Homework 3

Greg DePaul

2023-03-28

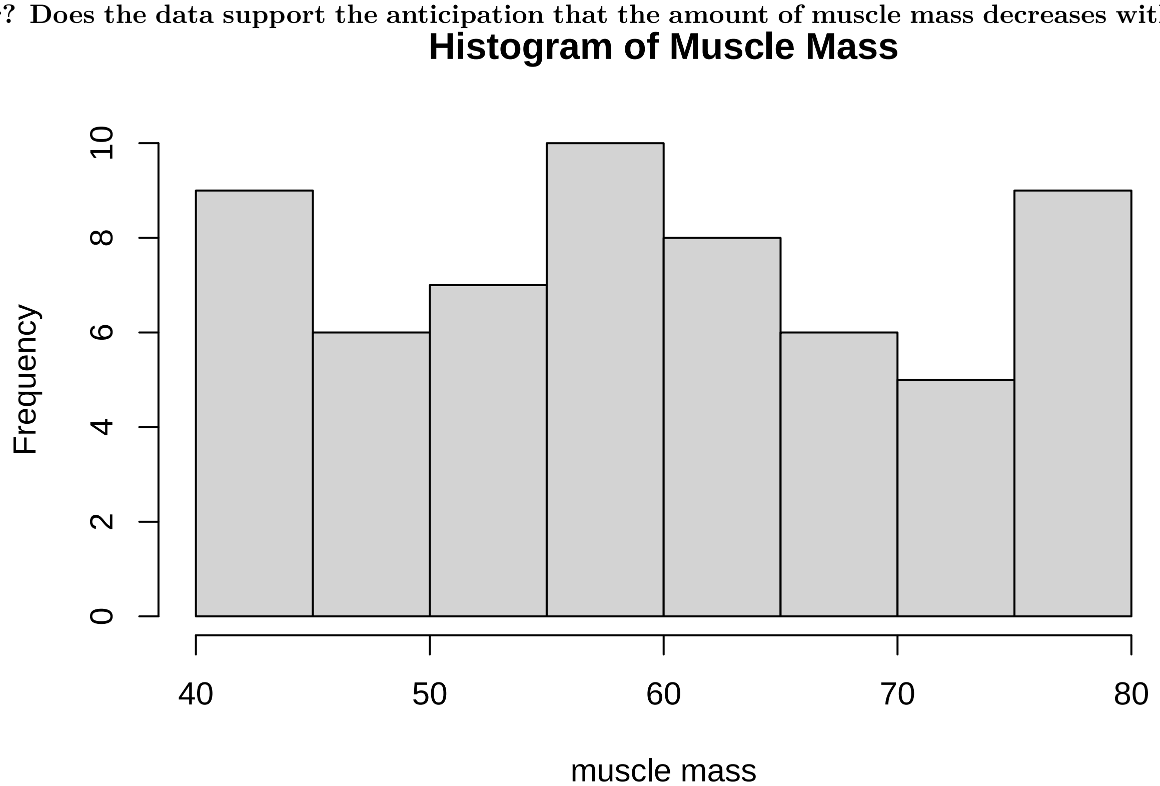
Problem 1 - A simple linear regression case study by R

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.

