Homework 1

PSTAT 126 Winter 2023

Due date: January 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)
```

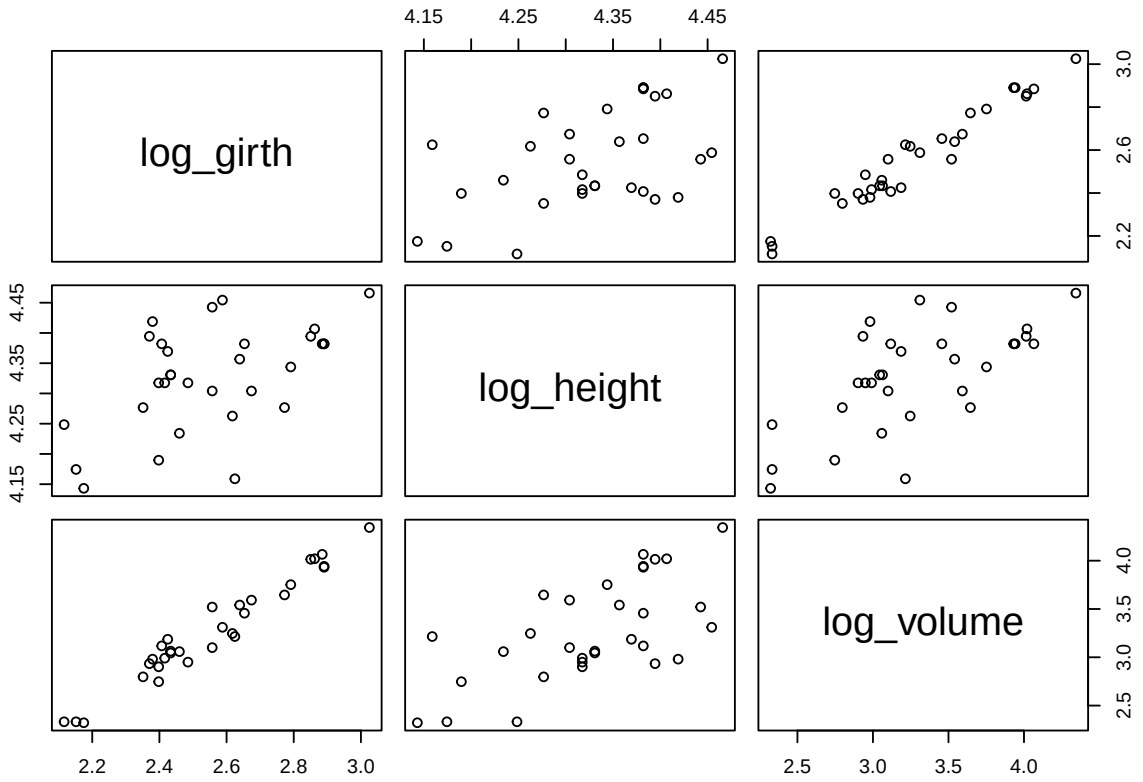
```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

   (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?

The data set trees provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. There consists 31 observations and 3 variables in the data set. The variables consist of "Girth", "Height", and "Volume."

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

```
log_girth <- log(trees$Girth)
log_height <- log(trees$Height)
log_volume <- log(trees$Volume)
Logs <- data.frame(log_girth,log_height,log_volume)
pairs(Logs)
```



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

```
cor(Logs)
```

```
##            log_girth log_height log_volume
## log_girth     1.0000     0.5302     0.9767
## log_height    0.5302     1.0000     0.6486
## log_volume    0.9767     0.6486     1.0000
```

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

When looking at the data set and producing the function summary(), I see that there are no values named "N/A"

```
summary(Logs)
```

```
##    log_girth      log_height     log_volume
##  Min.   :2.12   Min.   :4.14   Min.   :2.32
##  1st Qu.:2.40   1st Qu.:4.28   1st Qu.:2.96
##  Median :2.56   Median :4.33   Median :3.19
##  Mean   :2.56   Mean   :4.33   Mean   :3.27
##  3rd Qu.:2.72   3rd Qu.:4.38   3rd Qu.:3.62
##  Max.   :3.03   Max.   :4.47   Max.   :4.34
```

(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.

```
fit <- lm(log_volume ~ log_girth + log_height, data = Logs)
summary(fit)
```

```
##
## Call:
## lm(formula = log_volume ~ log_girth + log_height, data = Logs)
##
## 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) (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: $\hat{\beta} = (X^T X)^{-1} X^T y$.

From the Coefficient vector I produced using the formula above, we can see that it has the same coefficients as the summary output.

