Homework10

Hyun Suk (Max) Ryoo (hr2ee)

11/11/2021

Set up

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(datasets)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
data <- swiss
head(data)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
```

```
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

1 - A

```
## Model Specification
model <- lm(Fertility ~ Education + Catholic + Infant.Mortality, data=data)

## Externally Studentized Residuals
ext.student.res<-rstudent(model)

## Critical Value
n<-dim(data)[1]
p<- 4
crit<-qt(1-0.05/(2*n), n-1-p)

##identify
ext.student.res[abs(ext.student.res)>crit]
```

```
## named numeric(0)
```

To find observations that are outlying we can utilize the externally studentized resiudals. We can check if there are outliers in the observations by checking if the any externally studentized residuals is greater than the critical value. The above code checks that criteria and outputs nothing, which means that are no outliers for the observations.

1 - B

```
##leverages
lev<-lm.influence(model)$hat
## boundary
print(2*p/n)
```

```
## [1] 0.1702128
```

```
##identify
lev[lev>2*p/n]
```

```
##    La Vallee V. De Geneve
##    0.2461056    0.4501392
```

We can find values in our dataset that have a high leverage utilizing the above code. Some text books have different thresholds, but for our clas any leverage points that is greater than 2*p/n will be classified as leverage points. From the above output we can see that there are two outputs that have a high leverage point (La Vallee, V. De Geneve). As we can see the boundary for classification of leverage points was 0.1702128, but La Vallee had a leverage of 0.2461056 and V. De Geneve had a leverage of 0.4501392. Therefore, we can state these two observations have a high leverage.

1 - C

```
print(2*sqrt(p/n))
```

```
## [1] 0.58346
```

```
DFFITS<-dffits(model)
DFFITS[abs(DFFITS)>2*sqrt(p/n)]
```

```
##  Porrentruy      Sierre Rive Gauche
##  -0.6400846   0.8551451  -0.7437332
```

There are three observaions that are classified as influential based on the DFFITS. The critera utilized was $2\sqrt{\frac{p}{n}} = 0.58346$. The Porrentruy, Sierre, and Rive Gauche were all influential because their statistic was $-0.6400846,0.8551451, and -0.7437332 $ respectivly.

```
print(qf(0.5,p,n-p))
```

```
## [1] 0.8525511
```

```
COOKS<-cooks.distance(model)
COOKS[COOKS>qf(0.5,p,n-p)]
```

```
## named numeric(0)
```

Intersetingly enough, the cooks distance method didn't show any influential observations. The criteria utilized for cooks distance was 0.8525511.

1 - D

Cook's distance measure is a deletion diagnostic, which means that the influence of the ith observation if it is removed form the sample. DFFITS is the number of standard deviations that the fitted value changes if the observation is removed. This can be summarized by saying DFFITS measures how a fitted value will change while the Cook's distance measures how all the fitted values will change.

2 - A

The externally studentized residual can be found with the following formula

$t_i = \frac{e_i}{\sqrt{S_i^2(1-h_{ii})}}, i = 1, 2, ..., n$

Since we are looking at observation 6, the equation will be like such. $t_i = \frac{e_6}{\sqrt{S_6^2(1-h_{ii})}}, i = 1, 2, ..., n$

Solving this equation will be shown below

$$\begin{aligned} t_i &= \frac{e_6}{\sqrt{S_6^2(1-h_{ii})}} \\ &= \frac{120.829070}{\sqrt{22.6^2(1-0.23960510)}} \\ &= \frac{120.829070}{\sqrt{388.3793}} \\ &= \frac{120.829070}{19.70734} \\ t_6 &= 6.131171 \end{aligned}$$

We can utilize the bonferroni-type approcach and compare with the value of $t_{(1-\frac{0.05}{38}),19-2-1} = 3.556242$. Since 6.131171 is greater than 3.556242 we can say observation is an outlier in the response.

2 - B

The leverage for observation 6 is given, which is 0.23960510. The criterion is $\frac{2p}{n} = \frac{2*2}{19} = 0.2105263$. Since the given is greater than the criteria, we can say this is a high leverage point and and outlier in the predictor.

2 - C

DFFITS can be found utilizing the following.

$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5} t_i$ where t_i is R-student.

$$\begin{aligned} DFFITS_i &= \left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5} t_i \\ &= \left(\frac{0.23960510}{1-0.23960510}\right)^{0.5} * 6.131171 \\ &= 0.3151061^{0.5} * 6.131171 \\ &= 0.5613431 * 6.131171 \\ &= 3.441691 \end{aligned}$$

The role in leverages can be clearly seen by how the DFFITS was calcualted. $(leverage/1 - leverage)^{0.5}$. This expression shows that if leverages increases the DFFITS also increases.

2 - D

Cook's distance of Observation i can be found like such.

$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$ where r_i is the studentized residuals.

Using this equation the cooks distance for observation 6 is the following.

$$\begin{aligned} D_i &= \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}} \\ &= \frac{\left(\frac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}\right)^2}{p}\frac{h_{66}}{1-h_{66}} \\ &= \frac{\left(\frac{120.829070}{\sqrt{40.13^2(1-0.23960510)}}\right)^2}{2}\frac{0.23960510}{1-0.23960510} \\ &= \frac{\left(\frac{120.829070}{34.99361}\right)^2}{2}0.3151061 \\ &= \frac{11.92245}{2} * 0.3151061 \\ &= 5.961225 * 0.3151061 \\ &= 1.878418 \end{aligned}$$

3

$$
\begin{aligned}
D_i &= \frac{\left(\hat{\beta}-\hat{\beta}_{(i)}\right)'(X'X)\left(\hat{\beta}-\hat{\beta}_{(i)}\right)}{pMSres} \\
&= \frac{\left((1-h_{ii})^{-1}(X'X)^{-1}X_ie_i\right)'(X'X)\left((1-h_{ii})^{-1}(X'X)^{-1}X_ie_i\right)}{pMSres} \\
&= \frac{(1-h_{ii})^{-2}\left(\left((X'X)^{-1}X_ie_i\right)'(X'X)\left((X'X)^{-1}X_ie_i\right)\right)}{pMSres} \\
&= \frac{(1-h_{ii})^{-2}\left(e_iX_i'(X'X)^{-1}(X'X)(X'X)^{-1}X_ie_i\right)}{pMSres} \\
&= \frac{(1-h_{ii})^{-2}\left(e_iX_i'(X'X)^{-1}X_ie_i\right)}{pMSres} \\
&= \frac{(1-h_{ii})^{-2}\left(e_i^2X_i'(X'X)^{-1}X_i\right)}{pMSres} \\
&= \frac{(1-h_{ii})^{-2}\left(e_i^2X_i'(X'X)^{-1}X_i\right)}{pMSres} \\
&= \frac{(1-h_{ii})^{-2}\left(e_i^2h_{ii}\right)}{pMSres} \text{ due to } h_{ii}=X_i'(X'X)^{-1}X_ie_i \\
&= \frac{e_i^2h_{ii}}{(1-h_{ii})^2pMSres} \\
&= \frac{(r_i\sqrt{MSres(1-h_{ii})})^2h_{ii}}{(1-h_{ii})^2pMSres} \text{ due to } r_i=\frac{e_i}{\sqrt{MSres(1-h_{ii}}} \\
&= \frac{r_i^2MSres(1-h_{ii})h_{ii}}{(1-h_{ii})^2pMSres} \\
&= \frac{r_i^2(1-h_{ii})h_{ii}}{(1-h_{ii})^2p} \\
&= \frac{r_i^2h_{ii}}{(1-h_{ii})p} \\
&= \frac{r_i^2h_{ii}}{p(1-h_{ii})} \\
&= \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}
\end{aligned}
$$