STT 4660
Homework #4

Timothy Stubblefield

September 30, 2020

2.15 (b)

1) Refer to the Airfreght Breakage dataset from Problem 1.21

2) Given a new shipment of two transfers, we want to find 99 percent confidence interval for this new shipment.

3) Since we have two transfers, we know that,

$$x_{new} = 2$$

4) And from Problem 1.21, we know that the linear regression model is,

$$\hat{y} = 10.2 + 4x$$

5) Thus, we can solve for $\hat{y}_{new}$ with the following formula,

$$\hat{y}_{new} = 10.2 + 4x_{new}$$

6) Substituting in the value for x_{new}, we get,

$$\hat{y}_{new} = 10.2 + 4(2)$$

$$\hat{y}_{new} = 10.2 + 8$$

$$\hat{y}_{new} = 18.2$$

7) Now, use the following formula to construct a 99 percent confidence interval for y_{new},

$$\hat{y}_{new} \pm t_{\alpha/2} * S_{pred}$$

8) First, we need to solve for S_{pred}, with the following formula,

$$S^2_{pred} = \hat{\sigma}^2[1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}]$$

9) From the previous results, we know that,

$$\hat{\sigma}^2 = MSE = 2.2$$

10) From the same problem, we also know that,

$$S_{xx} = \sum (x_i - \bar{x})^2 = 10$$

And, that $n = 10$ $\qquad \bar{x} = 1$

11) Substituting in these values, we can calculate S_{pred}^2 as such,

$$S_{pred}^2 = 2.2[1 + \frac{1}{10} + \frac{(2-1)^2}{10}]$$

$$S_{pred}^2 = 2.2[1 + \frac{1}{10} + \frac{1}{10}]$$

$$S_{pred}^2 = 2.2[\frac{5}{5} + \frac{1}{5}]$$

$$S_{pred}^2 = 2.2(\frac{6}{5})$$

$$S_{pred}^2 = 2.64$$

12) Now, compute S_{pred} by taking square root of S_{pred}^2,

$$S_{pred} = \sqrt{2.64}$$

$$S_{pred} = 1.624807681$$

13) Since we want a 99 percent confidence interval, we use $\alpha = 0.01$.

14) Now we need to compute the t-value, that is,

$$t_{\alpha/2}(n-2) = t_{0.01/2}(10-2)$$

$$t_{0.005}(8) = 3.355387$$

15) Now, we can construct the 99 percent confidence interval for y_{new} by substituting in the appropriate values,

$$18.2 \pm 3.355387 * 1.624807681$$

$$18.2 \pm 5.774355273$$

16) Thus, the 99 percent confidence interval for y_{new} is,

$$(12.42564473, 23.97435527)$$

17) Therefore, we are 99 percent confident that y_{new} lies in the range of (12.42564473, 23.97435527).

2.17

1) Given the F-test of

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

The analyst concluded that the p-value = 0.033.

2) For the F-test, the analyst concluded $H_a : \beta_1 \neq 0$.

3) Since H_a was concluded, that means that H_0 was rejected. Therefore, that means p-value = $0.033 < \alpha$.

4) Thus, the the α is greater than 0.033.

5) Now, if the analyst used an $\alpha = 0.01$, then,

$$p - value = 0.033 > \alpha = 0.01$$

6) Therefore, with $\alpha = 0.01$, the analyst would not reject H_0 and would thus have evidence that $\beta_1 = 0$.

2.25

a) Referring to the Airfreight Breakage data of Problem 1.21, we want to set up the ANOVA table.

   1) Here is the ANOVA Table from SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	160.00000	160.00000	72.73	<.0001
Error	8	17.60000	2.20000		
Corrected Total	9	177.60000			

b) Now, we want to construct an F-test with $\alpha = 0.05$.

   1) For the F-test, we are testing the hypotheses,

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

   2) Well, we new that,

$$F_{obs} = \frac{MSR}{MSE} = \frac{160}{2.2}$$
$$F_{obs} = 72.727272$$

3) Use the p-value method to find,

$$p - value = P(F > 72.727272)$$

4) From R, we determined that p-value = 0.00002784
5) Thus, we have the result,

$$p - value = 0.00002784 < \alpha = 0.05$$

6) Therefore, reject H_0 and conclude H_a.
7) Thus, we can say that there is evidence to show that $\beta_1 \neq 0$ and subsequently, that there is a linear relationship between X and Y.

c) Determine the t* statistic and show that it is numerically equivalent to the F-test.

