STAT 5385: Homework 2

Willliam Ofosu Agyapong

13/02/2022

Problem 2.3:

Superficially, the student's interpretation appears to be consistent with the estimated regression equation since the negative slope coefficient indicates an inverse relationship between advertising and sales. However, the problem here is that the student neglected the implication of the reported two-sided p-value associated with the estimated slope. The two-sided p-value for the estimated slope is actually testing the hypotheses:

$$H_0 : \beta_1 = 0 \text{ (no linear relationship)}$$
$$H_a : \beta_1 \neq 0 \text{ (linear relationship}$$

And since the p-value of **0.91** is reasonably large we conclude H_0, that there is not enough evidence of a linear relationship between advertising expenditure and sales. Therefore, the student's observation is not correct.

Problem 2.5: Copier maintenance (Problem 1.20)

```
# Importing the dataset
copier <- read.table("../Data Sets/Chapter  1 Data Sets/CH01PR20.txt",
                     header = F,
                     col.names = c("service_time", "copiers_serviced")
                    )
# Displaying first few observations
# head(copier) %>%
#   kable(caption = "First 6 obervations from the data")
# dim(copier)
```

(a)

```
# Fit a linear regression model and extracts CI for the slope.
mdl_copier <- lm(service_time~copiers_serviced, data =  copier)

mdl_copier %>%
  tidy(conf.int = TRUE, conf.level = 0.90) %>%
  filter(term == "copiers_serviced") %>%
  mutate(term = "$\\beta_1$") %>%
  kable(caption = "Parameter Estimates with 90% Confidence Interval")
```

Table 1: Parameter Estimates with 90% Confidence Interval

term	estimate	std.error	statistic	p.value	conf.low	conf.high
β_1	15.04	0.4831	31.12	0	14.22	15.85

The 90% confidence interval for β_1 is: $14.223 \leq \beta_1 \leq 15.847$.

Thus, with confidence coefficient 0.90, we estimate that the mean service time increases by somewhere between **14.223** and **15.847** hours for each additional unit in the number of copiers serviced.

(b)

Let β_1 represents the true change in service time for every one unit increase in the number of copiers serviced. We are interested in testing whether or not there is a linear association between service time and the number of copiers serviced. This results in the following two alternative hypotheses:

$$H_0 : \beta_1 = 0 \text{ (no linear relationship)}$$
$$H_a : \beta_1 \neq 0 \text{ (linear relationship)}$$

-Significance level, $\alpha = 0.10$.

```
# obtain test statistic and critical value
mdl_copier %>%
  tidy(conf.int = TRUE, conf.level = 0.90) %>%
  filter(term == "copiers_serviced") %>%
  mutate(t = qt(1 - 0.1/2, nrow(copier) - 2), .after = "statistic") %>%
  rename(t_statistic = statistic) %>%
  mutate(term = "$\\beta_1$") %>%
  kable(align = "c", caption = "Testing the significance of the slope")
```

Table 2: Testing the significance of the slope

term	estimate	std.error	t_statistic	t	p.value	conf.low	conf.high
β_1	15.04	0.4831	31.12	1.681	0	14.22	15.85

The decision rule is as follows:

$$\text{If } |\text{t statistic}| \leq t = 1.681, \text{ conclude } H_0, \text{ otherwise conclude } H_a.$$

Therefore, since `t_statistic = 31.12 > 1.681` we conclude H_a, that, at 10% significance level, there is sufficient evidence of a linear relationship between service time and the number of copiers serviced.

- From the table above, the two-sided p-value is zero (**0**), which leads us to the same conclusion as before, since it is less than $\alpha = 0.10$.

(c)

Yes, the results in parts (a) and (b) are consistent. This is because we are able to reach the same conclusion in part (a) by using the confidence intervals constructed in part (b), since this interval does not include 0.

