**STAT 501 – Homework10 Solutions**

1. (**15x3 = 45 points**) Use the “FemaleBears” dataset. Data from *n* = 19 female bears of varying ages are used to develop an equation for estimating *Y* = female bear's weight from *X* = female bear's neck circumference.
   1. Fit a simple linear regression model with *Y* = female bear's weight and *X* = female bear's neck circumference. Click the “Storage” button in the Minitab Regression Dialog and select each of the items in the left-hand list (i.e., Fits, Residuals, Standardized residuals, Deleted residuals, Leverages, Cook’s distance, DFITS). Write down the estimated regression equation and the MSE for this model.

Y-hat = –158.8 + 16.95(Neck). MSE = 1610.

* 1. Which bear number has the highest leverage and what is that leverage? [Leverages are in the column labeled “HI1”]

Observation #13 with leverage 0.286265.

* 1. Is the leverage in the previous part higher than the threshold 3(*p*/*n*)?

It is not higher than the threshold, 3(2/19) = 6/19 = 0.316.

* 1. Use the estimated regression equation from part (a) to calculate the fitted value for bear #6. [You can check your answer with the one Minitab provides in the column labeled “FITS1”.]

–158.8 + 16.95(10.5) = 19.175 (Minitab answer: 19.171).

* 1. Use your answer from the previous part together with the actual weight of bear #6 to calculate the residual for this bear. [You can check your answer with the one Minitab provides in the column labeled “RESI1”.]

140 – 19.175 = 120.825 (Minitab answer: 120.829).

* 1. What is the leverage for bear #6?

0.239605.

* 1. Use the residual from part (e), the MSE from part (a), and the leverage from part (f) to calculate the internally studentized residual for bear #6. [You can check your answer with the one Minitab provides in the column labeled “SRES1” – remember Minitab calls these “Standardized residuals.”]

120.825 / √(1610(1–0.239605)) = 3.453 (Minitab answer: 3.45320).

* 1. Delete bear #6 from the dataset as follows: select Data > Subset Worksheet, click “Specify which rows to exclude,” click “Row numbers,” and type “6” into the adjoining box. Then refit the simple linear regression model with *Y* = female bear's weight and *X* = female bear's neck circumference. Write down the estimated regression equation and the MSE for this model.

Y-hat = –234.6 + 20.54(Neck). MSE = 511.

* 1. Use the residual from part (e), the MSE from part (h), and the leverage from part (f) to calculate the externally studentized residual for bear #6. [You can check your answer with the one Minitab provides in the column labeled “TRES1” in the original worksheet – remember Minitab calls these simply “Deleted residuals.”]

120.825 / √(511(1–0.239605)) = 6.130 (Minitab answer: 6.13120).

* 1. Use the estimated regression equation from part (h) to calculate the predicted value for bear #6 (i.e., based on the model fit to the subset worksheet excluding bear #6). [Note: the answer won’t make a whole lot of sense, but don’t worry about this since we’re simply going to use this predicted value for part (k).]

–234.6 + 20.54(10.5) = –18.93.

* 1. Use the fitted value from part (d), the predicted value from part (j), the MSE from part (h), and the leverage from part (f) to calculate the DFFITS for bear #6. [You can check your answer with the one Minitab provides in the column labeled “DFIT1” in the original worksheet.]

(19.175–(–18.93)) / √(511(0.239605)) = 3.444 (Minitab answer: 3.44171).

* 1. Is the absolute value of DFFITS in the previous part higher than the threshold given in the online notes, ?

It is higher than the threshold,

* 1. Use the residual from part (e), the MSE from part (a), and the leverage from part (f) to calculate the Cook’s distance for bear #6. [You can check your answer with the one Minitab provides in the column labeled “COOK1” in the original worksheet.]

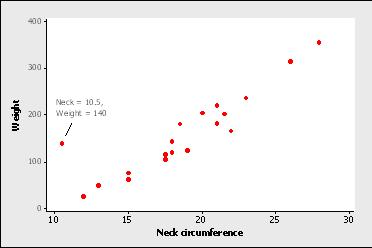
(120.8252/(2(1610)))(0.239605/(1–0.239605)2) = 1.879 (Minitab answer: 1.87875).

* 1. Is the Cook’s distance from the previous part higher than the upper threshold given in the notes, 1?

It is higher than the threshold, 1.

* 1. Briefly summarize your findings with respect to bear #6. You might want to consider graphical evidence too!