```
y= cbind(log_volume)
Ones = rep(1, 31)
X = cbind(Ones, log_girth, log_height)
X
```

```
##       Ones log_girth log_height
##  [1,]    1     2.116      4.248
##  [2,]    1     2.152      4.174
##  [3,]    1     2.175      4.143
##  [4,]    1     2.351      4.277
##  [5,]    1     2.370      4.394
##  [6,]    1     2.380      4.419
##  [7,]    1     2.398      4.190
##  [8,]    1     2.398      4.317
##  [9,]    1     2.407      4.382
## [10,]    1     2.416      4.317
## [11,]    1     2.425      4.369
## [12,]    1     2.434      4.331
```

```
## [13,]    1     2.434      4.331
## [14,]    1     2.460      4.234
## [15,]    1     2.485      4.317
## [16,]    1     2.557      4.304
## [17,]    1     2.557      4.443
## [18,]    1     2.588      4.454
## [19,]    1     2.617      4.263
## [20,]    1     2.625      4.159
## [21,]    1     2.639      4.357
## [22,]    1     2.653      4.382
## [23,]    1     2.674      4.304
## [24,]    1     2.773      4.277
## [25,]    1     2.791      4.344
## [26,]    1     2.851      4.394
## [27,]    1     2.862      4.407
## [28,]    1     2.885      4.382
## [29,]    1     2.890      4.382
## [30,]    1     2.890      4.382
## [31,]    1     3.025      4.466
```

```
Transposed_X = t(X)
inv_XT_times_X = solve((Transposed_X %*% X))
inv_times_XT = (inv_XT_times_X %*% Transposed_X)
Coefficients = inv_times_XT %*% y
Coefficients
```

```
##            log_volume
## Ones           -6.632
## log_girth       1.983
## log_height      1.117
```

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

```
Response_Values = X %*% Coefficients
Response_Values
```

```
##       log_volume
##  [1,]      2.310
##  [2,]      2.298
##  [3,]      2.309
##  [4,]      2.808
##  [5,]      2.977
##  [6,]      3.023
##  [7,]      2.803
##  [8,]      2.946
##  [9,]      3.036
## [10,]      2.981
## [11,]      3.057
## [12,]      3.031
## [13,]      3.031
## [14,]      2.975
```

```
## [15,]      3.118
## [16,]      3.247
## [17,]      3.401
## [18,]      3.475
## [19,]      3.320
## [20,]      3.218
## [21,]      3.468
## [22,]      3.524
## [23,]      3.478
## [24,]      3.643
## [25,]      3.755
## [26,]      3.929
## [27,]      3.966
## [28,]      3.983
## [29,]      3.994
## [30,]      3.994
## [31,]      4.355
```

```
Residual_Values = y - Response_Values
Residual_Values
```

```
##       log_volume
##  [1,]   0.021874
##  [2,]   0.034264
##  [3,]   0.013841
##  [4,]  -0.010619
##  [5,]  -0.043031
##  [6,]  -0.041961
##  [7,]  -0.055660
##  [8,]  -0.044315
##  [9,]   0.082173
## [10,]   0.009259
## [11,]   0.129223
## [12,]   0.013173
## [13,]   0.032041
## [14,]   0.083801
## [15,]  -0.168561
## [16,]  -0.146549
## [17,]   0.119002
## [18,]  -0.164525
## [19,]  -0.073211
## [20,]  -0.003299
## [21,]   0.073269
## [22,]  -0.067780
## [23,]   0.113363
## [24,]   0.002431
## [25,]  -0.002998
## [26,]   0.085102
## [27,]   0.054006
## [28,]   0.082405
## [29,]  -0.052661
## [30,]  -0.062417
## [31,]  -0.011641
```