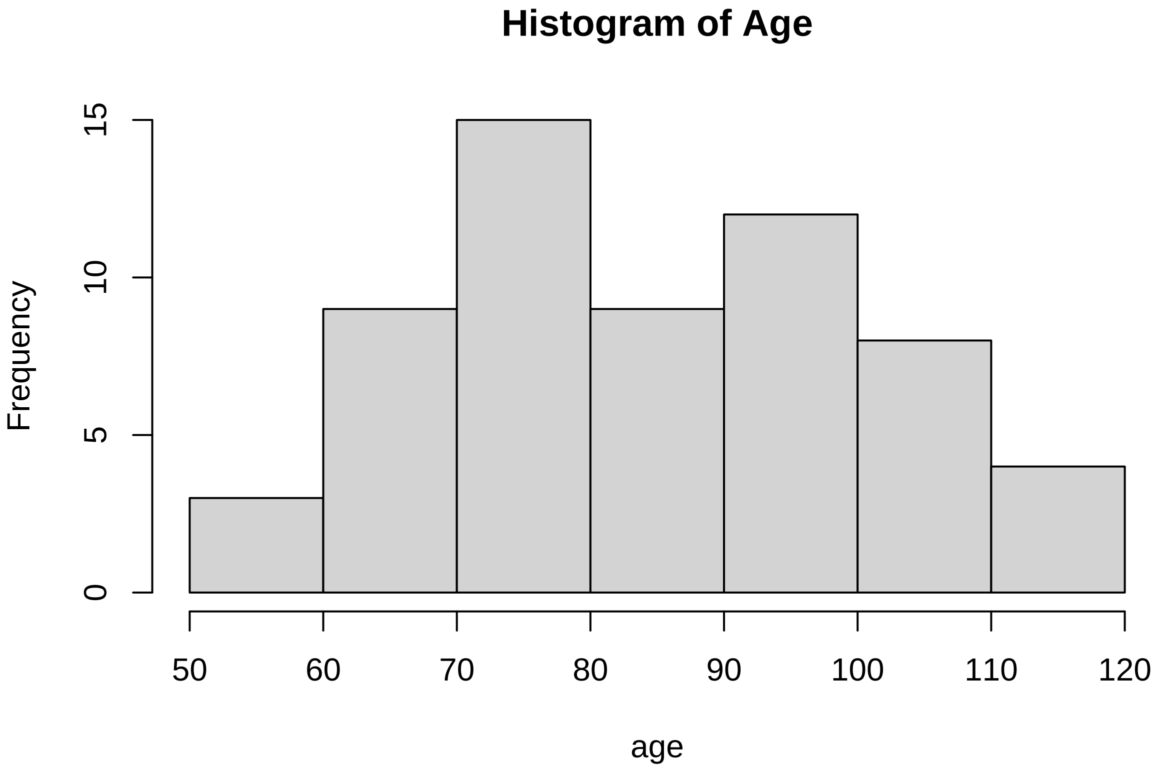
A person's muscle is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each of the four 10-year age groups, beginning with age 40 and ending with age 79. Two variables being measured are: age (X) and the amount of muscle mass (Y). Data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

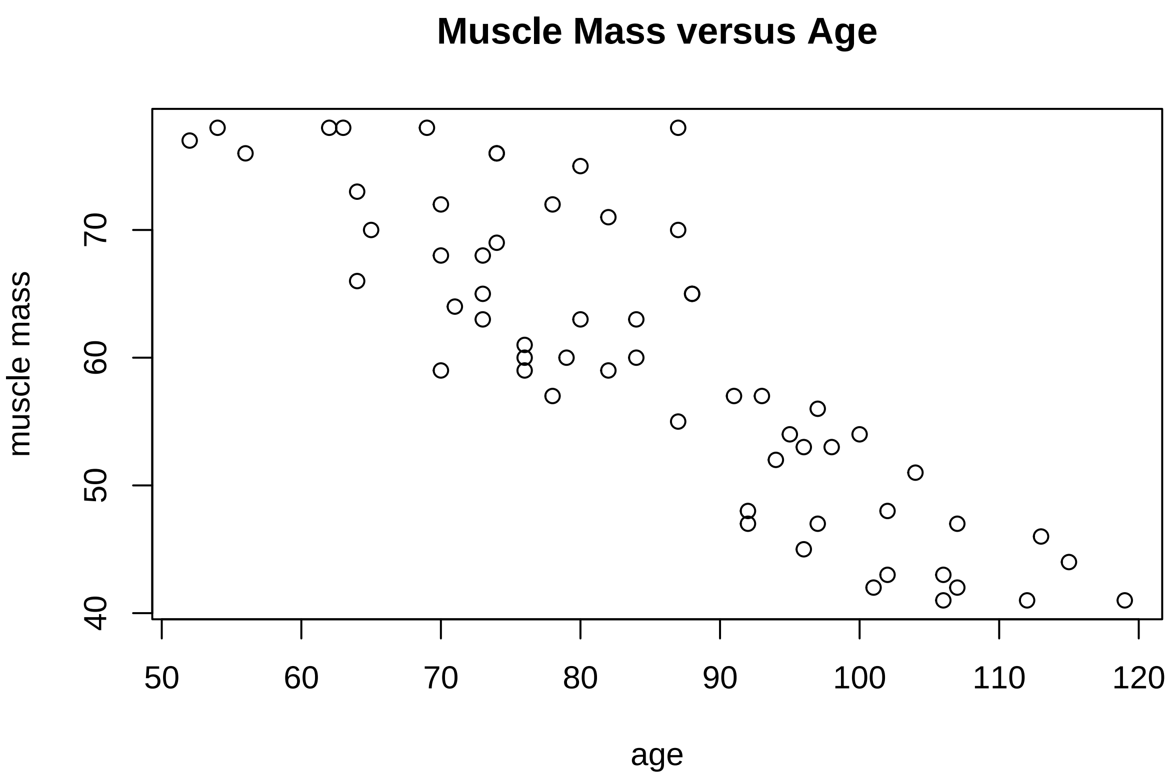
(a) Read data into R. Draw histogram for muscle mass and age, respectively. Comment on their distributions. Draw the scatter plot of muscle mass versus age. Do you think their relation is linear? Does the data support the anticipation that the amount of muscle mass decreases with age?



