uwa6xv_M10_HW

Alanna Hazlett

2024-04-09

Problem 1

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.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
Data<-swiss
result<-lm(Fertility~Education+Catholic+Infant.Mortality,data=Data)
```

(a)
Are there any observations that are outlying? Be sure to show your work and explain how you arrived at your answer.

```
#residuals
round(sort(abs(result$res)),4)
```

```
##      Payerne     Avenches       Morges      Conthey V. De Geneve        Broye
##       0.2461       0.5143       0.5823       0.8092       0.8352       0.9288
##     Lausanne      Aubonne      Veveyse ValdeTravers Paysd'enhaut        Aigle
##       0.9363       1.4369       1.7497       1.8682       2.0241       2.3289
##      Gruyere        Rolle La Chauxdfnd      Monthey     Grandson         Oron
##       2.4358       2.4535       2.5212       2.6171       2.8569       2.8675
##       Lavaux       Herens       Sarine       Boudry     Delemont        Vevey
##       2.9397       3.1847       3.6886       3.9822       4.3314       4.8143
##        Nyone    La Vallee      Yverdon         Orbe     Cossonay         Sion
##       6.0662       6.6029       6.9524       6.9561       7.6897       7.7308
##   Val de Ruz        Glane     Martigwy     Le Locle    Neuchatel    Echallens
##       7.8372       8.1890       8.2736       8.3181       8.5819       8.6617
##   St Maurice    Entremont  Rive Droite   Porrentruy       Moudon   Courtelary
##       9.4316      10.0612      10.3935      10.4406      10.8673      10.9026
```

```
##   Neuveville Franches-Mnt      Moutier  Rive Gauche       Sierre
##      12.4153      12.4644      12.8817      14.4781      15.1187
```

```
#studentized residuals
round(sort(abs(rstandard(result))),4)
```

```
##      Payerne     Avenches       Morges      Conthey        Broye     Lausanne
##       0.0344       0.0709       0.0794       0.1178       0.1290       0.1312
## V. De Geneve      Aubonne      Veveyse ValdeTravers Paysd'enhaut        Aigle
##       0.1501       0.1962       0.2461       0.2542       0.2794       0.3199
##      Gruyere        Rolle La Chauxdfnd      Monthey     Grandson         Oron
##       0.3352       0.3381       0.3416       0.3619       0.3893       0.3973
##       Lavaux       Herens       Sarine       Boudry     Delemont        Vevey
##       0.4003       0.4455       0.5173       0.5412       0.5930       0.6565
##        Nyone      Yverdon         Orbe    La Vallee     Cossonay   Val de Ruz
##       0.8303       0.9570       0.9757       1.0132       1.0539       1.0684
##         Sion     Le Locle     Martigwy        Glane    Echallens    Neuchatel
##       1.0737       1.1275       1.1429       1.1557       1.1847       1.2513
##   St Maurice    Entremont  Rive Droite   Courtelary   Porrentruy       Moudon
##       1.3136       1.3889       1.4671       1.4919       1.5077       1.5080
##   Neuveville Franches-Mnt      Moutier  Rive Gauche       Sierre
##       1.6908       1.7130       1.7396       2.0437       2.1544
```

```
#externally studentized residuals
round(sort(abs(rstudent(result))),4)
```

```
##      Payerne     Avenches       Morges      Conthey        Broye     Lausanne
##       0.0340       0.0700       0.0785       0.1164       0.1275       0.1297
## V. De Geneve      Aubonne      Veveyse ValdeTravers Paysd'enhaut        Aigle
##       0.1483       0.1940       0.2434       0.2514       0.2764       0.3166
##      Gruyere        Rolle La Chauxdfnd      Monthey     Grandson         Oron
##       0.3317       0.3346       0.3380       0.3582       0.3855       0.3934
##       Lavaux       Herens       Sarine       Boudry     Delemont        Vevey
##       0.3964       0.4413       0.5128       0.5367       0.5885       0.6521
##        Nyone      Yverdon         Orbe    La Vallee     Cossonay   Val de Ruz
##       0.8273       0.9561       0.9752       1.0135       1.0553       1.0702
##         Sion     Le Locle     Martigwy        Glane    Echallens    Neuchatel
##       1.0756       1.1312       1.1471       1.1604       1.1905       1.2599
##   St Maurice    Entremont  Rive Droite   Courtelary   Porrentruy       Moudon
##       1.3251       1.4045       1.4876       1.5141       1.5311       1.5314
##   Neuveville Franches-Mnt      Moutier  Rive Gauche       Sierre
##       1.7295       1.7539       1.7832       2.1257       2.2544
```

- Based on the residuals it appears that potentially Neuveville, Franches-Mnt, Moutier, Rive Gauche, and Sierre are outliers, as their values are larger than the other observations. There is about a 2 unit increase from Moudon to NeuThere. There is stronger evidentce still that Rive Gauche and Sierre are outliers as they are additional 2 units larger than Moutier. However the residuals are difficult to determine what is considered a large enough value to be an outlier as their units reflect the response variable.

- Based on the studentized residuals it appears that Rive Gauche and Sierre are outliers, as their studentized residuals are greater than 2. This is larger than the next closest Moutier of 1.74. This indicates that these observations are greater than 2 standard deviations their predicted response is from their actual response.

- Based on the externally studentized residuals it appears that while Rive Gauche and Sierre are less than 3, our guideline for this statistic, they are notably larger than the other observations.

In conclusion from these three statistics it appears that Rive Gauce and Sierre are both outliers.

(b)

Are there any observations that have high leverage? Be sure to show your work and explain how you arrived at your answer.

