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)* $+ \epsilon_i$ *(error)*

$y = 100 + 20X_i + \epsilon_i$

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

The probability cannot be calculated because with 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.

---

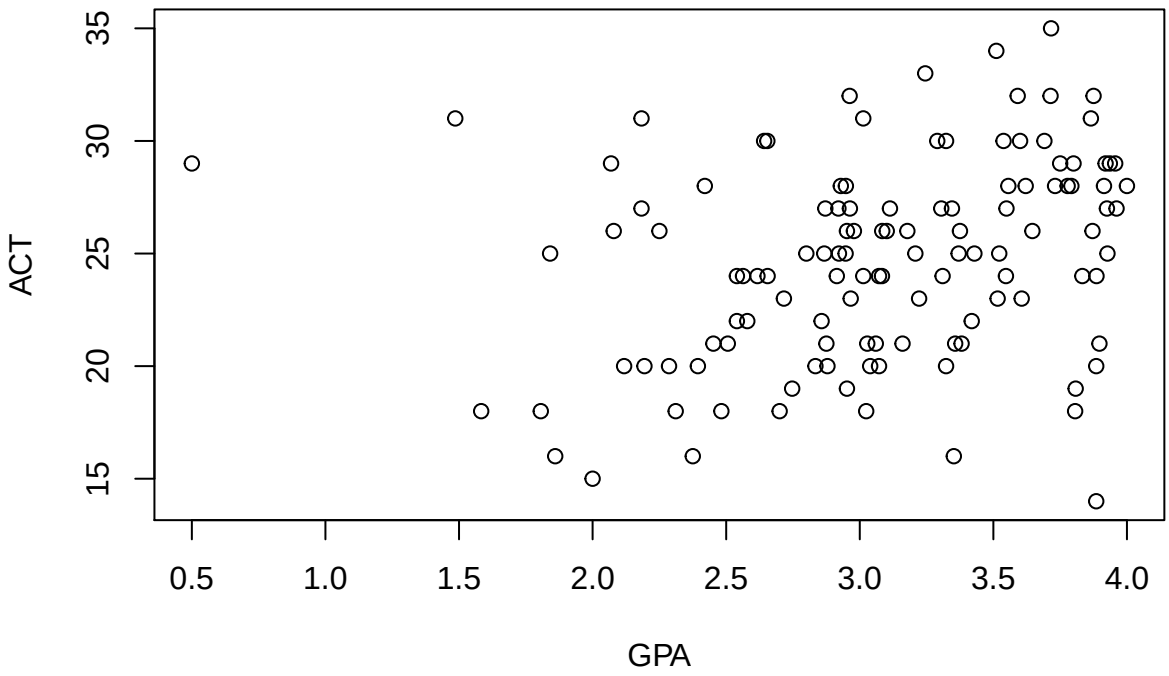
2. Grade Point Average Problem (Use R)

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 Ass


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

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

```
##     GPA ACT
## 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
ACT = gpa_data$ACT #X
GPA = gpa_data$GPA #Y


#scatterplot
plot(gpa_data)
```



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

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

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa_data)
##
## Coefficients:
## (Intercept)          ACT
##     2.11405      0.03883
```

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

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

```
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## ACT          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 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:

β_0 = **2.11405** ***(intercept)***

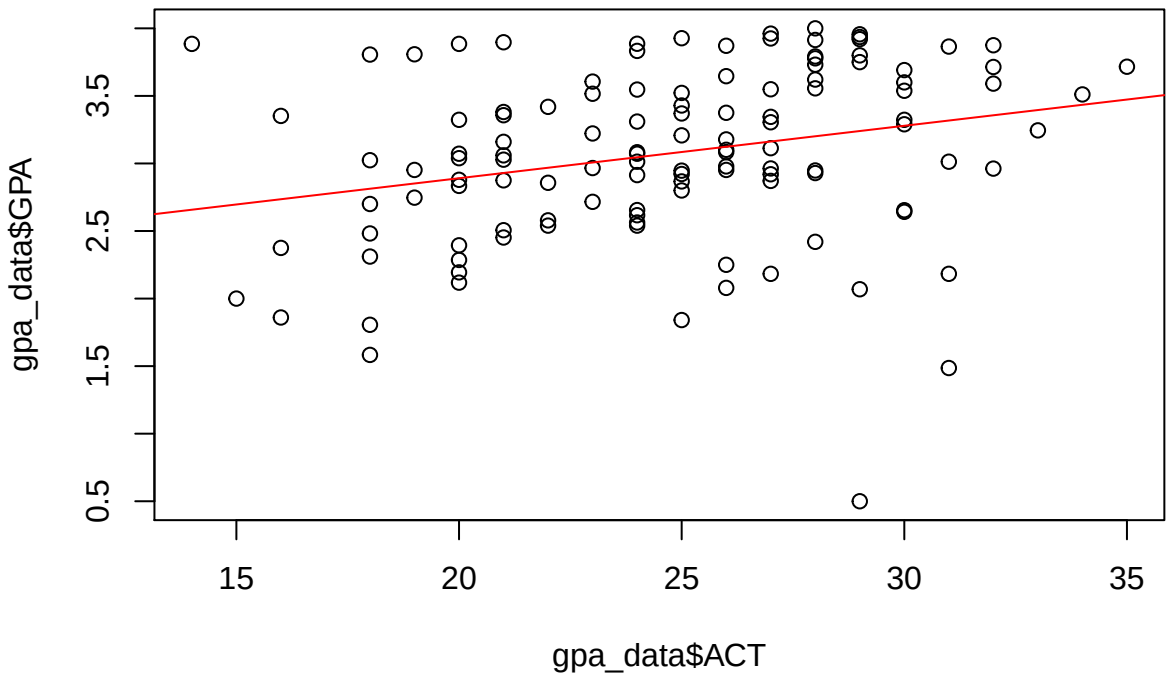
β_1 = **0.03883** ***(slope)***

and the estimated regression equation to be:

$\hat{Y} = 2.11405 + 0.03883X$

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

```r
plot(gpa_data$ACT, gpa_data$GPA)
gpa_lm = lm(GPA ~ ACT, data = gpa_data)
abline(gpa_lm, col = "red")
```



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

Using $\hat{Y} = 2.11405 + 0.03883X$ Solve for point estimate when x = 30

```
#2.11405 + 0.03883*30 = Y hat
2.11405 + 0.03883*30
```

```
## [1] 3.27895
```

$\hat{Y} = 3.27895$

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

The change in Y can be found using this formula: $\delta y = \hat{\beta}_1 * \delta x$

```
#Delta_y = beta1_hat * delta_X

#Change in y equals the slope multiplied by X in this case x = 1
DeltaY_gpa <- 0.03883*1

DeltaY_gpa
```

```
## [1] 0.03883
```

The change in mean response is 0.03883

Estimated change in $\hat{Y}$ (GPA) when ACT score increases by one point is 0.03883 .

---

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

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

```
#Finding predicted values of X using estimated regression equation
yhat <- 2.11405 + 0.03883*ACT

#Finding residual by doing observed minus predicted
e1_hat <- GPA - yhat
e1_hat
```

```
##   [1]  0.96752  1.22733  0.57671 -0.42831  0.09852  0.54722 -0.39461  0.79854
##   [9] -2.74012  0.05437  0.26403  0.25905  0.03703 -0.03297 -0.15044 -0.19946
##  [17]  0.43720 -0.30478 -0.13780 -0.77265 -0.48297  0.42752  0.52971  0.76254
##  [25]  0.35471 -0.02263 -0.78129 -0.38931  0.74737  0.13052  0.84220 -0.36033
##  [33] -0.27229  0.25137 -0.11131  0.02603  0.45152  0.01105  0.38654  0.52237
##  [41] -0.14563 -0.62495 -0.50597 -0.87363 -1.17112 -0.42897 -1.13478 -0.69650
##  [49]  0.10018  0.99301 -0.29146  0.61667  0.14254 -0.17163  0.50103  0.41205
##  [57]  0.23052 -0.69665  0.04405  0.69588 -0.16280 -0.29114  0.28520  0.59886
##  [65] -0.63695 -0.47748 -0.39097  0.35739 -1.00699  0.50886  0.14835 -0.04114
##  [73] -0.33099 -0.11299  0.67988 -0.05665  0.21486 -0.03963  0.79871  0.07673
##  [81]  0.43235  0.18135 -1.04463  0.51839  0.12320 -0.24246  0.18254  0.71588
##  [89]  0.95618 -0.42348  0.84003 -0.97946  0.34420  0.21101  0.50988  0.78703
##  [97] -0.04946 -0.05448 -0.10482 -0.50199 -1.24380 -1.22999 -0.01165  0.23439
## [105] -0.13197  0.24290 -0.28480  0.41971  0.59071 -0.21780  0.45069  0.32105
## [113] -0.49665 -0.60465 -1.83178  0.99435  0.55988  0.71271 -0.87533 -0.25329
```

```
#Summing residuals to see if adds up to 0
sum(e1_hat)
```

```
## [1] -0.00861
```

The residuals do not add up exactly to 0 but instead -0.00861

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

```
#Squaring the residual standard error from our summary to get error variance
err_var <- 0.6231^2
err_var
```

```
## [1] 0.3882536
```

σ is expressed in units of standard deviation and in this case in terms of grade point averages (GPA).

---

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.

$\hat{\beta}_0 = 3.07405$ is a coefficient, specifically, our y-intercept. It shows us where our response variable (GPA) is located when our predictor(ACT Scores) is 0.

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

$\hat{\beta}_1$ is also a coefficient which indicates 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.

```
#sample mean of x = 24.725
mean(ACT)
```

```
## [1] 24.725
```

```
#sample mean of y = 3.07405
mean(GPA)
```

```
## [1] 3.07405
```

```
#So then this regression line formula should hold true if it intersects sample mean of x and y

#plugging in sample mean into estimated regression equation

#3.07405 = 2.11405 + 0.03883 * mean(ACT)
 2.11405 + 0.03883 * mean(ACT)
```

```
## [1] 3.074122
```

```
 mean(GPA)
```

```
## [1] 3.07405
```

---

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. _

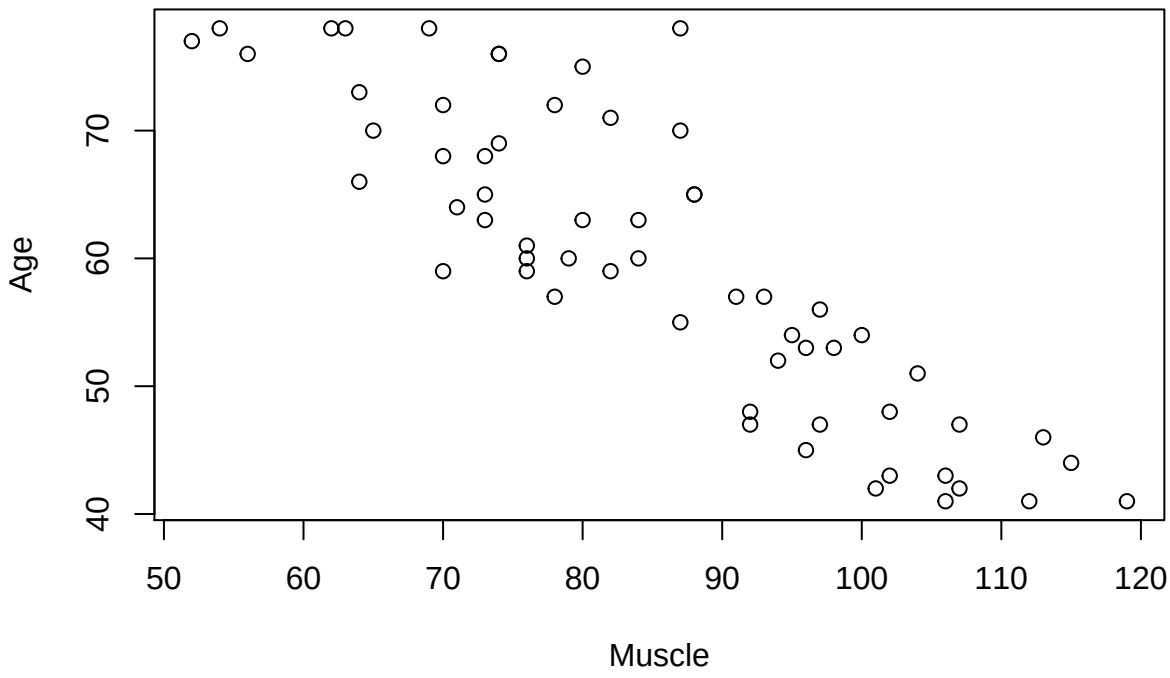
```
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 1/STAT511 As


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", "Age")
head(muscle_data)
```

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

```
#Defining dependent and independent vars
Age = muscle_data$Age #X
Muscle = muscle_data$Muscle #Y


#scatterplot
plot(muscle_data)
```



(a). Obtain the estimated regression equation.

