Homework 1

Alex Wako

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)
```

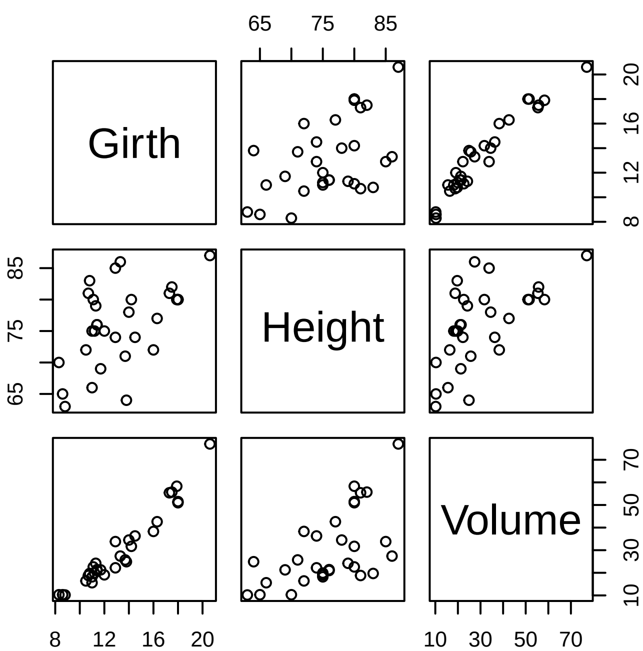
```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

(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?

```
## [1] 31  3
```

The trees data set has 31 rows and 3 variables. The names of the variables 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.



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

```
##        Girth Height Volume
## Girth  1.0000 0.5193 0.9671
## Height 0.5193 1.0000 0.5982
## Volume 0.9671 0.5982 1.0000
```

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

```
## [1] 0
```

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.

```
##
## Call:
## lm(formula = log(y) ~ log(x1) + log(x2))
##
## 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(x1)        1.983      0.075   26.43  < 2e-16 ***
## log(x2)        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: $\hat{\beta} = (X^T X)^{-1} X^T y$.

```
##    (Intercept) log(x1) log(x2)
## 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
## attr(,"assign")
## [1] 0 1 2
```

