Homework 10

Roly Vicaría
STAT501 Fall 2015

November 8, 2015

0.0.1 Question 1

a) $Weight = -158.8 +\ 16.95\ Neck$; MSE = 1610

b) Bear number 13 has the highest leverage: 0.286265

c) $3(p/n) = 3(2/19) = 0.316$; The leverage from the part (b) is less than the threshold.

d) $Weight = -158.8 + 16.95(10.5) = 19.175$; Minitab FITS1 value: 19.171

e) $Residual = 140 - 19.175 = 120.825$; Minitab RESI1 value: 120.829

f) Leverage of bear #6 = 0.239605

g) $r_6 = \frac{e_6}{\sqrt{MSE(1-h_{66})}} = \frac{120.825}{\sqrt{1610(1-0.239605)}} = 3.453$; Minitab SRES1 value: 3.45320

h) $Weight = -234.6 +\ 20.54\ Neck$; MSE = 511

i) $t_6 = \frac{e_6}{\sqrt{MSE(1-h_{66})}} = \frac{120.825}{\sqrt{511(1-0.239605)}} = 6.130$; Minitab TRES1 value: 6.13120

j) $Weight = -234.6 + 20.54(10.5) = -18.93$

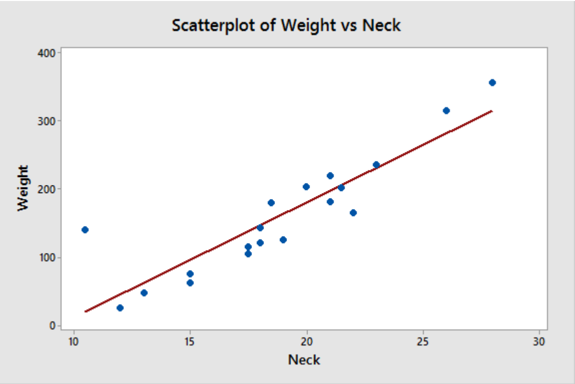
k) $DFITS_6 = \frac{\hat{y}_6 - \hat{y}_{(6)}}{\sqrt{MSE_{(6)}h_{66}}} = \frac{19.175-(-18.93)}{\sqrt{511(0.239605)}} = 3.444$; Minitab DFIT1 value: 3.44171

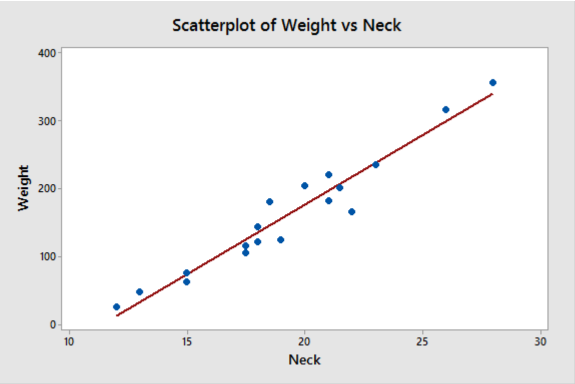
l) $2\sqrt{\frac{p+1}{n-p-1}} = 2\sqrt{\frac{2+1}{19-2-1}} = 0.866$; Yes, the absolute value of the value from part (k) is greater than the threshold given in the online notes.

m) $D_6 = \frac{(y_6-\hat{y}_6)^2}{p \times MSE}\left[\frac{h_{66}}{(1-h_{66})^2}\right] = \frac{120.825^2}{2\times 1610}\left[\frac{0.239605}{(1-0.239605)^2}\right] = 1.879$; Minitab COOK1 value: 1.87875

n) Yes, the Cook's distance is greater than 1 which indicates a high likelihood of being an influential point. It is also much greater than the Cook's distance of any other point which further supports the likelihood of it being influential.

o) The Cook's distance and the DFITS values for bear #6 both indicate that it's an outlier and influential point. The model fitted excluding bear 6 showed a much better MSE. It looks like deleting bear 6 was the correct action to take. It narrowed down the scope of the model to bears with weight greater than 12.





0.0.2 Question 2

a) $Gpa = -7.22 + 0.1263\, Verb + 0.1170\, Math - 0.001130\, Verb^2 - 0.001063\, Math^2 + 0.000878\, Verb * Math$

b) Student 28 has the highest absolute externally studentized residual: -3.03046

c) Yes, the absolute value of the result from part (b) is greater than 3. Therefore we would call this observation an outlier.

d) Student 4 has the highest leverage: 0.563070

e) $3(p/n) = 3(6/40) = 0.45$; Yes, the leverage from part (d) is greater than the threshold.

f) The range of observed Verbal scores is [39, 100] and the range of observed Math scores is [49, 99]. Student 4 has a Verbal score of 100 and a Math score of 49. They are the upper extreme of Verbal scores and the lower extreme of Math scores.

g) Student 9 has the highest Cook's distance: 0.308919

h) No, the Cook's distance from part (g) is not higher than the threshold given in the notes.

i) With all students:

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)

0.203390  93.05%     92.03%      89.94%

Coefficients

Term            Coef   SE Coef  T-Value  P-Value     VIF
Constant       -7.22      1.47    -4.91    0.000
Verb          0.1263    0.0231     5.47    0.000  130.31
Math          0.1170    0.0291     4.03    0.000  137.59
Verb*Verb  -0.001130  0.000127    -8.87    0.000   82.80
Math*Math  -0.001063  0.000173    -6.14    0.000  107.87
Verb*Math   0.000878  0.000156     5.61    0.000   47.03

Regression Equation

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

Excluding student 28:

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.182598  94.52%     93.69%      91.31%

Coefficients

Term            Coef   SE Coef  T-Value  P-Value     VIF
Constant       -7.28      1.32    -5.50    0.000
Verb          0.1156    0.0210     5.50    0.000  124.95
Math          0.1275    0.0263     4.85    0.000  138.96
Verb*Verb  -0.001033  0.000119    -8.69    0.000   81.81
Math*Math  -0.001123  0.000157    -7.16    0.000  108.62
Verb*Math   0.000854  0.000141     6.07    0.000   46.23

Regression Equation

Gpa = -7.28 + 0.1156 Verb + 0.1275 Math - 0.001033 Verb*Verb - 0.001123 Math*Math
         + 0.000854 Verb*Math
```

Removing observation for student 28 does not affect the R^2 very much. It's still a strong relationship. The standard error of all coefficients decreased marginally. The most notable difference is that the MSE decreased from 0.04137 to 0.03334. Therefore confidence/prediction intervals will be narrower in the model excluding 28.

Excluding student 4:

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.206400  92.20%     91.01%      87.00%

Coefficients

Term            Coef   SE Coef  T-Value  P-Value     VIF
Constant       -7.20      1.50    -4.79    0.000
Verb          0.1265    0.0235     5.38    0.000  120.94
Math          0.1162    0.0303     3.83    0.001  131.55
Verb*Verb  -0.001125  0.000135    -8.32    0.000   81.40
Math*Math  -0.001052  0.000196    -5.36    0.000  124.84
Verb*Math   0.000866  0.000186     4.67    0.000   64.09

Regression Equation

Gpa = -7.20 + 0.1265 Verb + 0.1162 Math - 0.001125 Verb*Verb - 0.001052 Math*Math
        + 0.000866 Verb*Math
```

Removing observation for student 4 has a small decrease in R^2, but still a strong relationship. All the coefficients show marginal increase, but nothing to affect p-value of any of them. The MSE actually increased a little bit, from 0.04137 to 0.04260. That means that confidence/prediction intervals based on this model will be slightly wider than the original model.

Excluding student 9:

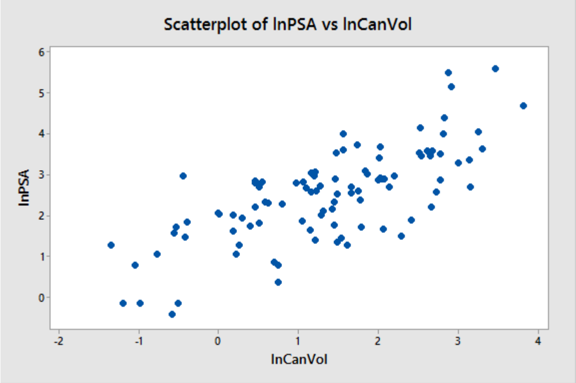
```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.186187  93.76%     92.82%      91.13%

Coefficients

Term            Coef   SE Coef  T-Value  P-Value     VIF
Constant       -5.81      1.44    -4.03    0.000
Verb          0.1163    0.0214     5.43    0.000  130.12
Math          0.0912    0.0282     3.23    0.003  145.29
Verb*Verb  -0.001127  0.000117    -9.66    0.000   80.44
Math*Math  -0.000957  0.000163    -5.86    0.000  108.39
Verb*Math   0.000994  0.000149     6.66    0.000   47.46

Regression Equation

Gpa = -5.81 + 0.1163 Verb + 0.0912 Math - 0.001127 Verb*Verb - 0.000957 Math*Math
        + 0.000994 Verb*Math
```

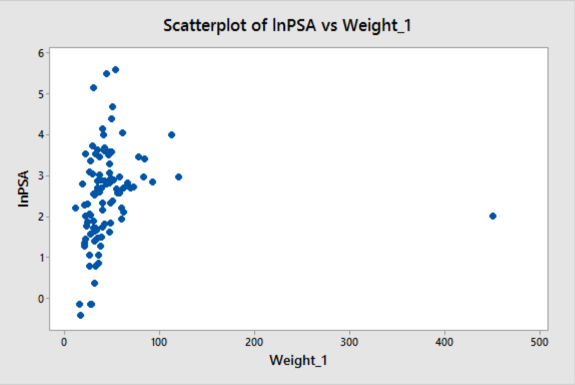
Removing observation for student 9 has very little effect on the R^2 value, still a strong relationship. It had a noticeable impact on the estimated coefficient b_0, from -7.22 in the original model to -5.81 in this model. The value for b_1 also saw a relatively significant decrease. The MSE of this model is lower than the original, 0.03467 compared to 0.04137 of the original. So this model would show narrower confidence/prediction intervals.

