STAT 511: Assignment #1

Rumil Legaspi

25 January 2021

Assignment Questions

---

1. KNN 4th Edition End of Chapter 1 Questions

In a regression model, β_0 = 100, β_1 = 20, and σ^2 = 25. An observation on Y variable will be made for X = 5.

$Y_i = \beta_0$ *(intercept)* + $\beta_1 X_i$ *(slope)(independent variable)* + ϵ_i *(error)*

y = 100 + 20X_i + ϵ_i

(a). Can we compute the exact probability that Y will fall between 195 and 205? Explain.

The probability cannot be calculated because the for a simple linear regression model the mean of ϵ_i should equal 0. Because ϵ_i is unspecified we are missing information and cannot compute the exact probability.

(b). If the normal error regression model is applicable, can we now compute the exact probability that Y will fall between 195 and 205? If so, compute it.

note: ϵ_i (error term) = 0 and follows a normal distribution

For this problem we recall:

- 1. The Z score formula since we are dealing with a normal distribution. $\frac{X-\mu}{\sigma}$
- 2. How to find the probability between 2 points given a normal distribution.

(aka find the z score which finds everything from the left and subtract it by the larger number to get the probability between a and b)

- 3. And that we are also given $\sigma^2 = 25$ *(variance)* and $\sigma = 5$ *(Standard deviation)*

SO: $P(195 \leq Y \leq 205) = P(\frac{195-200}{5} \leq \frac{X-\mu}{\sigma} \leq \frac{205-200}{5})$

$= P(-1 \leq z \leq 1)$

$= P(z < 1) - P(z < -1)$ *bigger number or b is P(z < 1)*

$= 0.841 - 0.158$ *converting numbers using pos/neg z table*

$= 0.683$

The probability that Y will fall between the 195 and 205 is roughly 0.683.

---

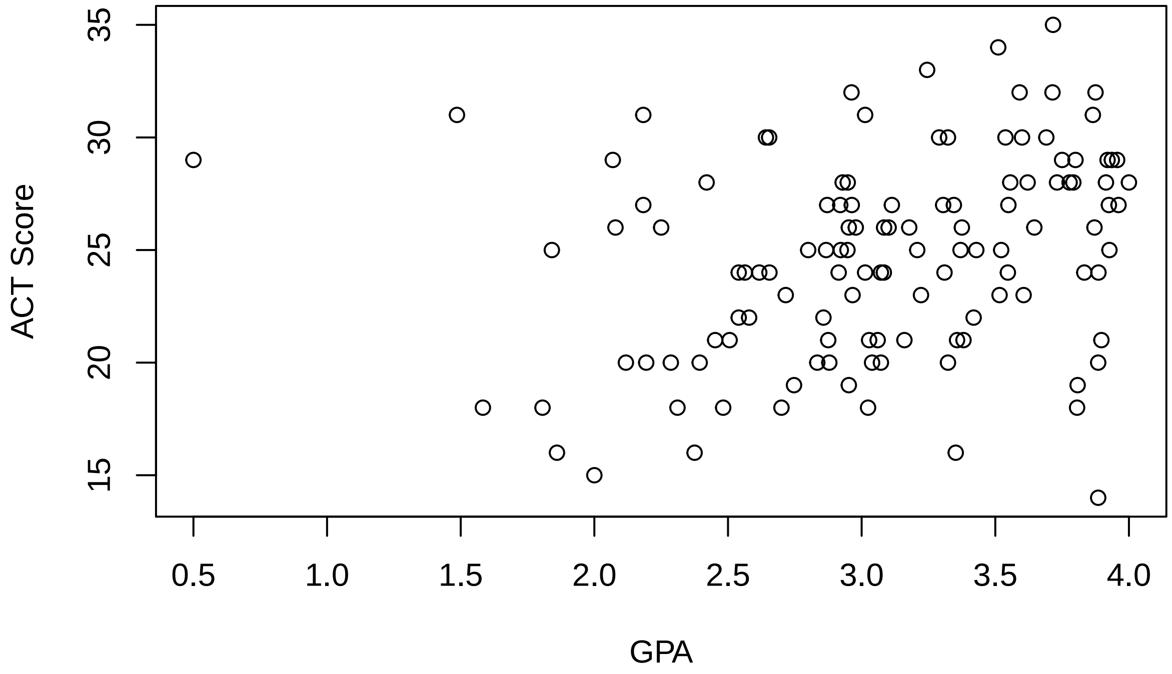
2. Grade Point Average Problem (Use R)

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). See the dataset "GPA.txt". The first column is GPA. The second column is ACT.

```
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 1/STAT511 As


gpa_data = read.table(file = "GPA.txt", header = FALSE, sep = "")

#Adding headers
names(gpa_data) <- c("GPA", "ACT Score")
head(gpa_data)
```

```
##     GPA ACT Score
## 1 3.897        21
## 2 3.885        14
## 3 3.778        28
## 4 2.540        22
## 5 3.028        21
## 6 3.865        31
```

```
#Defining dependent and independent vars
GPA = gpa_data$GPA
ACT = gpa_data$`ACT Score`

#scatterplot
plot(gpa_data)
```



