STAT 5385: Homework 1

Willliam Ofosu Agyapong

31/01/2022

Problem 1.5:

No, I do not agree with the student's answer because the correct form of the simple linear regression model does not involve the expectation of the response variable, Y, and it is written as:

$$Y_i = \beta_0 + \beta_1 Xi + \epsilon_i,$$

where ϵ_i denotes the random error terms with mean $= 0$ and variance $= \sigma^2$.

This should not be confused with the expectation of the response ($E\{Y_i\}$), called the estimated regression function, which comes with no addition of the error terms since $E\{\epsilon_i\} = 0$.

Problem 1.9:

Indeed, regression is a very powerful tool! And yes, we can use it to isolate fixed and variable costs, where the estimated intercept ($\hat{\beta}_0$) can be assumed to reasonably capture the fixed costs while the remaining parameters, such as $\hat{\beta}_1$, in the case of simple linear regression, can help determine the variable costs. However, I disagree with the last part of the statement since we cannot fit a linear regression model without data, irrespective of the size of the lot, whether small or large.

Problem 1.13

(a)

The data used by the seminar leader were `observational` data. This is because the explanatory variable, time spent in class preparation, was not controlled. In other words, class preparation times were not randomly assigned to the participants. Instead, the data were simply obtained from the participants' records.

(b)

I will question the validity of the conclusion reached by the seminar leader, since the analysis was solely based on observational data from which a positive linear relation between productivity level and the number of hours spent in class preparation may not imply that productivity level is a direct consequence of the time spent in class preparation. With how the study was conducted, there may be several other alternative (confounding) variables that could be responsible for the observed relationship (cause-and-effect) between participants' productivity levels and time spent in class preparation that were not accounted for.

As observed by Kutner et al. (2005), regression analysis by itself provides no information about causal patterns and must be supplemented by additional analyses to obtain insights about causal relations.

(c)

1. Participants' prior knowledge or experience in the area or topics for assessment.
2. Aptitude
3. Education level

(d)

One can actually conduct an experiment where participants will be randomly assigned to different class preparation times. For instance, the participants could be divided into groups and each group selected at random to different number of hours of class preparation. This will then results in experimental data which can provide much stronger information about cause-and-effect relationship between class preparation time and employee productivity level. As a result, the random assignment will balance out the effects of other variables, such as the effects of the variables listed in part (c), that might affect the dependent variable.

Problem 1.19: Grade Point Average.

```
# Importing the dataset
gpa_act <- read.table("../Data Sets/Chapter  1 Data Sets/CH01PR19.txt",
                      header = F,
                      col.names = c("GPA", "ACT")
                      )
# Viewing first few rows
head(gpa_act) %>%
  kable(caption = "First 6 observations from the data")
```

Table 1: First 6 observations from the data

GPA	ACT
3.897	21
3.885	14
3.778	28
2.540	22
3.028	21
3.865	31

```
# gpa_act %>% dfSummary() %>% view()
```

(a)

```
# Fit a linear regression model and display estimates
mdl_gpa_vs_act <- lm(GPA ~ ACT, data = gpa_act)

tidy(mdl_gpa_vs_act) %>%
  kable(caption = "GPA vs. ACT Parameter Estimates")
```

Table 2: GPA vs. ACT Parameter Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	2.1140	0.3209	6.588	0.0000
ACT	0.0388	0.0128	3.040	0.0029

