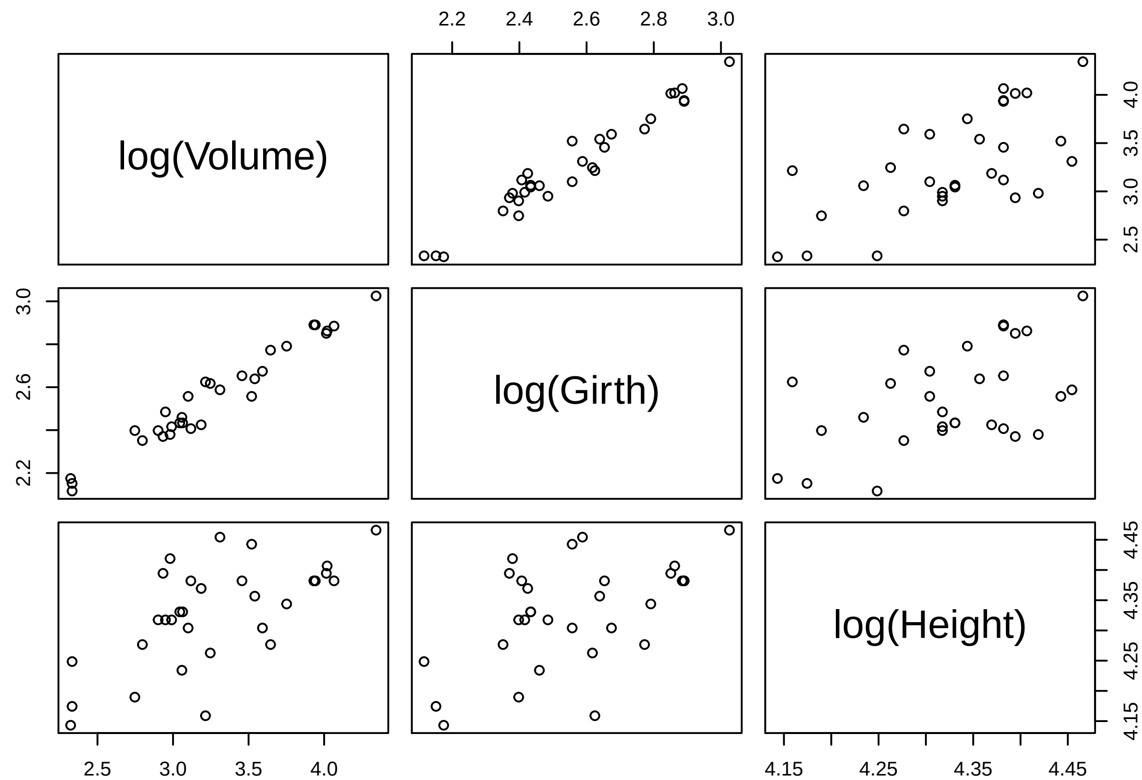
```r
trees
```

```
##    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
## 7   11.0     66   15.6
## 8   11.0     75   18.2
## 9   11.1     80   22.6
## 10  11.2     75   19.9
## 11  11.3     79   24.2
## 12  11.4     76   21.0
## 13  11.4     76   21.4
## 14  11.7     69   21.3
## 15  12.0     75   19.1
## 16  12.9     74   22.2
## 17  12.9     85   33.8
## 18  13.3     86   27.4
## 19  13.7     71   25.7
## 20  13.8     64   24.9
## 21  14.0     78   34.5
## 22  14.2     80   31.7
## 23  14.5     74   36.3
## 24  16.0     72   38.3
## 25  16.3     77   42.6
## 26  17.3     81   55.4
## 27  17.5     82   55.7
## 28  17.9     80   58.3
## 29  18.0     80   51.5
## 30  18.0     80   51.0
## 31  20.6     87   77.0
```

(a) There are 31 rows and 3 columns in the dataset "trees". The variable names are "Girth", "Height" and "Volume".

(b) Here is the scatterplot:

```r
pairs(formula = log(Volume) ~ log(Girth) + log(Height), data = trees)
```



(c)

```r
cor(log(trees), method = c("pearson", "kendall", "spearman"))
```

```
##            Girth Height Volume
## Girth  1.0000 0.5302 0.9767
## Height 0.5302 1.0000 0.6486
## Volume 0.9767 0.6486 1.0000
```

(d) No, there are no missing values.

```r
is.na(cor(log(trees)))
```

```
##          Girth Height Volume
## Girth  FALSE  FALSE  FALSE
## Height FALSE  FALSE  FALSE
## Volume FALSE  FALSE  FALSE
```

(e)

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

```
##
## Call:
## lm(formula = log(Volume) ~ log(Girth) + log(Height), data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16856 -0.04849  0.00243  0.06364  0.12922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.632      0.800   -8.29  5.1e-09 ***
## log(Girth)     1.983      0.075   26.43  < 2e-16 ***
## log(Height)    1.117      0.204    5.46  7.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0814 on 28 degrees of freedom
## Multiple R-squared:  0.978,  Adjusted R-squared:  0.976
## F-statistic:  613 on 2 and 28 DF,  p-value: <2e-16
```

(f)

```r
mod$coefficients
```

```
## (Intercept)  log(Girth) log(Height)
##      -6.632       1.983       1.117
```

```r
X = model.matrix(mod)
Y = log(trees)$Volume
(beta_hat = solve(t(X) %*% X) %*% t(X) %*% Y)
```

```
##               [,1]
## (Intercept) -6.632
## log(Girth)   1.983
## log(Height)  1.117
```

These two results are the same above.

(g)

```r
hat_y <- X %*% beta_hat
head(hat_y)
```

```
##    [,1]
## 1 2.310
## 2 2.298
## 3 2.309
## 4 2.808
## 5 2.977
## 6 3.023
```

```r
head(mod$residuals)
```

```
##        1        2        3        4        5        6
##  0.02187  0.03426  0.01384 -0.01062 -0.04303 -0.04196
```

```r
head((t(mod$residuals) %*% mod$residuals/28))
```

```
##          [,1]
## [1,] 0.006624
```

Q3 The procedure is below:

$$\mathrm{SS}R = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The derivative of SSR with respect to β_0 is

$$\frac{d}{d\hat{\beta}_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = -2\sum_{i=1}^{n} y_i + 2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i = -2n\bar{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\bar{x}$$

Set this derivative $= 0$, we can get

$$-2n\bar{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\bar{x} = 0.$$

So

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

The derivative of SSR with respect to β_1 is

$$\frac{d}{d\hat{\beta}_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = -2x_i \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2\sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2$$

Set this derivative $= 0$, we can get

$$-2\sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = 0.$$

So

$$-\sum_{i=1}^{n} x_i y_i + (\bar{y} - \hat{\beta}_1\bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\hat{\beta}_1 = \frac{\bar{\mathrm{y}} \sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n} x_i y_i}{\bar{x} \sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n} x_i^2} = \frac{n\bar{x}\bar{\mathrm{y}} - \sum\limits_{i=1}^{n} x_i y_i}{n\bar{x}^2 - \sum\limits_{i=1}^{n} x_i^2} = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$$

Proved