(d)

The two new alternative hypotheses are to be tested are:

$$H_0 : \beta_1 \leq 14$$
$$H_a : \beta_1 > 14$$

```r
# obtain test statistic and critical value
mdl_copier %>%
  tidy(conf.level = 0.95) %>%
  filter(term == "copiers_serviced") %>%
  mutate(statistic = ((estimate - 14) / std.error),
    t = qt(1 - 0.05, df = (nrow(copier) - 2)),
    p.value = pt(statistic, df = (nrow(copier) - 2), lower.tail = F),
    .after = "statistic") %>%
  mutate(term = "$\\beta_1$") %>%
  kable(align = "c", caption = "Testing the manufacturing standard")
```

Table 3: Testing the manufacturing standard

term	estimate	std.error	statistic	t	p.value
β_1	15.04	0.4831	2.143	1.681	0.0189

The decision rule is as follows:

$$\text{If t statistic} \leq t = 1.681, \text{ conclude } H_0, \text{ otherwise conclude } H_a.$$

Therefore, since `t_statistic = 2.143 > 1.681` we conclude H_a, that, at 5% significance level, $\beta_1 > 14$.

- From the table above, the one-sided p-value is **0.0189**, which leads us to the same conclusion as before, since it is less than $\alpha = 0.05$.

(e)

```r
# obtain the estimate of the intercept
mdl_copier %>%
  tidy() %>%
  filter(term == "(Intercept)") %>%
  mutate(term = "$\\beta_0$") %>%
  kable(align = "c", caption = "Estimate of the intercept")
```

Table 4: Estimate of the intercept

term	estimate	std.error	statistic	p.value
β_0	-0.5802	2.804	-0.2069	0.8371

No, because we are dealing with a dependent variable which measures time and cannot be negative but $\beta_0 = -0.5802$ means the expected service time is negative when no copiers are serviced.

Problem 2.6: Airfreight breakage (Problem 1.21)

```r
# 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")
```

(a)

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

mdl_airfreight %>%
  tidy(conf.int = TRUE, conf.level = 0.95) %>%
  mutate(t = qt(1-0.05/2, nrow(airfreight)-2), .after="statistic") %>%
  mutate(term = c("$\\beta_0$", "$\\beta_1$")) %>%
  kable(caption = "Parameter Estimates with 95% Confidence Interval")
```

Table 5: Parameter Estimates with 95% Confidence Interval

term	estimate	std.error	statistic	t	p.value	conf.low	conf.high
β_0	10.2	0.6633	15.377	2.306	0	8.670	11.730
β_1	4.0	0.4690	8.528	2.306	0	2.918	5.082

The 95% confidence interval for β_1 is: $2.918 \leq \beta_1 \leq 5.082$. This means, with 95% confidence level, we estimate that the mean number of broken ampules increases by somewhere between **2.918** and **5.082** for each additional unit in the number of transfers made.

(b)

We wish to test whether or not there is a linear association between number of times a carton is transferred (X) and number of broken ampules (Y). Below are the two alternative hypotheses:

$$H_0 : \beta_1 = 0 \text{ (no linear relationship)}$$
$$H_a : \beta_1 \neq 0 \text{ (linear relationship)}$$

, where β_1 is the true change in the number of broken ampules for every one unit increase in transfers made.

- Signification level, $\alpha = 0.05$.

```
# obtain test statistic and critical value
mdl_airfreight %>%
  tidy(conf.level = 0.95) %>%
  filter(term == "transfer_made") %>%
  mutate(t = qt(1 - 0.05/2, nrow(airfreight) - 2), .after = "statistic") %>%
  mutate(term = "$\\beta_1$") %>%
  kable(align = "c", caption = "Testing the significance of the slope")
```

Table 6: Testing the significance of the slope

term	estimate	std.error	statistic	t	p.value
β_1	4	0.469	8.528	2.306	0

The decision rule is as follows:

$$\text{If t statistic} \leq t = 2.306, \text{ conclude } H_0, \text{ otherwise conclude } H_a.$$

Therefore, since `t statistic = 8.528 > 2.306` we conclude H_a, that, at 5% significance level, $\beta_1 \neq 0$, or there is enough evidence that a linear association exists between number of times a carton is transferred (X) and number of broken ampules (Y).

- From the above table, the two-sided p-value is approximately **0**, which leads us to the same conclusion as before, since it is less than $\alpha = 0.05$.

(c)

From Table 5, a 95% confidence interval for β_0 is given by $8.670 \leq \beta_0 \leq 11.730$, which means we can be 95% confident that the mean number of ampules broken when no transfers of the shipment are made is estimated to lie somewhere between **8.670** and **11.730**.

(d)

According to the consultant's suggestion, the following two alternative hypotheses can be formulated:

$$H_0 : \beta_0 \leq 9.0$$
$$H_a : \beta_0 > 9.0$$

With prespecified significance level, $\alpha = 0.025$.