1) Let's perform a T-test on β using the following formula,

$$t_{obs} = \frac{b_1}{S_{b_1}}$$

2) From Problem 2.6 in HW3, we know $S_{b_1} = 0.469041576$
3) Substituting in appropriate values, we get,

$$t_{obs} = \frac{4}{0.469041576}$$
$$t_{obs} = 8.528028654$$

4) Now compute the p-value for the t-test using R or SAS, and we get,

$$p - value = 0.00002784$$

5) Notice, the p-value for F-test and the T-test are the same. Thus, the two tests are equivalent.

d) Finally, lets calculatte R^2 and r.

(a) To calculate the coefficient of determination, R^2, use the forumula,

$$R^2 = \frac{SSR}{SSTO}$$

(b) Filling in the values from the ANOVA Table, we have,

$$R^2 = \frac{160}{177.6}$$
$$R^2 = 0.900900900$$

(c) Therefore, around 90 percent of the variation in Y is explained by introducing X into the regression model.

2.30

a) Using the Crime Rate Data from 1.28, use a t-test with $\alpha = 0.01$ to determine if there is a linear relationship between crime rate and percentage of high school graduates.

1) Since we are testing to see if X and Y are linearly related, the hypothesis situation will be,

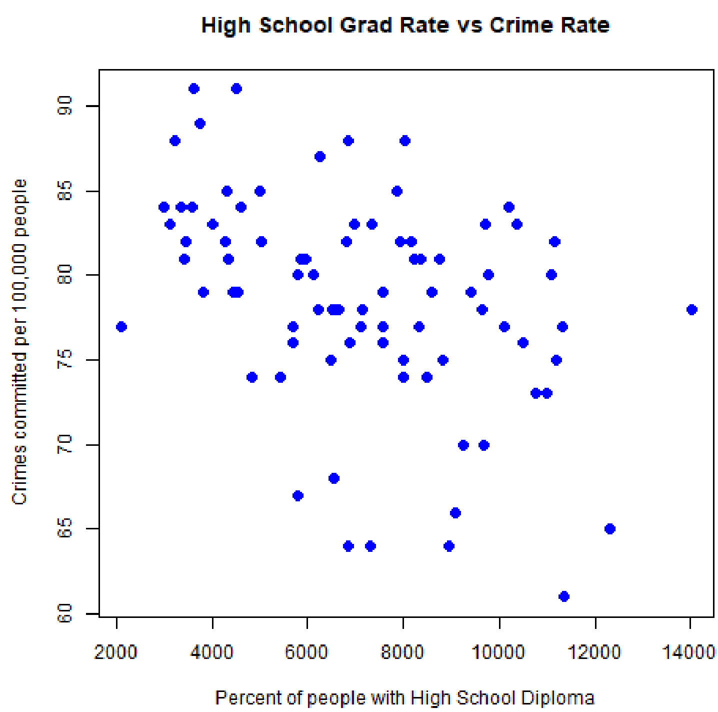
$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

2) From the SAS output, the linear regression function is,

$$\hat{y} = 20518 - 170.57519x$$

3) Here is the scatterplot of the data,



4) Now, we need to calculate the t_{obs} using the formula,

$$t_{obs} = \frac{b_1 - \beta_{10}}{S_{b_1}}$$

5) So, we know from the SAS output that,

$$b_1 = -170.57519, \qquad \beta_{10} = 0, \qquad S_{b_1} = 41.57433$$

6) Substituting in these values, we get,

$$t_{obs} = \frac{-170.57519 - 0}{41.57433}$$

$$t_{obs} = \frac{-170.57519}{41.57433}$$

$$t_{obs} = -4.102896908$$

7) From the notes, we know that

$$t = \frac{b_1}{S_{b_1}} \sim t(n-2) \qquad under \qquad H_0 : \beta_1 = 0$$

8) Now, use R or SAS to compute the p-value, that is,

$$p - value = 0.001202528$$

9) So, compare p-value to α, as such,

$$p - value = 0.001202528 < 0.01 = \alpha$$

10) Since p-value $< \alpha$, then we reject H_0 and conclude H_a.

11) Therefore, we have evidence that shows $\beta_1 \neq 0$, and subsequently, that there is a linear relationship between high school graduation rate and crime rate.

b) Now, lets's estimate β_1 with a 99 percent confidence interval.

1) A 99 percent confidence interval for β_1 is,

$$b_1 \pm t_{\alpha/2}(n-2) * S_{b_1}$$

And, $\alpha = 0.01$, so $\frac{\alpha}{2} = 0.005$

2) First, we need to compute $t_{\alpha/2}(n-2)$, as such,

$$t_{\alpha/2}(n-2) = t_{0.005}(84-2)$$

3) Using a two-tailed T-test, the evaluated t-value from R is,

$$t_{0.005}(82) = 2.637123$$

4) Now, from previous results we know that $b_1 = -170.57519$ and $S_{b_1} = 41.57433$.

