**STAT 501 – Homework 10 (covering Lesson 11) – Due Nov 8**

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question.

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.

Weight = -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”]

Bear in row number 13 (0.286)

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

No, Threshold = 3\*(2/19) = 0.3158 > 0.286

* 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”.]

The fitted value: Weight = -158.8 + 16.95 \* 10.5 = 19.175 (FITS1 has 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”.]

Actual weight = 140

Residual = 140 - 19.175 = 120.825 (RESI1 has 120.829)

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

Leverage is 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.”]



r6 = 120.825 / sqrt(1610\*(1-0.239605)) = 3.453 (SRES1 has 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.

Weight = -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.”]

Residual: 120.825, MSE: 511, Leverage: 0.239605



t6 = 120.825 / sqrt(511\*(1-0.239605)) = 6.1295 (TRES1 has 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).]

Weight = -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.]

Fitted value: 19.175, predicted value = -18.93, MSE: 511, Leverage: 0.239605



DFITS6 = (19.175-(-18.93))/sqrt(511\*0.239605) = 3.444 (DFIT1 has 3.44171)

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

Yes, since threshold = 2 \* (sqrt(3/(19-2-1))) = 0.866 < 3.444

* 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.]

Residual = 120.825, MSE = 1610, Leverage = 0.239605



D6 = ((140-19.175)^2/(2\*1610)) \* (0.239605/(1-0.239605)^2) = 1.87878 (COOK1 has 1.87875)

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

Comparing to the notes we find that the cook’s distance was both greater than 1 and also sticks out like a sore thumb from the other Di values. It is therefore certainly influential.

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

|  |  |  |
| --- | --- | --- |
|  | Row #6 included | Row #6 excluded |
|  |  |  |
| Eq  p-value  T stat | Weight = -158.8 + 16.95 Neck  p-value for b1 < 0.001  T stat = 8.07 | Weight = -234.6 + 20.54 Neck  p-value for b1 < 0.001  T stat = 15.56 |
| MSE | 1610 | 511 |
| R-sq, R-sq(adj) | 79.30%, 78.09% | 93.80%, 93.42% |
| S | 40.1264 | 22.5999 |

The above shows the following impacts of including the bear #6

* The R2 value has decreased substantially from 93.80% to 79.30%. If we include the point, we conclude that the relationship between y and x is moderately strong, whereas if we exclude the point, we conclude that the relationship between y and x is very strong.
* The standard error of b1 is almost 2 times larger when the point is included. This increase would have a substantial effect on the width of our confidence interval for β1.
* In each case, the P-value for testing H0: β1 = 0 is less than 0.001. In both cases, we can conclude that there is sufficient evidence at the 0.05 level to conclude that, in the population, x is related to y. Note, however, that the t-statistic halves from 15.56 to 8.07 upon inclusion of the bear #6

Therefore in summary the bear# 6 is an influential point as it has moderate influence on the various stats of interest as explained above.

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.

Gpa = -7.22 + 0.1263 Verb + 0.1170 Math - 0.001130 Verb\*Verb - 0.001063 Math\*Math + 0.000878 Verb\*Math

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

Student with ID: 28 has the largest absolute externally studentized residual = -3.03046

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

Yes it is greater than 3. If an observation has an externally studentized residual that is larger than 3 (in absolute value) we can call it an outlier.

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

Student with ID: 4 has the highest leverage = 0.563070

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

Yes, since threshold = 3 \* (3/40) = 0.225 < 0.563

* 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.)

The scores for this student are atypical. The following plots are interesting

|  |  |
| --- | --- |
|  |  |
|  |  |

We know that the leverage hii quantifies how far away the ith x value is from the rest of the x values. If the ith x value is far away, the leverage hii will be large; and otherwise not. The graphs above indicate that the X values for 4th student are far from their respective mean values.

Verb: Mean = 72.10, StdDev = 16.10, student ID #4 score = 100 (more than 2 StdDev away from mean)

Math: Mean = 74, StdDev = 13.15, student ID #4 score = 49 (almost 2 StdDev away from mean)

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

The student with ID: 9 has the highest Cook’s distance = 0.308919

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

No.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All data | #28 excluded | #4 excluded | #28 and #4 excluded |
|  | Gpa = -7.22 + 0.1263 Verb + 0.1170 Math - 0.001130 Verb\*Verb - 0.001063 Math\*Math + 0.000878 Verb\*Math  All terms p-value < 0.001 | Gpa = -7.28 + 0.1156 Verb + 0.1275 Math - 0.001033 Verb\*Verb - 0.001123 Math\*Math + 0.000854 Verb\*Math  All terms p-value < 0.001 | Gpa = -7.20 + 0.1265 Verb + 0.1162 Math - 0.001125 Verb\*Verb - 0.001052 Math\*Math + 0.000866 Verb\*Math  All terms p-value <= 0.001 | Gpa = -7.11 + 0.1165 Verb + 0.1219 Math - 0.000985 Verb\*Verb - 0.001046 Math\*Math + 0.000760 Verb\*Math  All terms p-value < 0.001 |
| S | 0.203390 | 0.18259 | 0.20640 | 0.182399 |
| MSE | 0.04137 | 0.03334 | 0.04260 | 0.03327 |
| R-Sq | 93.05% | 94.52% | 92.20% | 94.02% |