```
round(sort(lm.influence(result)$hat),4)
```

```
##      Moutier La Chauxdfnd     Le Locle       Boudry ValdeTravers       Lavaux
##       0.0266       0.0329       0.0338       0.0387       0.0409       0.0428
##   Neuveville     Grandson   Val de Ruz       Morges        Vevey      Aubonne
##       0.0429       0.0442       0.0447       0.0450       0.0454       0.0477
##    Echallens   Courtelary        Nyone     Delemont     Cossonay        Aigle
##       0.0511       0.0519       0.0524       0.0530       0.0549       0.0592
## Franches-Mnt      Gruyere      Yverdon        Rolle     Avenches    Entremont
##       0.0601       0.0624       0.0632       0.0650       0.0654       0.0684
## Paysd'enhaut     Martigwy      Monthey         Oron       Moudon        Broye
##       0.0684       0.0697       0.0714       0.0754       0.0781       0.0793
##         Sion   St Maurice      Payerne       Herens     Lausanne       Sarine
##       0.0797       0.0848       0.0887       0.0929       0.0959       0.0973
##         Orbe      Veveyse        Glane  Rive Droite  Rive Gauche       Sierre
##       0.0977       0.1028       0.1087       0.1090       0.1091       0.1258
##   Porrentruy      Conthey    Neuchatel    La Vallee V. De Geneve
##       0.1488       0.1625       0.1650       0.2461       0.4501
```

```
#guideline
n<-nrow(Data)
p<-4
hii<-(2*p/n)
print(hii)
```

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

- Our guideline for leverage (hii) is 0.17. We see that La Vallee and V. De Geneve are over this guideline, additionally they are pretty distinctly higher than the other observations.

(c)

Are there any influential observations based on DFFITs and Cook's Distance?

```
round(sort(abs(dffits(result))),4)
```

```
##      Payerne       Morges     Avenches        Broye     Lausanne      Aubonne
##       0.0106       0.0170       0.0185       0.0374       0.0422       0.0434
##      Conthey ValdeTravers La Chauxdfnd Paysd'enhaut        Aigle      Veveyse
##       0.0513       0.0519       0.0623       0.0749       0.0794       0.0824
##     Grandson       Lavaux      Gruyere        Rolle      Monthey       Boudry
##       0.0828       0.0839       0.0856       0.0882       0.0993       0.1077
##         Oron V. De Geneve     Delemont       Herens        Vevey       Sarine
##       0.1124       0.1342       0.1392       0.1412       0.1423       0.1684
##        Nyone     Le Locle   Val de Ruz      Yverdon     Cossonay    Echallens
##       0.1946       0.2117       0.2315       0.2482       0.2544       0.2763
##      Moutier     Martigwy         Sion         Orbe   Courtelary   Neuveville
##       0.2949       0.3141       0.3164       0.3209       0.3544       0.3659
##    Entremont   St Maurice        Glane Franches-Mnt       Moudon  Rive Droite
##       0.3806       0.4034       0.4053       0.4435       0.4458       0.5203
```

```
##    Neuchatel    La Vallee   Porrentruy  Rive Gauche       Sierre
##       0.5601       0.5791       0.6401       0.7437       0.8551
```

```
dffits_guideline<-2*(sqrt(p/n))
print(dffits_guideline)
```

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

* Based on our guideline for dffits we see that Porrentruy, Rive Gauche, and Sierre are influential.

```
#get rid of scientific notation
options(scipen=999)
round(sort(cooks.distance(result)),4)
```