```
# obtain test statistic and critical value
mdl_airfreight %>%
  tidy(conf.level = 0.975) %>%
  filter(term == "(Intercept)") %>%
  mutate(statistic = ((estimate - 9) / std.error),
    t = qt(1 - 0.025, df = (nrow(airfreight) - 2)),
    p.value = pt(statistic, df = (nrow(airfreight) - 2), lower.tail = F),
    .after = "statistic") %>%
  mutate(term = "$\\beta_0$") %>%
  kable(align = "c", caption = "Testing the consultant's claim about the intercept")
```

Table 7: Testing the consultant's claim about the intercept

term	estimate	std.error	statistic	t	p.value
β_0	10.2	0.6633	1.809	2.306	0.054

The decision rule is as follows:

$$\text{If t statistic} \leq t = 2.306, \text{ conclude } H_0, \text{ otherwise conclude } H_a.$$

Therefore, since `t statistic = 1.809 < 2.306` we conclude H_0, that, at 2.5% significance level, $\beta_0 \leq 9$, or the mean number of broken ampules should not exceed 9.0 when no transfers are made.

- From the table above, the one-sided p-value is **0.054**, which leads us to the same conclusion as before, since it is greater than $\alpha = 0.025$.

(e)

- In part (b), we tested $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Given that actual $\beta_1 = 2.0$, $\sigma\{b_1\} = 0.50$, and $\alpha = 0.05$. The associated *noncentrality measure* is computed as

$$\delta = \frac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}} = \frac{|2 - 0|}{0.5} = \frac{2}{0.5} = 4.$$

From **Table B.5** in the textbook (Kutner, et al., 2004), for $\alpha = 0.05$ and 8 degrees of freedom, the corresponding power is **0.94**. Thus, if $\beta_1 = 2.0$, the probability would be about 94% that led to concluding H_a.

Next, we attempt to achieve the same results in the `R Statistical Software` with the following codes:

```
n <- nrow(airfreight)
sd <- 0.5
ncp <- (2-0)/sd
t <- qt(0.975, df = n-2)
(power <- 1 - ((pt(t,df=n-2,ncp=ncp)-pt(-t,df=n-2,ncp=ncp))))
```

```
## [1] 0.9367
```

From the above output, the power of the test is given as 0.9367, which coincides with the result of our computation by hand.

- Again, in part (d), we tested $H_0 : \beta_0 \leq 9$ versus $H_a : \beta_0 > 9$. Given that the actual $\beta_0 = 11$, $\sigma\{b_0\} = 0.75$, and $\alpha = 0.025$. The associated *noncentrality measure* is computed as

$$\delta = \frac{|\beta_0 - \beta_{00}|}{\sigma\{b_0\}} = \frac{|11 - 9|}{0.75} = \frac{2}{0.75} = 2.667.$$

The power table referenced above does not have direct values for $\alpha = 0.025$. But one-half of 0.05 is equal to 0.025, so we can read from values under $\alpha = 0.05$ for our one-sided test as suggested by (Kutner, et al., 2004). Therefore, for $\alpha = 0.05$ and 8 degrees of freedom, we interpolate linearly between $\delta = 2$ and $\delta = 3$, since **2.67** lies between 2 and 3. Thus, the power of the test if $\beta_0 = 11$ is given by:

$$\text{Power} = 0.42 + \frac{2.67 - 2}{3 - 2}(0.75 - 0.42) = 0.6411 \approx 0.64$$

```
n <- nrow(airfreight)
sd <- 0.75
ncp <- (11-9)/sd
t <- qt(0.975, df = n-2, lower.tail = T)
(power <- pt(t,df=n-2,ncp=ncp, lower.tail = F))
```

```
## [1] 0.648
```

According to the above output, the power associated with the test in part (d) is **0.648**, which is close to our result by linear interpolation.

Problem 2.14: Copier maintenance (Problem 1.20)

(a)

