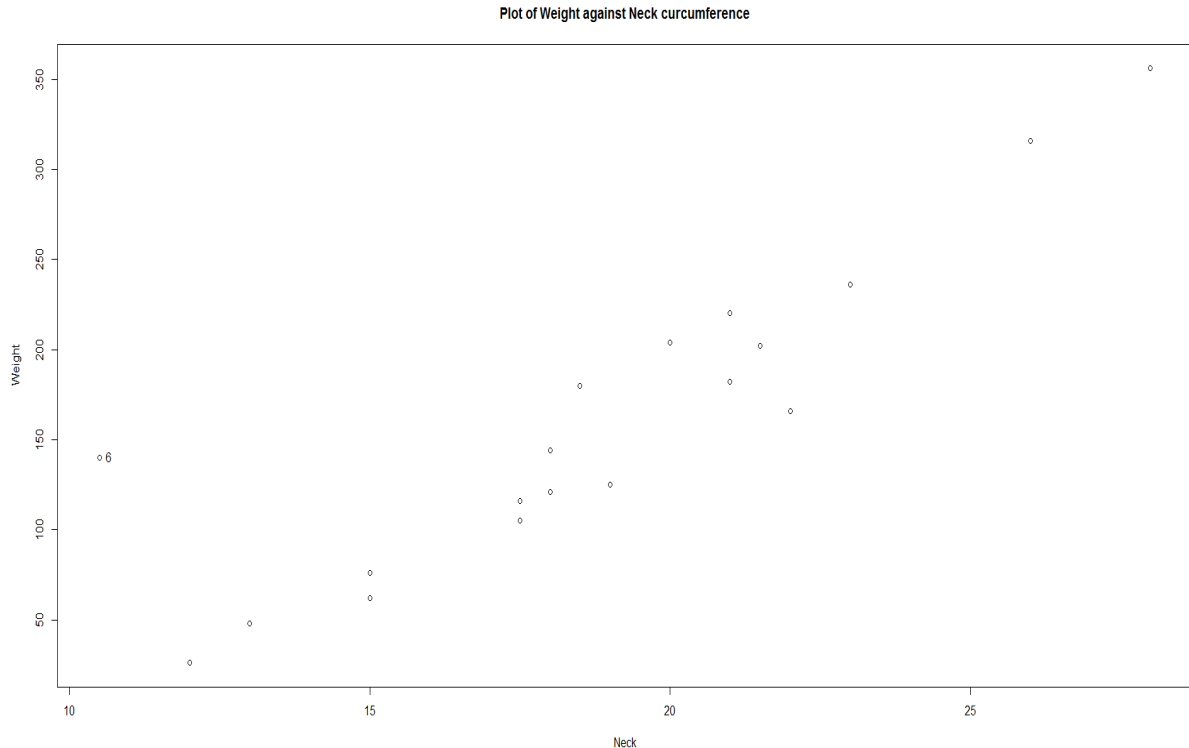


Stat 6021: Homework Set 10

1. In this question, you will revisit the `swiss` data set that you worked on in Homeworks 6 and 7. The data set contains information regarding a standardized fertility measure and socio-economic indicators for each of the 47 French-speaking provinces of Switzerland around the year 1888. In Homework 5, you found that the model with just three predictors: *Education*, *Catholic*, and *Infant Mortality* was preferred to a model with all the predictors. Fit the model with the three predictors, and answer the following questions.
 - (a) Are there any observations that are outlying? Be sure to show your work and explain how you arrived at your answer.
 - (b) Are there any observations that have high leverage? Be sure to show your work and explain how you arrived at your answer.
 - (c) Are there any influential observations based on DFFITs and Cook's Distance?
 - (d) Briefly describe the difference in what DFFITS and Cook's distance are measuring.

2. (You may only use R as a simple calculator or to find p-values or critical values)
Data from $n = 19$ bears of varying ages are used to develop an equation for estimating *Weight* from *Neck* circumference. From a visual inspection of the scatterplot, it appears observation 6 may be an outlier.



