Homework 3

Greg DePaul

2023-03-28

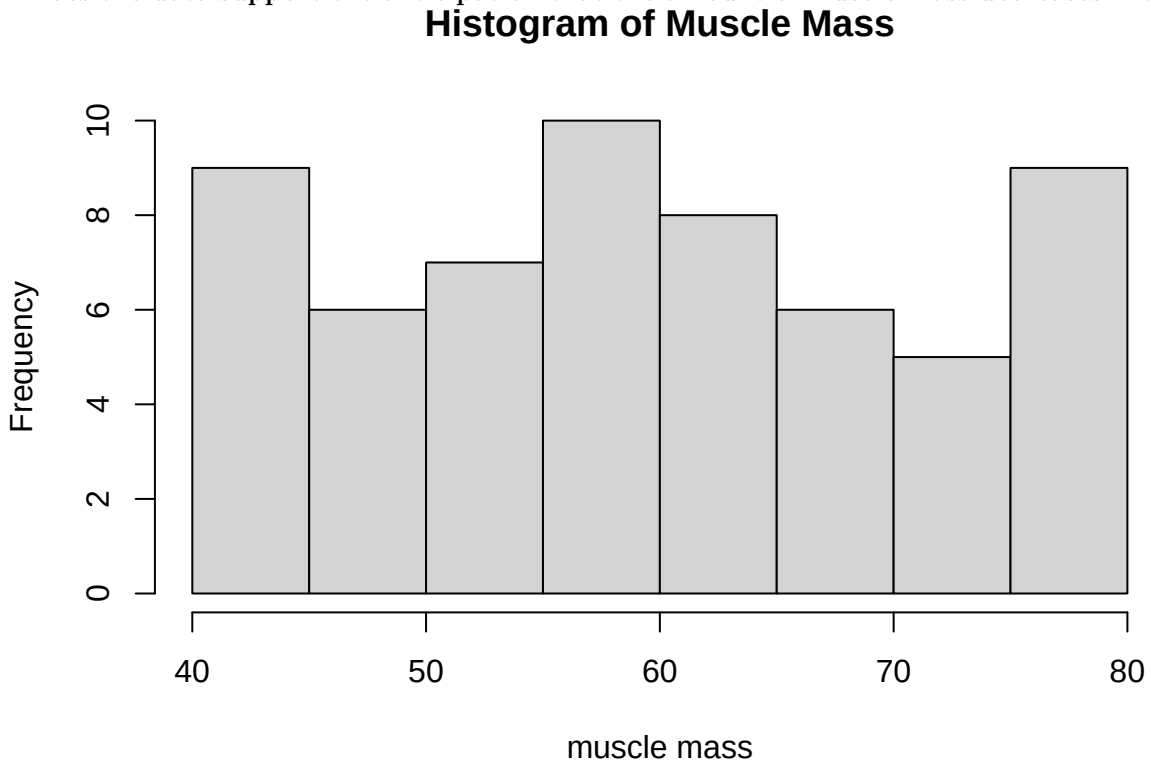
Problem 1 - A simple linear regression case study by R

You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file and its corresponding .html file.

A person's muscle is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each of the four 10-year age groups, beginning with age 40 and ending with age 79. Two variables being measured are: age (X) and the amount of muscle mass (Y). Data are stored in the file "muscle.txt".