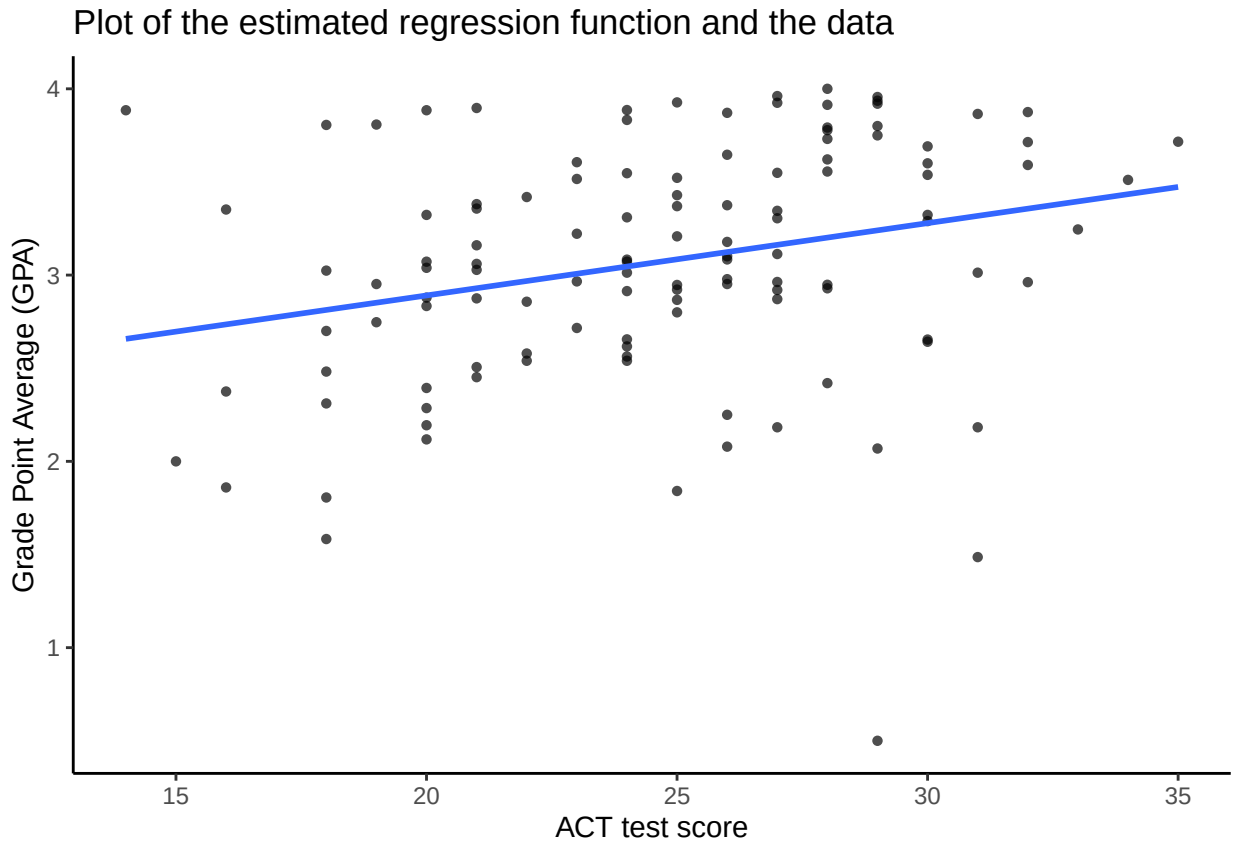
From the above table of results, the estimated regression model for a student's average (GPA) at the end of the freshman year (Y) is given by

$$\hat{Y} = 2.114 + 0.0388X$$

, where X is the ACT test score. This suggests a positive linear relationship between ACT test score and GPA.

(b)

```
# Create the plot
ggplot(gpa_act, aes(ACT, GPA)) +
  geom_point(alpha = 0.7, shape = 16) +
  labs(x="ACT test score", y="Grade Point Average (GPA)",
       title = "Plot of the estimated regression function and the data") +
  geom_smooth(formula = y~x, method = "lm", se=FALSE)
```



We can observe that the estimated regression function appears to fit the data reasonably well, since the fitted regression line appears to lie at the center of the data points and there is no obvious curvature overall in the data points. However, the presence of some extreme data points (possible outliers) may cause some issues.

(c)

The point estimate of the mean freshman GPA for students with ACT test score $X = 30$ is given by

```
# Compute the mean response for X = 30.
X <- 30
mean_gpa <- coef(mdl_gpa_vs_act)[1] + coef(mdl_gpa_vs_act)[2] * X
```

$$\bar{Y}_h = 2.114 + 0.0388 * 30 = 3.2789$$

(d)

The point estimate of the change in the mean response when the entrance test score increases by one point is $\hat{\beta}_1 = 0.0388$.