```
Error_Variance = (sum(t(Residual_Values) %*% Residual_Values))/28
Error_Variance
```

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

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?

The interpretation of this assumption is that the regression line will start at 0. This is because y_i will equal 0 when $\beta_1 x_i + \epsilon_i = 0$. So when $x_i = 0$, $y_i = 0$. We can also assume that $\bar{y} = \hat{\beta}_1 \bar{x}$

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

We derive the LS estimate of β_1 using this formula: $0 = \frac{d}{d\hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. So when $\hat{\beta}_0 = 0$, we have $0 = \frac{d}{d\hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2$. This simplifies to $0 = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 x_i)$ which then simplifies to $0 = \sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$. From here we can see that $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$.

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

We can introduce this assumption within the *lm* function by listing the observations as so:

lm(y ~ x - 1, data = dataset). We can get ride of an intercept by subtracting the predictor by 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?

The interpretation of this assumption is that the relationship between x_i and y_i is not linear. The line in our model will be horizontal with no slope. A change in x does not correspond to a change in y

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

We will start at a similar spot as we did in part b of this question since we are minimizing the sum of the squared residuals. Except this time we will take the derivative with respect to $\hat{\beta}_0$. This looks like: $0 = \frac{d}{d\hat{\beta}_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ which we can reduce into $0 = \frac{d}{d\hat{\beta}_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2$ since $\hat{\beta}_1 = 0$. This reduces to $0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0)$ and finally, $0 = (n\bar{y} - n\hat{\beta}_0)$. From here, we can see that $\hat{\beta}_0 = \bar{y}$ when $\beta_1 = 0$.

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

We would implement this into the *lm* function by using the syntax: lm(y ~ 0 + x, data = dataset) so that we have an exact horizontal line.

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.

Using the LS estimation general result $\hat{\beta} = (X^T X)^{-1} X^T y$, we can find our values of $\hat{\beta}_0$ and $\hat{\beta}_1$ by using matrix multiplication. First we would need to design our X matrix with ones in the first column and then our respective predictors in the next columns. Then use the functions in R to find the inverse of $X^T X$ and use the matrix multiplication function to find our final vector of $\hat{\beta}$. For a more explicit reference, part f of question 1 uses these exact tendencies to find $\hat{\beta}$. The final outcome should be equal to a vector of quantities which represent our $\hat{\beta}$. The number of quantities in this final vector is equal to the amount of variables we decide to use for our model. I will provide the code I used in part f of question 1 to show how we would use the formula to find the explicit estimates for β_0 and β_1

```
y= cbind(log_volume)
Ones = rep(1, 31)
X = cbind(Ones, log_girth, log_height)
Transposed_X = t(X)
inv_XT_times_X = solve((Transposed_X %*% X))
inv_times_XT = (inv_XT_times_X %*% Transposed_X)
Coefficients = inv_times_XT %*% y
Coefficients
```

```
##            log_volume
## Ones           -6.632
## log_girth       1.983
## log_height      1.117
```

Thus, entry $a1$ of value -6.632 corresponds to $\hat{\beta}_0$ and so on. This is how we would explicitly calculate our $\hat{\beta}$ vector using the general formula $\hat{\beta} = (X^T X)^{-1} X^T y$.

(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 $\hat{\beta}_0$, we have the equation: $\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{y} - \hat{\beta}_1 \bar{x}]$ which is then equal to $\beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$ which is simply just β_0. Thus, $\mathbb{E}[\hat{\beta}_0] = \beta_0$ and we can say that $\hat{\beta}_0$ is an unbiased estimate for β_0

For $\hat{\beta}_1$, we have the equation: $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\frac{\sum_{i=1}^{n}(x_i \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$. Then we can simplify this expression into the form: $\frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}[y_i]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. We know that $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$ so we can simplify our expression more to achieve the final form of $(\beta_1)\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ which is just equal to β_1. This shows that $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and thus, that $\hat{\beta}_1$ is an unbiased estimate for β_1.