Stat 6021: Exam

Hyun Suk (Max) Ryoo (hr2ee)

11/18/2021

Pledge

On My honor, I pledge that I have neither given nor recieved any help on this assignment.

Signed: Hyun Suk (Max) Ryoo

Question 1

From figure 1 we can see that the data looks like the X and Y variables have a linear relationship with one another, which is an indicator that we should be able to use a linear regression model to give more insights about the relationship. Also from this plot we can see also see that the data points are evenly scattered around the fitted line. From the residual plot in Figure 2 we can see that the residuals are not evenly scattered around 0 at random, which doesn't indicate much transformation. We see a fanning out residual plot where as the x values get bigger the residuals are further away from 0 in the residual plot when compared to the lower x values. The residual plot is an empirical way to evaluate the assumptions however, the Box Cox method is an analytical way to evaluate the constant variance and normailty assumtptions.The Box Cox Plot in Figure 2 shows the lambda being a value between 0.2 and 0.7. The middle points towards a lambda value of about 0.4. If the box cox 95% confidence interval had a value of 1 then the response variable doesn't need to transform, but since that is not the case we would want to try transforming the response variable by a power of 0.7 (optimal value based on Box Cox) and recheck the residual plot to see if our model will still meet the assumptions. If the phenomena still exists then more transformations might need to be done.

Question 2 - A

The estimated slop is 23.687. In context the interpretation of will be that the for every increase of the Gleason score (score variable), the antigen level (PSA) will increase by 1 unit, which in this case is mg/ml.

Question 2 - B

Source of Variation	DF	SS	MS	F
Regression	1	$SST - SSE = 159651.3 - 130195.6 = 29455.7$	29455.7	$\frac{MSR}{MSE} = \frac{29455.7}{1370.48} = 21.49298$
Error	95	$s^2 * DF = 37.02^2 * 95 = 130195.6$	$37.02^2 = 1370.48$	$***$
Total	96	$SST = \frac{SSE}{1-r^2} = \frac{130195.6}{1-0.1845} = 159651.3$	$***$	$***$

Question 2 - C

$$H_0 : \beta_1 = 15$$

$$H_A : \beta_1 > 15$$

The t stat is

$$t = \frac{23.687 - 15}{5.109} = \frac{8.687}{5.109} = 1.700333$$

With this t stat we can calculate the p - value, which will be $1 - pt(1.700333, 95) = 1 - 0.95383 = 0.04617$. The critical value will be $qt(0.95, 95) = 1.661052$.

The p-value is below the threshold value of 0.05 and the t -statisic is greater than the crticical value so we can reject the null hypothesis. This means that the data does show support for the memeber's belief that the prostate-specific antigen level increases on average by more than 15mg.

Question 2 - D

The 95% Confidence Interval Equation is as follows

$$\hat{\mu_0} - t_{alpha,n} SE_{res} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{\mu_0} + t_{alpha,n} SE_{res} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

The 95% Prediction Interval Equation is as follows

$$\hat{y_0} - t_{alpha,n} SE_{res} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{y_0} - t_{alpha,n} SE_{res} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

The main information we are missing that we need to find is the quantity inside the sqaure root notation of $\frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}$. We can derive this from the 95% confidence interval since the interval is given to us. The upper bound is set to , 34.22774

$$\begin{aligned}
34.22774 &= \hat{\mu_0} - t_{alpha,n} SE_{res} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\
34.22774 &= (-139.150 + 23.687 * 7) + t_{0.975,95} * 37.02 \sqrt{\frac{1}{97} + x} \\
34.22774 &= 26.659 + 1.985251 * 37.02 * \sqrt{0.01030928 + x} \\
7.56874 &= 1.985251 * 37.02 * \sqrt{0.01030928 + x} \\
7.56874 &= 73.49399 * \sqrt{0.01030928 + x} \\
0.1029845 &= \sqrt{0.01030928 + x} \\
0.01060581 &= 0.01030928 + x \\
0.0002965316 &= x
\end{aligned}$$