(a). Obtain the least squares estimates of β_0 and β_1. Write down the estimated regression equation.

```
lm(`ACT Score`~ GPA, data = gpa_data)
```

```
##
## Call:
## lm(formula = `ACT Score` ~ GPA, data = gpa_data)
##
## Coefficients:
## (Intercept)          GPA
##       18.98         1.87
```

```
#ACT Score is our response, GPA is our explanatory.
# in other words `ACT Score`~ GPA says, ACT is explained by GPA
gpa_lm = lm(`ACT Score`~ GPA, data = gpa_data)
summary(gpa_lm)
```

```
##
## Call:
## lm(formula = `ACT Score` ~ GPA, data = gpa_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -12.242  -3.276   0.218   2.657   9.245 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  18.9754     1.9322   9.821  < 2e-16 ***
## GPA           1.8704     0.6153   3.040  0.00292 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.325 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476 
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

From the lm() we get:

$\beta_0 =$ **18.9754** ***(intercept)***

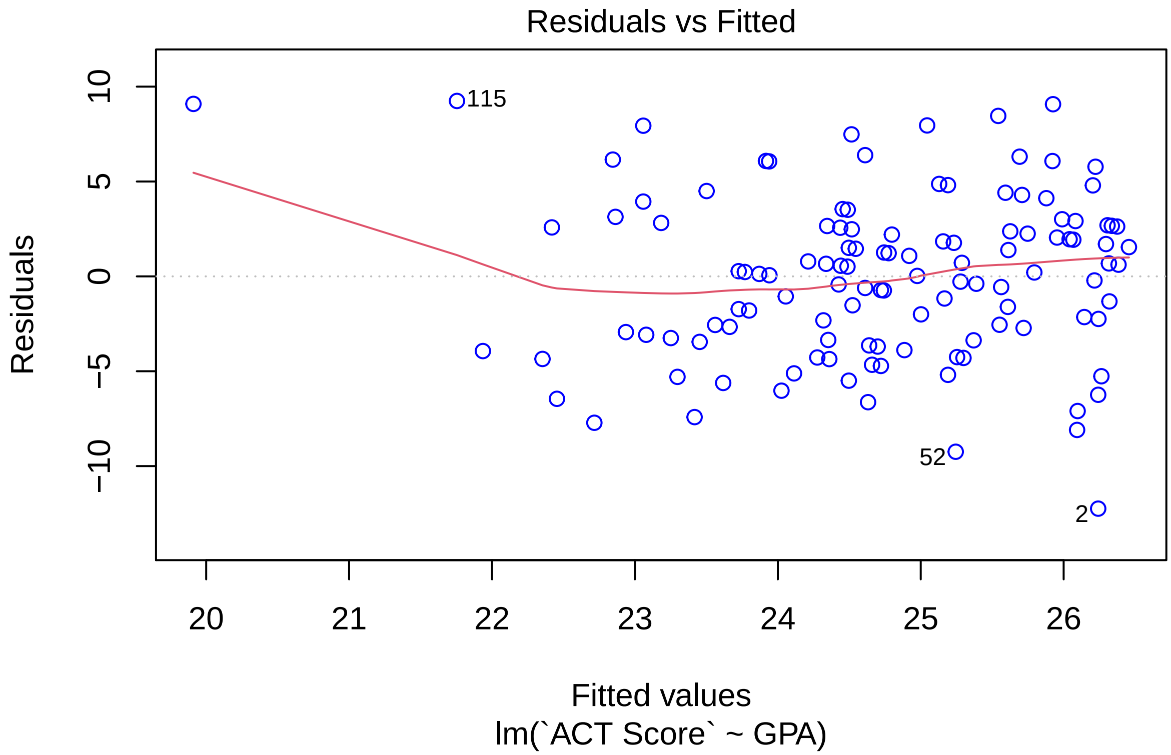
$\beta_1 =$ **1.87** ***(slope)***

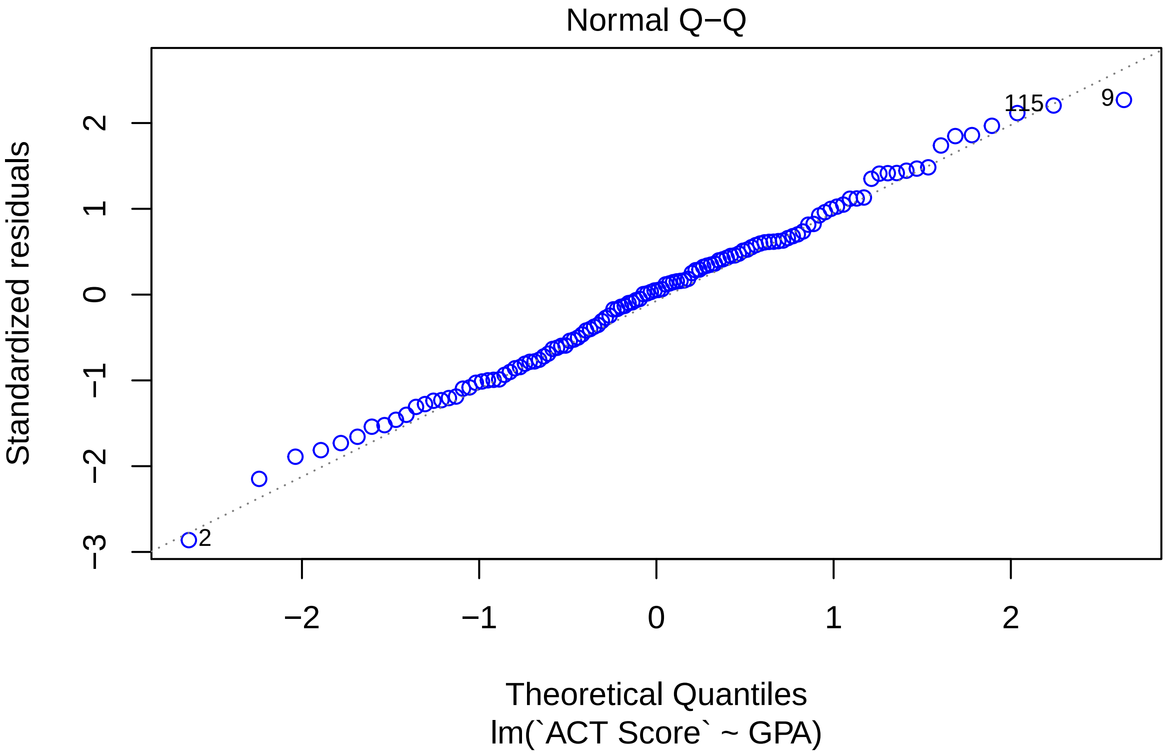
and the estimated regression equation to be:

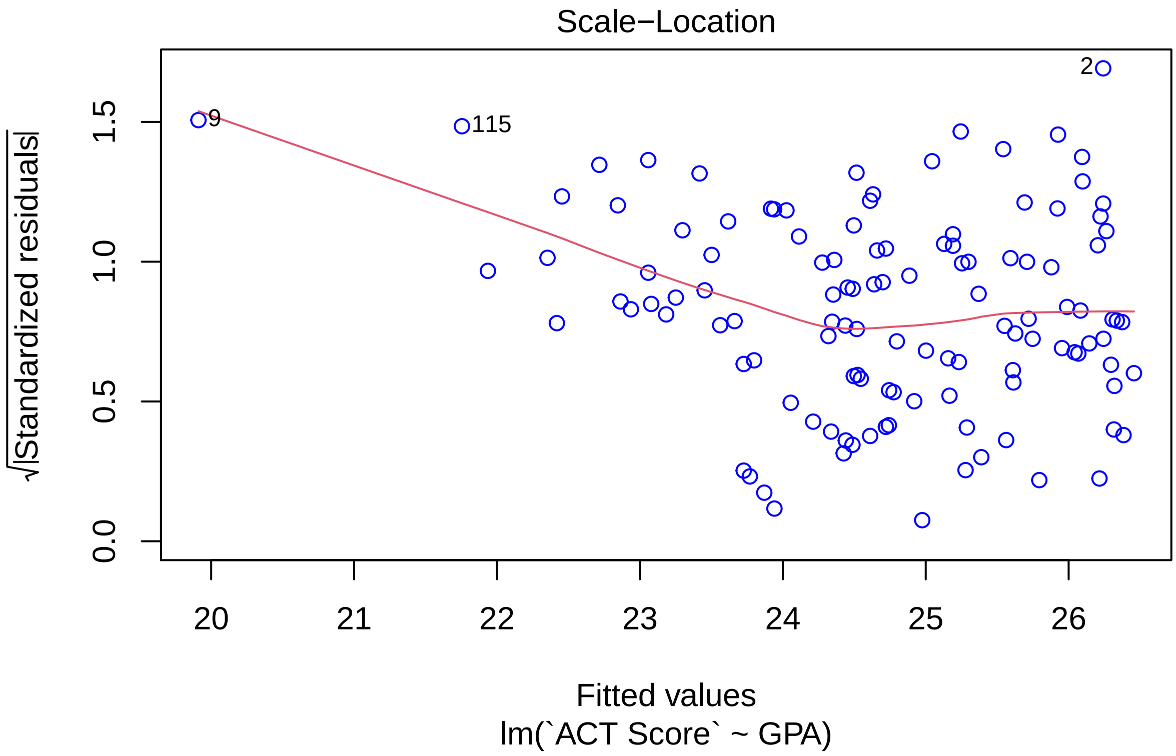
$\hat{Y} = 18.97 + 2.1X$

(b). Plot the estimated regression line and the data points. Does the estimated regression function appear to fit the data well?

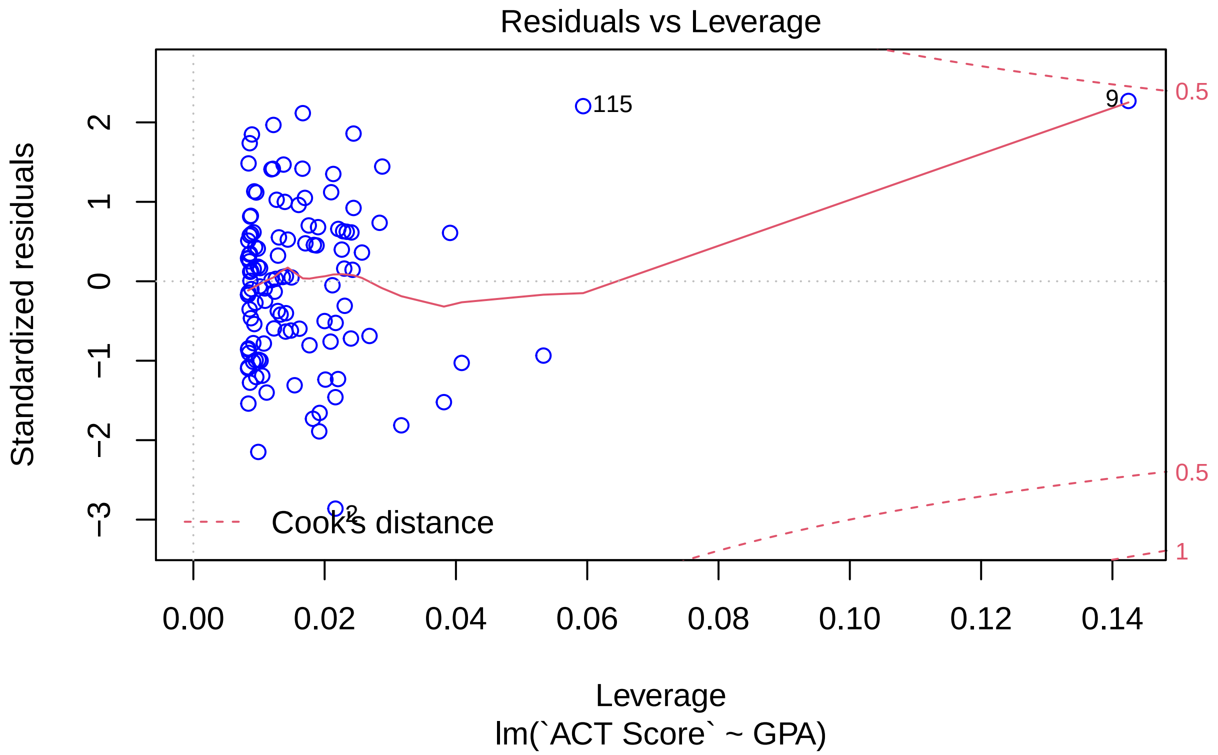
```r
plot(gpa_lm,
     col = "blue")
```


Residuals vs Fitted
Residuals
10
5
0
−5
−10
20
21
22
23
24
25
26
Fitted values
lm(`ACT Score` ~ GPA)
115
52
2


Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
−2
−1
0
1
2
2
115
9
Theoretical Quantiles
lm(`ACT Score` ~ GPA)