5) Thus, substitute in the values to the equation from (1), to find the confidence interval,

$$-170.57519 \pm 2.637123 * 41.57433$$

$$-170.57519 \pm 109.6366219$$

6) Thus, the 99 percent confidence interval for β_1 is,

$$(-280.2118119, -60.9385681)$$

7) Therefore, we are 99 percent confident that the true value of β_1 lies in the range of (-280.2118119, -60.9385681).

2.32

a) Referring to the Crime Rate data use in 2.30, we want to find the full and reduced regression models.
Here is the ANOVA Table for Crime Rate Data using SAS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	93462942	93462942	16.83	<.0001
Error	82	455273165	5552112		
Corrected Total	83	548736108			

1) From Problem 2.30, we determined the regression model to be

$$\hat{y} = 20518 - 170.57519x$$

2) Since we are using the Simple Linear Regression (SLR) model, the full model is the regression function with $\beta_1 \neq 0$ and the reduced model is the intercept only model of the regression function where $\beta_1 = 0$.

3) Thus, the full and reduced models are as follows,

$$FullModel: \quad \hat{y} = 20518 - 170.57519x$$

$$ReducedModel: \quad \hat{y} = 20518$$

b) Now, let's obtain some useful statistics.

1) As detailed in the notes, the $SSE(F)$ uses the full model so,

$$SSE(F) = SSE$$

Looking at the ANOVA Table from (a), we have,

$$SSE(F) = 5552112$$

2) Now, we want to find $SSE(R)$. The notes specify that in the SLR model,

$$SSE(R) = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - \bar{y})^2 = SSTO$$

From the ANOVA Table, we can see that,

$$SSE(R) = SSTO = 548736108$$

3) Next, we need to calculate df_F From the book, we can see that

$$df_F = n - 2$$

Also, we know n =84. Thus, substituting in values, we get,

$$df_F = 84 - 2$$

$$df_F = 82$$

4) Now, let's calculae df_R From the book, we can determine that

$$df_R = n - 1$$

Thus, substituting in the value, n=84, we get,

$$df_R = 84 - 1$$

$$df_R = 83$$

5) Find the statistic F* for the general linear test. The formula for F* for the general linear test is,

$$F* = \frac{SSLF/1}{SSE(F)/(n-2)}$$

First, we need to compute the SSLF, using the following formula,

$$SSLF = SSE(R) - SSE(F)$$

Substitute in the appropriate values to determine SSLF,

$$SSLF = 548736108 - 5552112$$

$$SSLF = 543183996$$

Now, substitute in the values to compute F*,

$$F* = \frac{543183996/1}{5552112/(84-2)}$$

$$F* = \frac{543183996}{5552112/82}$$

$$F* = \frac{543183996}{67708.68293}$$

$$F* = 8022.368365$$

6) Determine the decision from the F-test. Construct the following hypothesis test scenario,

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Now, $F* \sim F(df(SSLF), df(SSE(F)))$.
Find $df(SSLF)$ using the formula,

$$df(SSLF) = df(SSE(R)) - df(SSE(F))$$

Since SSE(F) = SSE, then from ANOVA Table,

$$df(SSE(F)) = df(SSE) = 82$$

Similarly, since SSE(R) = SSTO, from the ANOVA Table,

$$df(SSE(R)) = df(SSTO) = 83$$

Evaluate using the correct values, yielding,

$$df(SSLF) = 83 - 82$$

$$df(SSLF) = 1$$

So, then we know,

$$F* \sim F(1, 82)$$

We need to find the p-value, which can be computed like so,

$$p - value = P(F > F*_{obs})$$

$$p - value = P(F > 8022.368365)$$

Using the following R code to calculate the p-value,

```
p_value <- pf(8022.368365,1,82, lower.tail = FALSE)
p_value
```

And so, the p-value for the F-test is,

$$p - value = 1.42806 * 10^{-83}$$

We are testing using $\alpha = 0.01$, but regardless of whatever α we used, we see that,

$$p - value = 1.42806 * 10^{-83} < \alpha$$

Therefore reject H_0 and conlcude H_a. Thus, there is reason to believe that crime rate and high school graduation rate a related.

c) Finally, we need to determine if the decision rule from using F* is the same as using the T-test in 2.30 (a).

1) While both tests ended up rejecting H_0 because the p-value $< \alpha$, the two p-values were different. The p-values are as follows,

$$p - value \quad for \quad T - test = 0.001202528$$

$$p - value \quad for \quad F - test = 1.42806 * 10^{-83}$$

2) Even though these p-values will reach the same conclusion for most values of α, they are not technically equivalent, and thus, the two tests are not truly equivalent either.