```
# Create a generic function to obtain interval estimates for a given linear model that can be reused.
# mdl_object: this is the fitted model object of class "lm".
# xh: this is the level of the predictor for which we wish to estimate the response.
# m is number of new observations. It defaults to 1 except for the prediction of m new obs.
get_interval <- function(mdl_object, xh, conf.level= 0.9, m = 1,
                         type = c("confidence", "prediction", "band"))
{
  if(!inherits(mdl_object,"lm")) stop('model object must be of class "lm"')

  x <- mdl_object$model[, 2] # get the predictor vector
  type   <- match.arg(type) # ensure valid entries
  c      <- ifelse(type == "prediction", 1, 0) # used to obtain appropriate standard error.
  newdf <- data.frame(x=xh)  # new data frame for predicted value
  colnames(newdf) <- colnames(mdl_object$model)[2] # ensure name matches between predictors
  yhat_h <- predict(mdl_object, newdf) # compute the fitted value
```

```
  MSE <- anova(mdl_object)["Residuals", 3]  # extract MSE

  #critical value
  n <- length(x)
  df <- n - 2 # denominator degrees of freedom
  f_critical <- qf(conf.level, 2, df, lower.tail=T)
  if (type == "band") {
    critical_val <- sqrt( 2 * qf(conf.level, 2, df, lower.tail=T) )
  } else {
    critical_val <- qt((1 - (1-conf.level)/2), df, lower.tail=T)
  }

  # compute the standard error
  s <- sqrt(MSE * (c/m + 1/n + (xh-mean(x))^2/(sum((x-mean(x))^2))))

  MoE <- critical_val * s # compute margin of error

  # compute the limits of the interval:
  lwr <- yhat_h - MoE
  upr <- yhat_h + MoE

  # return outputs
  return(data.frame(fitted = yhat_h, std.error=s, critical.value=critical_val, lwr = lwr, upr = upr))
}
```

```
# construct 90% CI for the mean response
kable(get_interval(mdl_copier, xh = 6, type = "confidence"), align = "c",
      caption = "90% confidence interval for the mean service time on calls in which 6 copiers are service
```

Table 8: 90% confidence interval for the mean service time on calls in which 6 copiers are serviced.

fitted	std.error	critical.value	lwr	upr
89.63	1.396	1.681	87.28	91.98

The 90% confidence interval is: $87.28 \leq E\{Y_h\} \leq 91.98$. Therefore, we conclude with 90% confidence that the mean service time on the next call in which 6 copiers are serviced is somewhere between **87.28** and **91.98** hours.

(b)

```
# construct 90% PI for the mean response
kable(get_interval(mdl_copier, xh = 6, type = "prediction"), align = "c",
      caption = "90% prediction interval for the service time on the next call in which 6 copiers are serv
```

Table 9: 90% prediction interval for the service time on the next call in which 6 copiers are serviced.

fitted	std.error	critical.value	lwr	upr
89.63	9.022	1.681	74.46	104.8

- The 90% prediction interval is: $74.46 \leq Y_{h(new)} \leq 104.8$.
- Thus, we predict with 90 confidence that the service time for the next call in which six copiers are serviced will be somewhere between **74.46** and **104.8** minutes.

- Yes, the prediction interval is wider than the corresponding confidence interval constructed in part (a). And yes this should be the case since the prediction interval has greater variability because it combines the variability for the confidence interval for the mean response (Y_h) to the variation within the probability distribution of ($\hat{Y}$).

(c)

We have to divide the limits of 90% CI in part (a) by 6. Thus, the appropriate 90% confidence interval for the mean service time per copier is:

$$(\frac{87.82}{6} \leq \text{mean service time per copier} \leq \frac{91.98}{6}) = (14.55 \leq \text{mean service time per copier} \leq 15.33)$$

This signifies that, at a 90% level of confidence, the true mean service time per copier lies somewhere between **14.55** and **15.33**.

(d)

```
# use above function to compute the confidence bands
kable(get_interval(mdl_copier, xh=6, type = "band"),
      caption = "Confidence band for the regression line ")
```

Table 10: Confidence band for the regression line

fitted	std.error	critical.value	lwr	upr
89.63	1.396	2.205	86.55	92.71

Yes, the confidence band is slightly wider than the confidence interval found in part (a). Yes, since the confidence band has a larger margin of error for the same estimated mean value of **89.63**.

Problem 2.15

(a)

For $X_h = 2$