Bear #6 does not have excessively high leverage, but it does have a large internally studentized residual, externally studentized residual, DFFITS, and Cook’s distance. It also sticks out like a sore thumb in a scatterplot of the data:



1. (**9x3 = 27 points**) Use the “CollegeGPA” dataset. Data from *n* = 40 college students are used to develop an equation for estimating *Y* = grade point average (GPA) from *X1* = verbal score on a college entrance exam (percentile) and *X2* = math score on a college entrance exam (percentile).
   1. Fit a “full quadratic” multiple linear regression model with *Y, X1, X2, X12, X22,* and *X1 X2*. [In Minitab: Select *Y* as the Response, *X1* and *X2* as the Continuous predictors, click “Model,” select both *X1* and *X2* together in the Predictors box and click the Add buttons next to “Interactions through order 2” and “Terms through order 2.”] Also click the “Storage” button in the Minitab Regression Dialog and select Deleted residuals, Leverages, and Cook’s distance. Write down the estimated regression equation.

Y-hat = -7.22 + 0.1263 X1 + 0.1170 X2 - 0.001130 X12 - 0.001063 X22 + 0.000878 X1 X2.

* 1. Which student has the largest absolute externally studentized residual and what is that externally studentized residual?

Student 28 has the largest absolute externally studentized residual and the externally studentized residual is –3.03.

* 1. Is the externally studentized residual from the previous part greater in absolute value than 3? What do we call such points?

Yes, the externally studentized residual from the previous part is greater in absolute value than 3, so it is an outlier.

* 1. Which student has the highest leverage and what is that leverage?

Student 4 has a leverage of 0.56.

* 1. Is the leverage from the previous part higher than the threshold 3(*p*/*n*)?

Yes, the leverage from the previous part is higher than the threshold 3(*p*/*n*) = 3(6/40) = 0.45.

* 1. What is it about the student identified in part (d) that gives him/her such a high leverage? (Hint: compare this student’s exam scores with other students’ scores.)

This student has a very high X1 value (100) but a very low X2 value (49). Thus they lie on the edge of the (X1, X2) sample distribution of students and are potentially influential.

* 1. Which student has the highest Cook’s distance and what is that Cook’s distance?

Student 9 has a Cook’s distance of 0.31.

* 1. Is the Cook’s distance from the previous part higher than the upper threshold given in the notes, 1?

No, this Cook’s distance is not higher than 1.

* 1. Investigate whether removing any of the observations identified in the previous parts dramatically alters the model results.

Removal of any of the observations identified in the previous parts does not dramatically alter the model results.

1. (**7x4 = 28 points**) Use the “ProstateCancer” dataset. The data are from *n* = 97 prostate cancer patients. The variables are:

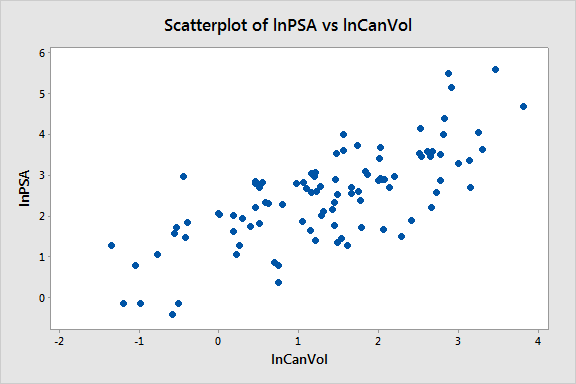
*Y* = lnPSA; natural log of the prostate specific antigen value, a blood chemistry measurement affected by the presence of prostate cancer

*X*1 = lnCanVol; natural log of the cancer volume (cc)

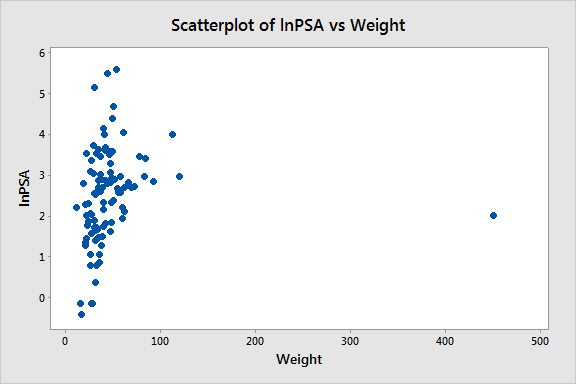
*X*2 = Weight; prostate weight (gm)

1. Create scatterplots of Y versus X1 and Y versus X2 and discuss noteworthy features of the data and the relationships.

There is a moderate positive linear association in the scatterplot of *Y* versus *X1* with no obvious outlying points.



There is one clear outlying point with a very high value of weight in the scatterplot of *Y* versus *X2*. It is difficult to discern any particularly strong patterns in the rest of the data.



1. Do a multiple regression to predict *Y* using *X1*  and *X2* as predictors. Store Cook's Di and DFFITS values (we will look at those in the next part). To store these in Minitab, use the Storage button within the regression dialog box and select then from under the Diagnostic measures. Complete the table below with values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Coefficient Value** | **Standard Error** | ***p*-value** |
| lnCanVol | 0.7183 | 0.0675 | 0.000 |
| Weight | 0.00307 | 0.00174 | 0.081 |

1. The table below is a list of the “Unusual Observations” that Minitab gives for the regression done in part (b). Discuss this list in terms of what data difficulties (or potential difficulties) may be indicated. As an aid to understanding why some observations may have been marked *X*, plot *X1* versus*X2*. Use that plot and the plots done in part (a) to guide your discussion.