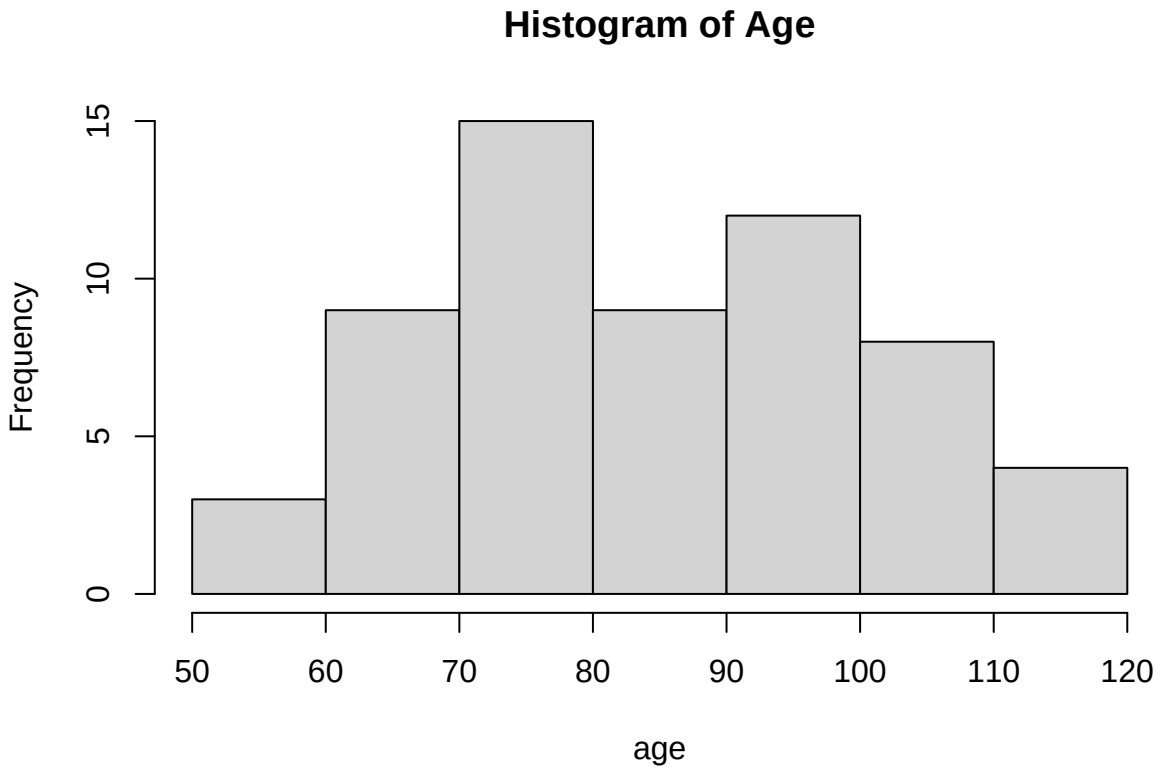
```
my_data <- read.table("muscle.txt", header=FALSE)
colnames(my_data) <- c('age','muscle_mass')
hist(my_data$muscle_mass, xlab='muscle mass', main='Histogram of Muscle Mass')
```

(a) Read data into R. Draw histogram for muscle mass and age, respectively. Comment on their distributions. Draw the scatter plot of muscle mass versus age. Do you think their relation is linear? Does the data support the anticipation that the amount of muscle mass decreases with age?



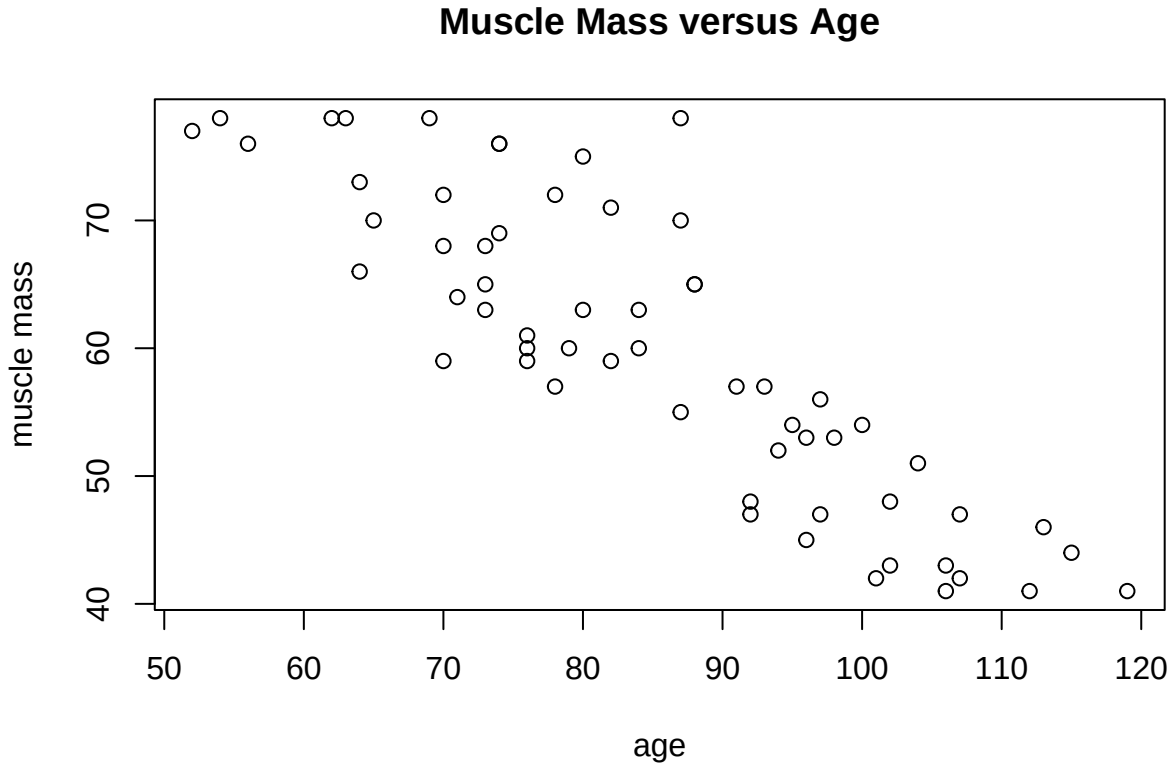
We see that this distribution is heavy tailed.

```
hist(my_data$age, xlab='age', main='Histogram of Age')
```