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

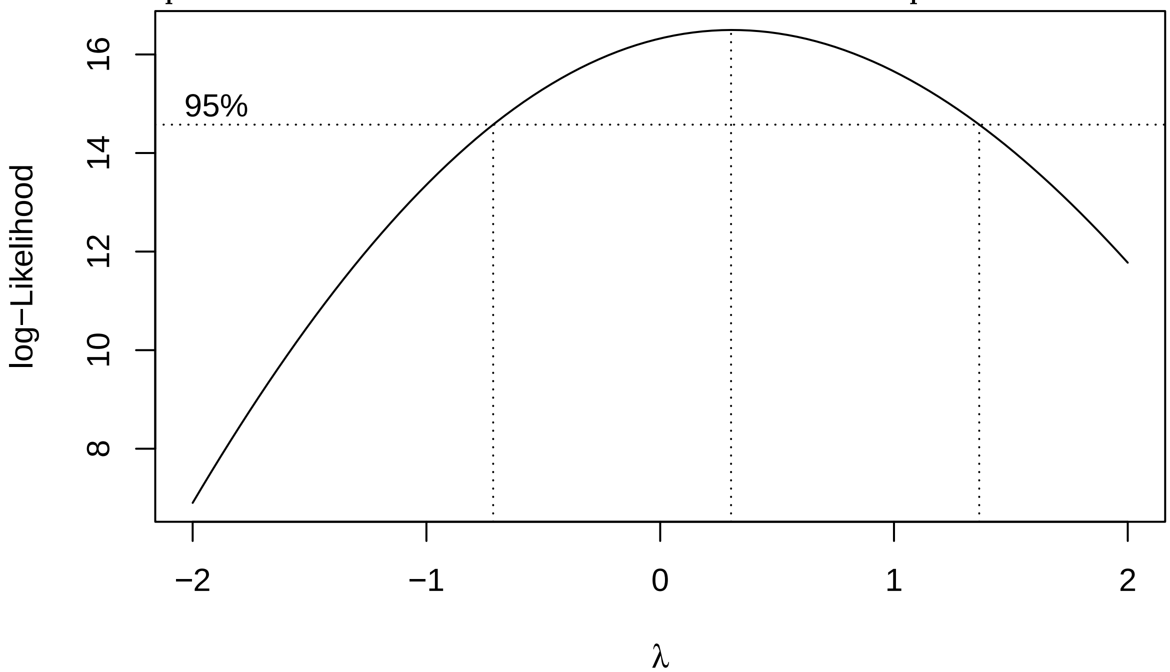



```r
plot(my_data$age, my_data$muscle_mass, xlab='age', ylab='muscle mass', main='Muscle Mass versus Age')
```



```
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.