Problem 1.21: Airfreight Breakage.

```
# Importing the dataset
airfreight <- read.table("../Data Sets/Chapter  1 Data Sets/CH01PR21.txt",
                         header = F,
                         col.names = c("broken_ampules", "transfer_made")
                       )

head(airfreight) %>%
  kable(caption = "First 6 observations from the data")
```

Table 3: First 6 observations from the data

broken_ampules	transfer_made
16	1
9	0
17	2
12	0
22	3
13	1

(a)

```
# Fit a linear regression model and display estimates
mdl_airfreight <- lm(broken_ampules~transfer_made, data =  airfreight)

tidy(mdl_airfreight) %>%
  kable(caption = "Parameter Estimates")
```

Table 4: Parameter Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	10.2	0.6633	15.377	0
transfer_made	4.0	0.4690	8.528	0

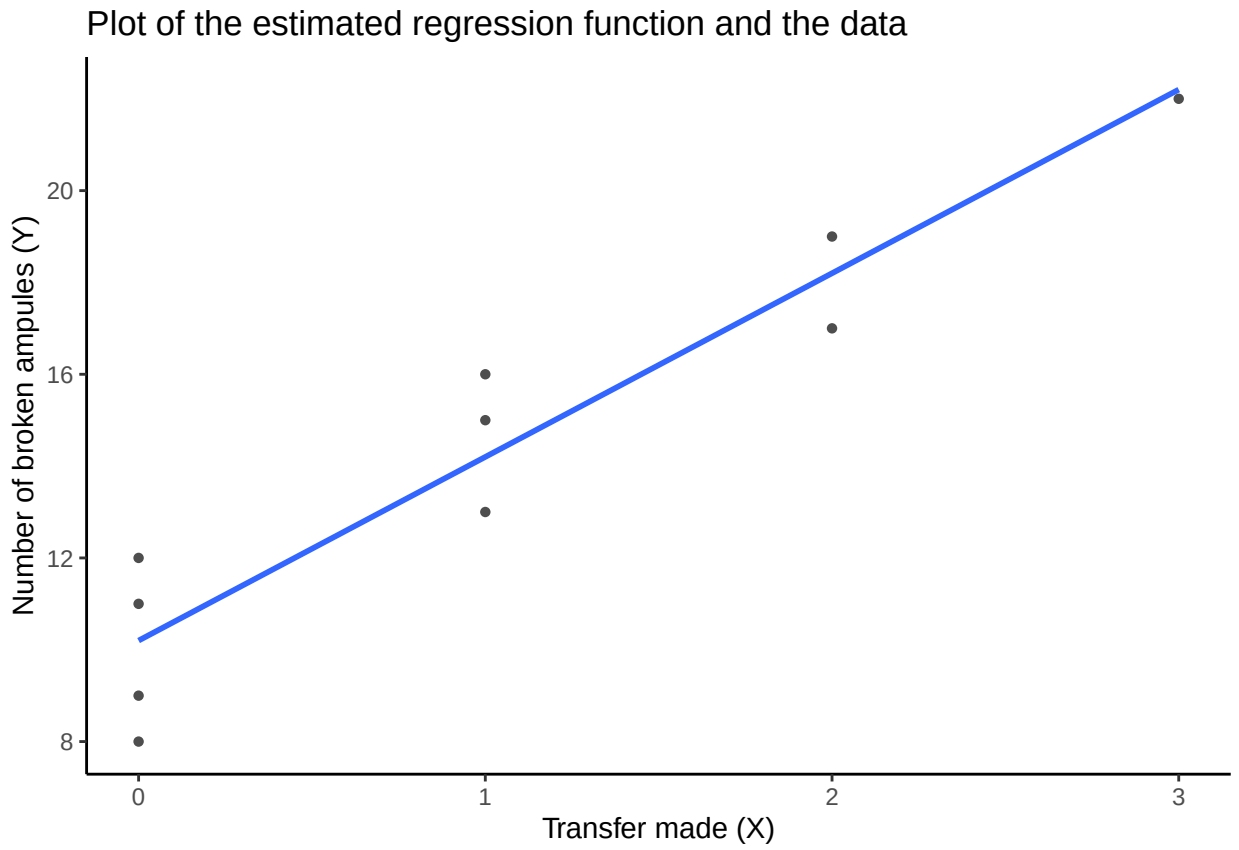
Below is the estimated regression function:

$$\hat{Y} = 10.2 + 4X,$$

where

- Y is the number of ampules found to be broken upon arrival.
- X is the number of times the carton was transferred from one aircraft to another over the shipment route. That is, there is a positive linear relationship between the transfers made and the number of ampules broken.

```
ggplot(airfreight, aes(transfer_made, broken_ampules)) +
  geom_point(alpha = 0.7, shape = 16) +
  labs( x = "Transfer made (X)", y = "Number of broken ampules (Y)",
      title = "Plot of the estimated regression function and the data") +
  theme_classic() +
  geom_smooth(formula = y~x, method = "lm", se=FALSE)
```



From the plot, a linear regression function does appear to give a good fit to the data since the fitted regression line lies exactly at the center of the data.