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

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

```
#Muscle is our response, Age is our explanatory.
#in other words Muscle ~ Age says, muscle mass is explained by Age

muscle_lm = lm(Muscle ~ Age, data = muscle_data)
summary(muscle_lm)
```

```
##
## Call:
## lm(formula = Muscle ~ 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.

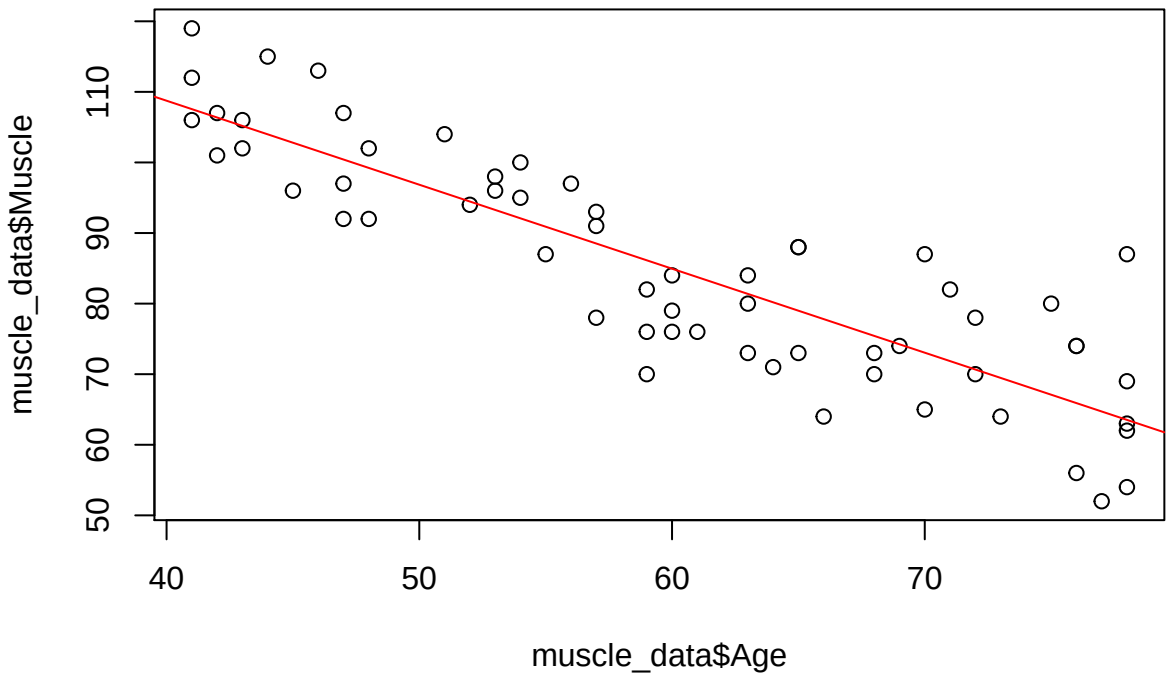
In this case $\beta_0 = 156.35$ is our y-intercept and shows us the where the value of our response variable *(muscle)* is located when X *(age)* is 0. At the moment the y-intercept looks to be off-chart and is unable to provide any relevant information.

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

$\hat{\beta}_1 = -1.19$ 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_data$Age, muscle_data$Muscle)
abline(muscle_lm, col = "red")
```



Yes, the estimated regression function appears to give a good fit with our data. By observing the line, we see that it does support the case that muscle mass decreases with age over time.

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

```
#Delta_y = beta1_hat * delta_X
#Change in y equals the slope multiplied by X in this case x = 1
DeltaY_muscle <- -1.19*1

DeltaY_muscle
```

```
## [1] -1.19
```

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
```

$\widehat{y} = 84.95$ 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 crosses the origin point (0,0).