```
abline(gpa_lm, col = "red")
```



(c). Obtain a point estimate of the mean freshman GPA for students with ACT test score = 30.

Using $\hat{Y} = 18.97 + 2.1X$ Solve for point estimate when x = 30

```
#18.97 + 2.1(30) = Y hat
18.97 + 2.1*30
```

```
## [1] 81.97
```

$\hat{Y} = 81.97$

(d). What is the estimated change in the mean response when the ACT score increases by one point?

In other words, when ACT score goes from 0 to 1 what is the change in $\hat{Y}$? Lets try when x =1 minus when x =0

```
ChangeinYhat <- 18.97 + 2.1*1 - 18.97 + 2.1*0

ChangeinYhat
```

```
## [1] 2.1
```

```
#*or is it simply?
# 1.87 our given slope?
```

Estimated change in $\hat{Y}$ (GPA) when ACT score increases by one point is 2.1 .

---

3. Refer to the GPA problem in Question 2. (Use R)

(a). Obtain the residuals $\hat{\epsilon_1}$. Do they sum to zero?

```
#Gives us each residual for every x value
residuals(gpa_lm)
```

```
##            1            2            3            4            5            6 
##  -5.26420719 -12.24176295   1.95836484  -1.72613786  -3.63887023   4.79564411 
##            7            8            9           10           11           12 
##   7.48457308   0.61609020   9.08938274   1.08057678  -1.16630985   4.40724962 
##           13           14           15           16           17           18 
##  -0.74173966  -0.61081494   7.95526312   2.48270273  -0.56282473   6.38918506 
##           19           20           21           22           23           24 
##   0.51262838  -2.93684879   0.23084402  -4.25421645   2.04627144   0.68342292 
##           25           26           27           28           29           30 
##   2.37358326   1.22459398   4.49830453  -1.79908163  -0.21557801  -3.69872154 
##           31           32           33           34           35           36 
##  -1.32031779  -7.41752957   3.54629474   0.71211719  -2.31903983  -0.72116578 
##           37           38           39           40           41           42 
##  -4.29910492   4.87109722   1.38667573   0.20525147   1.45464743   6.06064188 
##           43           44           45           46           47           48 
##   0.27386214   2.81626458   6.15479852   0.12984494   7.94157825  -7.71614711 
##           49           50           51           52           53           54 
##  -5.49672339  -8.09400505   2.65477522  -9.24486468   1.84304192   1.50327661 
##           55           56           57           58           59           60 
##  -1.60958356   4.12108557  -3.88575686  -3.07899564   4.80937556   2.66284903 
##           61           62           63           64           65           66 
##   0.55938721  -1.05532003  -0.27853104  -2.71993440   6.08308611  -2.56154677 
##           67           68           69           70           71           72 
##   0.05877152   6.07806745  -4.35329858  -2.55160261  -4.65944412  -1.52290833 
##           73           74           75           76           77           78 
##  -5.61765737  -6.02539437   2.69277468  -4.27602171  -2.00171876   1.25638999 
##           79           80           81           82           83           84 
##   1.54314642   8.45774916  -5.19062444  -4.72116578   3.13609498   5.77694058 
##           85           86           87           88           89           90 
##   0.02446618   2.56312792   1.76822779   2.62544197  -7.09774575  -2.66254585 
##           91           92           93           94           95           96 
##  -2.24363331   3.94157825  -0.38888188  -6.63138882   3.01073473  -2.14450459 
##           97           98           99          100          101          102 
##   2.20214974  -3.35270619  -5.11330098  -5.29782697   2.58123905  -3.93620981 
##          103          104          105          106          107          108 
##  -4.36018760   6.30812090  -0.42564997   9.07432674   0.78757030   2.25201030 
##          109          110          111          112          113          114 
##   1.93217990   0.66225664  -3.37017835   4.29128772  -3.45306628  -3.25106814
##          115          116          117          118          119          120
##   9.24521445  -6.24176295   2.91721707   1.70399680  -6.45429766   3.51075802
```