```
# construct 99% CI for the mean response for x = 2
kable(get_interval(mdl_airfreight, xh=2, conf.level = 0.99, type = "confidence"), align = "c",
      caption = "99% confidence interval for the mean breakage.")
```

Table 11: 99% confidence interval for the mean breakage.

fitted	std.error	critical.value	lwr	upr
18.2	0.6633	3.355	15.97	20.43

The estimated mean breakage for 2 transfers ($X = 2$) is **18.2**, with confidence interval, $15.97 \leq EY_h \leq 20.43$

We conclude with confidence coefficient 0.99 that the mean breakage for 2 transfers is estimated somewhere between **15.97** and **20.43**.

For $X_h = 4$

```
# construct 90% CI for the mean response for x = 4
kable(get_interval(mdl_airfreight, xh = 4, conf.level = 0.99, type = "confidence"), align = "c",
      caption = "99% confidence interval for the mean breakage.")
```

Table 12: 99% confidence interval for the mean breakage.

fitted	std.error	critical.value	lwr	upr
26.2	1.483	3.355	21.22	31.18

The estimated mean breakage for 4 transfers ($X = 4$) is **26.2**, with confidence interval, $21.22 \leq EY_h \leq 31.18$

We conclude with confidence coefficient 0.99 that the mean breakage for 2 transfers is estimated somewhere between **21.22** and **31.18**.

(b)

```
# 99% PI for the mean response when x = 2
kable(get_interval(mdl_airfreight, xh = 2, conf.level = 0.99, type = "prediction"), align = "c",
      caption = "99% prediction interval for the number of broken ampules with 2 transfers")
```

Table 13: 99% prediction interval for the number of broken ampules with 2 transfers

fitted	std.error	critical.value	lwr	upr
18.2	1.625	3.355	12.75	23.65

The 99% prediction interval is: $12.75 \leq Y_{h(new)} \leq 23.65$

Therefore, with confidence coefficient 0.99, we predict that the number of broken ampules for the next 2 shipment transfers will be somewhere between **12.75** and **23.65**.

(c)

```
kable(get_interval(mdl_airfreight, xh = 2, m=3, conf.level = 0.99, type = "prediction"), align = "c",
      caption = "99% prediction interval for 3 independent shipments entailing 2 transfers.")
```

Table 14: 99% prediction interval for 3 independent shipments entailing 2 transfers.

fitted	std.error	critical.value	lwr	upr
18.2	1.083	3.355	14.57	21.83

The 99% prediction interval for the 3 shipments is: $14.57 \leq Y_{h(new)} \leq 21.83$

Therefore, it can be predicted with 99 confidence that between **14.57** and **21.83** ampules will be broken in the three shipments entailing two transfers.

```
# PI for total number of broken ampules
intvals <- get_interval(mdl_airfreight, xh = 2, m=3, conf.level = 0.99, type = "prediction")
(total.lwr <- 3 * intvals$lwr)
```

```
## [1] 43.7
```

```
(total.upr <- 3 * intvals$upr)
```

```
## [1] 65.5
```

The 99% prediction interval for the total number of broken ampules is:

$$43.7 \leq \text{total number of broken ampules} \leq 65.5.$$

Thus, it can be predicted with 99 confidence that between **43.7** and **65.5** total ampules will be broken in the three shipments entailing two transfers.

(d)

For $X_h = 2$

```
# use the function created in problem 2.14 (d) to compute the confidence bands
kable(get_interval(mdl_airfreight, xh = 2, conf.level = 0.99, type = "band"), align = "c",
      caption = "99% confidence band for the mean breakage.")
```

Table 15: 99% confidence band for the mean breakage.

fitted	std.error	critical.value	lwr	upr
18.2	0.6633	4.159	15.44	20.96

From the above output, the boundary values of the confidence band for the regression line at $X_h = 2$ is

$$15.44 \leq (\beta_0 + \beta_1 X_h) \leq 20.96$$

For $X_h = 4$

