Page 1

October 14, 2019 2:54 PM

Hair Alberto Parra Barrera
26073861G

MATH 533
HMW 2

Question 1

Suppose that Y is a random variable and

$$Y = \beta_0 + \varepsilon$$

where $\mathbb{E}[\varepsilon]=0$ and $\text{Var}[\varepsilon]=\sigma^2$.
Suppose that we know the distribution of Y.
We would like to predict Y. If we require that the prediction is just a constant, denote c.
Now we ask the question "What is the optimal prediction constant c we can get?", i.e.

$$\boxed{c^* = \arg\min_c \mathbb{E}_Y[(Y-c)^2]}$$

Prove that the optimal solution is

$$c^* = \mathbb{E}[Y] = \beta_0$$

Solution

For any r.v. X, we have

$$\text{Var}[X] = \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2$$
$$\Leftrightarrow \mathbb{E}[X^2] = \text{Var}[X] + \{\mathbb{E}[X]\}^2$$

So that for a constant c, letting $X = Y - c$

$$\mathbb{E}[(Y-c)^2] = \text{Var}[Y-c] + \{\mathbb{E}[Y-c]\}^2$$
(MSE = Variance + Bias^2)

$$= \text{Var}[Y] + (\mathbb{E}[Y]-c)^2$$
$$= \text{Var}[\beta_0+\varepsilon] + (\mathbb{E}[\beta_0+\varepsilon]-c)^2$$
$$= \text{Var}[\varepsilon] + (\mathbb{E}[\varepsilon]+\beta_0-c)^2$$
$$= \sigma^2 + (0+\beta_0-c)^2$$
$$\Rightarrow \frac{\partial \mathbb{E}_Y[(Y-c)^2]}{\partial c} = 0 - 2\beta_0 - 2c := 0$$
$$\Rightarrow \boxed{c^* = \beta_0 = \mathbb{E}[Y]} \quad \blacksquare$$

Question 2

Suppose that in the following intercept-only model

$$Y_i = \beta_0 + \varepsilon_i, \quad i = 1, \dots, n$$

where $\mathbb{E}[\varepsilon_i]=0$ and $\text{Var}[\varepsilon_i]=\sigma^2$

(a) Use the result from Q1 to derive the plug-in estimate for β_0.

Sol

From (1), $\beta_0 = \mathbb{E}[Y] \Rightarrow \boxed{\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^n y_i}$

(b) Find the least squares estimate $\hat{\beta}_0$.

Sol From (1),

$$MSE(\hat{\beta}_0) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_0)^2$$
$$\Rightarrow \frac{\partial MSE(\hat{\beta}_0)}{\partial \beta_0} = -\frac{2}{n}\sum_{i=1}^n (y_i - \beta_0) := 0$$
$$= \bar{y} - \hat{\beta}_0 = 0 \Rightarrow \boxed{\hat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i}$$

(c) Show that $\hat{\beta}_0$ is unbiased.

Proof

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Y_i\right]$$
$$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_i] = \frac{1}{n}\sum_{i=1}^n \beta_0 + \mathbb{E}[\varepsilon_i] = \boxed{\beta_0}$$

Question 3

To be attached later

Question 4

Suppose that for the SLR model, we use the estimator

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n c_i^2$$

What is the amount of bias in the estimator?

(Turn the page)

Solution NOTE: – In what follows, the summation under and superscripts are understood.
– We consider the Y_i's as r.v's.

$\hat{\sigma}^2 = \frac{1}{n}\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$ (*) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$

$= \frac{1}{n}\sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1}))^2$

$\overset{(*)}{=} \frac{1}{n}\sum (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_1 X_{i1})^2$

$= \frac{1}{n}\sum [(Y_i - \bar{Y}) - (\hat{\beta}_1 (X_{i1} - \bar{X}_1))]^2$

(1) $= \boxed{\frac{1}{n}\left[\sum (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum (X_{i1} - \bar{X}_1)(Y_i - \bar{Y}) + \hat{\beta}_1^2 \sum (X_{i1} - \bar{X}_1)^2\right]}$

But,

(2) $\boxed{\sum (X_{i1} - \bar{X}_1)(Y_i - \bar{Y}) = \sum (X_{i1} - \bar{X}_1)^2 \hat{\beta}_1}$

since $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \Rightarrow S_{xy} = S_{xx}\hat{\beta}_1$,

and also (3) $\boxed{\sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2}$

so (1) becomes

$= \frac{1}{n}\left[\sum Y_i^2 - n\bar{Y}^2 - 2\hat{\beta}_1^2 \sum (X_{i1} - \bar{X}_1)^2 + \hat{\beta}_1^2 \sum (X_{i1} - \bar{X}_1)^2\right]$

$= \boxed{\frac{1}{n}\left[\sum Y_i^2 - n\bar{Y}^2 - \hat{\beta}_1^2 S_{xx}\right]\Big|_{X=x}}$ (4)

$\Rightarrow E_{Y|X=x}[\hat{\sigma}^2]$

$= \boxed{\frac{1}{n}\left[\sum E[Y_i^2] - nE[\bar{Y}^2] - S_{xx} E[\hat{\beta}_1^2]\right]}$ (5)

Now, we solve $E(Y_i^2)$, $E[\bar{Y}^2]$ and $E[\hat{\beta}_1^2]$ sequentially,

- $\boxed{E[\hat{\beta}_1^2]} = \mathrm{Var}[\hat{\beta}_1] + \{E[\hat{\beta}_1]\}^2$
$= \boxed{\frac{\sigma^2}{S_{xx}} + \beta_1^2}$

- $\boxed{E[Y_i^2]} = \mathrm{Var}[Y_i] + \{E[Y_i]\}^2$
$= \boxed{\sigma^2 + (\beta_0 + \beta_1 x_{i1})^2}$

- $\boxed{E[\bar{Y}^2]} = \mathrm{Var}[\bar{Y}] + \{E[\bar{Y}]\}^2$
$= \mathrm{Var}\left[\frac{1}{n}\sum Y_i\right] + \left\{E\left[\frac{1}{n}\sum Y_i\right]\right\}^2$

$= \frac{1}{n^2}\left[\sum_i \mathrm{Var}[Y_i] - 2\sum_{1 \le i < j \le n} \mathrm{Cov}[Y_i, Y_j]\right] + \left\{\frac{1}{n}\sum E[Y_i]\right\}^2$

($= 0$, as $Y_i \perp\!\!\!\perp Y_j$)

$= \frac{\sigma^2}{n} + \left(\frac{1}{n}\sum (\beta_0 + \beta_1 x_{i1})\right)^2$

$= \boxed{\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x}_1)^2}$

So plugging these into (5) yields

$$\boxed{E[\hat{\sigma}^2] = \frac{1}{n}\left[\sum \left(\sigma^2 + (\beta_0 + \beta_1 x_{i1})^2\right) - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x}_1)^2\right) - S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right)\right]}$$

$= \frac{1}{n}\left[n\sigma^2 + \sum (\beta_0 + \beta_1 x_{i1})^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x}_1)^2 - \sigma^2 - S_{xx}\beta_1^2\right]$

$= \frac{1}{n}\left[n\sigma^2 - 2\sigma^2 + \sum (\beta_0^2 + 2\beta_0\beta_1 x_{i1} + \beta_1^2 x_{i1}^2) - n(\beta_0^2 + 2\beta_0\beta_1 \bar{x}_1 + \beta_1^2 \bar{x}_1^2) - S_{xx}\beta_1^2\right]$

$= \frac{1}{n}\left[n\sigma^2 - 2\sigma^2 + n\beta_0^2 + 2n\beta_0\beta_1 \bar{x}_1 + \beta_1^2 \sum x_{i1}^2 - n\beta_0^2 - 2n\beta_0\beta_1 \bar{x}_1 - \beta_1^2 n\bar{x}_1^2 - S_{xx}\beta_1^2\right]$

October 14, 2019 6:54 PM

$$= \frac{1}{n}\left[n\sigma^2 - 2\sigma^2 + n\beta_0^2 + 2n\beta_0\beta_1\bar{x}_1 + \beta_1^2\sum x_{i1}^2 - n\beta_0^2 - 2n\beta_0\beta_1\bar{x}_1 - \beta_1^2 n\bar{x}_1^2 - \beta_1^2 S_{xx}\right]$$

As (*) $\boxed{\sum (x_{i1} - \bar{x}_1)^2 = \sum x_{i1}^2 - n(\bar{x}_1)^2}$

$$= \frac{1}{n}\left[\sigma^2(n-2) + \beta_1^2\left(\sum x_{i1}^2 - n\bar{x}_1^2\right) - \beta_1^2 S_{xx}\right]$$

$$= \frac{1}{n}\left[\sigma^2(n-2) + \beta_1^2 S_{xx} - \beta_1^2 S_{xx}\right] = \left(\frac{n-2}{n}\right)\sigma^2$$

So we have,

$$\boxed{E\left[\tilde{\sigma}^2\right] = \left(\frac{n-2}{n}\right)\sigma^2}, \text{ so}$$

$$\boxed{\text{Bias}(\tilde{\sigma}^2)} = E[\tilde{\sigma}^2] - \sigma^2 = \left(\frac{n-2}{n}\right)\sigma^2 - \sigma^2$$

$$= \boxed{\sigma^2\left(\frac{n-2}{n} - 1\right)}$$

■

d Question 7 (MATH 533)

d Study Question 2.3.1

Figure 2.5 A directed graph for demonstrating conditional independence (error terms are not shown explicitly)

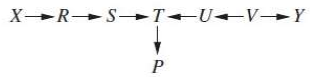
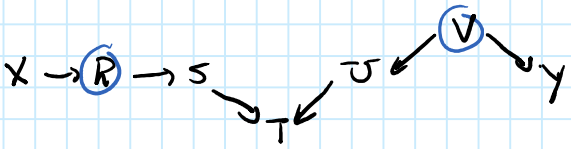


Figure 2.6 A directed graph in which P is a descendant of a collider

(a) List all pairs of variables in figure 2.5 that are independent conditional on $Z = \{R, V\}$

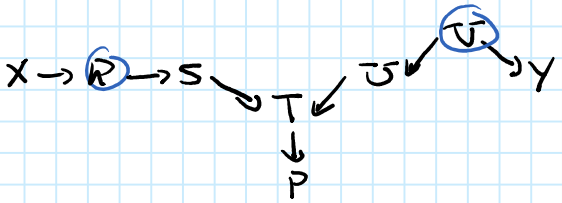


- $X \perp S \mid Z$
- $X \perp T \mid Z$
- $S \perp U$
- $S \perp Y$
- $S \perp Y \mid Z$
- $U \perp Y \mid Z$
- $T \perp Y \mid Z$
- $X \perp U$
- $X \perp U \mid Z$
- $X \perp Y$
- $S \perp U \mid Z$
- $X \perp Y \mid Z$?

(b) For each pair of unconfounded variables in Figure 2.5, give a set of variables that, when conditioned on, renders that pair independent.

- $X \perp S \mid \{R\}$
- $X \perp T \mid \{R, S\}$
- $X \perp U \mid \{R, S\}$
- $X \perp V \mid \{R, S, U\}$
- $X \perp V \mid \{T, U\}$
- $S \perp U \mid \{\emptyset\}$
- $S \perp U \mid \{T, U\}$
- $T \perp Y \mid \{U, V\}$
- $X \perp Y \mid \{T, U\}$
- $R \perp T \mid \{S\}$
- $R \perp U \mid \{S, T\}$
- $R \perp V \mid \{U, T\}$
- $R \perp Y \mid \{S, T, U\}$
- $S \perp Y \mid \{T, U\}$
- $T \perp V \mid \{U\}$
- $U \perp Y \mid \{V\}$

(c) List all pairs of variables in figure 2.6 that are independent conditional on the set $Z = \{R, P\}$.