```
#We then square and sum
sum(residuals(gpa_lm)^2)
```

```
## [1] 2207.094
```

```
 SSE <- sum((gpa_lm$residuals)^2)
```

(b). Estimate the error variance σ^2 and standard deviation σ. In what units is σ expressed?

Using the sum of squared errors from the previous question, we can use it to find the variance of errors, σ^2.

We can simply divide by $n - 2$

n being 120 (total # of observations)

```
#Mean squared error (variance of errors)
MSE <- SSE/ 118 - 2
MSE
```

```
## [1] 16.70418
```

Then square root to get error standard deviation σ. AKA root mean squared error

```
#Root mean squared error (Standard deviation of errors)
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 4.087075
```

---

4. Refer to the GPA problem in Question 2.

(a). Interpret $\hat{\beta}_0$ in your estimated regression function. Does $\hat{\beta}_0$ provide any relevant information here? Explain.

$\hat{\beta}_0 = 18.97$ is a coefficent, specifically, our y-intercept. It shows us where our response variable (ACT Scores) is located when our predictor(GPA) is 0.

(b). Interpret $\hat{\beta}_1$ in your estimated regression function.

$\hat{\beta}_1$ is also a coefficient that is the slope. This slope can help us indicate at which direction and what rate our regression line is going.

(c). Verify that your fitted regression line goes through the point ($\bar{X}$, $\bar{Y}$). (Use R)

We plug in values for x and y in our regression line formula to test.

First I'll try x = 3.897 to test if the response is 21, which is our the first observation in our dataset

```
first_obs <- 18.97 + 2.1*3.897
first_obs
```

```
## [1] 27.1537
```

*DNE 21?

---

5. Muscle Mass Problem (Use R)

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each l0-year age group, beginning with age 40 and ending with age 79. is age, and Y is a measure of muscle mass. See the dataset "Muscle.txt". The first column is muscle mass. The second column is women's age. _

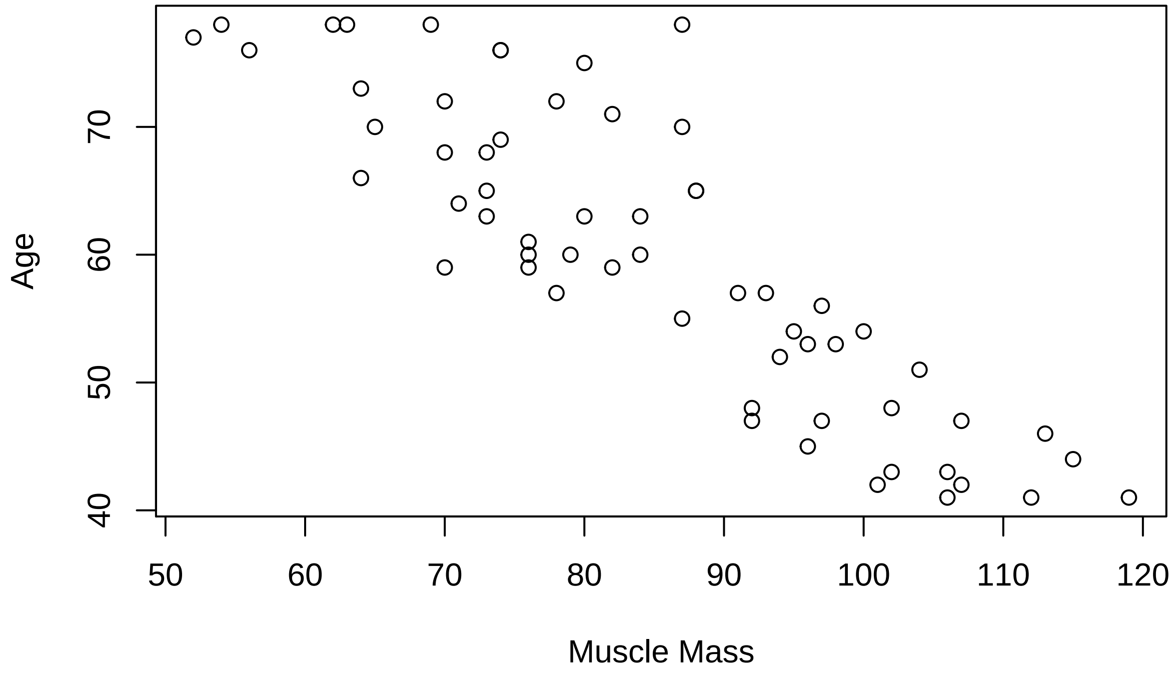
```
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 1/STAT511 As


muscle_data = read.table(file = "Muscle.txt", header = FALSE, sep = "")
head(muscle_data)
```