This distribution is bimodal.

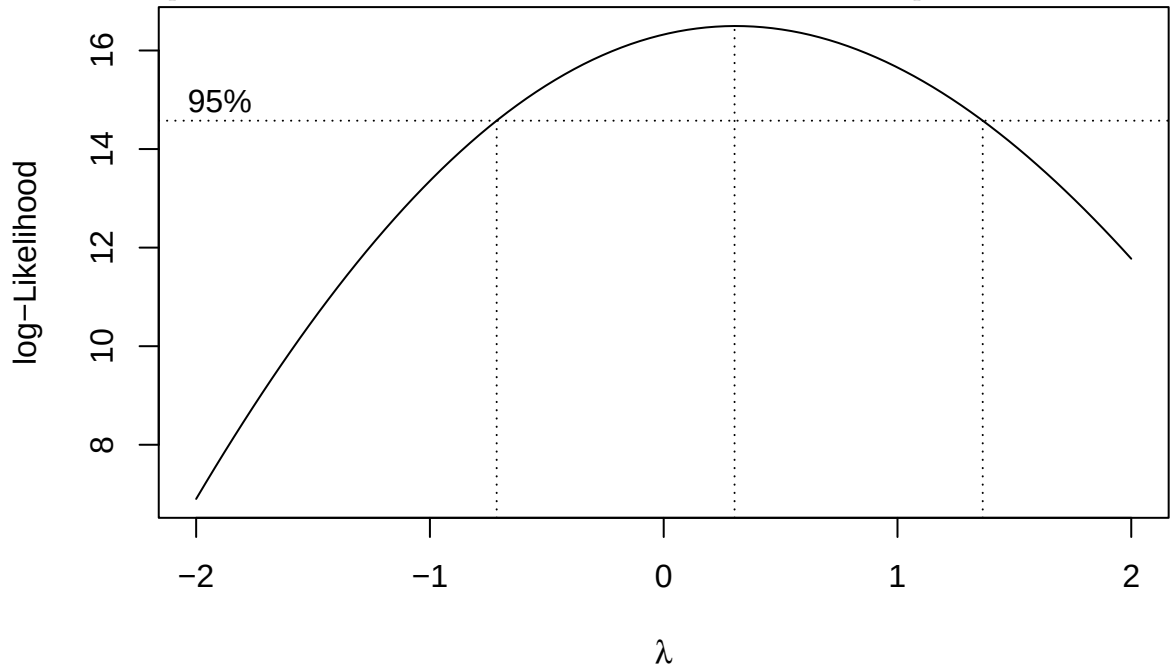
```
plot(my_data$age, my_data$muscle_mass, xlab='age', ylab='muscle mass', main='Muscle Mass versus Age')
```



There is clearly mutual information between the two variables. Therefore, we do expect some function to exist.

```
library(MASS)
X <- my_data$age
Y <- my_data$muscle_mass
bc <- boxcox(Y ~ X)
```

(b) Use the Box-Cox procedure to decide whether a transformation of the response variable is needed.