- $X \perp S \mid Z$
- $X \perp T \mid Z$
- $X \perp Y \mid Z$
- $S \perp U$
- $S \perp Y$
- $U \perp Y \mid Z$
- $X \perp P \mid Z$
- $X \perp U$
- $S \perp U$
- $S \perp Y \mid Z$
- $X \perp U \mid Z$
- $S \perp U \mid Z$
- $T \perp Y \mid Z$

October 23, 2019 1:46 PM

d) For each pair of non-adjacent variables in Fig 2.6, give a set of variables that, when conditioned on, renders that pair independent.

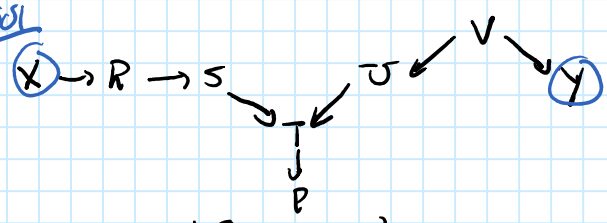
- $X \perp S \mid \{R\}$
- $X \perp P \mid \{R, S, T\}$
- $X \perp T \mid \{R, S\}$
- $X \perp U \mid \{R, S\}$
- $X \perp U \mid \phi$
- $X \perp Y \mid \{R\}$
- $R \perp T \mid \{S\}$
- $R \perp P \mid \{S, T\}$
- $R \perp U \mid \{S\}$
- $R \perp U \mid \{U\}$
- $R \perp Y \mid \{U, V\}$
- $S \perp U \mid \{X, Y\}$
- $S \perp U \mid \{U, Y\}$
- $S \perp Y \mid \{U\}$
- $T \perp V \mid \{U\}$
- $T \mid Y \mid \{U, V\}$
- $P \perp U \mid \{R, S\}$
- $P \perp U \mid \{U\}$
- $P \perp Y \mid \{U\}$
- $U \perp Y \mid \{V\}$

e) Suppose we generate data by the model described in Fig 2.5, and we fit the equation

$$Y = a + bX + cZ$$

What of the variables in the model may be chosen as Z so as to guarantee that the slope b would be equal to zero?

Sol



ans: $Z = \{R, S, U, V\}$

f) Suppose we fit the equation

$$Y = a + bX + cR + dS + eT + fP$$

which of the coefficients would be zero?

Answer: b, c,

Study question 2.4.1

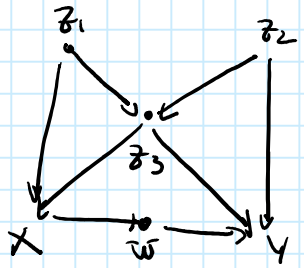


Fig 2.4

(a) For each pair of nonadjacent nodes, find a set of variables that d-separates that pair. What does this list tells us about the independencies in the data? (unconditionally)

Sol $(Z_1, Z_2) : \{Z_3\}$ $(Z_1, Y) : \{Z_3, Z_2\}$

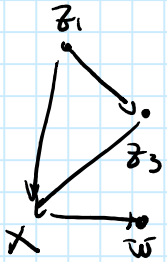
$(Z_1, X) : \{Z_3, Y, W\}$ $(Z_2, W) : \{Y\}$

$(Z_1, W) : \{Z_3, Y\}$ $(Z_3, W) : \{Y\}$

$(Z_2, X) : \{Y, W\}$ $(X, Y) : \{Z_1, Z_2, Z_3\}$

The list tells us what variables are independent about which pairs, and upon which other variables they would become dependent if conditioned on.

(b) Unconditionally d-separated

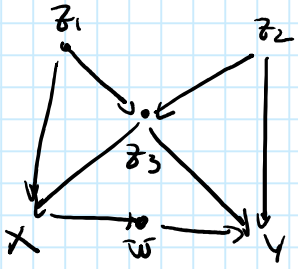


$(Z_1, W) : \phi$

$(Z_3, W) : \phi$

None of them are d-separated.

(c) For each pair of nonadjacent nodes in the graph, determine whether they are indeed conditioned on all other variables. Let $A = \{\text{all other variables}\}$



$Z_1 \perp W \mid A$ $Z_2 \perp X \mid A$

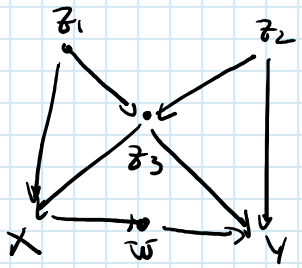
$Z_1 \not\perp Z_2 \mid A$ $Z_2 \not\perp W \mid A$

$Z_1 \perp Y \mid A$ $Z_3 \perp W \mid A$

$X \perp Y \mid A$.

October 23, 2019 3:04 PM

d) For every variable V in the graph, find a minimal set of nodes that renders V indep of all other variables in the graph when conditioned on?

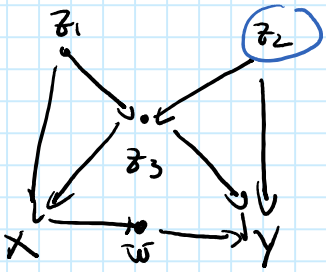


I'm assuming this since the question is not clear

- Z_1 : $\{X\}$
- Z_2 : $\{Z_3\}$
- Z_3 : $\{Z_1, Z_2, X\}$
- X : $\{Z_1, Z_3, W\}$
- W : $\{X, Y\}$
- Y : $\{W, Z_3, Z_2\}$

e) Suppose we wish to estimate the value of Y from measurements taken on all other variables. Smallest set just as good estimate?

Sol



$Y = a + b\cancel{X} + cW + dZ_1 + eZ_2 + fZ_3$

f) Same but for Z_2

$Z_2 = a + b\cancel{X} + cW + dZ_1 + eY + fZ_3$

g) Yes, we know from the model and from question f) that $Z_2 \not\perp Z_3$ and in the regression line, the coefficient of W cannot be 0, so it is statistically significant.

MATH 423/533: ASSIGNMENT 2

Hair Parra

October 14, 2019

Hair ALbeiro Parra Barrera

260738619

MATH 423/533: ASSG1

Assignment 2

Question 3

1. Fit the data and summarize the results

```
# Create the data
X_1 <- c(4,3,5,7)
Y <- c(6,2,4,11)

# fit a linear model with it
m0 <- lm(Y~X_1)

# display a summary of the model
summary(m0)
```

```
##
## Call:
## lm(formula = Y ~ X_1)
##
## Residuals:
##       1       2       3       4
##  1.7714 -0.2000 -2.2571  0.6857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.8857     3.5164  -1.105    0.384
## X_1           2.0286     0.7068   2.870    0.103
##
## Residual standard error: 2.091 on 2 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.7069
## F-statistic: 8.237 on 1 and 2 DF,  p-value: 0.103
```

2. Compute $\mathbf{X}^T\mathbf{X}$ and $(\mathbf{X}^T\mathbf{X})^{-1}$ and then compute $\hat{\beta}$, using formula $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and confirm that you got the same answer as you got from R using function `lm`.

```
n <- length(X_1)
X<-cbind(rep(1,n),X_1) # column cconcat a vector of ones
```

$\mathbf{X} =$

```
print(X)
```

```
##        X_1
## [1,] 1   4
## [2,] 1   3
## [3,] 1   5
## [4,] 1   7
```

$\mathbf{X}^T\mathbf{X}$, $(\mathbf{X}^T\mathbf{X})^{-1}$ and $\mathbf{X}^T\mathbf{Y}$:

```
(XtX<-t(X)%*%X) # XtX
```

```
##        X_1
##      4  19
## X_1 19  99
```

```
(XtX_inv <- solve(XtX)) # inverse of XtX
```

```
##                       X_1
##      2.8285714 -0.5428571
## X_1 -0.5428571  0.1142857
```

```
(XtY <- t(X) %*% Y) # Xty
```

```
##     [,1]
##       23
## X_1  127
```

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

```
(beta.hat <- solve(XtX, XtY)) # solve XtX (Beta)= Xty
```

```
##          [,1]
##     -3.885714
## X_1  2.028571
```

So we see that these are the same as with the R function lm.

3. Compute the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

```
(H =  X %*% XtX_inv %*% t(X)) # X (XtX)^{-1} X^T
```

```
##            [,1] [,2]      [,3]        [,4]
## [1,] 0.31428571  0.4 0.2285714  0.05714286
## [2,] 0.40000000  0.6 0.2000000 -0.20000000
## [3,] 0.22857143  0.2 0.2571429  0.31428571
## [4,] 0.05714286 -0.2 0.3142857  0.82857143
```

4. Compute

$$\hat{\sigma}^2 = MS_{Res} = \frac{SS_{res}}{n-2} = \frac{(\mathbf{Y}-\mathbf{HY})^T(\mathbf{Y}-\mathbf{HY})^T}{n-2} = \frac{\mathbf{Y}^T(\mathbf{I}_n-\mathbf{H})\mathbf{Y}}{n-2}$$

```
# Calculate sum of squares with formula
SS.Res <- (t(Y) %*% (diag(1,n) - H) %*% Y )[1,1]
# Calculate mean sum of squares
MS.Res <- SS.Res/(n-2)
# Calculate sima hat
sigma.hat <- sqrt(MS.Res)

sprintf("SS.Res = %.3f", SS.Res)
```

```
## [1] "SS.Res = 8.743"
```

```
sprintf("Ms.Res = %.3f", MS.Res)
```

```
## [1] "Ms.Res = 4.371"
```

```
sprintf("Sigma hat = %.3f", sigma.hat)
```

```
## [1] "Sigma hat = 2.091"
```

5. Compute

$$\widehat{Var}(\hat{\beta}) = \begin{pmatrix} \widehat{Var}(\hat{\beta}_0|\mathbf{X}) & \widehat{Cov}(\hat{\beta}_0,\hat{\beta}_1|\mathbf{X}) \\ \widehat{Cov}(\hat{\beta}_1,\hat{\beta}_0|\mathbf{X}) & \widehat{Var}(\hat{\beta}_1|\mathbf{X}) \end{pmatrix} = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$$

```
# Compute the variance-covariance matrix with the previous results
(var_beta = MS.Res*XtX_inv)
```

```
##                       X_1
##     12.364898 -2.3730612
## X_1 -2.373061  0.4995918
```

6. Compute $ese(\hat{\beta}_0)$ and $ese(\hat{\beta}_1)$

```
# access the var-cov matrix and square
ese_beta0 <- sqrt(var_beta[1,1])
ese_beta1 <- sqrt(var_beta[2,2])
sprintf("ese_beta0 = %.3f ese_beta1 = %.3f", ese_beta0, ese_beta1)
```

```
## [1] "ese_beta0 = 3.516 ese_beta1 = 0.707"
```

Question 5 (new)

Data is provided on average pulic teacher annual salayr in dollars.

```r
salary <- read.csv("C:/Users/jairp/Desktop/BackUP/McGill-20180719T015111Z-001/McGill/7. Fall 2019/MATH 4
x1 <- salary$SPENDING/1000
y <- salary$SALARY
lm.salary <- lm(y ~ x1)
summary(lm.salary)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3848.0 -1844.6  -217.5  1660.0  5529.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12129.4     1197.4   10.13 1.31e-13 ***
## x1            3307.6      311.7   10.61 2.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2325 on 49 degrees of freedom
## Multiple R-squared:  0.6968, Adjusted R-squared:  0.6906
## F-statistic: 112.6 on 1 and 49 DF,  p-value: 2.707e-14
```

1.

Write R code to verify the calculation of the entries in the `Estimate` column, and show that your code produces the correct results.

```r
n <- length(x1) # obtain total number of observations
X <- cbind(rep(1,n), x1) # attach a column of ones
Y <- y # rename y
print(head(X)) # print the first 5 observations
```

```
##           x1
## [1,] 1 3.346
## [2,] 1 3.114
## [3,] 1 3.554
## [4,] 1 4.642
## [5,] 1 4.669
## [6,] 1 4.888
```

The following code does the following calculates X^TX, $(X^TX)^{-1}$ and X^TY respectively.

```r
(XtX<-t(X)%*%X) # XtX
```

```
##                 x1
##     51.000 188.5270
## x1 188.527 752.5364
```

```
(XtX_inv <- solve(XtX)) # inverse of XtX
```

```
##                           x1
##     0.26526473 -0.06645468
## x1 -0.06645468  0.01797721
```

```
(XtY <- t(X) %*% Y) # Xty
```

```
##       [,1]
##    1242167
## x1 4775792
```

Using these quantities, we can calculate $\hat{\beta} \in \mathbb{R}^2$ as

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

$

```
(beta_hat<-solve(XtX,XtY))
```

```
##          [,1]
##    12129.371
## x1  3307.585
```

```
# store independently for later
beta_0_hat <- beta_hat[1]
beta_1_hat <- beta_hat[2]
```

So we see that our estimates $\hat{\beta_0} = 12129.371$ and $\hat{\beta_1} = 3307.585$ are the same as the ones from the `lm` function.