```
##    V1 V2
## 1 106 43
## 2 106 41
## 3  97 47
## 4 113 46
## 5  96 45
## 6 119 41
```

```
#No headers, so we add

names(muscle_data) <- c("Muscle Mass", "Age")
head(muscle_data)
```

```
##   Muscle Mass Age
## 1         106  43
## 2         106  41
## 3          97  47
## 4         113  46
## 5          96  45
## 6         119  41
```

```
#scatterplot
plot(muscle_data)
```



(a). Obtain the estimated regression equation.

```
lm(`Muscle Mass`~ Age, data = muscle_data)
```

```
##
## Call:
## lm(formula = `Muscle Mass` ~ Age, data = muscle_data)
##
## Coefficients:
## (Intercept)          Age
##      156.35        -1.19
```

```
#Muscle Mass is our response, Age is our explanatory.
# in other words `Muscle Mass`~ Age says, muscle mass is explained by Age
muscle_lm = lm(`Muscle Mass`~ Age, data = muscle_data)
summary(muscle_lm)
```

```
##
## Call:
```

```
## lm(formula = `Muscle Mass` ~ Age, data = muscle_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## Age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 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
```

(b). Interpret β_0 in your estimated regression function. Does β_0 provide any relevant information here? Explain.

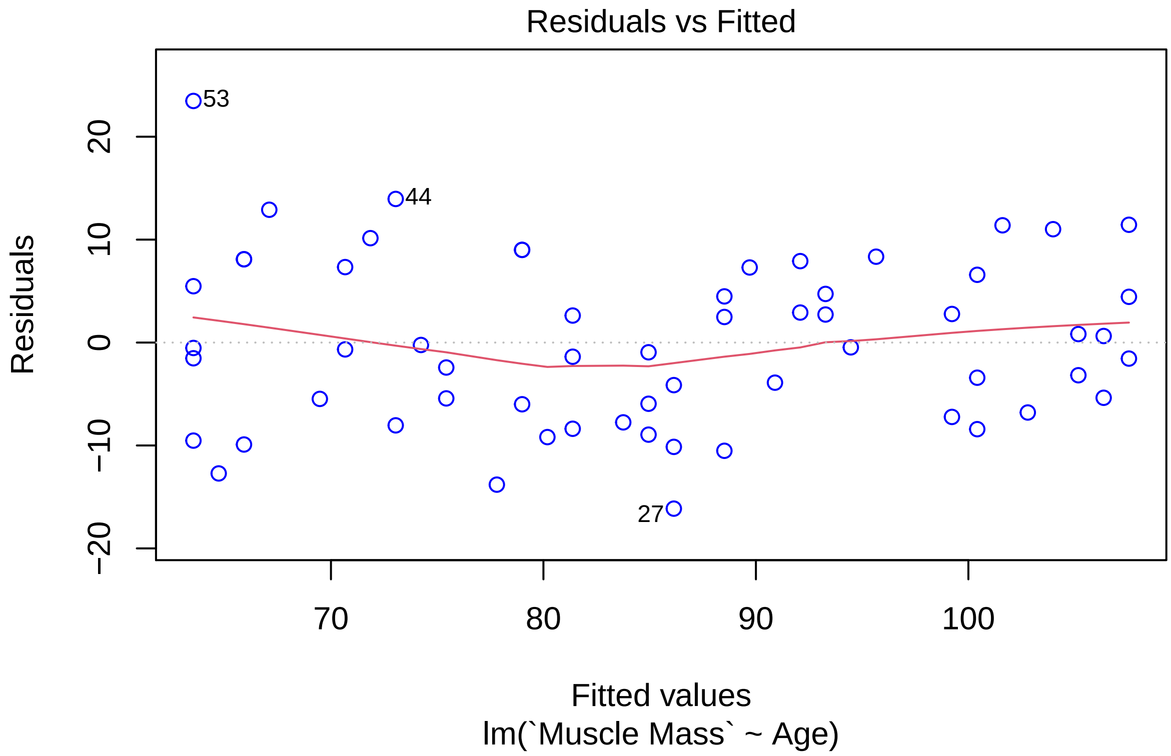
In this case $\beta_0 = 156.35$ is our y-intercept and shows us the where the value of our response variable *(muscle mass)* is located when X *(age)* is 0. At the moment the y-intercept looks to be off-chart and is unable to provide any relevant information.

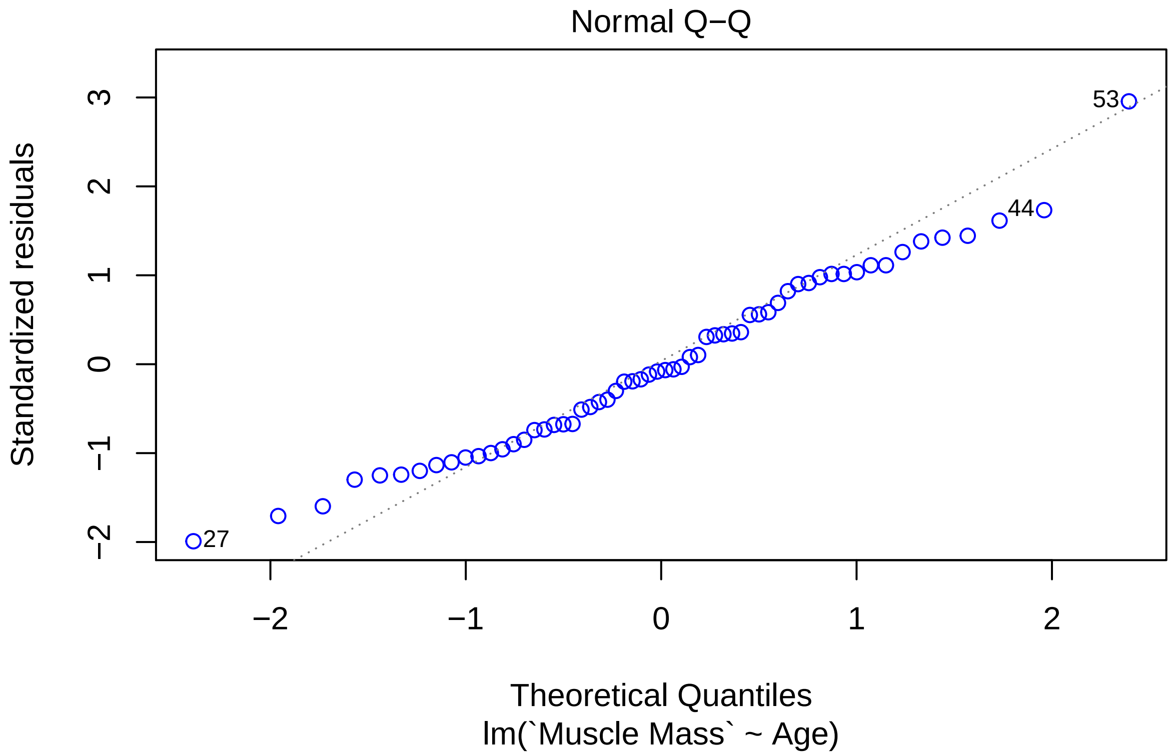
(c). Interpret $\widehat{\beta}_1$ in your estimated regression function.

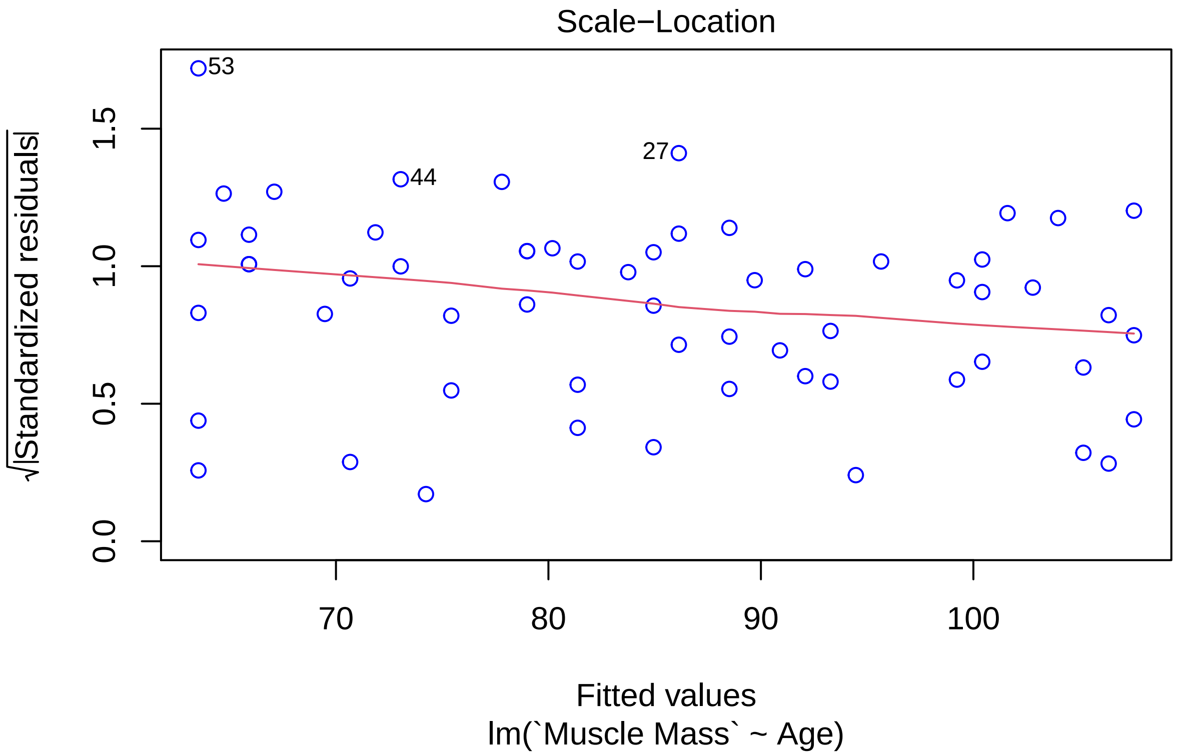
$\widehat{\beta}_1 = -1.19$ is our slope and indicates that our estimated line is moving in a downward fashion.

(d). Plot the estimated regression function and the data points. Does a linear regression function appear to give a good fit here? Does your plot support that muscle mass decreases with age?

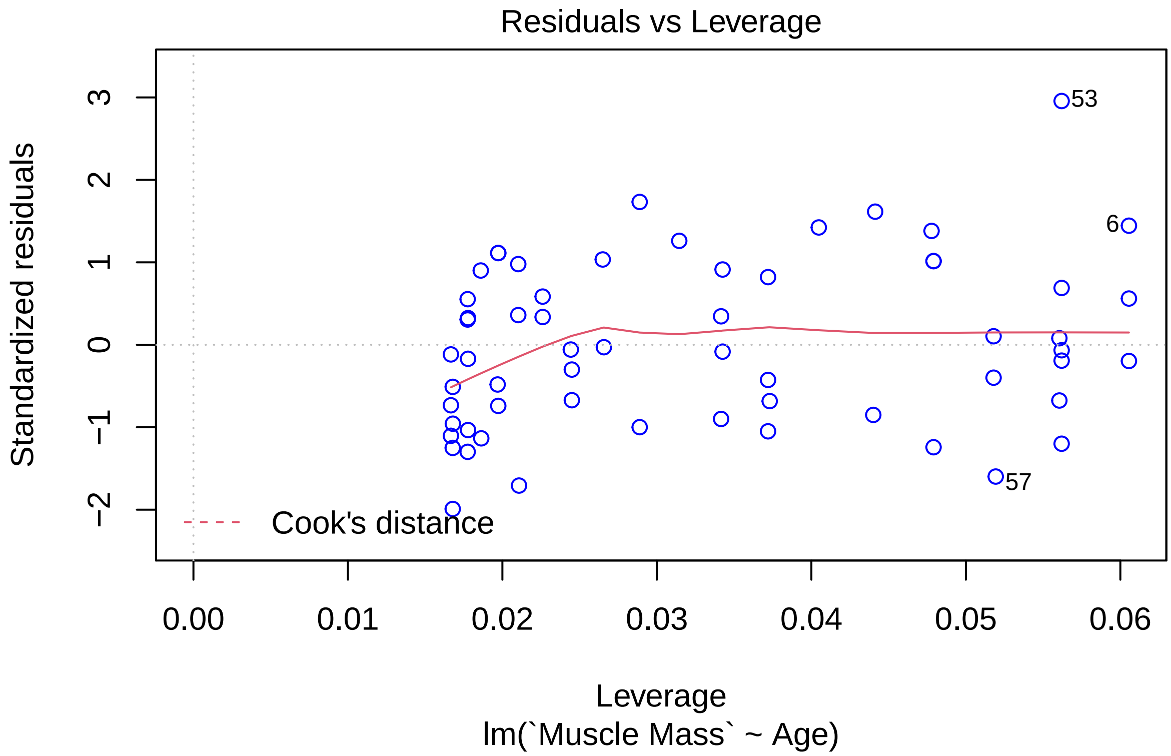
```r
plot(muscle_lm,
     col = "blue")
```


Residuals vs Fitted
Residuals
Fitted values
lm(`Muscle Mass` ~ Age)
−20
−10
0
10
20
70
80
90
100
53
44
27


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
53
44
27
−2
−1
0
1
2
Theoretical Quantiles
lm(`Muscle Mass` ~ Age)



```r
abline(muscle_lm, col = "red")
```



Yes, the estimated regression function appears to give a good fit with our data. Also, the first plot “Residuals vs Fitted” indicates that the line measuring the fit of residual is somewhat identical to the line measuring fit across our data points. Therefore we can say our regression function supports the case the muscle mass decreases with age.

(e). Obtain a point estimate of the difference in the mean muscle mass for women differing in age by one year.

Our slope $\hat{\beta}_1 = -1.19$ indicates the change in Muscle Mass when adjusting age by one year

(f). Obtain a point estimate of the mean muscle mass for women aged $= 60$ years.

$\hat{y} = 156.35 + -1.19x$

```
whenAgeIs60 <- 156.35 + -1.19*60
whenAgeIs60
```

```
## [1] 84.95
```

$\hat{y} = 84.95$ when x $= 60$

(g). Find the estimate of error variance σ^2

From our summary we are given: Residual standard error as $\sigma = 8.173$ and can conclude that:

Point estimate for error variance: $\sigma^2 = 66.79$

6. Special regression models

(a). What is the implication for the regression model $Y_i = \beta_0 + \epsilon_i$? How does it plot on a graph?

When $\beta_1 X_i = 0$ and not accounted for, this tells us the slope is 0 and there is no change in our response variable.

Furthermore, in this case our dependent variable Y is actually now an independent variable and would make our regression line a straight horizontal line.

(b). What is the implication for the regression model $Y_i = \beta_1 X_i + \epsilon_i$? How does it plot on a graph?

When $\beta_0 = 0$ *(y-intercept)* (the regression line would be would also cross the origin point (0,0).