Notice that $\lambda = 1$ is within the 95% confidence interval. Therefore, no transformation is needed.

```
fit <- lm(Y ~ X)
summary(fit)
```

(c) Perform linear regression of the amount of muscle mass on age and obtain a summary. From the summary, obtain the estimated regression coefficients and their standard errors, the mean squared error (MSE) and its degrees of freedom.

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4170  -4.2031  -0.3957   3.1333  19.2983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 113.53873    4.13130   27.48   <2e-16 ***
## X           -0.63031    0.04778  -13.19   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.948 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

$$\hat{\beta}_0 = 113.53873 \quad S\{\hat{\beta}_0\} = 4.13130$$
$$\hat{\beta}_1 = -0.63031, \quad S\{\hat{\beta}_1\} = 0.04778$$

```
SSE <- sum((fitted(fit) - mean(my_data$muscle_mass))^2)
SSR <- sum((fitted(fit) - my_data$muscle_mass)^2)
n <- length(X)
MSR <- SSR
MSE <- SSE / (n-2)
print(MSE)
```

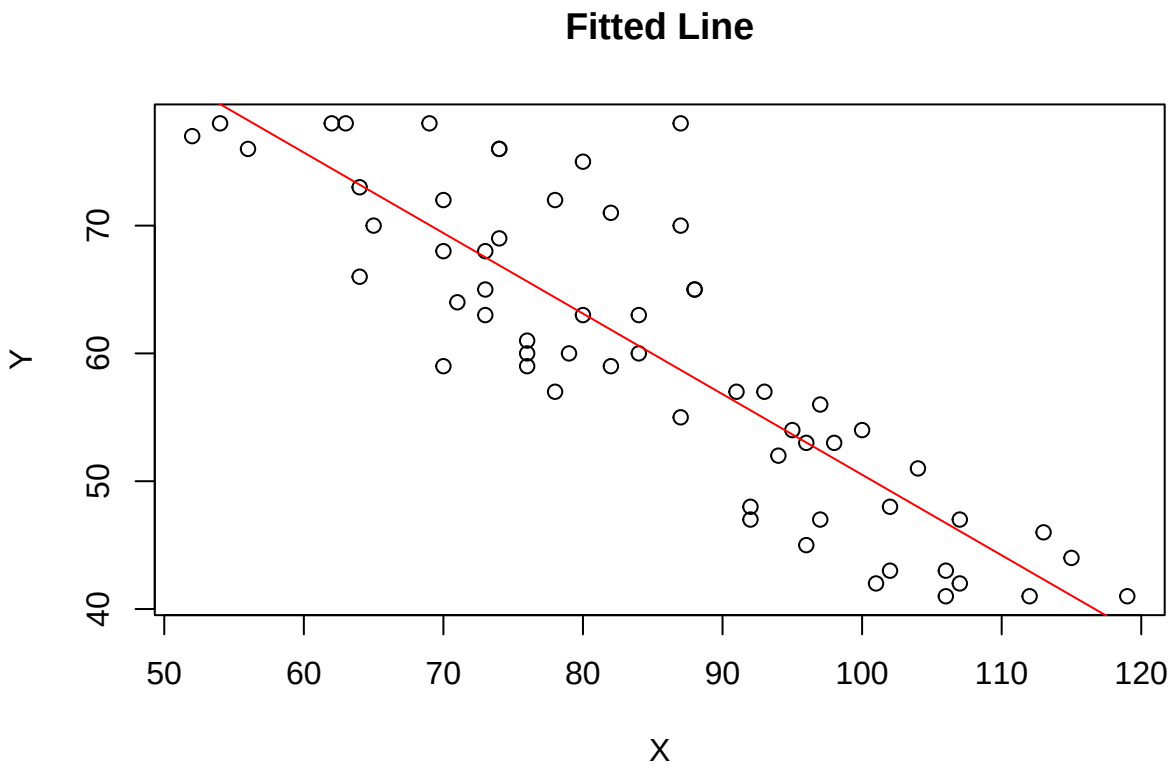
```
## [1] 106.186
```

$$MSE = 106.186 \quad \text{degrees of freedom} = 58$$

(d) Write down the fitted regression line. Add the fitted regression line to the scatter plot. Does it appear to fit the data well? The fitted line on the transformed data, based off the coefficients found will be:

$$\hat{Y} = 192.10419 - 0.63724X$$

```
plot(X, Y, main = "Fitted Line")
abline(fit, col='red')
```



```
print(fit$residuals[6])
```

(e) Obtain the fitted values and residuals for the 6th and 16th cases in the data set.

```
##        6
## 2.468238
```

```
print(fit$fitted.values[6])
```

```
##        6
## 38.53176
```

```
print(fit$residuals[16])
```

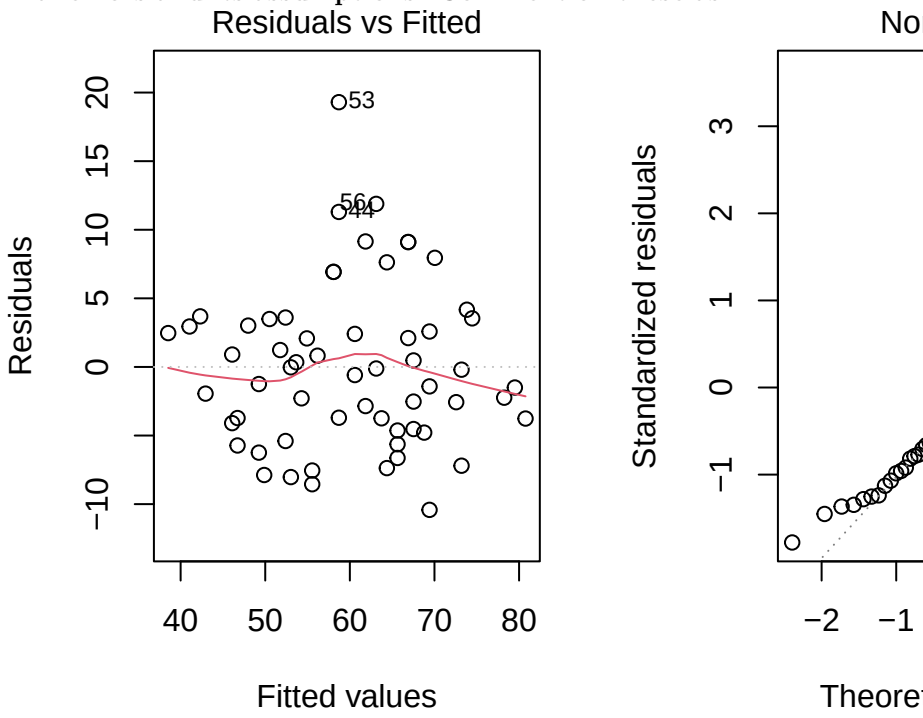
```
##        16
## -3.701702
```

```
print(fit$fitted.values[16])
```

```
##      16
## 58.7017
```

```
par(mfrow = c(1,2))
plot(fit, which=1)
plot(fit, which=2)
```

(f) Draw the residuals vs. fitted values plot and the residuals Normal Q-Q plot. Write down the simple linear regression model with Normal errors and its assumptions. Comment on these assumptions based on the residual plots.



(g) Construct a 99% confidence interval for the estimated regression intercept. Interpret your confidence interval. Further, we know that for a given $1-\alpha$ confidence interval for β_0, we can estimate it using

$$CI_{99\%}(\beta_0) = \hat{\beta}_0 \pm t(1-\frac{\alpha}{2}, n-2)SE(\hat{\beta}_0) = \hat{\beta}_0 \pm t_{58}(0.995)SE(\hat{\beta}_0)$$

```
betas <- fit$coefficients
beta_0 <- betas[1]
s_beta_0 <- summary(fit)$coefficients["(Intercept)","Std. Error"]
crit_val  <- qt(1 - 0.01 / 2, df = 58)
crit_val <- qt(1 - 0.01 / 2, df = 58)
left_val <- beta_0 - s_beta_0*crit_val
print(left_val)
```

```
## (Intercept)
##    102.5359
```

```
right_val <- beta_0 + s_beta_0*crit_val
print(right_val)
```

```
## (Intercept)
##    124.5416
```

So we get the condidence interval to be:

$$CI_{99\%}(\beta_0) = (102.5359, 124.5416)$$

These values of course doesn't make sense, which suggests that we can really generalize or extrapolate around the intercept.

