Homework Assignment 1

Zejie Gao

2023-02-01

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.

```
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) 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
```

```
ls(trees)
```

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

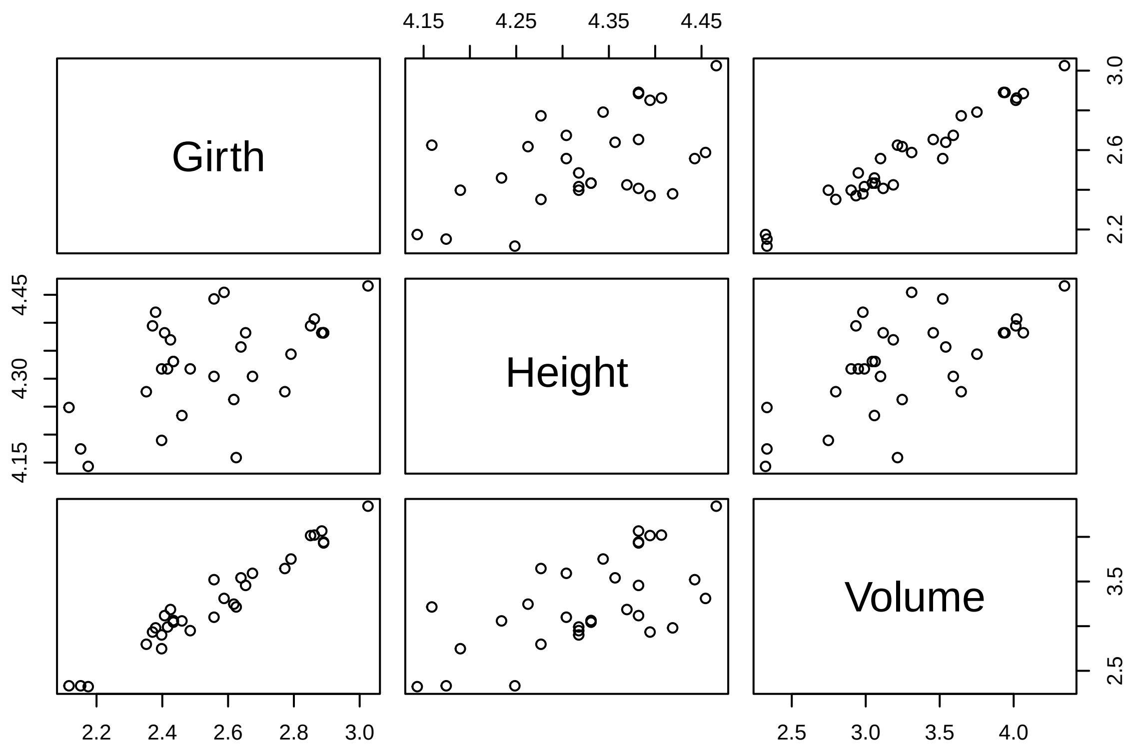
```
# There are 31 rows and 3 column in the data set, with three variables named "Girth","Height", and  "Vo
```

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
log_trees <- trees |>
  mutate(Girth = log(Girth)) |>
  mutate(Height = log(Height))|>
  mutate(Volume = log(Volume))
pairs(log_trees)
```



(c) 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) Are there missing values?

```r
is.na(log_trees)
```

```
##       Girth Height Volume
##  [1,] FALSE  FALSE  FALSE
##  [2,] FALSE  FALSE  FALSE
##  [3,] FALSE  FALSE  FALSE
##  [4,] FALSE  FALSE  FALSE
##  [5,] FALSE  FALSE  FALSE
##  [6,] FALSE  FALSE  FALSE
##  [7,] FALSE  FALSE  FALSE
##  [8,] FALSE  FALSE  FALSE
##  [9,] FALSE  FALSE  FALSE
## [10,] FALSE  FALSE  FALSE
## [11,] FALSE  FALSE  FALSE
## [12,] FALSE  FALSE  FALSE
## [13,] FALSE  FALSE  FALSE
## [14,] FALSE  FALSE  FALSE
## [15,] FALSE  FALSE  FALSE
## [16,] FALSE  FALSE  FALSE
## [17,] FALSE  FALSE  FALSE
## [18,] FALSE  FALSE  FALSE
## [19,] FALSE  FALSE  FALSE
## [20,] FALSE  FALSE  FALSE
## [21,] FALSE  FALSE  FALSE
## [22,] FALSE  FALSE  FALSE
## [23,] FALSE  FALSE  FALSE
## [24,] FALSE  FALSE  FALSE
## [25,] FALSE  FALSE  FALSE
## [26,] FALSE  FALSE  FALSE
## [27,] FALSE  FALSE  FALSE
## [28,] FALSE  FALSE  FALSE
## [29,] FALSE  FALSE  FALSE
## [30,] FALSE  FALSE  FALSE
## [31,] FALSE  FALSE  FALSE
```

```r
sum(is.na(log_trees))
```

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

```r
# No, there is no missing values
```

(e) Use the lm function in R to fit the multiple regression model: log(Volumei) = beta0 + beta1 log(Girthi) + beta2 log(Heighti) + ei and print out the summary of the model fit.

```r
fit <- lm(Volume ~ Girth + Height, data = log_trees)
fit
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = log_trees)
##
```

```
## Coefficients:
## (Intercept)        Girth       Height
##      -6.632        1.983        1.117
```

```r
y <- log_trees$Volume
x1 <- log_trees$Girth
x2 <- log_trees$Height
R.2 <- 1 - sum((fit$residuals^2))/ (sum((y - mean(y))^2))
R.2
```

```
## [1] 0.9776784
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = log_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 ***
## Girth        1.98265    0.07501  26.432  < 2e-16 ***
## 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
```

```r
# Estimation of the lm: log(Volumei) = -6.63 + 1.98 log(Girthi) + 1.12 log(Heighti) + ei
# Since R^2 (0.9777 or 0.978 on the summary) is very close to 1, the model better fits the data.
```

(f)

```r
v1 <- log_trees$Girth
v2 <- log_trees$Height
X <- cbind(rep(1, times=nrow(log_trees)),v1,v2)
y <- matrix(log_trees$Volume)
beta_hat <- solve(t(X)%*%(X))%*%(t(X)%*%(y))
beta_hat
```

```
##         [,1]
##    -6.631617
## v1  1.982650
## v2  1.117123
```

```r
# The beta_hat matrix match the output I got in (e)
```

(g) Compute the predicted response values from the fitted regression model, the residuals, and an estimate of the error variance.

```r
y_hat <- X%*%beta_hat
y_hat
```

```
##           [,1]
##  [1,] 2.310270
##  [2,] 2.297879
##  [3,] 2.308547
##  [4,] 2.807900
##  [5,] 2.976888
##  [6,] 3.022580
##  [7,] 2.802931
##  [8,] 2.945736
##  [9,] 3.035777
## [10,] 2.981461
## [11,] 3.057130
## [12,] 3.031349
## [13,] 3.031349
## [14,] 2.974906
## [15,] 3.118250
## [16,] 3.246641
## [17,] 3.401459
## [18,] 3.475068
## [19,] 3.319702
## [20,] 3.218167
## [21,] 3.467691
## [22,] 3.524097
## [23,] 3.478455
## [24,] 3.643019
## [25,] 3.754853
## [26,] 3.929478
## [27,] 3.965974
## [28,] 3.983197
## [29,] 3.994242
## [30,] 3.994242
## [31,] 4.355446
```

```r
Res <- fit$residuals
sigma2.hat <- sum(Res^2) / fit$df.residual
sigma2.hat
```

```
## [1] 0.006623692
```

```r
SSR <- sum(fit$residuals^2)
SSR
```

```
## [1] 0.1854634
```

```
# SSR = 0.1855
# Var(ei) = 0.006624
```

2. Consider the simple linear regression model:

(a) Assume beta0 = 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 assumption represent that the yi will have a high likelihood to be zero when xi equals to zero.
# It implicate the y intercept of linear regression model is zero.
# The regression line start from coordinate (0,0).
```

(b) Derive the LS estimate of beta1 when beta0 = 0.

```
# beta1 = sum((y-mean(y))*(x-mean(x)))/sum((x-mean(x))^2)
# beta1 does not influenced by beta0 value
```

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

```
# lm(y ~ x-1, data = dataset)
# Based on the assumption beta0 = 0, beta0 from lm function can be deleted.Then, the lm function will b
```

(d) For the same model, assume beta1 = 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 assumption represent that the value of yi does not affected by the value of xi. The predictor x1
# It implicate the linear regression model wasn't able to find a linear relationship between the yi and
# The plot will only have a horizontal line which is y = constant value (beta0 + ei) or so called the m
```

(e) Derive the LS estimate of beta0 when beta1 = 0.

```
# beta0 = mean(y)
```

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

```
# lm(y ~ 1, data = dataset)
# Based on the assumption beta1 = 0, beta1 part from lm function can be deleted.Then, the lm function w
```

3. Consider the simple linear regression model:

(a) Use the LS estimation general result beta_hat = to find the explicit estimates for beta0 and beta1.

```
# on the pdf
```

(b) Show that the LS estimates beta0_hat and beta1_hat are unbiased estimates for beta0 and beta1 respectively.

```
# on the pdf
```