```
lambda <- bc$x[which.max(bc$y)]
print(lambda)
```

```
## [1] 0.3030303
```

Recall, the box cox procedure tells us an appropriate transformation for the response variable Y. Specifically,

$$Y_{\text{transformed}} = \begin{cases} \frac{K_1}{\lambda}(Y^{\lambda} - 1) & \lambda \neq 0 \\ K_2 \log(Y) & \text{otherwise} \end{cases}$$

```
K_2 <- prod(Y)^(1/length(Y))
K_1 <- 1/(K_2^(lambda - 1))
print(K_1)
```

```
## [1] 17.10951
```

```
Y_transform <- K_1/lambda * (Y^lambda - 1)
inverse_transform <- function(z) {
  return((lambda*z / K_1 + 1)^(1/lambda))
}
```

So from this, we see that we should select the transformation:

$$Y_{transform} = \frac{17.10951}{0.3030303}(Y^{0.3030303} - 1)$$

```r
fit <- lm(Y_transform ~ 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_transform ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7036 -4.2937 -0.1852  3.0178 18.2781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 192.10419    4.07380   47.16   <2e-16 ***
## X            -0.63724    0.04711  -13.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.866 on 58 degrees of freedom
## Multiple R-squared:  0.7593, Adjusted R-squared:  0.7552
## F-statistic:   183 on 1 and 58 DF,  p-value: < 2.2e-16
```

$$\hat{\beta}_0 = 192.10419 \quad S\{\hat{\beta}_0\} = 4.07380$$
$$\hat{\beta}_1 = -0.63724, \quad S\{\hat{\beta}_1\} = 0.04711$$

```r
MSE <- sum(fit$residuals^2)/fit$df.residual
print(MSE)
```

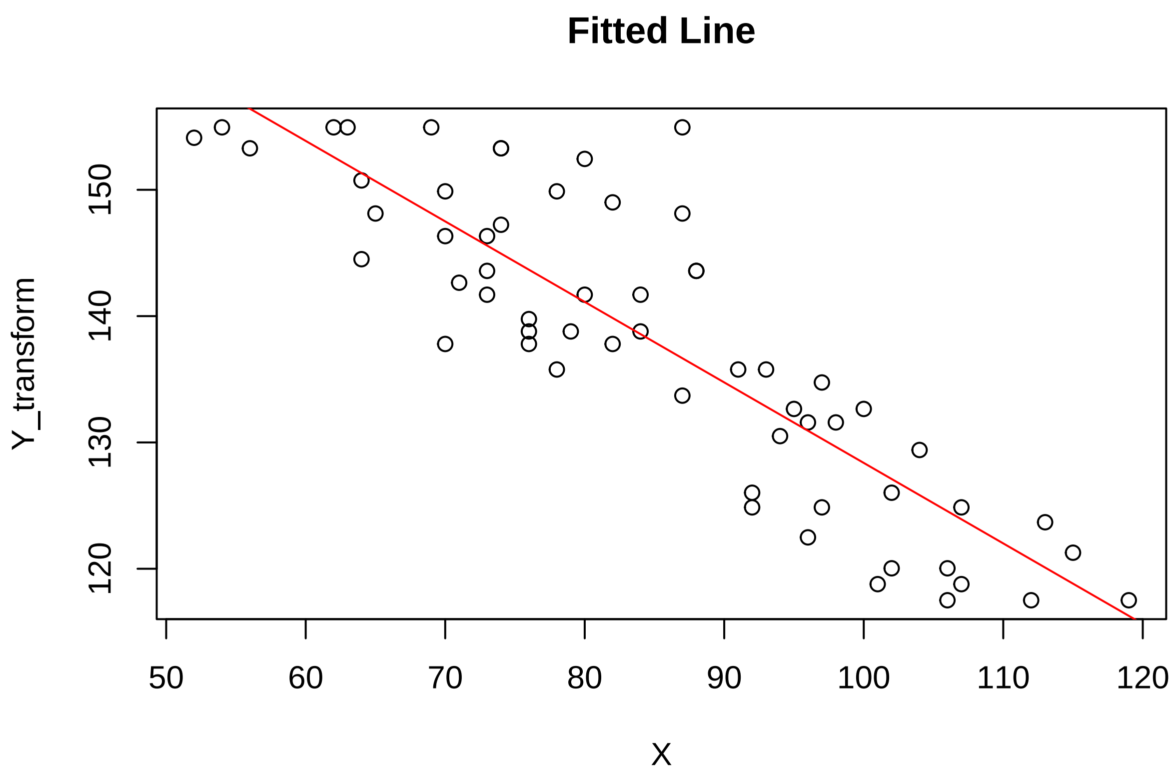
```
## [1] 34.40473
```

$$MSE = 33.25791 \quad \text{degrees of freedom} = 58$$

(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:

$$\hat{Y}_{\text{transformed}} = 192.10419 - 0.63724X$$