```
##        [,1]
## [1,] -6.632
## [2,]  1.983
## [3,]  1.117
```

The least squares coefficient given in the summary output matches the least squares coefficient found through $\hat{\beta} = (X^T X)^{-1} X^T y$.

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

Predicted response values: 2.3103, 2.2979, 2.3085, 2.8079, 2.9769, 3.0226, 2.8029, 2.9457, 3.0358, 2.9815, 3.0571, 3.0313, 3.0313, 2.9749, 3.1182, 3.2466, 3.4015, 3.4751, 3.3197, 3.2182, 3.4677, 3.5241, 3.4785, 3.643, 3.7549, 3.9295, 3.966, 3.9832, 3.9942, 3.9942, 4.3554

The residuals: 0.0219, 0.0343, 0.0138, -0.0106, -0.043, -0.042, -0.0557, -0.0443, 0.0822, 0.0093, 0.1292, 0.0132, 0.032, 0.0838, -0.1686, -0.1465, 0.119, -0.1645, -0.0732, -0.0033, 0.0733, -0.0678, 0.1134, 0.0024, -0.003, 0.0851, 0.054, 0.0824, -0.0527, -0.0624, -0.0116

Estimate of error variance: 0.0066

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?

When $\beta_0 = 0$, we can interpret the model as having no intercept. The errors are unobservable random variables with mean of 0 and variance of σ^2. The mean of y_i would be $\beta_1 x_i$, the variance of y_i would be σ^2, and the covariance would be 0. Therefore, the plot of the model would be a slope starting at the origin. The new model would be $y_i = \beta_1 x_i + \epsilon_i$ regression line.

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

arg min_{β_0} SSR $= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
$\dfrac{\partial}{\partial \beta_1}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
$= \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$
Since $\beta_0 = 0$:
$= \sum_{i=1}^{n} x_i(y_i - \beta_1 x_i)$
$= \sum_{i=1}^{n} x_i y_i - \beta_1 x_i^2)$
$= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \beta_1 x_i^2$
$= \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2$
$=> \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$
$=> \sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} x_i^2$
$=> \beta_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

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

We can introduce the assumption within the lm function by adding 0 into the formula to indicate that a constant does not exist in the model.

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?

When $\beta_1 = 1$, we can interpret the model as not having a slope. The errors are still unobservable random variables with mean of 0 and variance of σ^2. The mean of y_i would be β_0, the variance of y_i would still be σ^2, and the covariance would still be 0. Therefore, the plot of the model would always be a constant horizontal line. The new model would be $y_i = \beta_0 + \epsilon_i$ regression line.

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

arg min_{β_0} SSR $= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
$\dfrac{\partial}{\partial \beta_0}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
$= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$
= n$\overline{y}$ - nβ_0 - n$\beta_1\overline{x} = 0$
Since $\beta_1 = 0$:
=> n$\overline{y}$ - n$\beta_0 = 0$
=> n$\overline{y}$ = nβ_0
$=> \overline{y} = \beta_0$

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

We can introduce the assumption within the lm function by creating a formula relating y_i to only β_0.

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.

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

We are trying to solve for β_0 and β_1, so for the purpose of this problem, let $\beta_0 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} = \frac{SS_{xy}}{SS_x}$.

$(X^TX)^{-1} = \frac{1}{n}\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1}$
$= \frac{1}{nSS_x}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum xi & n \end{bmatrix}$

$(X^Ty) = \begin{bmatrix} 1 & ... & 1 \\ x_1 & ... & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ ... \\ y_n \end{bmatrix}$
$= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{bmatrix}$

$(X^TX)^{-1}X^Ty = \frac{1}{nSS_x}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum xi & n \end{bmatrix}\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_iy_i \end{bmatrix}$
$= \frac{1}{nSS_x}\begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_iy_i \\ -\sum x_i \sum y_i + n\sum x_iy_i \end{bmatrix}$
$= \frac{1}{SS_x}\begin{bmatrix} \overline{y}\sum x_i^2 - \overline{x}\sum x_iy_i \\ \sum x_iy_i - n\overline{xy} \end{bmatrix}$
$= \frac{1}{SS_x}\begin{bmatrix} \overline{y}\sum x_i^2 - \overline{y}n\overline{x}^2 + \overline{x}n\overline{xy} - \overline{x}\sum x_iy_i \\ SS_{xy} \end{bmatrix}$
$= \frac{1}{SS_x}\begin{bmatrix} \overline{y}SS_x - SS_{xy}\overline{x} \\ SS_{xy} \end{bmatrix}$
$= \begin{bmatrix} \overline{y} - \frac{SS_{xy}}{SS_x}\overline{x} \\ \frac{SS_{xy}}{SS_x} \end{bmatrix} => \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

(b) (5pts) Show that the LS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates for β_0 and β_1 respectively.

$Bias[\hat{\beta}_1] = E[\hat{\beta}_1] - \beta_1$
$E[\hat{\beta}_1] = E[\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}]$
$= \frac{E[\sum(X_i-\overline{X})Y_i - \overline{Y}\sum(X_i-\overline{X})]}{\sum(X_i-\overline{X})^2}$
$= \frac{\sum(X_i-\overline{X})E[Y_i]}{\sum(X_i-\overline{X})^2}$
$= \frac{\sum(X_i-\overline{X})(\beta_0-\beta_1X_i)}{\sum(X_i-\overline{X})^2}$
$= \frac{\beta_0\sum(X_i-\overline{X})+\beta_1\sum(X_i-\overline{X})X_i}{\sum(X_i-\overline{X})^2}$
$= \beta_1\frac{\sum(X_i-\overline{X})(X_i-\overline{X}+\overline{X})}{\sum(X_i-\overline{X})^2}$
$= \beta_1\frac{\sum(X_i-\overline{X})^2}{\sum(X_i-\overline{X})^2}$
$= \beta_1$
$Bias[\hat{\beta}_1] = \beta_1 - \beta_1 = 0$

$Bias[\hat{\beta}_0] = E[\hat{\beta}_0] - \beta_0$
$E[\hat{\beta}_0] = E[\overline{Y} - \hat{\beta}_1\overline{X}]$
$= E[\frac{1}{n}\sum Y_i - \hat{\beta}_1\overline{X}]$
$= \frac{1}{n}\sum E[Y_i] - E[\hat{\beta}_1]\overline{X}$
$= \frac{1}{n}\sum(\beta_0 + \beta_1X_i) - \beta_1\overline{X}$
$= \beta_0 + \beta_1\overline{X} - \beta_1\overline{X}$
$= \beta_0$
$Bias[\hat{\beta}_0] = \beta_0 - \beta_0 = 0$

Appendix

```r
library(knitr)
library(MASS)
# set global chunk options: images will be 7x5 inches
knitr::opts_chunk$set(fig.width=7, fig.height=5)
options(digits = 4)
# the data set "trees" is contained in the R package "datasets"
require(datasets)
head(trees)
# The dimensions of the tree data set
(dim(trees))
# Scatter plot matrix of the three variables
pairs(trees)
# Correlation matrix of the three variables
cor(trees)
# Number of NA values
sum(is.na(trees))
# Creating variables representing y, x1, and x2
y <- trees$Volume
x1 <- trees$Girth
x2 <- trees$Height

# Fitting a linear model to the tree data using the given formula
lm_tree <- lm(log(y) ~ log(x1) + log(x2))
summary(lm_tree)
model_matrix <- model.matrix(lm_tree)
model_matrix
beta_hat <- ginv(t(model_matrix) %*% model_matrix) %*% t(model_matrix) %*% log(y)
beta_hat
beta0_hat <- -6.631617
beta1_hat <- 1.982650
beta2_hat <- 1.117123
y_hat <- beta0_hat + beta1_hat * log(x1) + beta2_hat * log(x2)
residual <- log(y) - y_hat
estimate_of_error <- sum((residual)^2)/28
```