Using this value, we can double check our work and solve for the lower bound to see if matches the given value of 19.09329

$$\begin{aligned}
19.09329 &= \hat{\mu_0} - t_{alpha,n} SE_{res} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\
&= (-139.150 + 23.687 * 7) - t_{0.975,95} * 37.02 \sqrt{\frac{1}{97} + 0.0002965316} \\
&= 26.659 - 1.985251 * 37.02 \sqrt{0.01030928 + 0.0002965316} \\
&= 26.659 - 1.985251 * 37.02 * 0.1029845 \\
&= 26.659 - 7.568742 \\
19.09329 &= 19.09026
\end{aligned}$$

The difference may have came from rounding issues in R, but it seems like we can utilze this for our 95% PI.

$$\hat{y_0} - t_{alpha,n} SE_{res} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{y_0} - t_{alpha,n} SE_{res} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$26.659 - t_{0.975,95} * 37.02 \sqrt{1 + \frac{1}{97} + 0.0002965316}, 26.659 + t_{0.975,95} * 37.02 \sqrt{1 + \frac{1}{97} + 0.0002965316}$$

$$26.659 - 73.49399 \sqrt{1 + \frac{1}{97} + 0.00028802}, 26.659 + 73.49399 \sqrt{1 + \frac{1}{97} + 0.00028802}$$

$$26.659 - 73.49399 \sqrt{1.010606}, 26.659 + 73.49399 \sqrt{1.010606}$$

$$26.659 - 73.49399 * 1.005289, 26.659 + 73.49399 * 1.005289$$

$$26.659 - 73.8827, 26.659 + 73.8827$$

$$-47.2237, 100.5417$$

Question 3 - A

For the testing of $H_0 : \beta_4 = 0$, $H_A : \beta_4 \neq 0$, the p-value can be found by first calculating the t-statistic.

$t = \frac{\text{Estimate}}{\text{Std. Error}} \rightarrow \frac{1.104}{1.325} \rightarrow 0.8332075$. The p-value was computed using the pt function in R.

```
(1 - pt(0.8332075,92)) * 2
```

```
## [1] 0.4068856
```

The p-value is greater than 0.05 so we fail to reject the null hyptohesis. Our data usggests that capsular is not useful in predicting the response of PSA when the other predictors are already in the model. However, I do not agree with the opinion of the classmate of 'capsular is not linearly associated with PSA' because of the reason that the t test in multiple linear regression does not test if a predictor is linearly related to the response, on its own.

Question 3 - B

The two models can be summarized into a full model (Model 2), and a reduced model (Model 1). We can conduct a partial F test to see which model we can utilize.

Full Model 2 : $PSA = \beta_0 + \beta_1 score + \beta_2 volume + \beta_3 invasion + \beta_4 capsular$

Reduced Model 1 : $PSA = \beta_0 + \beta_1 score$

The null and hypothesis are as follows.

$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ H_A : at least one of the coefficients in H_0 is non zero

$$\begin{aligned} F - stat &= \frac{\frac{SS_R(x1,x2,x3,x4) - SS_R(x1)}{r}}{\frac{SS_{res}(x1,x2,x3,x4)}{n-p}} \\ &= \frac{\frac{SS_R(x1,x2,x3,x4) - SS_R(x1)}{r}}{\frac{SS_{res}(x1,x2,x3,x4)}{n-p}} \\ &= \frac{\frac{36200 + 4985 + 666}{3}}{\frac{88354}{92}} \\ &= \frac{\frac{41851}{3}}{\frac{88354}{92}} \\ &= \frac{13950.33}{960.3696} \\ &= 14.526 \end{aligned}$$

The p-value is found by using $1 - pf(14.526, 3, 92) = 8.014422e - 08$

```
1-pf(14.526, 3, 92)
```

```
## [1] 8.014422e-08
```

The critical value is the following.

```
qf(0.95, 3, 92)
```

```
## [1] 2.703594
```

The critical value is 2.70359 and, which is lower than our partial F statistic. Therefore, we reject the null hypothesis. Data suggests that the for the coefficients of volumne, invasion, and capsular at least one is not zero, which means that we should use the full model.

Question 3 - C

We can also utilize the partial F test for this question as well.

The null and hypothesis are as follows.