```
# use the function created in problem 2.14 (d) to compute the confidence bands
kable(get_interval(mdl_airfreight, xh = 4, conf.level = 0.99, type = "band"), align = "c",
      caption = "99% confidence band for the mean breakage.")
```

Table 16: 99% confidence band for the mean breakage.

fitted	std.error	critical.value	lwr	upr
26.2	1.483	4.159	20.03	32.37

- Similarly, for $X_h = 4$ the boundary values of the confidence band for the regression line is

$$20.03 \leq (\beta_0 + \beta_1 X_h) \leq 32.37$$

- Yes, the confidence bands are wider at the two points, $X_h = 2$, $X_h = 4$, than the corresponding confidence intervals in part (a). And yes, this is expected because confidence bands are constructed for the entire regression and not just the mean response at a certain level of the predictor.

Problem 2.24

(a): ANOVA Tables

```
# Create a generic function for setting up the 2 types of ANOVA tables for SLR models.
# mdl_object: this is the fitted model object of class "lm".
```

```r
# type: a character string, either 'basic' or 'modified'.
my_anova <- function(mdl_object, type = "basic")
{
  if(!inherits(mdl_object, "lm")) stop('model object must be of class "lm"')

  # get the underlying data
  x <- mdl_object$model[, 2]
  y <- mdl_object$model[, 1]
  yhat <- fitted(mdl_object)
  n <- length(y)

  # compute sums of squares
  SSR <- sum((yhat - mean(y))^2)
  SSE <- sum((y - yhat)^2)
  SSTO <- sum((y - mean(y))^2)

  # compute degrees of freedom
  df1 <- 1
  df2 <- n - 2
  df3 <- n - 1

  # Compute mean squares
  MSR <- SSR/df1
  MSE <- SSE/df2

  # Make the ANOVA table
  if(type == 'basic') {
      anova_tbl <- data.frame(
        `Source of variation` = c("Regression", "Error", "Total"),
        SS = c(SSR, SSE, SSTO),
        df = c(df1, df2, df3),
        MS = c(MSR, MSE, NA),
        check.names = F
    )
  } else if(type == 'modified') {
      # compute additional sums of squares
      SS_correct <- n * mean(y)^2
      SSTOU <- sum((y^2))

      anova_tbl <- data.frame(
      `Source of variation` = c("Regression", "Error", "Total",
                                "Correction for mean", "Total, uncorrected"),
      SS = c(SSR, SSE, SSTO, SS_correct, SSTOU),
      df = c(df1, df2, df3, 1, n),
      MS = c(MSR, MSE, NA, NA, NA),
      check.names = F
  )
  } else stop("Invalid ANOVA type value provided!")

  return(anova_tbl)
}
```

```r
# Set up the basic ANOVA table using our self-defined function.
my_anova(mdl_copier, type = 'basic') %>%
  kable(caption = "Basic ANOVA Table for Simple Linear Regression", align = "lccc")
```

Table 17: Basic ANOVA Table for Simple Linear Regression

Source of variation	SS	df	MS
Regression	76960	1	76960.42
Error	3416	43	79.45
Total	80377	44	NA

```
# Set up the modified ANOVA table using our self-defined function.
my_anova(mdl_copier, type = 'modified') %>%
  kable(caption = "Modified ANOVA Table for Simple Linear Regression", align = "lccc")
```

Table 18: Modified ANOVA Table for Simple Linear Regression

Source of variation	SS	df	MS
Regression	76960	1	76960.42
Error	3416	43	79.45
Total	80377	44	NA
Correction for mean	261747	1	NA
Total, uncorrected	342124	45	NA

The modified ANOVA table has one additional element of decomposition, called `correction sum of square for the mean`, and `total uncorrected sum of square`.

(b): F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

```
# compute necessary measures for the test.
anova_tbl <- my_anova(mdl_copier, type = 'basic')
(F_stat <-  anova_tbl$MS[1] / anova_tbl$MS[2])
```