The above table indicates that removing any of the observations identified in the previous parts doesn’t dramatically alter the model results. The points in that sense don’t seem to be influential points. The student with ID #28 has high externally studentized residual but it doesn’t stick out like a sore thumb. The next highest absolute value is 2.75195 which isn’t very far from 3.030457843

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.

Plot for Y versus X1



This indicates the following:

* All of the data points follow the general trend of the rest of the data, so there are no outliers
* There is a linear relationship and the equal variance assumption seems to hold

Plot for Y versus X1



This indicates the following:

* All of the data points follow the general trend of the rest of the data except the one on the extreme right. This data point also seems to be influential as it is able to pull the trend in its direction.
* There seems to be an overall linear relationship when outlier is removed
  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**

The plot for X1 versus X2:



* It is clear from the plot that row 32 which is the rightmost point in the above plot (and also in the Y vs X2) that the point is an outlier. It is also having a high leverage and is able to influence the regression output towards itself. This is what the minitab output is indicating with the labels R and X for this data point. This particular point is both:
  + Leverage point,since it is extreme in the x-direction
  + Outliers (large residuals), since it is extreme in the y-direction relative to the fitted regression line
* The output is also indicating that there are many more outliers in this data. However they are not influencing the regression output like the row 32 is doing. This is also apparent from the three plots. Therefore minitab’s output with R label for all these points is in alignment with our expectation.
* It is also to be noted that Minitab is conservative and flags any observation with an internally studentized residual that is larger than 2. Typically other do this when internally studentized residual that is larger than 3
* Most of the points marked R (not X) can be traced (in Minitab) as being furthest in this plot: 

This indicates that their residual is high as is apparent from the above picture as the distance between the point and the line.

* 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).

DFITS threshold = 2 \* sqrt((3+1)/(97-3-1)) = 0.415

Unusually large values for absolute DFITS > 0.415

|  |  |
| --- | --- |
| ID\_1 | DFIT1 |
| 32 | -6.55862 |
| 96 | 0.433025 |
| 97 | 0.447282 |
| 69 | 0.461492 |

Unusually large values for Cook’s Di values > 1

|  |  |
| --- | --- |
| ID\_1 | COOK1 |
| 32 | 13.30227067 |

Unusually large values for both

|  |  |  |
| --- | --- | --- |
| ID\_1 | COOK1 | DFIT1 |
| 32 | 13.30227067 | -6.55862 |

We will remove the data row with ID=32 and do the analysis for the weight range between 10 and 120. The row 32 has weight = 450.339

Reason for exclusion: analysis for weight range between 10 and 120

* 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.

We see the following differences:

* The coefficient values have changed for both X1 and X2
* Weight was not significant (at alpha=0.05) previously but is signifance after the deletion of row 32
  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.



Following observations can be made:

* Some points still have high residual (at the extremes). However other than that the plot depicts good characteristics in as far as linearity and equal variance assumptions are concerned.
* Most of the points marked R in part c (not X) can be traced in the above plot as being furthest in this plot. That is in line with our expecation of high residuals.
  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).]

DFITS threshold = 2 \* sqrt((3+1)/(96-3-1)) = 0.417

Unusually large values for absolute DFITS > 0.417

|  |  |
| --- | --- |
| ID\_1 | DFIT1 |
| 97 | 0.455408 |
| 96 | 0.464041 |
| 95 | 0.486745 |
| 69 | 0.540437 |

Unusually large values for Cook’s Di values > 1 : None

Unusually large values for both : None

Row with the largest Cook’s distance:

|  |  |
| --- | --- |
| ID\_1 | Cook’s Di |
| 69 | 0.0958 |

While the Cook’s Di is not > 1, we will try to perform regression by deleting this row:

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Coefficient Value** | **Standard Error** | ***p*-value** |
| lnCanVol | 0.6947 (0.6711) | 0.0681 (0.0670) | 0.000 (0.000) |
| Weight | 0.01213 (0.01393) | 0.00424 (0.00412) | 0.005 (0.001) |

Older values from part e in brackets.

We can see from the regression that the values have not deviated from the part e significantly. Therefore there is no point in deleting any more rows. The residual vs fits plot hasn’t changed as illustrated below:

