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 3.

```
> ## 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
>
> ex.student.res[abs(ex.student.res)>3]
named numeric(0)
```

   (b) There are two observations with high leverages: La Vallee and V. de Geneve. Observations with leverage higher than $\frac{2p}{n} = \frac{2\times 4}{47} = 0.1702128$ have high leverages.

```
> ## Obtain leverages
> lev<-lm.influence(result)$hat
>
> n<-nrow(data)
> p<-4
>
> ## Find the observations with high leverage along with their index and row na
> lev[lev>2*p/n]
   La Vallee V. De Geneve
   0.2461056    0.4501392
> 2*p/n
[1] 0.1702128
```

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

```
> DFFITS<-dffits(result)
> DFFITS[abs(DFFITS)>2*sqrt(p/n)]
 Porrentruy      Sierre Rive Gauche
 -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. We deem observations with Cook's distance greater than 1 to be influential.

```
> COOKS[COOKS>1]
named numeric(0)
```

(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 equations (14) with (16) from the notes.

2. (a) There are a few ways to derive the externally studentized residual for observation 6, t_6:
Version 1: Use equation (10) from course notes

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

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: Sub in equation (11) from course notes

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

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 3, 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 (15) from course notes

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

Version 2: Use equation (14) from course notes

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

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

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)

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

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 equations (17) and (7) from Module 10 course notes.

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

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

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