```
## [1] 968.7
```

```
(F_critical <- qf(0.9, anova_tbl$df[1], anova_tbl$df[2]))
```

```
## [1] 2.826
```

- Desired significance level, $\alpha = 0.10$.
- We are interested in testing the following two alternative hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

- **F Statistic, F*** $= \frac{\text{MSR}}{MSE} = \frac{76960.42}{79.45} = 968.6572$.
- **F Critical, F** $= F((1 - 0.10); 1, 43) = 2.826$
- **Decision rule:**

$$\text{If } F^* \leq F, \text{ conclude } H_0$$
$$\text{If } F^* > F, \text{ conclude } H_a$$

- Thus, since $F^* = 968.6572 > F = 2.826$ we conclude H_a, that there is a linear association between time spent and number of copiers serviced, at 10% significance level.

(c)

```
(R_squared <- mdl_copier %>% glance() %>% pull(r.squared))
```

```
## [1] 0.9575
```

The above output shows that the total variation in the number of minutes spent on a call is reduced by `95.75%` when the number of copiers serviced is introduced into the analysis. This reduction in the total variation in the number of minutes spent on a call is **relatively large**. Such a measure is called **coefficient of determination**.

(d)

```
(r <- sqrt(R_squared))
```

```
## [1] 0.9785
```

$$r = +0.9785; \text{this is a positive value since the slope coefficient } (\hat{\beta}_1 = 4) \text{ is positive.}$$

(e)

R^2 has the more clear-cut operational interpretation.

Problem 2.25

(a)

```
# Set up the ANOVA table using our self-defined function.
my_anova(mdl_airfreight, type = 'basic') %>%
  kable(caption = "ANOVA Table for Simple Linear Regression", align = "lccc")
```

Table 19: ANOVA Table for Simple Linear Regression

Source of variation	SS	df	MS
Regression	160.0	1	160.0
Error	17.6	8	2.2
Total	177.6	9	NA

Sum of squares and degrees of freedom are the additive elements. As can be seen, the regression sum of square and error sum of square add up to the total sum of square, and so is the degrees of freedom column.

(b): F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

```
# compute necessary measures for the test.
anova_tbl <- my_anova(mdl_airfreight, type = 'basic')
(F_stat <-  anova_tbl$MS[1] / anova_tbl$MS[2])
```

```
## [1] 72.73
```

```
(F_critical <- qf(0.95, anova_tbl$df[1], anova_tbl$df[2]))
```

```
## [1] 5.318
```

- Desired significance level, $\alpha = 0.05$.
- Let β_1 denotes the true coefficient associated with the number of times a carton is transferred. Then, we are interested in testing the following two alternative hypotheses:

$$H_0 : \beta_1 = 0 \text{ (no linear association)}$$
$$H_a : \beta_1 \neq 0 \text{ (linear association)}$$

- **F Statistic, F*** $= \frac{\text{MSR}}{MSE} = \frac{160}{2.2} = 72.7273$.
- **F Critical, F** $= F((1 - 0.10); 1, 43) = 5.3177$
- **Decision rule:**

$$\text{If } \ F^* \leq F, \text{ conclude } H_0$$
$$\text{If } F^* > F, \text{ conclude } H_a$$

- Here, we conclude H_a since $F^* = 72.7273 > F = 5.3177$. Therefore, at 5% significance level, there is a significant evidence that a linear association exists between the number of times a carton is transferred and the number of broken ampules.

(c)

```
mdl_airfreight %>% tidy() %>% kable()
```

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

From the above table, the t^* statistic for the test in part(b) is **8.528**, where the following establishes its equivalence to the F^* statistic, using equation (2.63) from the Textbook (Ketner, et al., 2005).

$$(t^*)^2 = 8.528^2 = 72.73 = F^*$$

(d)

```
(R_squared <- mdl_airfreight %>% glance() %>% pull(r.squared))
```

```
## [1] 0.9009
```

```
(r <- sqrt(R_squared))
```

```
## [1] 0.9492
```

From the above output, $R^2 = 0.9009$, and $r = 0.9492$. Here, the value of r is positive because the slope coefficient of the underlying regression model is positive ($\hat{\beta}_1 = 4$). The value of R^2 implies that **90.09%** of the total variation in the number of broken ampules (Y) is accounted for by introducing the number of times a carton is transferred variable (X) into the regression model.

Reference

- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4, pp. 563-568). New York: McGraw-Hill/Irwin..
- RPubs site for ALSR (Inferences in Regression and Correlation Analysis): https://rpubs.com/bryangoodrich/5216.