2.

We know $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta} =$

```
Y_hat <- X %*% beta_hat
```

and we want to verify

$$sum(\mathbb{Y} - \hat{\mathbb{Y}}) = sum(\mathbb{Y} - \mathbb{X}\beta) = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

Which is calculated as

```
res <- Y - Y_hat # Y - Y_hat
(sum_res <- sum(res))
```

```
## [1] -5.456968e-11
```

so

$$\sum_{i=1}^{n} e_i = -0.00000000005456968 \approx 0$$

.

Then we have

$$e^T(x_1 - \bar{x_1}) = \sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0$$

which we calculate as

```
(identity2 <- t(res) %*% (x1 - mean(x1)))
```

```
##               [,1]
## [1,] 1.478284e-09
```

so that

$$\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0.000000001478284 \approx 0$$

checks out.

Finally, we have

$$e^T\hat{\mathbb{Y}} = \sum_{i=1}^{n} e_i\hat{y}_i = \sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

```
(identity3 <- t(res) %*% Y_hat)
```

```
##               [,1]
## [1,] 3.769994e-06
```

so that

$$\sum_{i=1}^{n} e_i\hat{y}_i = 0.000003769994 \approx 0$$

also checks out.

3.

Compute the value of the entry in the Std. Error column on line 17 using entries already given in the table, and then using the data directly. The (estimated) Std. Error of β_0 is given as

$$ese(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x_1}^2}{S_{xx}}\right)}$$

but we also have in matrix format that

$$\hat{\Sigma} = \widehat{Var}(\hat{\beta}) = \begin{pmatrix} \widehat{Var}(\hat{\beta}_0|\mathbf{X}) & \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1|\mathbf{X}) \\ \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_0|\mathbf{X}) & \widehat{Var}(\hat{\beta}_1|\mathbf{X}) \end{pmatrix} = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$$

which means that $\sqrt{\hat{\Sigma}_{1,1}} = \sqrt{\widehat{Var}(\hat{\beta}_0|\mathbf{X})} = ese(\hat{\beta}_0)$, so we first compute

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

```
H =  X %*% XtX_inv %*% t(X) # X (XtX)^{-1} X^
dim(H)
```

```
## [1] 51 51
```

Then we compute

$$\hat{\sigma}^2 = MS_{Res} = \frac{SS_{res}}{n-2} = \frac{(\mathbf{Y}-\mathbf{HY})^T(\mathbf{Y}-\mathbf{HY})^T}{n-2} = \frac{\mathbf{Y}^T(\mathbf{I}_n-\mathbf{H})\mathbf{Y}}{n-2}$$

```
# Calculate sum of squares with formula
SS.Res <- (t(Y) %*% (diag(1,n) - H) %*% Y )[1,1]
# Calculate mean sum of squares
MS.Res <- SS.Res/(n-2)
# Calculate sima hat
sigma.hat <- sqrt(MS.Res)

sprintf("SS.Res = %.3f", SS.Res)
```

```
## [1] "SS.Res = 264825249.995"
```

```
sprintf("MS.Res = %.3f", MS.Res)
```

```
## [1] "MS.Res = 5404596.939"
```

```
sprintf("Sigma hat = %.3f", sigma.hat)
```

```
## [1] "Sigma hat = 2324.779"
```

Then we can calculate $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$ as

```
# Calculate variance covariance matrix of bet_hat
(var_beta_hat = MS.Res*XtX_inv)
```

```
##                       x1
##    1433648.9 -359160.75
## x1 -359160.8   97159.55
```

```
# extract the estimated variances
var_beta0 <- var_beta_hat[1,1]
var_beta1 <- var_beta_hat[2,2]

# obtain the ese's for each parameter
ese_beta0 <- sqrt(var_beta0)
ese_beta1 <- sqrt(var_beta1)

# display the results
sprintf("ese_beta0 = %.4f", ese_beta0)
```

```
## [1] "ese_beta0 = 1197.3508"
```

```
sprintf("ese_beta1 = %.4f", ese_beta1)
```

```
## [1] "ese_beta1 = 311.7043"
```

Then, $ese(\hat{\beta}_0) = 1197.3508$ and $ese(\hat{\beta}_1) = 311.7043$.

The previous was using the data directly, but we can also calculate it as

$$ese(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x_1}^2}{S_{xx}} \right)}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_{i1} - \bar{x_1})^2 = (X_1 - \bar{X_1})^T (X_1 - \bar{X_1})$$

```
x1_mean <- mean(x1) # x1_bar
Sxx <- t(x1 - mean(x1)) %*% (x1 - mean(x1))
ese_beta0 <- sqrt(MS.Res * ((1/n) + (x1_mean**2) / Sxx ) )
print(ese_beta0)
```

```
##          [,1]
## [1,] 1197.351
```

And so we obtain approximatedly the same result.

4.

Write `R` code to compute the value of the omitted entry for the `Residual standard error` on line 22.

NOTE: We already did so in the previous question, and so it is given as

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{MS_{Res}} = \sqrt{\frac{SS_{Res}}{n-2}} = 2324.779$$

5.

We want to test for a linear relationship between `SALARY` and `SPENDING`, using $\alpha = 0.05$.

The hypothesis we want to test is

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

If H_0 is true and $\beta_1 = 0$, then

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{ese(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{ese(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$$

So that in which case we have

```
(T_1 <- beta_1_hat / sqrt(( MS.Res / Sxx )) )
```

```
##          [,1]
## [1,] 10.61129
```

so $T_1 = 10.61129.$, but we see that $|t_{\alpha/2}|$ equals

```
# P()
(t_alpha_half <- abs(qt(0.05/2, 49)))
```

```
## [1] 2.009575
```

i.e. $T_1 < |t_{\alpha/2}| = 2.009575$ and so we **reject** the null hypothesis that $\beta_0 = 0$ and conclude that $\beta_1 \neq 0$, i.e., it is **statistically significant**.

The p-value is obtained as

$$p\text{-}val \ = 2\mathbb{P}(|t_{\alpha/2}| \leq T_1)$$

```
# calculate p-value for alpha=0.05 and n-2 = 51-2 = 49 degrees of freedom
p_val <- 2*pt(-abs(T_1), df=49)
print(p_val)
```

```
##              [,1]
## [1,] 2.706871e-14
```

So we see that p-value$= 2.706871e-14 < \alpha = 0.05$, so indeed, we **reject** $H_0 : \beta_1 = 0$ at the $\alpha = 0.05$ confidence level.

6.

We have that for the sums-of-squares decomposition

$$SS_T = SS_{Res} + SS_R$$

the F-statistic uses the formula

$$F = \frac{SS_R/(p-1)}{SS_{Res}/(n-p)}$$

where $p = 2$ for SLR. We have that

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\hat{Y} - \bar{Y})^T(\hat{Y} - \bar{Y})$$

```
y_hat_mean <- mean(Y_hat) # x1_bar
SS_R <- t(Y_hat - mean(Y_hat)) %*% (Y_hat - mean(Y_hat))
F_stat <- (SS_R / (2-1)) / (SS.Res / (n-2)) # n = 51
sprintf("F-stat = %.5f", F_stat)
```

```
## [1] "F-stat = 112.59952"
```

i.e., here $F = 112.59952$.

7.

We verify numerically that

$$SS_T = SS_{Res} + SS_R \iff \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

First we calculate SS_T as

```
y_mean <- mean(y) # x1_bar
SS_T <- t(y - mean(y)) %*% (y - mean(y))
sprintf("SS_T = %.5f", SS_T)
```

```
## [1] "SS_T = 873380264.62745"
```

```
sprintf("SS_Res = %.5f", SS.Res)
```

```
## [1] "SS_Res = 264825249.99461"
```

```
sprintf("SS_R = %.5f", SS_R)
```

```
## [1] "SS_R = 608555014.63282"
```

```
sprintf("SS_Res + SS_R = %.5f", SS.Res + SS_R)
```

```
## [1] "SS_Res + SS_R = 873380264.62743"
```

$$\begin{cases} SS_T = 873380264.62745 \\ SS_{Res} = 264825249.99461 \\ SS_R = 608555014.63282 \\ SS_{Res} + SS_R = 873380264.62743 \end{cases}$$

So we see that indeed $SS_T = SS_{Res} + SS_R$. (NOTE: The slight difference in the last decimals is due to floating point overflow in computations).

8.

Finding a 90% confidence interval for β_1.

We can compute the $100(1-\alpha)\%$ confidence interval as $CI(\beta_1) = \hat{\beta}_1 \pm t_{\alpha/2} ese(\hat{\beta}_1)$ interval from

$$\mathbb{P}\left(-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{ese(\hat{\beta}_1)} \leq t_{\alpha/2}\right) = \mathbb{P}\left(\hat{\beta}_1 - t_{\alpha/2} ese(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} ese(\hat{\beta}_1)\right) = 1 - \alpha$$

```
(t_alpha_half <- qt(.95, 49)) # .95th quantile
```

```
## [1] 1.676551
```

```r
beta_1_low <- beta_1_hat - t_alpha_half*ese_beta1 # lower bound
beta_1_high <- beta_1_hat + t_alpha_half*ese_beta1 # upper bouund
CI_beta_1 <- c(beta_1_low, beta_1_high)
print("90% confidence interval for beta_1: ")
```

```
## [1] "90% confidence interval for beta_1: "
```

```r
print(CI_beta_1)
```

```
## [1] 2784.997 3830.173
```

```r
sprintf("Beta_1 estimate: %.5f", beta_1_hat)
```

```
## [1] "Beta_1 estimate: 3307.58500"
```

Interpretation: If we repeat the experiment $k \to \infty$ many times, 90% of the time, our interval will trap the true β_1. In other words, we are 90% confident that the given interval contains β_1.

9.

Using the fited model , predict what the average public teacher annual salary would be in a state where the spending per pupil is $4800.

We now that

$$\mathbb{Y}_{pred} = \mathbb{X}_{new}\beta$$

```r
# Create a new value to predict.
X_new <- cbind(seq(1,1), c(4.8))
Y_pred <- X_new %*% beta_hat
print(Y_pred)
```

```
##          [,1]
## [1,] 28005.78
```

So we predict that the average public teacher annual salary would be around 28005.78$ when the spending per pupil is $4800.

10.

The prediction at an arbitrary new x value, x_1^{new} can be written in terms of the estimates $\widehat{\beta}$ as

$$\hat{y}^{new} = \mathbb{X}_1^{new}\widehat{\beta} = [1x_1^{new}]\widehat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{new}$$

with $\widehat{\beta}$ the least squares **estimate**. Compute the *estimated standard prediction error* for $\hat{y}^{new}$, that is, the square root of the estimated variance of the corresponding random variable

$$\widehat{Y}^{new} = \mathbb{X}_1^{new}\widehat{\beta} = [1x_1^{new}]\widehat{\beta}$$

now with $\widehat{\beta}$ the least squares **estimator** , if x_1^{new} is \$4800.

Now, if we take into account the **ese for prediction for the mean**, this is given by

$$ese(\hat{y}_{pred}) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_1^{new} - \bar{x_1})^2}{S_{xx}} \right)}$$

in which case we would have

```
x_new <- X_new[1,2]
ese_y_hat <- sqrt(MS.Res * ( (1/n) + (x_new - x1_mean)**2 / Sxx ) )
print(ese_y_hat)
```

```
##          [,1]
## [1,] 473.5628
```

However we have that the **estimated standard error of prediction** for a new observation x_1^{new} is given by

$$ese(\hat{y}_{pred}) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_1^{new} - \bar{x_1})^2}{S_{xx}} \right)}$$

So here we can calculate this for $x_1^{new} = 4.8$ (since it is in thousands)

```
x_new <- X_new[1,2]
ese_y_hat <- sqrt(MS.Res * ( 1+ (1/n) + (x_new - x1_mean)**2 / Sxx ) )
print(ese_y_hat)
```

```
##          [,1]
## [1,] 2372.522
```

So we estimate

$$ese(\hat{y}_{pred}) = 2372.522$$

END OF ACTUAL questions

OLD QUESTIONS

PREVIOUS Question 5

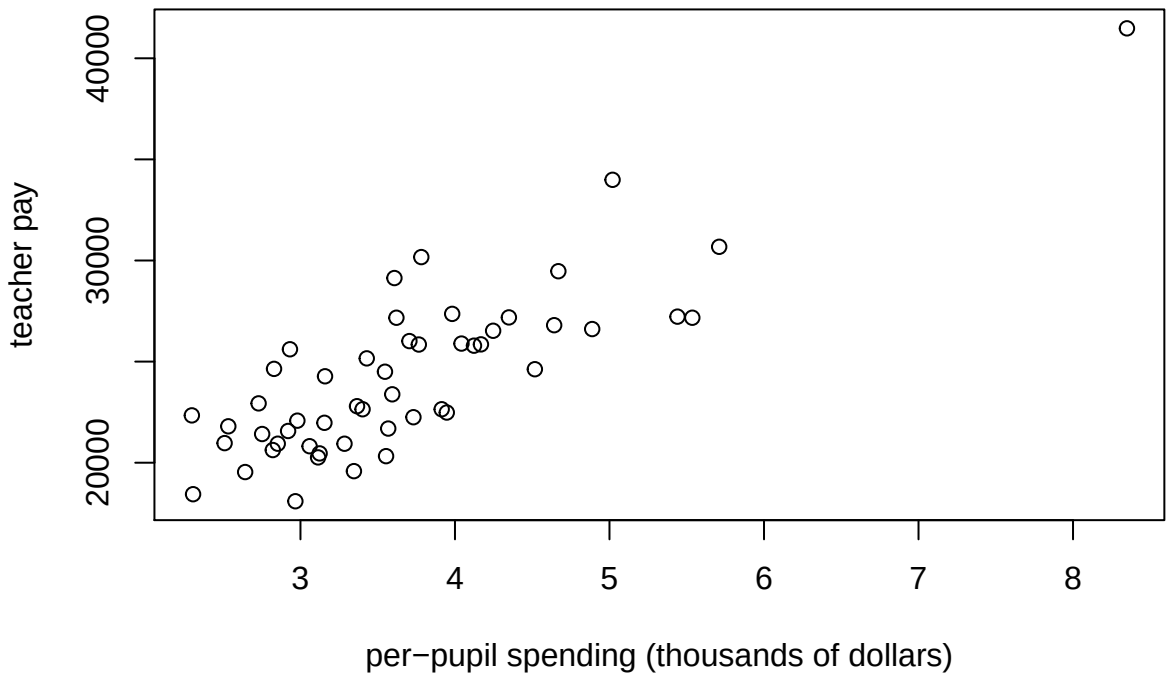
The following data gives data on average public teacher annual salary in dollars, recorded in the data frame `salary` as the variable `SALARY`, and spending (`SPENDING`) per pupil (in thousands of dollars) on public schools in 1985 in the 50 US states and the District of Columbia.

The objective ot the anlysis is to understand whether there is a relationship between teacher pay, y, and per-pupil spending x.

```
file1 <- "http://www.math.mcgill.ca/yyang/regression/data/salary.csv"
salary <- read.csv(file1, header = TRUE)
x1 <- salary$SPENDING/1000
y <- salary$SALARY
lm.salary <- lm(y ~ x1) # fit the linear model
summary(lm.salary)
```

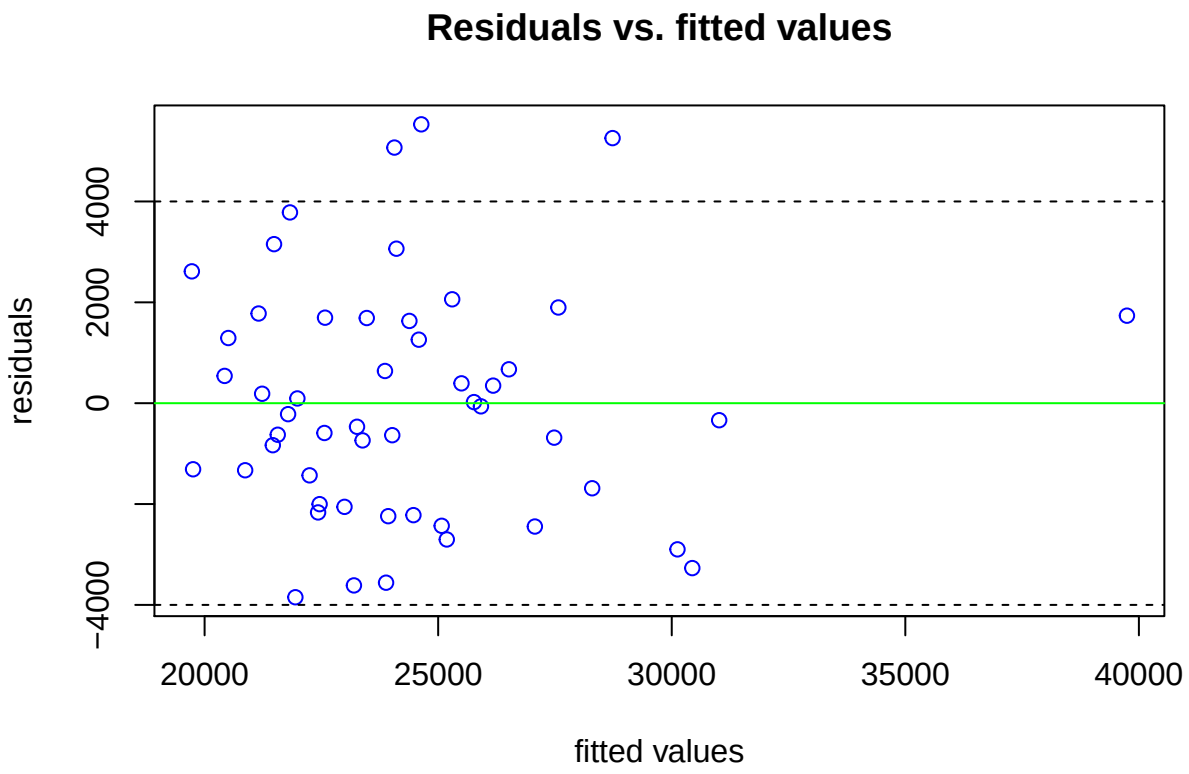
```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3848.0 -1844.6  -217.5  1660.0  5529.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12129.4     1197.4   10.13 1.31e-13 ***
## x1            3307.6      311.7   10.61 2.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2325 on 49 degrees of freedom
## Multiple R-squared:  0.6968, Adjusted R-squared:  0.6906
## F-statistic: 112.6 on 1 and 49 DF,  p-value: 2.707e-14
```

```
# Plot of the variables only
plot(x1, y, xlab = "per-pupil spending (thousands of dollars)", ylab = "teacher pay")
```



1. Make a residual plot of e_i versus the fitted values. What does the plot suggest about the lienarity assumption of the regression model?

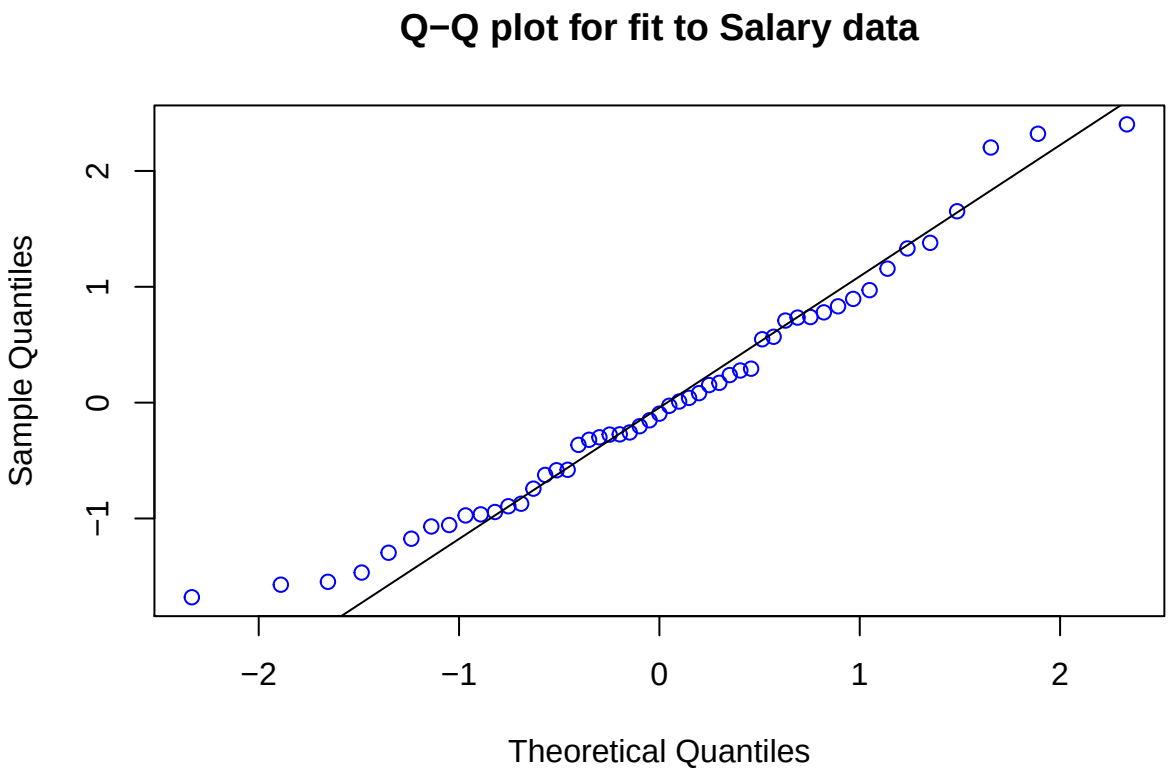
```
res.salary <- residuals(lm.salary) # extract the residuals
fitted.salary <- lm.salary$fitted.values # extract the fitted values
plot(fitted.salary, res.salary, xlab= "fitted values" , ylab = "residuals",
     main="Residuals vs. fitted values", col = "blue") # plot the residuals vs fitted values
abline(h=0, col="green")
abline(h=c(-4000,4000), lty=2)
```



Except for a couple outliers, there doesn't seem to be a specific pattern around the center (0). Since the residuals (except for the outlier) seems to be evenly and randomly distributed with respect to the center line, this suggests that *the assumption of linear relationship is reasonable*. We also observe that most of the points lie within the an approximatedly equal "band", which suggests that the the variance of the error terms are mostly equal.

2. Prepare a qq-plot fo the residuals. Do the residuals appear to be Normally distributed?

```
## QQ-plot to test for normality
res.salary <- rstandard(lm.salary) # obtain the standard residuals
qqnorm(res.salary, main="Q-Q plot for fit to Salary data", col="blue")  # plot the qq-plot
qqline(res.salary)
```



The residuals appear to be mostly normally distributed between -1 and 1, so we could somehow assume that the data is approximatedly normally distributed.