```r
plot(X, Y_transform, 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
## 1.235252
```

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

```
##        6
## 116.2732
```

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

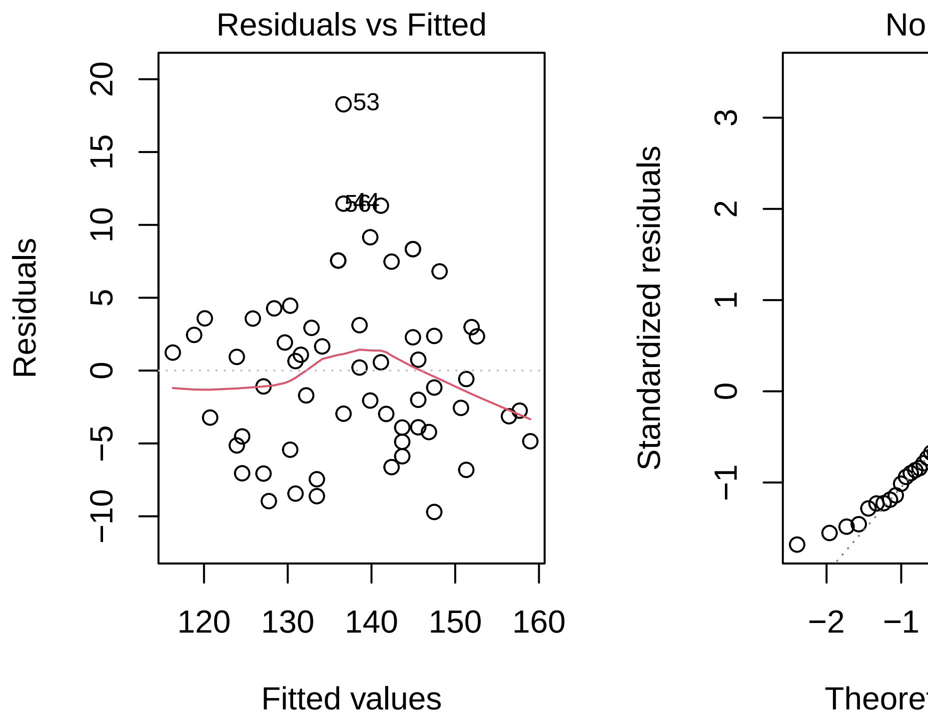
```
##        16
## -2.959528
```

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

```
##       16
## 136.6647
```

```
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 as-



sumptions 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)
##    181.2545
```

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

```
## (Intercept)
##    202.9539
```

So we get the condidence interval to be:

$$CI_{99\%}(\beta_0) = (181.2545, 202.9539)$$

When we take the inverse transform, we get the confidence interval bounds to be:

```
print(inverse_transform(left_val))
```

```
## (Intercept)
##    114.8714
```

```
print(inverse_transform(right_val))
```

```
## (Intercept)
##    153.2517
```

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
```

```
n <- length(X)
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)
##    141.7996
```

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

```
## (Intercept)
##    165.9405
```

(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.

(k) What proportion of the total variation in muscle mass is "explained" by age? What is the correlation coefficient between muscle mass and age?