Stat 6021: Homework Set 10 Solutions

1. (a) There are no outliers in the response variable, since none of the externally studentized residuals is greater (in magnitude) than the critical value of $t_{1-0.05/(2n),n-1-p} = 3.516461$.

```
> ## Fit the regression model
> result<-lm(Fertility~ Education+ Catholic+ Infant.Mortality)
>
> ## Obtain externally studentized residuals
> ex.student.res<-rstudent(result)
>
> ## Find the observations that are outlying in response
> n<-length(Infant.Mortality)
> n
[1] 47
> p<-4
> ex.student.res[abs(ex.student.res)>qt(1-0.05/(2*n), n-1-p)]
named numeric(0) ##No observation has ti greater than critical
##value
> ##critical value using Bonferroni procedure
> qt(1-0.05/(2*n), n-1-p)
[1] 3.516461
```

   (b) There are two observations with high leverages: observations 19 and 45. Observations with leverage higher than $\frac{2p}{n} = \frac{2\times4}{47} = 0.1702128$ have high leverages.

```
> lev<-lm.influence(result)$hat
>
> ## Find the observations with high leverage
> lev[lev>2*p/n]
       19        45
0.2461056 0.4501392
> 2*p/n
[1] 0.1702128
```

   (c) There are three observations with high DFFITS: observations 6, 37, 47. DFFITS greater than $2\sqrt{p/n} = 2\sqrt{4/47} = 0.58346$, in magnitude, are influential.

```
> DFFITS[abs(DFFITS)>2*sqrt(p/n)]
         6          37          47
-0.6400846  0.8551451 -0.7437332
> 2*sqrt(p/n)
[1] 0.58346
```

There are no observations that are influential based on Cook's distance. The critical value is $F_{0.5,4,43} = 0.8525511$.

```
> COOKS[COOKS>qf(0.5,p,n-p)]
named numeric(0)
> qf(0.5,p,n-p)
[1] 0.8525511
```

(d) DFFITS measures how removing an observation changes its predicted value. Cook's distance measures how removing an observation changes the predicted values for all the observations. Compare (6.6) with (6.9) from the textbook.

2. (a) There are a few ways to derive the externally studentized residual for observation 6, t_6:

Version 1: Use equation (4.13) from textbook

$$\begin{aligned} t_6 &= \frac{e_6}{\sqrt{S^2_{(6)}(1-h_{ii})}} \\ &= \frac{120.829070}{\sqrt{22.6^2(1-0.23960510)}} \\ &= 6.131171 \end{aligned}$$

Note that $S^2_{(6)}$ is the residual standard error squared of the model with observation 6 removed, so this is 22.6^2.

Version 2: Use equation (4.12) from textbook to find $S^2_{(6)}$.

$$\begin{aligned} S^2_{(6)} &= \frac{(n-p)MSres - e_6^2/(1-h_{66})}{n-p-1} \\ &= \frac{(19-2)40.13^2 - 120.829070^2/(1-0.23960510)}{19-2-1} \\ &= 511.0612. \end{aligned}$$

MSres is the residual standard error squared of the model with all observations, 40.13^2. Therefore,

$$\begin{aligned} t_6 &= \frac{e_6}{\sqrt{S^2_{(6)}(1-h_{ii})}} \\ &= \frac{120.829070}{\sqrt{511.0612(1-0.23960510)}} \\ &= 6.129363 \end{aligned}$$

Version 3: Sub in equation (4.12) into (4.13), and get

$$\begin{aligned} t_6 &= e_6 \left[\frac{n-1-p}{SS_{res}(1-h_{66}) - e_6^2} \right]^{1/2} \\ &= 120.829070 \left[\frac{19-1-2}{27377.09(1-0.23960510) - 120.829070^2} \right]^{1/2} \\ &= 6.129. \end{aligned}$$

since

$$\begin{aligned} SS_{res} &= (n-p)MSres \\ &= (19-2) \times 40.13^2 \\ &= 27377.09. \end{aligned}$$

Differences in final numerical answers due to rounding off in output from R.

t_6 is greater than $t_{1-0.05/38;19-1-2} = 3.556242$, in magnitude. So observation 6 is an outlier in the response.

(b) The leverage, h_{66}, is 0.23960510. Since $\frac{2p}{n} = \frac{2\times 2}{19} = 0.2105263$, it's an outlier in the predictor.

(c) Two ways to get the answer for $(\text{DFFITS})_6$:
Version 1: Use equation (6.10) from textbook

$$\begin{aligned} (\text{DFFITS})_6 &= t_6 \left(\frac{h_{66}}{1-h_{66}} \right)^{1/2} \\ &= 6.129 \times \left(\frac{0.23960510}{1-0.23960510} \right)^{1/2} \\ &= 3.440472. \end{aligned}$$

Version 2: Use equation (6.9) from textbook

$$
\begin{aligned}
(\text{DFFITS})_6 &= \frac{\hat{y}_6 - \hat{y}_{(6)}}{\sqrt{S^2_{(6)} h_{66}}} \\
&= \frac{(-158.78 + 16.96 \times 10.5) - (-234.60 + 20.54 \times 10.5)}{\sqrt{22.6^2 \times 0.23960510}} \\
&= 3.446302.
\end{aligned}
$$

Differences in final numerical answers due to rounding off in output from R.

Looking at (6.10), as leverage increases, DFFITS increases. This means that if an observation is far away from the center of the predictors, the larger the difference in predicted values with and without that observation in the regression model.

(d) Two ways to find Cook's distance for observation 6, D_6:

Version 1: Use equation (6.5) from textbook

$$
\begin{aligned}
D_6 &= \frac{r_6^2}{p} \frac{h_{66}}{1 - h_{66}} \\
&= \frac{3.452889^2}{2} \frac{0.23960510}{1 - 0.23960510} \\
&= 1.878418
\end{aligned}
$$

where

$$
\begin{aligned}
r_6 &= \frac{e_6}{\sqrt{MSres(1 - h_{66})}} \\
&= \frac{120.829070}{\sqrt{40.13^2(1 - 0.23960510)}} \\
&= 3.452889
\end{aligned}
$$

using equation (4.8).

Version 2: Plug in (4.8) into (6.5) and get

$$
\begin{aligned}
D_6 &= \frac{e_6^2}{pMS_{res}} \left[\frac{h_{66}}{(1 - h_{66})^2} \right] \\
&= \frac{120.829070^2}{2 * 40.13^2} \frac{0.23960510}{(1 - 0.23960510)^2} \\
&= 1.878.
\end{aligned}
$$

3. Recall leverage

$$h_{ii} = \boldsymbol{X_i}' \left(\boldsymbol{X'X} \right)^{-1} \boldsymbol{X_i}$$

and the leave-one-out formula

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\boldsymbol{i})} = (1 - h_{ii})^{-1} (\boldsymbol{X'X})^{-1} \boldsymbol{X_i} e_i.$$

Therefore, Cook's distance is

$$\begin{aligned}
D_i &= \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\boldsymbol{i})}\right)' (\boldsymbol{X'X}) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\boldsymbol{i})}\right)}{p\text{MSres}} \\
&= (1 - h_{ii})^{-2} \left[(\boldsymbol{X'X})^{-1} \boldsymbol{X_i} e_i \right]' (\boldsymbol{X'X}) \left[(\boldsymbol{X'X})^{-1} \boldsymbol{X_i} e_i \right] / p\text{MSres} \\
&= (1 - h_{ii})^{-2} \left[e_i \boldsymbol{X_i}' (\boldsymbol{X'X})^{-1} (\boldsymbol{X'X}) (\boldsymbol{X'X})^{-1} \boldsymbol{X_i} e_i \right] / p\text{MSres} \\
&= (1 - h_{ii})^{-2} \left[e_i \boldsymbol{X_i}' (\boldsymbol{X'X})^{-1} \boldsymbol{X_i} e_i \right] / p\text{MSres} \\
&= (1 - h_{ii})^{-2} e_i^2 h_{ii} / p\text{MSres} \\
&= \frac{e_i^2}{p\text{MSres}} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right] \\
&= \frac{h_{ii} r_i^2 \text{MSres}(1 - h_{ii})}{p\text{MSres}(1 - h_{ii})^2} \\
&= \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}.
\end{aligned}$$

since $e_i = r_i \times \sqrt{\text{MSres}(1 - h_{ii})}$ and $h_{ii} = \boldsymbol{X_i}' (\boldsymbol{X'X})^{-1} \boldsymbol{X_i}$.

4. **Note: Optional question**

   (a) i. First we establish the following:

$$\sum_{i=1}^{n} x_i = n\bar{x}.$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

$$\begin{aligned}
n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 &= n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2 \\
&= n \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \\
&= n \sum_{i=1}^{n} (x_i - \bar{x})^2.
\end{aligned}$$

For $p = 2$,

$$\begin{aligned}
\boldsymbol{X'X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \cdot \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\
&= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.
\end{aligned}$$

Therefore,

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

We see that

$$\begin{aligned}
\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} &= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2 + n\bar{x}^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

$$\begin{aligned}
\frac{-\sum_{i=1}^n x_i}{n\sum_{i=1}^n (x_i - \bar{x})^2} &= \frac{-n\bar{x}}{n\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

$$\frac{n}{n\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Therefore,

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

Leverages are

$$
\begin{aligned}
h_{ii} &= \left[\begin{array}{cc} 1 & x_i \end{array}\right] . \left[\begin{array}{cc} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{array}\right] . \left[\begin{array}{c} 1 \\ x_i \end{array}\right] \\
&= \frac{1}{n} + \frac{\bar{x}^2 - \bar{x}x_i - \bar{x}x_i + x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}
$$

Summing up the leverages when $p = 2$,

$$
\begin{aligned}
\sum_{i=1}^n h_{ii} &= \sum_{i=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= 1 + \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= 2.
\end{aligned}
$$

ii. We know that $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$, and that $\sigma^2\{\boldsymbol{Y}\} = \sigma^2\boldsymbol{I}$, so $\sigma^2\{\boldsymbol{HY}\} = \boldsymbol{H'H}\sigma^2 = \boldsymbol{H}\sigma^2$, since $\boldsymbol{H}$ is idempotent. Therefore, $\sigma^2\{\hat{\boldsymbol{Y}}\} = \boldsymbol{H}\sigma^2$, and $\sigma^2\{\hat{y}_i\} = h_{ii}\sigma^2$.

So $\sum_{i=1}^n \sigma^2\{\hat{y}_i\} = \sum_{i=1}^n h_{ii}\sigma^2 = \sigma^2 \sum_{i=1}^n h_{ii} = p\sigma^2$.