(b)

A point estimate of the expected number of broken ampules when $X = 1$ transfer is made can be calculated by using the estimated regression function as follows:

```
# Compute the mean response for X = 1.
X <- 1
mean_airfreight <- coef(mdl_airfreight)[1] + coef(mdl_airfreight)[2] * X
```

$$\hat{Y}_h = 10.2 + 4 * 1 = 14.2.$$

(c)

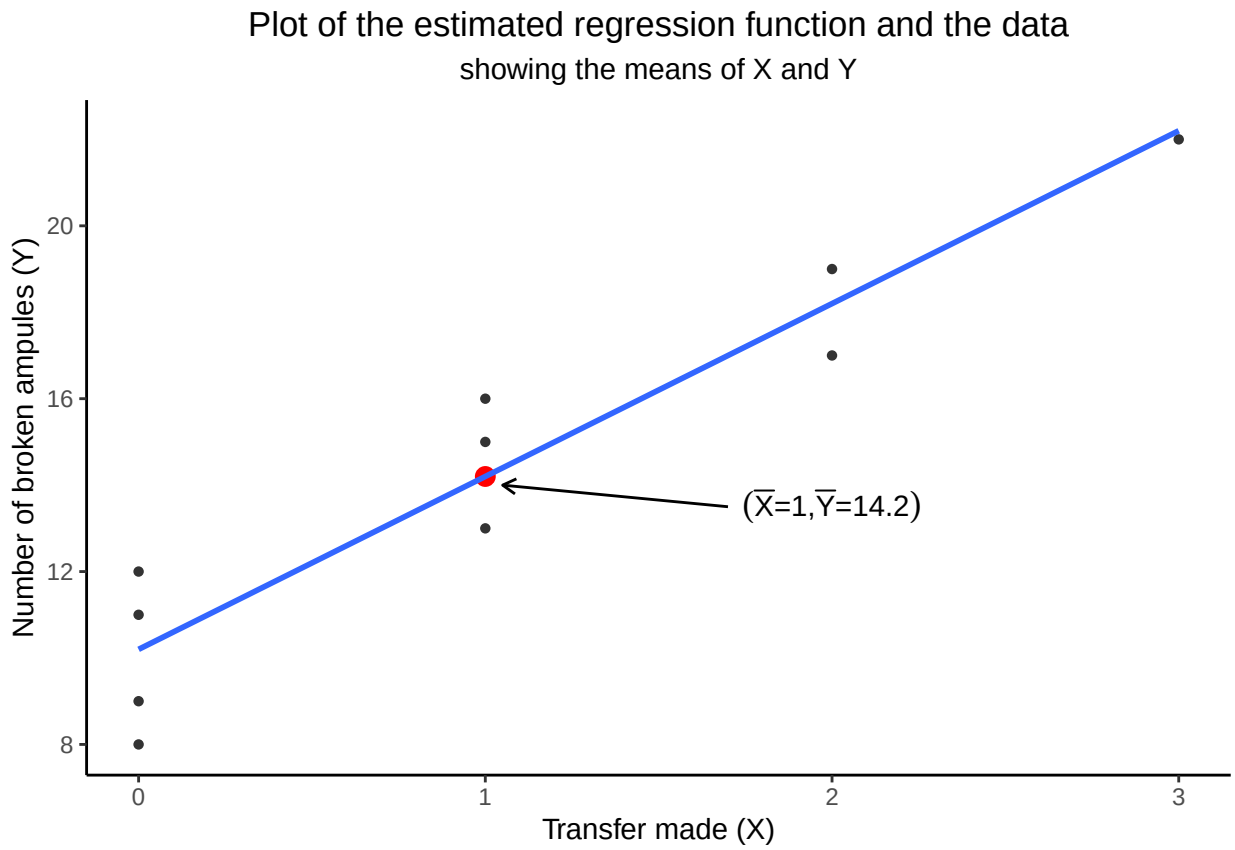
Because the increase in the number of transfer made is only 1 unit (i.e. $2 - 1 = 1$), the estimated increase in the expected number of ampules broken is equivalent to the value of $\hat{\beta}_1$. Thus, the expected number of ampules broken increases by 4 when there are 2 transfers as compared to 1 transfer.

(d)

We can verify that the fitted regression line goes through the point $(\bar{X}, \bar{Y})$ by plotting the means of the two variables on the same plot with the regression line. The point, $(\bar{X}, \bar{Y})$, is shown in red. It can be seen that the blue fitted regression line passes through the red point, hence verified.

```
# Compute the sample means
x_bar <- mean(airfreight$transfer_made)
y_bar <- mean(airfreight$broken_ampules)
```

```
# Plot the data, the means, and the fitted regression line.
ggplot(airfreight, aes(transfer_made, broken_ampules)) +
  geom_point(alpha = 0.8, shape = 16) +
  annotate("point", x = x_bar, y= y_bar, color = "red", size = 3) +
  annotate("segment", x = 1.7, xend = 1.05, y = 13.5, yend = 14,
  colour = "black", arrow = arrow(length = unit(.2,"cm"))) +
  annotate("text", x = 2, y = 13.5, label = "(bar(X)*'=1'*','*bar(Y)*'=14.2')",
  parse = TRUE) +
  labs( x = "Transfer made (X)", y = "Number of broken ampules (Y)",
      title = "Plot of the estimated regression function and the data",
      subtitle = "showing the means of X and Y") +
  theme_classic() +
  geom_smooth(formula = y~x, method = "lm", se=FALSE) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



It can also be verified by substituting the means of X and Y into the fitted regression function as follows:

$$\begin{aligned} \bar{Y} &= 10.2 + 4 * \bar{X} \\ 14.2 &= 10.2 + 4 * (1) \\ &= 14.2 \end{aligned}$$

Problem 1.24: Copier Maintenance

```
# Importing the dataset
copier <- read.table("../Data Sets/Chapter  1 Data Sets/CH01PR20.txt",
                     header = F,
```

```
                    col.names = c("service_time", "copiers_serviced")
                  )

head(copier) %>%
  kable(caption = "First 6 obervations from the data")
```

Table 5: First 6 obervations from the data

service_time	copiers_serviced
20	2
60	4
46	3
41	2
12	1
137	10

```
# Fit a linear regression model and display estimates.
mdl_copier <- lm(service_time~copiers_serviced, data =  copier)

tidy(mdl_copier) %>%
  kable(caption = "Parameter Estimates")
```

Table 6: Parameter Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	-0.5802	2.8039	-0.2069	0.8371
copiers_serviced	15.0352	0.4831	31.1233	0.0000

We obtain the estimated regression function as follows:

$$\hat{Y} = -0.5802 + 15.0352X,$$

where

- Y is the total number of minutes spent by the service person.
- X is the number of copiers serviced. There is a negative relationship between the service time in minutes and the number of the number of copiers serviced.

(a)

```
# Use the fitted model to obtain the residuals and their squared values.
resid_df <- data.frame(i = 1:nrow(copier),
                       resid = resid(mdl_copier),
                       resid_sq = resid(mdl_copier)^2
                       )
names(resid_df) <- c("$i$", "$e_i$", "$e^2_i$")
kable(resid_df, escape = F, align = "ccc",
      caption = "Residuals obtained from the fitted model",)
```

Table 7: Residuals obtained from the fitted model

i	e_i	e_i^2
1	-9.4903	90.0665

i	e_i	e_i^2
2	0.4392	0.1929
3	1.4744	2.1739
4	11.5097	132.4723
5	-2.4551	6.0275
6	-12.7723	163.1323
7	-6.5961	43.5083
8	14.4039	207.4728
9	-10.4551	109.3089
10	2.5097	6.2984
11	9.2629	85.8018
12	6.2277	38.7840
13	3.3687	11.3479
14	-8.5256	72.6856
15	12.4392	154.7328
16	-19.7018	388.1620
17	0.3334	0.1112
18	11.2982	127.6487
19	-22.7723	518.5787
20	-2.5608	6.5579
21	-8.5961	73.8927
22	-3.6666	13.4438
23	4.3334	18.7785
24	-0.5961	0.3553
25	-0.7371	0.5433
26	7.3334	53.7791
27	-11.4903	132.0279
28	-1.5961	2.5475
29	6.3334	40.1122
30	6.3687	40.5599
31	3.2982	10.8779
32	15.4039	237.2806
33	-9.4903	90.0665
34	-1.4903	2.2211
35	-11.4551	131.2191
36	-2.5608	6.5579
37	11.4039	130.0493
38	-2.7371	7.4916
39	7.3334	53.7791
40	12.5449	157.3747
41	-3.7371	13.9657
42	4.5097	20.3370
43	-2.4903	6.2018
44	1.4392	2.0712
45	2.4039	5.7788

Summing the squared residuals' column gives us the sum of the squared residuals, $\sum e_i^2 = 3416.377$

Here, the sum of the squared residuals obtained is the minimum value of the quantity Q in (1.8). Thus,

$$\sum e_i^2 = 3416.377 = \text{Min}Q$$

(b)