(h) Conduct a test at level 0.01 to decide whether or not there is a negative linear association between the amount of muscle mass and age. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. (Hint: Which form of alternatives should you use?)

- $H_0 : \beta_1 \geq 0$
- $H_A : \beta_1 < 0$
- Test Statistic: $T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}}$
- Null distribution of T^* is $t_{n-2} = t_{58}$
- Rule: Reject if $T^* < t_{58}(\alpha)$

```
beta_1 <- betas[2]
s_beta_1 <- summary(fit)$coefficients["X","Std. Error"]
crit_val  <- qt(0.01, df = 58)
T_star <- beta_1 / s_beta_1
print(T_star < crit_val)
```

```
##    X
## TRUE
```

Therefore, we reject the hypothesis that $\beta_1 \geq 0$.

```
mean_x <- mean(X)
sum_x_squared <- sum(X^2)

X_pred <- 60
Y_pred <- beta_0 + beta_1*X_pred
crit_val <- qt(1 - 0.05 / 2, df = n - 2)
```

```
s_Y_pred <- sqrt(MSE * (1 + 1/n + (X_pred - mean_x)^2 / (sum_x_squared - n * mean_x^2)))
left_val <- Y_pred - crit_val*s_Y_pred
print(left_val)
```

(i) Construct a 95% prediction interval for the muscle mass of a woman aged at 60. Interpret your prediction interval.

```
## (Intercept)
##    54.51459
```

```
right_val <- Y_pred + crit_val*s_Y_pred
print(right_val)
```

```
## (Intercept)
##    96.92559
```

So we get the condidence interval to be:

$$CI_{95\%}(Y_{pred}(60)) = (63.47927, 87.96091)$$

(j) Obtain the ANOVA table for this data. Test whether or not there is a linear association between the amount of muscle mass and age by an F test at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. Printing the necessary values,

```
print(n)
```

```
## [1] 60
```

```
print(SSE)
```

```
## [1] 6158.786
```

```
print(SSR)
```

```
## [1] 2052.197
```

```
print(SSR + SSE)
```

```
## [1] 8210.983
```

```
print(MSE)
```

```
## [1] 106.186
```

```
print(MSR)
```

```
## [1] 2052.197
```

we get the ANOVA table:

	SS	df	MS
Regression	SSR = 2052.197	1	MSR = 2052.197
Error	SSE = 6158.786	58	MSE = 106.186
Total	SSTO = 8210.983	59	

Under the normal error model, we know $SSE \sim \sigma^2 \chi^2_{82}$. Therefore,

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$
- Test Statistic: $F^* = \frac{MSR}{MSE}$
- Null distribution of F^* is $F_{1,58}$

We use the one-side F-test. The rule becomes is we reject H_0 if the following comparison is true:

$$F^* > F(1-\alpha, 1, 58) = F(0.99, 1, 58)$$

```
F_star <- MSR / MSE
crit_val <- qf(0.99, 1, n-2, lower.tail = TRUE)
print(F_star > crit_val)
```

```
## [1] TRUE
```

Therefore, we reject the claim that there is no linear relationship between these two values.

(k) What proportion of the total variation in muscle mass is "explained" by age? What is the correlation coefficient between muscle mass and age? By definition, the coefficient of determination, R^2, is the proportion of the variation that explained by the regression line.

```
R_squared <- 1 - (SSR/(SSE + SSR))
print(R_squared)
```

```
## [1] 0.7500668
```

and so we get

$$R^2 = 0.7500668$$

whereas, for correlation, we get:

```
print(cor(X,Y))
```

```
## [1] -0.866064
```

$$r = -0.866064$$