3. Verify numerically the orthogonality results concerning the residuals, that is,

$$\sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

```
beta_0_hat <- coefficients(lm.salary)[1]
beta_1_hat <- coefficients(lm.salary)[2]

# 1.
sum_res = sum(res.salary)
# 2.
centered_x1 <- x1-mean(x1) # x_1 - X_bar
ecentered_x1 <- res.salary * centered_x1 # e_i (x_i - x)
sum_ecentered_x1 <- sum(ecentered_x1)
# 3.
```

```r
ey_fitted <- res.salary * (beta_0_hat + beta_1_hat*x1) # e_i * yfit_i
sum_ey_fitted = sum(ey_fitted)



sprintf("1) %.3f  2) %.3f  3)%.3f", sum_res, sum_ecentered_x1, sum_ey_fitted)
```

```
## [1] "1) 0.173  2) 0.860  3)7067.592"
```

We observe that the two first of these are very close to 0, while the third one greatly deviates from it. However, relative to the scale of the data, perhaps we would assume it is "close". One of the reasons this could be happening is because of the precense of outliers.

4. Report the value of the estimated intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$

```r
# Obtain intercepts from the model
sprintf("1) beta_0_hat=%.3f  2) beta_1_hat=%.3f", beta_0_hat, beta_1_hat)
```

```
## [1] "1) beta_0_hat=12129.371  2) beta_1_hat=3307.585"
```

We observe that $\hat{\beta}_0 = 12129.371$ and $\hat{\beta}_1 = 3307.585$. The first tells us that when there is not pupil spending, the average public teacher annual salary is around 12129.371\$, and the second one tells us that for each increase in one unit (thousand of dollars) of per-pupil spending, the average teacher salary approximatedly increases by 3307.585\$ dollars.

5. Test whether or not there is a linear association between `SALARY` and `SPENDING`, using $alpha = 0.05$. State the alternative hypothesis, decision rule and conclusion. What is the *p-value* for the test?

6. **Hypothesis** we wish to test for the following hypothesis:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

If H_0 is true and $\beta_1 = 0$, then

$$t = \frac{\hat{\beta}_1 - \beta_1}{ese(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$$

So we test testing with $\alpha = 0.05$ and performing a two-tailed test, we observe that $|t_{\alpha/2,n-2}| = 2.71e^{-14} < \frac{\alpha}{2} = 0.025$, and so we **reject** the null hypothesis at the 0.05-level confidence. The attained significance level (p-value) is $2.71e^{-14}$. Equivalently, we can run the Pearson's product-moment correlation test:

```r
cor.test(x1, y, alternative = "two.sided")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x1 and y
## t = 10.611, df = 49, p-value = 2.707e-14
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7262065 0.9026688
## sample estimates:
##       cor
## 0.8347343
```

We can see that in fact there is apositive correlation between the two variables.

6. Find a 90% confidence interval for β_1. How do you interprete it?

In this case, for a 90% CI, $\alpha = 0.1$, and so we have for $t \sim t_{n-2=49}$, and $ese(\hat{\beta}_1) = 311.7$

```
(t_alpha_half <- qt(.95, 49)) # .95th quantile
```

```
## [1] 1.676551
```

```
beta_1_low <- beta_1_hat - t_alpha_half*311.7 # lower bound
beta_1_high <- beta_1_hat + t_alpha_half*311.7 # upper bound
CI_beta_1 <- c(beta_1_low, beta_1_high)
print("90% confidence interval for beta_1: ")
```

```
## [1] "90% confidence interval for beta_1: "
```

```
print(CI_beta_1)
```

```
##       x1       x1
## 2785.004 3830.166
```

Interpretation: the confidence interval (2785.004, 3830.166) traps β_1 with probability equal to 0.9.

7. Using the fitted model, predict whate the average public teacher annual salary would be in a state where the spending per pupil is $4800.

```
# Create new observation
new_x = data.frame(x1=4.8)
predict(lm.salary, newdata = new_x, interval="prediction")
```

```
##        fit      lwr      upr
## 1 28005.78 23238.02 32773.54
```

Therefore we predict that the average public teacher annual salary would be around 23238.02$ with a 4800$ spending per pupil.

PREVIOUS Question 6: Data Analysis Practice

```r
# Load the datasett
file1 <- "http://www.math.mcgill.ca/yyang/regression/data/abalone.csv"
abalone <- read.csv(file1, header = TRUE)
head(abalone)
```

```
##   Height Rings
## 1  0.095    15
## 2  0.090     7
## 3  0.135     9
## 4  0.125    10
## 5  0.080     7
## 6  0.095     8
```

```r
nrow(abalone) # number of observations = 4177
```

```
## [1] 4177
```

Research Problem: Abalones, also called ear-shells or sea ears, are one type of reef-dwelling marine snails. The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked in a variety of cultures. It is difficult to tell the ages of abalones because their shell sizes not only depend on how old they are, but also depend on the availability of food.The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. A research group are interested in using some of abalones' physical measurements, especially the height measurement to predict their ages. The research group believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and theirages, and particularly, that a larger height is associated with an older age.

Research Problem

We wish to investigate the relationship between between the height of abalones and their ages, assuming the simple linear regression model with normal error assumption is appropriate. As this is is related to the number of rings thet have, we hypothesize there exists a linear relationship between the variables "Height" and "Rings". (Reseach problem formulation). If there is such a relationship, we could especulate the simulatenous change of these.

Variables basic statistics

We examine the variables independly, providing summary statistics as a part of basic data exploration.

Height variable

```r
# Obtain minium and maximum values, 1st and 3rd quantiles, mean and median of both vairables
print("Summary statistics: ")
```

```
## [1] "Summary statistics: "
```

```r
height <- abalone$Height
summary(height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1150  0.1400  0.1395  0.1650  1.1300
```

```r
height.mean <- mean(height)
height.variance <- var(height)
height.std <- sqrt(height.variance)

sprintf("Mean of the Height variable: %.3f", height.mean)
```

```
## [1] "Mean of the Height variable: 0.140"
```

```r
sprintf("Variance of the Height variable: %.3f", height.variance)
```

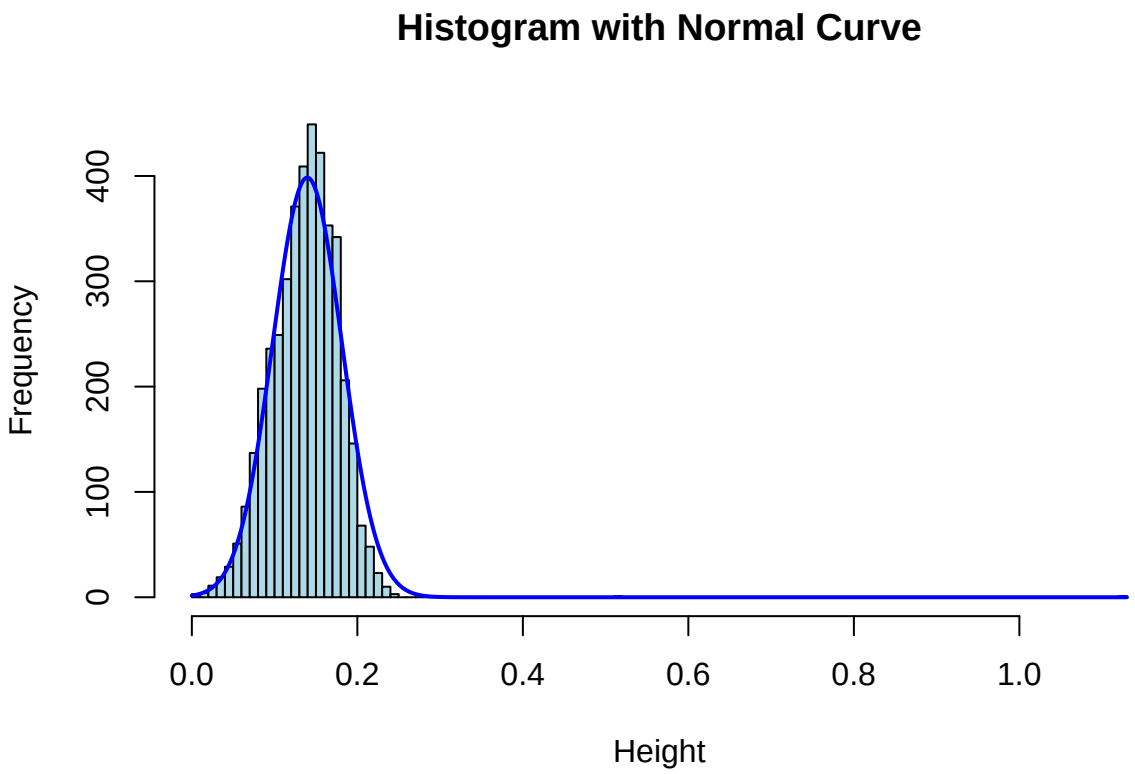
```
## [1] "Variance of the Height variable: 0.002"
```

```r
sprintf("Standard deviation of the Height variable: %.3f", height.std)
```

```
## [1] "Standard deviation of the Height variable: 0.042"
```

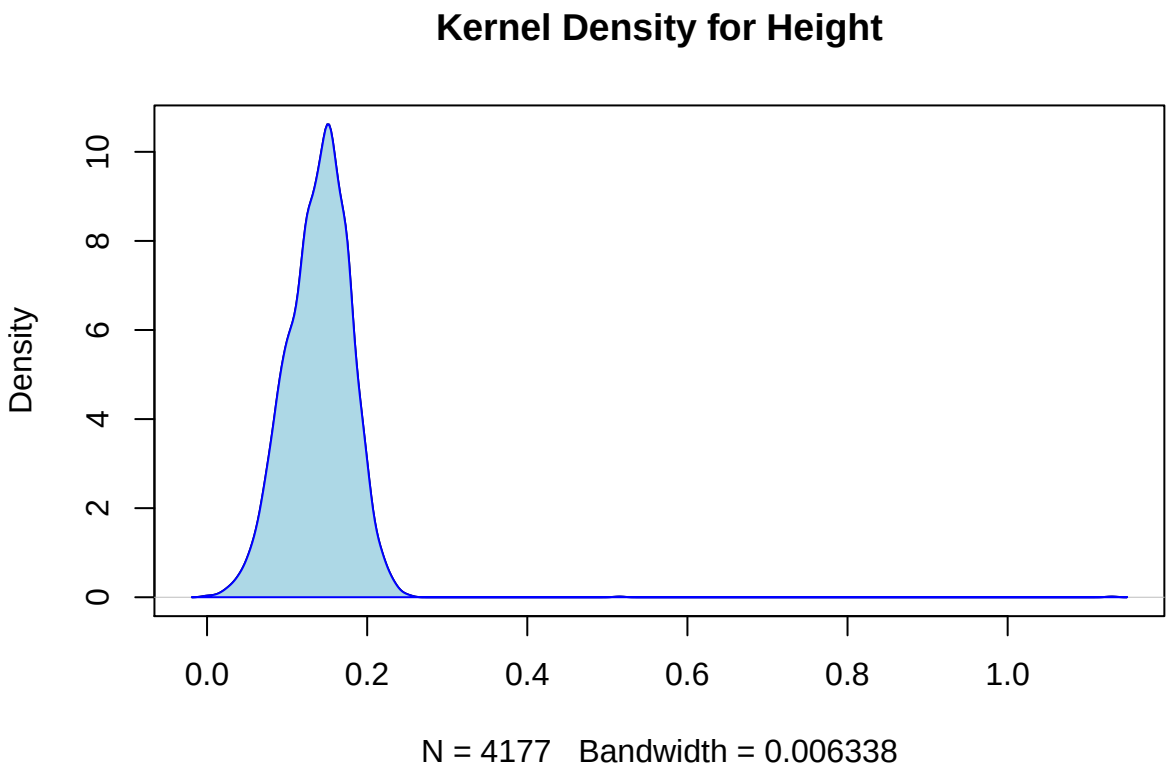
We observe that there is quite some difference between the min and max if we take into account the scale. However this, along with the fact that the measurements were done with a microscope seems to indicate these measures were performed in centimeters or a smaller scale.

```r
# plot a histogram along with a normal curve
h <- hist(abalone$Height, breaks = 100 , col="light blue", xlab="Height", main="Histogram with Normal Cu
xfit <- seq(min(height), max(height),length = 4177)
yfit <- dnorm(xfit, mean=height.mean, sd=height.std)
yfit <- yfit*diff(h$mids[1:2])*length(height)
lines(xfit, yfit, col="blue", lwd=2)
```