$H_0 : \beta_2 = \beta_3 = 0$ H_A : at least one of the coefficients in H_0 is non zero

$$\begin{aligned} F - stat &= \frac{\frac{SS_R(x1,x2,x3) - SS_R(x1)}{r}}{\frac{SS_{res}(x1,x2,x3)}{n-p}} \\ &= \frac{\frac{SS_R(x1,x2,x3) - SS_R(x1)}{r}}{\frac{SS_{res}(x1,x2,x3)}{n-p}} \\ &= \frac{\frac{36200+4985}{2}}{\frac{88354+666}{97-4}} \\ &= \frac{\frac{41185}{2}}{\frac{89020}{93}} \\ &= \frac{20592.5}{957.2043} \\ &= 21.51317 \end{aligned}$$

The p-value is found by using $1 - pf(21.51317, 2, 93) = 2.09405e - 08$

```
1-pf(21.51317, 2, 93)
```

```
## [1] 2.09405e-08
```

The critical value is the following.

```
qf(0.95, 2, 93)
```

```
## [1] 3.094337
```

The critical value is 3.094337 and, which is lower than our partial F statistic. Therefore, we reject the null hypothesis. Data suggests that the for the coefficients of volumne and invasion at least one is not zero.

Question 4 - A

We know from the provided information that the HIspanic Race is the reference class for race and Also, male is coded as 0. Therefore we can reeduce the regresssion equation for hispanic males to the following.

$$
\begin{aligned}
E(y) &= \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 F + \beta_5 x_1 \\
&= 19.29178 + 7.34591 I_1 + 0.66265 I_2 + 3.36449 I_3 + 5.26062 F + 0.53047 x_1 \\
&= 19.29178 + 7.34591(0) + 0.66265(0) + 3.36449(0) + 5.26062(0) + 0.53047 x_1 \\
&= 19.29178 + 0.53047 x_1
\end{aligned}
$$

Question 4 - B

The value for $\hat{\beta}_4$ is 5.26062, which means that for females the student's score on a standardized writing test (write) is 5.26062 greater than males given the same race and read varaibles.

Question 4 - C

If there are no significant interaction terms in this regression, that means when looking at the data with all variables of race, gender, and read score, the race and gender do not play a significant role in determining the score for write. This does not mean that there is no relationship individually rather it means that if all variables are present, which was what the analysis was done for, the race and gender do not play a significant role.

Question 4 - D

The 95% confidence interval for the true mean difference in mean score can be calculated with the Bonferroni procedure by $\hat{\beta}_j \pm t_{\frac{\alpha}{2p}, n-p} se(\hat{\beta}_j)$ The $qt(1 - 0.05/8, 195) = 2.521089$ function was used to find the multiplier

- Caucasion and hispanic students for a given level of gender and given value of read.

$$
\begin{aligned}
\hat{\beta}_j \pm t_{\frac{\alpha}{2p}, n-p} se(\hat{\beta}_j) \\
3.36449 \pm t_{1-\frac{0.05}{2*4}, 195} 1.58778 \\
3.36449 \pm 2.521089 * 1.58778 \\
3.36449 \pm 4.002935 \\
(-0.638445, 7.367425)
\end{aligned}
$$

We are 95% confident that the difference in write scores (response variable) for caucasian and hispanic students falls in the range of (-0.638445, 7.367425).

Question 5 - A

False, one advantae of adjusted R^2 is that the addition of predictors that do not help fruther eplain the response vairable will lead to a decrease in adjusted R^2 while R^2 will always increase, regardless of whether the predictor helps further epxlain the variance in the response variable or not.

Question 5 - B

False, the model may differ depending on which statistic you are utilizing, r-squared, AIC, BIC , Cp, etc.. Depending on the statistic being used, the selection process can output different models.

Question 5 - C

True

Question 5 - D

False, the equation of DFFITS is $DFFITS_i = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5}$. We can see from this equaation that as leverage increases DFFITS increases, which means that if an observation is far away from the center of the predictors, the large the difference in predicted values with and without that observation in the regression model. From an initial analysis it seems like Case 3 is further away from the center of the predictors, so the case 3 point will have a higher DFFITS value.