The output below comes from fitting the linear regression model on the data.

```
##with all 19 bears
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-158.78	40.46	-3.924	0.00109	**
Neck	16.95	2.10	8.071	3.24e-07	***

```
Residual standard error: 40.13 on 17 degrees of freedom
```

```
Multiple R-squared: 0.793, Adjusted R-squared: 0.7809
```

```
F-statistic: 65.14 on 1 and 17 DF, p-value: 3.235e-07
```

The output below comes from fitting the linear regression model on the data, with the outlier removed.

```
##with outlier removed, so 18 bears
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-234.60	25.93	-9.049	1.08e-07	***
Neck	20.54	1.32	15.562	4.39e-11	***

Residual standard error: 22.6 on 16 degrees of freedom
Multiple R-squared: 0.938, Adjusted R-squared: 0.9342
F-statistic: 242.2 on 1 and 16 DF, p-value: 4.394e-11

The output below displays the values of the predictor and response for the 6th observation.

```
> data[6,]
      Neck Weight
6  10.5      140
```

Some additional information from R, regarding ordinary residuals, e_i , and leverages, h_{ii} shown below, from the full data.

```
> result$residuals ##residuals
      1      2      3      4      5      6      7
-25.276933 -48.066801 22.880666 23.828133 -2.276933 120.829070 -32.803200
      8      9     10     11     12     13     14
-18.592131 -38.224400 25.249333 -21.803200 -15.119334 40.248397 34.143331
     15     16     17     18     19
-3.593068 -33.434532 4.985732 -19.434532 -13.539598

> tmp$hat ##leverages
      1      2      3      4      5      6      7
0.05422642 0.08132161 0.06633278 0.05682064 0.05422642 0.23960510 0.05700079
      8      9     10     11     12     13     14
0.17788427 0.05278518 0.05282121 0.05700079 0.06633278 0.28626504 0.19604381
     15     16     17     18     19
0.07314261 0.09141025 0.10178713 0.09141025 0.14358291
```

- Calculate the externally studentized residual, t_i , for observation 6. Will this be considered outlying in the response?
- What is the leverage for observation 6? Based on the criterion that leverages greater than $\frac{2p}{n}$ are considered outlying in the predictor(s), is this observation high leverage?
- Calculate the DFFITS for observation 6. Briefly describe the role of leverages in DFFITS.
- Calculate Cook's distance for observation 6.

3. (No R Required) Cook's distance has the equivalent formulae

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' (\mathbf{X}' \mathbf{X}) (\hat{\beta} - \hat{\beta}_{(i)})}{p \text{MSres}} \quad (1)$$

$$= \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}. \quad (2)$$

where r_i denotes studentized residuals. Show that (1) and (2) are equivalent. You may use the following without proof:

$$\hat{\beta} - \hat{\beta}_{(i)} = (1 - h_{ii})^{-1} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_i e_i. \quad (3)$$

4. **This question is optional** (No R Required) Recall that leverage, h_{ii} , is defined by

$$h_{ii} = \mathbf{X}_i' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}_i,$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ 1 & x_{2,1} & \cdots & x_{2,k} \\ \vdots & & & \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{bmatrix}$$

and

$$\mathbf{X}_i = \begin{bmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,k} \end{bmatrix}.$$

\mathbf{X} is called the design matrix, containing the values of the k predictors for all observations, with 1s appended in the first column for the intercept, and \mathbf{X}_i is a column vector that contains the values of the k predictors for observation i , with 1 appended in the first entry. It turns out that the sum of the leverages is

$$\sum_{i=1}^n h_{ii} = p, \quad (4)$$

where p denotes the number of parameters in the regression model.

(a) Show that (4) is true in simple linear regression, i.e. when $p = 2$. In case you forgot, the inverse of a 2×2 matrix is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Hint: In your intermediate step, you will need to show that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}$$

(b) Show that

$$\sum_{i=1}^n \sigma^2 \{\hat{y}_i\} = p\sigma^2. \quad (5)$$

In other words, that the sum of the variances of \hat{y}_i is $p\sigma^2$. Hint: how are \hat{y}_i and y_i related? You may also use (4).