We observe that although the distribution seems to be approximatedly normal, the huge range with concentration towards the left reflects the presence of outliers, so we should be aware of this. The following Kernel Density plot confirms this.

```
# Filled Denisty plot
d <- density(height)
plot(d, main="Kernel Density for Height")
polygon(d, col="light blue", border = "blue")
```



Rings variable

```
# Obtain minium and maximum values, 1st and 3rd quantiles, mean and median of both vairables
print("Summary statistics: ")
```

```
## [1] "Summary statistics: "
```

```
rings <- abalone$Rings
summary(rings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   8.000   9.000   9.934  11.000  29.000
```

```
rings.mean <- mean(rings)
rings.variance <- var(rings)
rings.std <- sqrt(rings.variance)

sprintf("Mean of the Rings variable: %.3f", rings.mean)
```

```
## [1] "Mean of the Rings variable: 9.934"
```

```
sprintf("Variance of the Rings variable: %.3f", rings.variance)
```

```
## [1] "Variance of the Rings variable: 10.395"
```

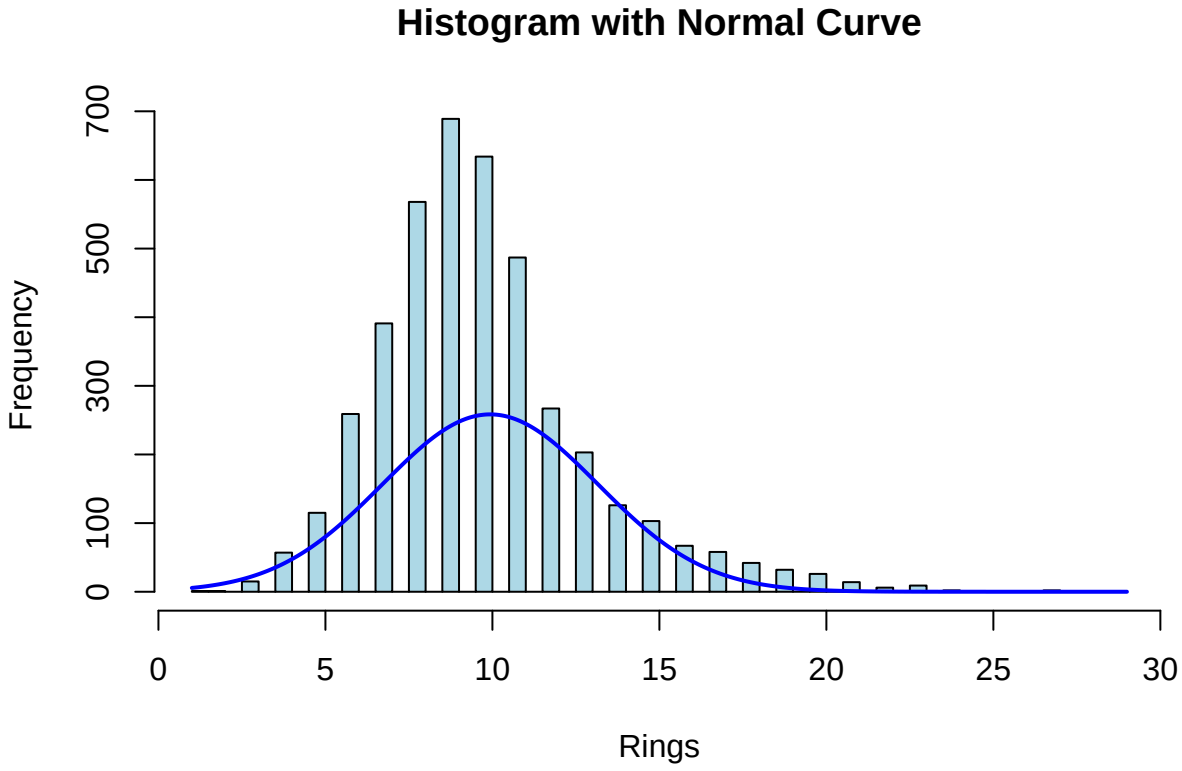
```
sprintf("Standard deviation of the Rings variable: %.3f", rings.std)
```

```
## [1] "Standard deviation of the Rings variable: 3.224"
```

We observe a range [1.0,29.0] , approximatedly equal mean and median , which indicates presence of relatively few outliers. Osserve that the variance of these data seems to be quite big in comparison to the previous one, so it might have an impact when fitting a linear model.

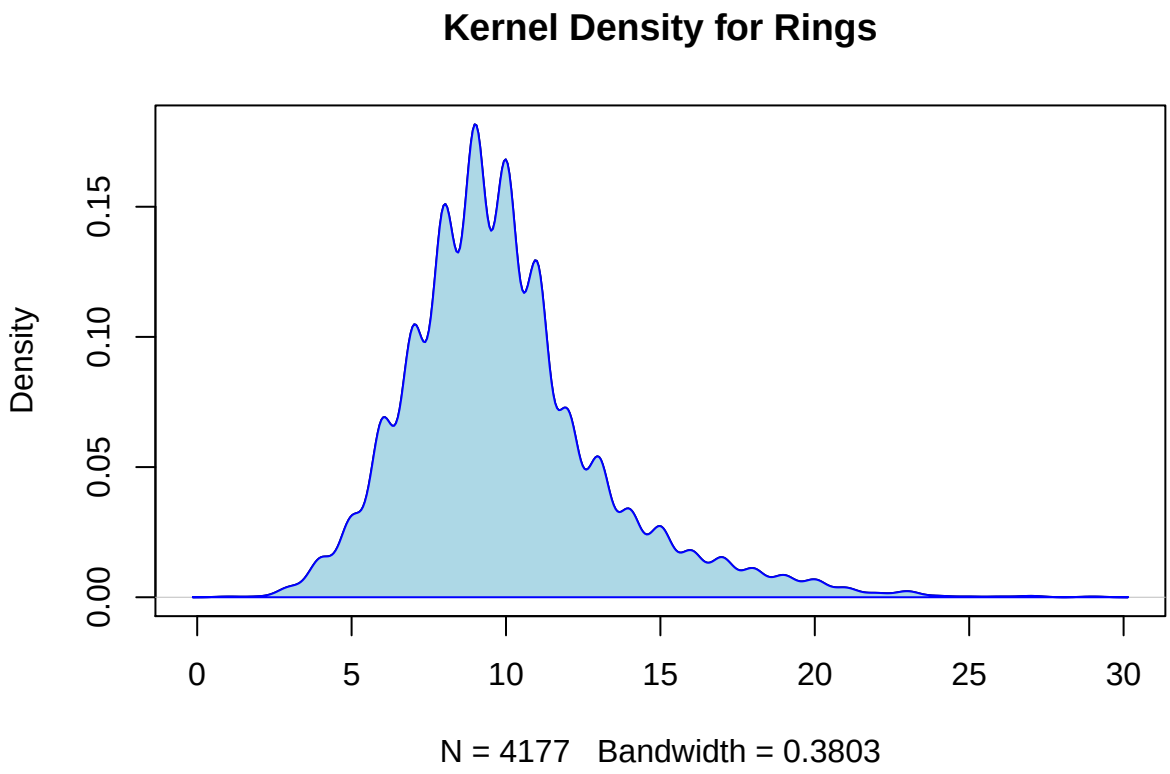
Now plot the density

```
# plot a histogram along with a normal curve
h <- hist(rings, breaks = 50 , col="light blue", xlab="Rings", main="Histogram with Normal Curve")
xfit <- seq(min(rings), max(rings),length = 4177)
yfit <- dnorm(xfit, mean=rings.mean, sd=rings.std)
yfit <- yfit*diff(h$mids[1:2])*length(rings)
lines(xfit, yfit, col="blue", lwd=2)
```



Although it looks to be somehow bell-shaped, we see that it deviates from the regular normal distribution with same sample mean and std. In fact, we suspect it could come from a χ^2 distribution instead. Next we draw the density plot:

```
# Filled Denisty plot
d <- density(rings)
plot(d, main="Kernel Density for Rings")
polygon(d, col="light blue", border = "blue")
```

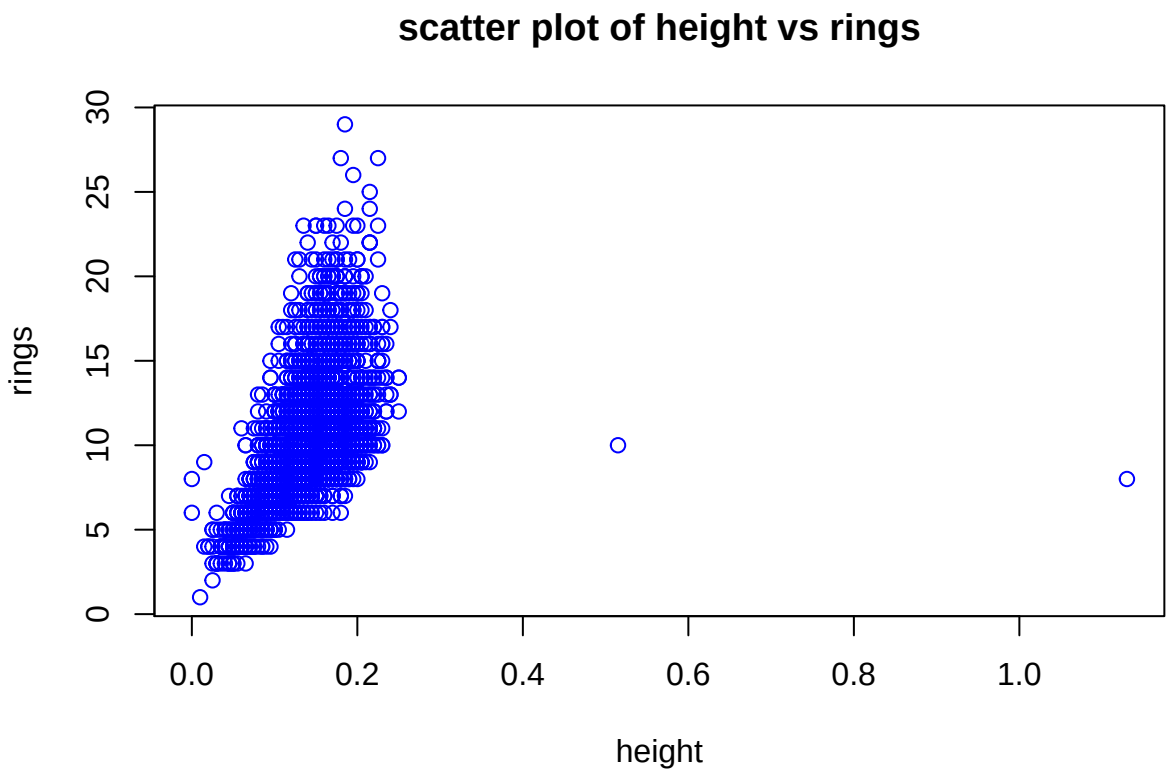


We can observe that indeed, the Rings random variable does not appear to be normally distributed.

Scatterplot

Now we bring our analysis to both variables in a scatterplot

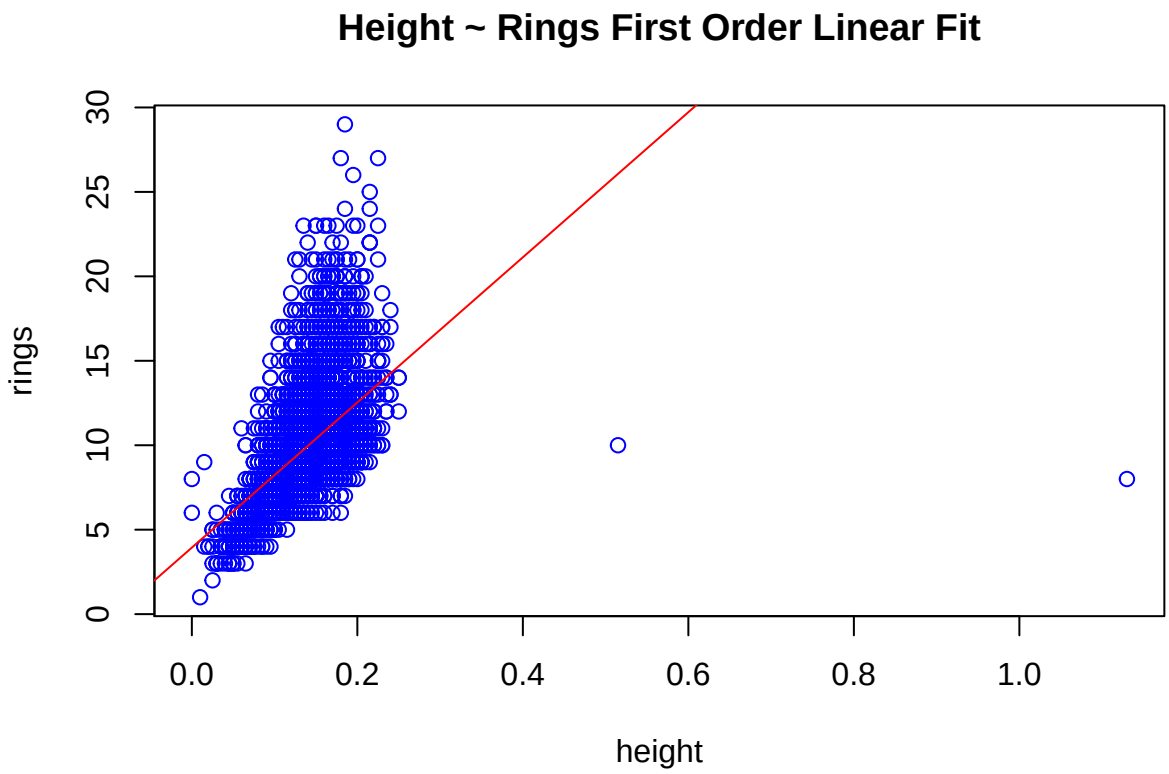
```
# Draw scatter plot
plot(height, rings, col = "blue", main= "scatter plot of height vs rings")
```



We observe a cluster of points which appear to have an positive (but not necessarily linear) correlation pattern. We also observe there are two significantly outliers, which do not follow the group trend. We suspect that a linear relationship might not be the most adequate in this case.

First order linear fit

```
# Fit a linear model, plot and then obtain summary.
lm.abalone <- lm(rings ~ height)
plot(height, rings, xlab="height", ylab="rings", main="Height ~ Rings First Order Linear Fit", col = "bl
abline(coef(lm.abalone), col="red")
```



We now examine the model's fit summary:

```
summary(lm.abalone)
```

```
##
## Call:
## lm(formula = rings ~ height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.496  -1.657  -0.607   0.839  17.112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9385     0.1443   27.30   <2e-16 ***
## height       42.9714     0.9904   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.677 on 4175 degrees of freedom
## Multiple R-squared:  0.3108, Adjusted R-squared:  0.3106
## F-statistic:  1882 on 1 and 4175 DF,  p-value: < 2.2e-16
```

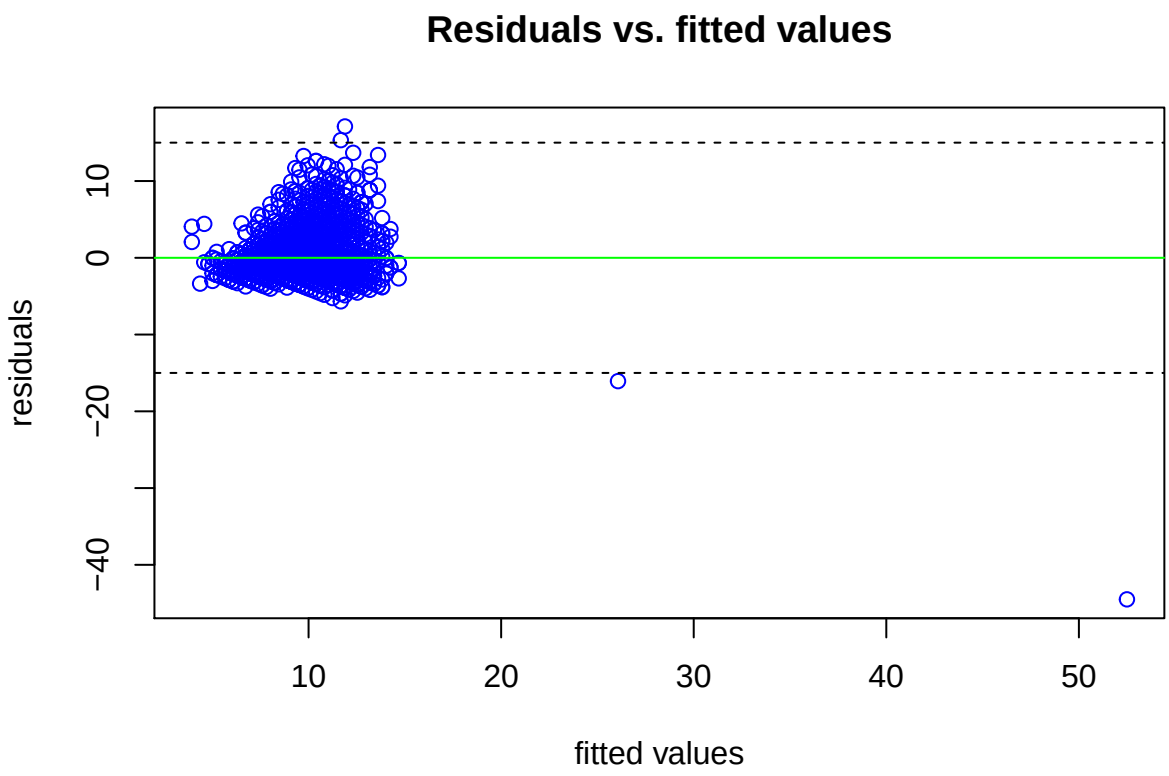
We notice that although the $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates both have significant p-values, looking at the model's Adjusted R-square we observe that the current fit only explains the variance in the data with about 31% precision, so it is overall a very bad fit.

Diagnostics

We now examine the model more closely but examining the residuals, as well as normality assumptions.

Residuals plot

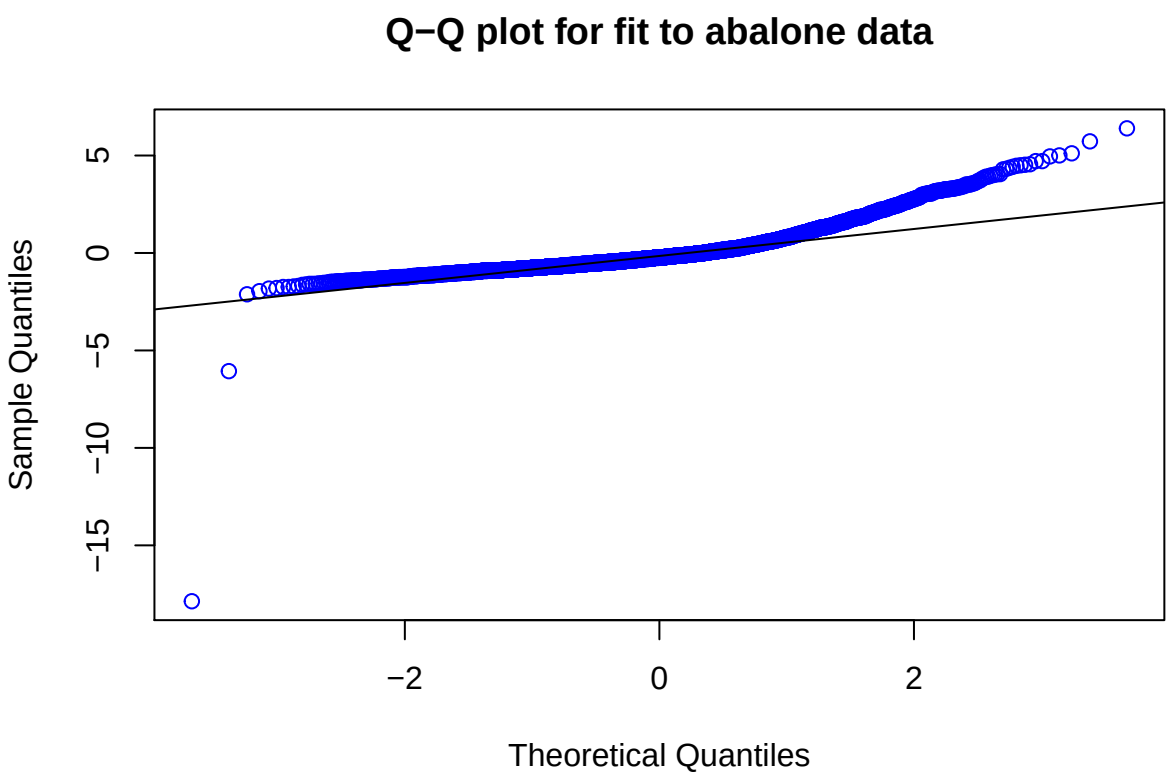
```
res.abalone <- residuals(lm.abalone) # extract the residuals
fitted.abalone <- lm.abalone$fitted.values # extract the fitted values
plot(fitted.abalone, res.abalone, xlab= "fitted values" , ylab = "residuals",
     main="Residuals vs. fitted values", col = "blue") # plot the residuals vs fitted values
abline(h=0, col="green")
abline(h=c(-15,15), lty=2)
```



Just as we suspected, we can observe two significant outliers as well as a clustering of points with a pattern that does not seem symmetrically and randomly distributed aout 0. This implies that the **linear relationship assumption might now be appropriate**. Further, this also implies the variance might be very different in both cases.

Q-Q Plot

```
## QQ-plot to test for normality
res.abalone <- rstandard(lm.abalone) # obtain the standard residuals
qqnorm(res.abalone, main="Q-Q plot for fit to abalone data", col="blue")  # plot the qq-plot
qqline(res.abalone)
```



We can observe from the Q-Q Plot that since most of the points are on top of the line and even follow a certain pattern, this indicates that the normality assumptions are not met.

Transforming the data

Although the assumptions are not directly met, we can modify the data to improve fit. We do so by first getting rid of outliers, and then applying an exponential transformation to the feature variable.

```
library(outliers)


df <- data.frame(height, rings)
# obtian outliers using z-score
outliers <- scores(height, type="chisq", prob=0.99)  # beyond 95th %ile based on z-scores
# remove the rows whose z-score is less than 0.99
df <- df[which(outliers == FALSE),]
sprintf("Previous number of observations: %d ", length(height))
```

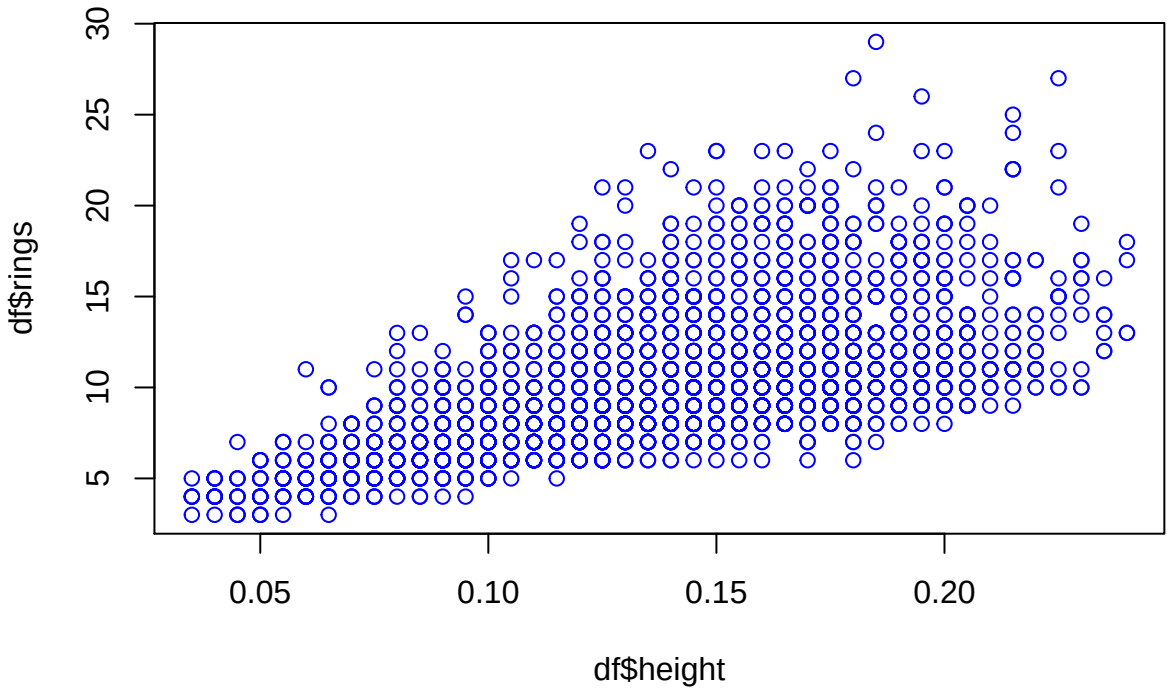
```
## [1] "Previous number of observations: 4177 "
```

```
sprintf("Current number of observations: %d", nrow(df))
```

```
## [1] "Current number of observations: 4154"
```

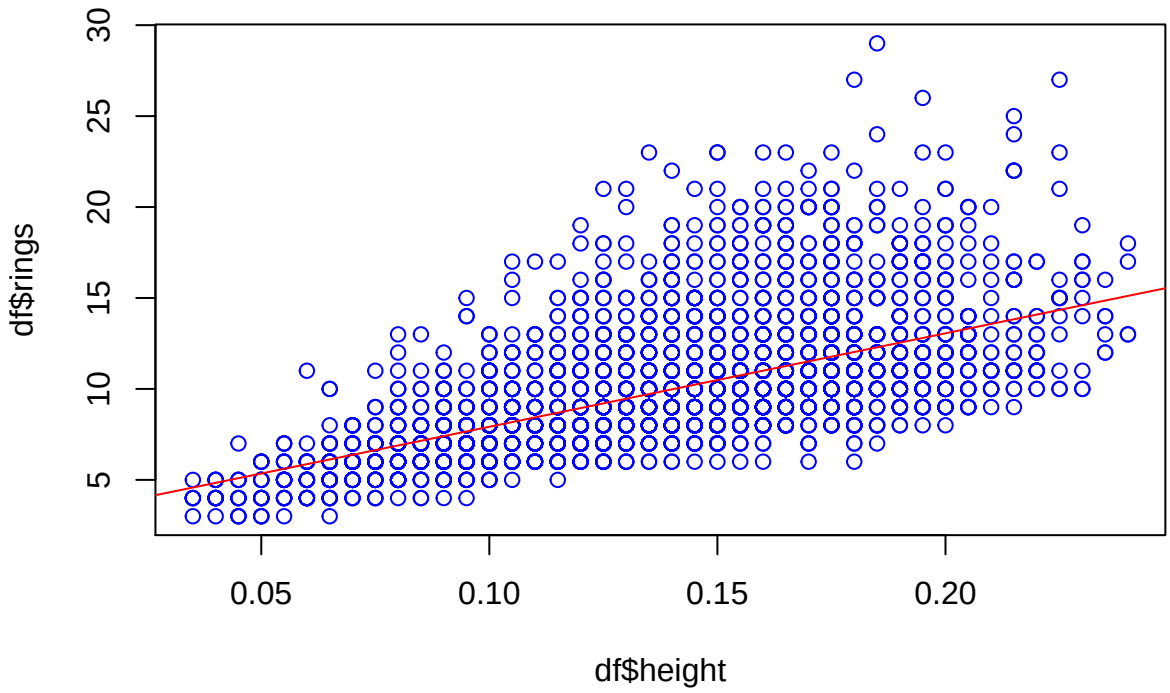
Recall that we originally had 4177 observations, after eliminating potential outliers, we are left with 4154. Even if some of them were not actual outliers, at least they were potential outliers, and since the amount of data we lost is relateively small to the size of the dataset, we can be sure we haven't loss that much relevant information. Let's plot the data and see how it looks:

```
plot(df$height, df$rings, col="blue")
```



Now we can concentrate on the big cluster of data without extreme outliers. We estimate that the relationship between the two given variables is in fact more complicated than just linear, so given the shape of the data, we will attemp to apply log transformations to both the rings variable and the height variable, and then fit a model of the log(log(rings)) as a function of the height and log(height). Notice that the variance is quite big with respect to the rings variable.

```
# straight regression fit ?
mod.abalone <- lm(rings ~ height, data = df)
plot(df$height, df$rings,  col="blue")
abline(coef(mod.abalone), col="red")
```

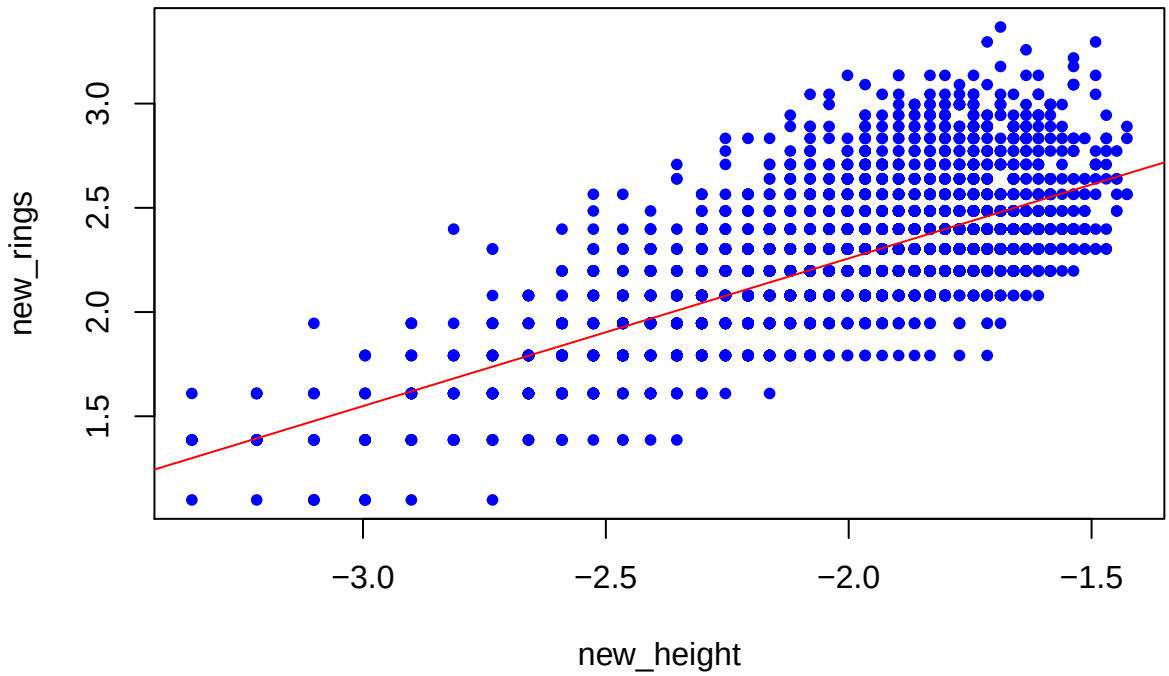



```
summary(mod.abalone)
```

```
##
## Call:
## lm(formula = rings ~ height, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0312 -1.6628 -0.5451  0.8233 16.7118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7805     0.1524   18.25   <2e-16 ***
## height       51.3929     1.0536   48.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.558 on 4152 degrees of freedom
## Multiple R-squared:  0.3643, Adjusted R-squared:  0.3642
## F-statistic:  2379 on 1 and 4152 DF,  p-value: < 2.2e-16
```

```
# otbain filtered variables
new_height <- log(df$height)
new_rings <- log(df$rings)
```

```r
plot(new_height, new_rings, col= "blue", pch=20)
lm_new.abalone <- lm(new_rings ~ new_height )
abline(coef(lm_new.abalone), col="red")
```




```r
# xp <- seq(0.01,0.30, by=0.01)
# m.fit <- cbind(rep(1,length(xp)), xp, log(xp)) %*% coef(lm_new.abalone)
# lines(xp,m.fit, col="red", lty=2)
# legend(0.25,0.4,c("Straight Line","Quadratic"),lty=c(1,2),col=c("black","red"))
```

It seems to be doing a decent job, although we observe the variance in the rings variable seems to be very big for the observations, so a better fit is very hard.

```r
summary(lm_new.abalone)
```

```
##
## Call:
## lm(formula = new_rings ~ new_height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66780 -0.15638 -0.03137  0.11642  0.88832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  3.67466    0.02308  159.25   <2e-16 ***
## new_height   0.70860    0.01134   62.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2248 on 4152 degrees of freedom
## Multiple R-squared:  0.4847, Adjusted R-squared:  0.4845
## F-statistic:  3905 on 1 and 4152 DF,  p-value: < 2.2e-16
```

```
summary(aov(lm_new.abalone))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## new_height     1  197.4  197.37    3905 <2e-16 ***
## Residuals   4152  209.9    0.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observing the two tables above we can see that the intercept, the filtered height and log of height attributes have very significant p-values. In addition, we see that the amount of variance explained in the Adjusted R-square statistic improved from 0.3106 in the raw , first order linear model to 0.4845 in the current model with logarithmic features and target variables adjustments. In addition, the F-statistic's p-value $< \alpha = 0.01$ indicates that the model's fit is significant.

Interpretation

Letting Y =number of rings and X_1 =height, we are assuming that the true model takes the form

$$log(Y) = \beta_0 + log(\beta_1 X_1) \iff Y = e^{\beta_0}\beta_1 X_1$$

Where $\hat{\beta}_0 = 3.67466$ and $\hat{\beta}_1 = 0.70860$. We can interpret these values as follows: when $X_1 = 0$, i.e. when height remain unchaged, say $X_1 = x$, then the number of rings is a multiple of $e^{\beta_0}\beta_1 = exp(3.67466)(0.70860)$ and per each unit of increase in height, the number increases by a factor of $e^{\beta_0}\beta_1 = exp(3.67466)(0.70860) = 27.94382$

Confidence intervals

We can obtain the confidence 95% confidence intevals for β_0, β_1 as

```
(confint(lm_new.abalone))
```

```
##                 2.5 %    97.5 %
## (Intercept) 3.6294218 3.7199023
## new_height  0.6863655 0.7308299
```

```
# COnfidence intervals for beta_0, beta_1 and beta_2

(t_alpha_half <- qt(.95, 49)) # .95th quantile
```

```
## [1] 1.676551
```

```r
beta_0_hat <- coef(lm_new.abalone)[1] # get beta 0  hat
beta_1_hat <- coef(lm_new.abalone)[2] # get beta 1 hat
```

Is there a statistically significant relationship between the height and the number of rings?

We test for the hypothesis that

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

At the $\alpha = 0.01$ significance level, we see that since we have p-value$< 2e - 16 < 0.01$, we **reject** the null hypothesis that $\beta_1 = 0$ and therefore conclude there is a significantly statistical relationship between height and the number of rings (and hence, the age) of abalones. In other words, knowing the predictor "height" is somehow informative of the actual number of rings of the abalone.

Predictions

We know find a point estimate and a 95% confidence interval for the average number of rings for abalones with height at 0.128.

```r
# Create new observation
new_data <- data.frame(new_height <- log(0.128))
exp(predict(lm_new.abalone, newdata = new_data, interval = "prediction")) # predict and convert back to
```

```
##        fit      lwr      upr
## 1 9.188752 5.912954 14.27935
```

So we predict that the average number of rings for abalones with height at 0.128 is X