**Obs lnPSA Fit Resid Std Resid**

**5 0.370 2.003 -1.633 -2.11 R**

**18 1.490 3.133 -1.643 -2.13 R**

**32 2.010 2.882 -0.872 -2.78 R X**

**69 2.960 1.299 1.661 2.18 R**

**95 5.140 3.552 1.588 2.07 R**

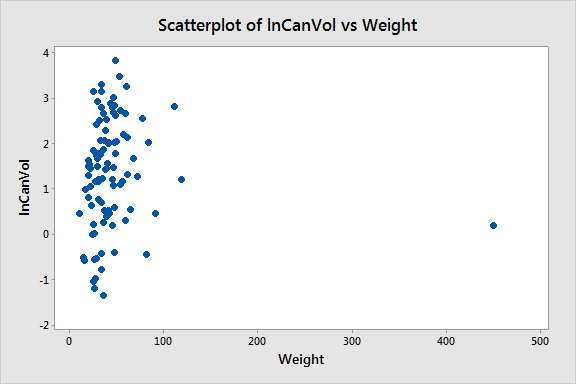
**96 5.480 3.571 1.909 2.48 R**

**97 5.580 4.025 1.555 2.04 R**

**R Large residual**

**X Unusual X**

Seven points are marked as having “large” residuals. That is a little high for this sample size, and one of the standardized residuals is close to –3 (that for observation 32). This shows up as the outlying point in two of the scatterplots. This observation is also marked as having unusual *x*-values (since it has such a high value for weight). The other six observations just marked R but not X can be located along the upper and lower edges of the band of points in the *Y* versus *X1* scatterplot.



1. Concerning the regression done in part (b), determine which data point(s) may have unusually large values for both the DFFITS values and the Cook's Di values. For DFITS use the “greater than *2√((p+1)/(n–p–1))* in absolute value” standard and for Cook's Di use a “greater than 1” standard. Delete any such data points from the dataset. Describe which observation(s) you're deleting and explain why you're doing the deletion(s).

The DFFITS for four of the previously identified observations (observations 32, 69, 96, and 97) exceed *2√((p+1)/(n–p–1))* = *2√((3+1)/(97–3–1))* = *2√(4/93)* = 0.415, but only observation 32 has a Cook’s distance greater than 1.

1. Do the multiple regression again using the new dataset after the deletion(s) of part (d). [Minitab: Data > Subset Worksheet.] Complete the table below with the resulting values.

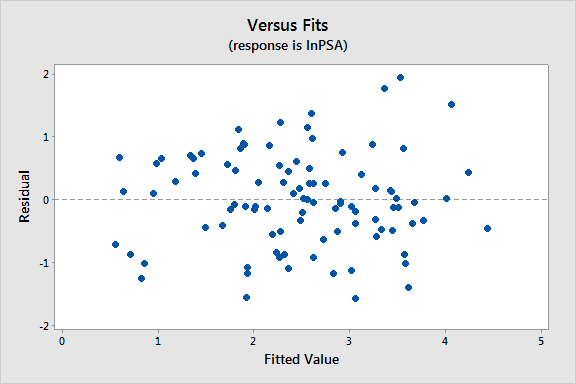
|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Coefficient Value** | **Standard Error** | ***p*-value** |
| lnCanVol | 0.6711 | 0.0670 | 0.000 |
| Weight | 0.01393 | 0.00412 | 0.001 |

Comment on the differences between the results in this part and part (b). The main point here is that one or two points may influence matters so much that their presence or absence can change conclusions.

Compared to the results in part (b), after the deletion of point 32 we see that Weight is now significant at the 0.05 significance level and its coefficient estimate is higher. The coefficient estimate for lnCanVol is slightly lower with a slightly smaller standard error.

1. For the data and model of part (e), create a plot of residuals versus fits. Discuss whether any difficulties with the model or the data are indicated.

There are no problems with the LINE conditions indicated by the residual plot.



1. Discuss whether you think any further data points should be deleted. Indicate which observations you would delete, if any, or say why you don’t think any more points should be deleted. [Hint: repeat what you did before in part (e); if there are no observations that exceed the thresholds for DFFITS or Cook’s distance then it’s unlikely any further data points should be deleted, but to be sure you can delete the observation with the largest Cook’s distance and see what effect this has on the values in part (e).]

Observation 68 (ID 69) has the highest DFFITS and Cook’s distance but since the Cook’s distance is only 0.0958 it is unlikely to have a large influence on the model. Indeed if we do delete it the model results change very little.