```
# Generate anova table to help answer question.
kable(anova(mdl_copier),
      caption = "Analysis of Variance Table (ANOVA")
```

Table 8: Analysis of Variance Table (ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
copiers_serviced	1	76960	76960.42	968.7	0
Residuals	43	3416	79.45	NA	NA

From the above table, σ^2 is estimated as MSE $= 79.45$, implying that the point estimate of σ, $\hat{\sigma} = \sqrt{79.45} = 8.91$. And σ is expressed in the same units as the response variable, hence, it is in minutes.

Problem 1.25: Airfreight Breakage (revisited)

(a)

```
# Obtain the residual for the first case
(resid_1 <- resid(mdl_airfreight)[1])
```

```
##   1
## 1.8
```

$$
\begin{aligned}
&\text{Error} = \text{Sampled value} - \text{Predicted Value} \\
&\Rightarrow e_1 = Y_1 - \hat{Y}_1 \\
&\qquad = 16 - (10.2 + 4 * 1) \\
&\qquad = 16 - 14.2 \\
&\qquad = 1.8
\end{aligned}
$$

Hence, the residual for the first case $(X = 1, Y = 16)$ is `1.8`, which can also be obtained in `R` with the `resid()` function like so: `resid(mdl_airfreight)[1]`. This residual, e_1, estimates the value of ϵ_1.

(b)

```
# Generate anova table to help answer question.
kable(anova(mdl_airfreight),
      caption = "Analysis of Variance Table (ANOVA")
```

Table 9: Analysis of Variance Table (ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
transfer_made	1	160.0	160.0	72.73	0
Residuals	8	17.6	2.2	NA	NA

From the ANOVA table, $\sum e_I^2 = 17.6$, and MSE $= \frac{160}{2.2} = 72.73$, where

MSE estimates σ^2, the variance of the error terms.

Problem 1.38: Airfreight Breakage (revisited)

From equation (1.8), the least squares criterion Q is given by

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

For convenience, we first create a function based on the above criterion.

```
# A function to compute the least square criterion for the Airfreight data
ls_criterion <- function(b0, b1)
{
  # obtain Y and X from the Airfreight data
  X <- airfreight$transfer_made
  Y <- airfreight$broken_ampules
  return(sum((Y - b0 - b1*X)^2))
}

# Use the function to compute Q
Q0 <- ls_criterion(b0 = 10.2, b1 = 4)
Q1 <- ls_criterion(b0 = 9, b1 = 3)
Q2 <- ls_criterion(b0 = 11, b1 = 5)

# Make a table of the results for easy comparison.
Q_tbl <- data.frame(beta0 = c(10.20, 9, 11),
                    beta1 = c(4, 3, 5),
                    Q = c(Q0, Q1, Q2)
                    )
names(Q_tbl) <- c("$b_0$", "$b_1$", "Q criterion")
kable(Q_tbl, escape = F, align = "ccc")
```

b_0	b_1	Q criterion
10.2	4	17.6
9.0	3	76.0
11.0	5	60.0

Clearly, we notice that the criterion Q is larger for the estimates ($b_0 = 9$, $b_1 = 11$) and ($b_0 = 11$, $b_1 = 5$) than for the least squares estimates ($b_0 = 10.2$, $b_1 = 4$). This confirms that the method of least squares yields estimates of β_0 and β_1 that minimize the criterion Q for the given sample observations.

Reference

- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models.