Overall, the impacts of removing these observations seem small. They show some improvement, but none that look very significant to me. The difference in MSE by removing observation 9 was the most significant difference in my opinion.

0.0.3 Question 3

a) The scatterplot of lnPSA vs lnCanVol shows a positive relationship between lnPSA and lnCanVol. All the data points seem to follow the same general trend. The scatterplot of lnPSA vs Weight also shows a positive relationship. This plot, however, shows a handful of points that appear to stray from the general trend. At least one definite outlier.

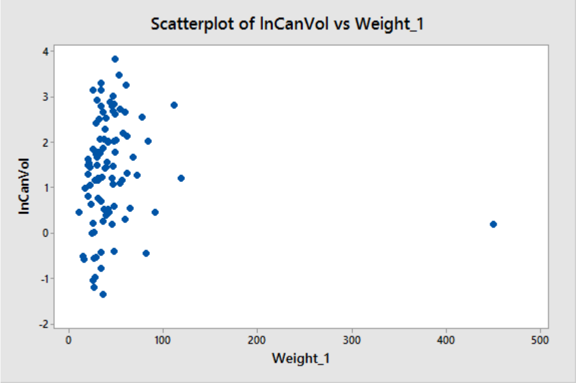




b)

Predictor	Coefficient value	Standard error	p-value
lnCanVol	0.7183	0.0675	0.000
Weight	0.00307	0.00174	0.081

c) Plot of X_1 vs X_2:



The majority of the observations flagged by Minitab are marked with 'R' indicating an internally studentized residual greater than 2. These are "unusual" y values that could potentially be outliers. The observation marked with an 'R' and an 'X' indicates an "usual" y value AND an "unusual" x value. Indeed, we can see in all the plots that this observation has a weight = 450 which is much larger than all other weight values. Of the flagged observations, the only clear outlier is observation 32. The others are potential outliers that require further investigation.

d) DFITS threshold: $2\sqrt{\frac{p+1}{n-p-1}} = 2\sqrt{\frac{3+1}{97-3-1}} = 0.4148$
Observations with Cook's distance greater than 1:

```
Observation 32
```

Observations with DFITS greater than 0.4148:

```
Observation 32
Observation 69
Observation 96
Observation 97
```

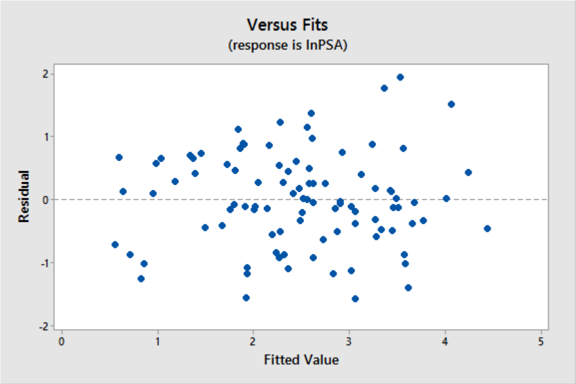
Of these observations, the only one I would consider deleting is observation 32. It is clearly an outlier by any measure. By deleting this observation, we would be narrowing down the scope of the model. The other observations with high DFITS values are over the threshold, but do not stick out like sore thumbs. There are other values just below that threshold as well, so I do not think they warrant any action.

e)

Predictor	Coefficient value	Standard error	p-value
lnCanVol	0.6711	0.0670	0.000
Weight	0.01393	0.00412	0.001

The model computed after removing observation 32 had a small effect on the value of the lnCanVol coefficient. The standard error and p-value remained the same, however. The change to the coefficient on Weight was more significant, from 0.00307 to 0.01393. The p-value for the Weight coefficient also changed significantly. Under the original model, the p-value would have indicated that $\beta_2 = 0$. The model excluding observation 32 has changed the outcome of that hypothesis test.

f) The plot of residuals vs fits for the model in part (e) does not indicate any difficulties. It has a good horizontal band shape around the residual = 0 line. The spread of points around the line indicate constant and even distribution of errors.



g) DFITS threshold: $2\sqrt{\frac{p+1}{n-p-1}} = 2\sqrt{\frac{3+1}{96-3-1}} = 0.417$

Observations with Cook's distance greater than 1:

```
None
```

Observations with DFITS greater than 0.417:

```
Observation 69
Observation 95
Observation 96
Observation 97
```

Of the observations with DFITS values over the threshold, I tried fitting a model removing observation 69, and it didn't seem to have a positive effect on the model. The t-value of the Weight coefficient actually went down a little. As I said in part (d), I don't think that any of the flagged observations warrant any action. The residuals vs fits plot doesn't indicate any difficulties with the model.