```
##      Payerne       Morges     Avenches        Broye     Lausanne      Aubonne
##       0.0000       0.0001       0.0001       0.0004       0.0005       0.0005
##      Conthey ValdeTravers La Chauxdfnd Paysd'enhaut        Aigle      Veveyse
##       0.0007       0.0007       0.0010       0.0014       0.0016       0.0017
##     Grandson       Lavaux      Gruyere        Rolle      Monthey       Boudry
##       0.0018       0.0018       0.0019       0.0020       0.0025       0.0029
##         Oron V. De Geneve     Delemont       Herens        Vevey       Sarine
##       0.0032       0.0046       0.0049       0.0051       0.0051       0.0072
##        Nyone     Le Locle   Val de Ruz      Yverdon     Cossonay    Echallens
##       0.0095       0.0111       0.0133       0.0154       0.0161       0.0189
##      Moutier     Martigwy         Sion         Orbe   Courtelary   Neuveville
##       0.0207       0.0245       0.0249       0.0258       0.0305       0.0320
##    Entremont   St Maurice        Glane Franches-Mnt       Moudon  Rive Droite
##       0.0354       0.0400       0.0407       0.0469       0.0482       0.0658
##    Neuchatel    La Vallee   Porrentruy  Rive Gauche       Sierre
##       0.0774       0.0838       0.0993       0.1278       0.1670
```

- Based on our Cook's distance none of our observations are considered influential, as none of them are over the value of one.

(d)

Briefly describe the difference in what DFFITS and Cook's distance are measuring.

For DFFITS the value indicates how many standard errors the predicted response changes when the model is estimated with and without the observation. It is the measure of influence of the observation on it's own fitted value. Cook's distance measures how the fitted values for all observations change if observation i is removed from the estimated model.

Problem 2

(a)

Calculate the externally studentized residual, ti, for observation 6. Will this be considered outlying in the response?

$$t_i = \frac{e_i}{\sqrt{MS_{res(i)}(1-h_{ii})}} = e_i\sqrt{\frac{n-1-p}{SSres(1-h_{ii})-e_i^2}}$$

$SS_{res} = MS_{res} * df_{res} = (40.13^2)(19-2) = 27377.0873$

$$t_6 = \frac{120.829070}{\sqrt{(22.6)^2(1-0.23960510)}} = 120.829070\sqrt{\frac{19-1-2}{27377.0873(1-0.23960510)-(120.829070^2)}} = 6.131170535$$

Generally our guideline for externally studentized residuals is any magnitude over 3 would be considered outlying. This value is over our guidleine and does appear to be outlying.

(b)
What is the leverage for observation 6? Based on the criterion that leverages greater than 2p/n are considered outlying in the predictor(s), is this observation high leverage? Observation 6 leverage is 0.23960510. Guideline = 2 * 2 / 19 = 0.2105263158. Observation 6's leverage is over our guideline and so it is considered high leverage.
(c)
Calculate the DFFITS for observation 6. Briefly describe the role of leverages in DFFITS.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{s^2_{(i)} h_{ii}}} \;=\; t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}} =$$

$\hat{y}_6 = -158.78 + (19.96 * 10.5) = 50.8 \quad \hat{y}_{6(6)} = -234.60 + (20.54 * 10.5) = -18.93$

$$DFFITS_{(6)} = \frac{50.8 - -18.93}{\sqrt{40.13^2 * 0.23960510}} \;=\; 6.131170535 \sqrt{\frac{0.23960510}{1 - 0.23960510}} = 3.441691$$

Observations that display high leverage (and are outlying) are more likely to have high values of DFFITS. This is noticeable in the right side formula, as an observation with high leverage will have a larger value for the square root component. The ti is the externally studentized residual, a measure we use to determine if an observation is an outlier. So if the observation is an outlier the value of ti will be larger as well. This combination will lead to a high value of DFFITS, where a high value is indicative of the observation being influential.
(d)
Calculate Cook's distance for observation 6.

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})}{p MS_{res}} = \frac{r_i^2 h_{ii}}{p(1 - h_{ii})}$$

Calculating the right side equation based on the information we have available from the output.

$$r_i = \frac{e_i}{\sqrt{MS_{res}(1 - h_{ii})}}$$

$$r_6 = \frac{120.829070}{\sqrt{40.13^2(1 - 0.23960510)}} = 3.452889$$

$$r_6^2 = 3.452889^2 = 11.922446$$

$$D_6 = \frac{11.922446(0.23960510)}{2(1 - 0.23960510)} = 1.878418$$

Problem 3

Cook's distance has the equivalent formulae

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})'(x'x)(\hat{\beta} - \hat{\beta}_{(i)})}{p\ MS_{res}} = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})}$$

Show that these are equivalent. You can utilize $\hat{\beta} - \hat{\beta}_{(i)} = (1 - h_{ii})^{-1}(x'x)^{-1} x_i e_i$
We know that r_i is the studentized residuals equal to $\frac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}$ so $r_i^2 = \frac{e_i^2}{MS_{res}(1-h_{ii})}$.

We also know that $h_{ii} = x_i'(x'x)^{-1}x_i$.
Step 1: Substitute equivalent $\hat{\beta} - \hat{\beta}_{(i)}$

$$= \frac{x_i'(x'x)^{-1}(x'x)(x'x)^{-1}x_i e_i^2}{(1-h_{ii})(1-h_{ii})\ p\ MS_{res}}$$

Step 2: Rewrite to separate e_i^2 *and* $(1-h_{ii})^2$

$$= \frac{e_i^2}{(1-h_{ii})^2}\ \frac{x_i'(x'x)^{-1}(x'x)(x'x)^{-1}x_i}{p\ MS_{res}}$$

Where $(x'x)(x'x)^{-1} = I$
Step 3: h_{ii} substitution and I elimination

$$= \left(\frac{e_i}{1-h_{ii}}\right)^2 \frac{h_{ii}}{p\ MS_{res}}$$

Step 4: Rewrite

$$= \frac{e_i^2}{MS_{res}(1-h_{ii})}\ \frac{h_{ii}}{(1-h_{ii})\ p}$$

$$= r_i^2\ *\ \frac{1}{p}\ *\ \frac{h_{ii}}{1-h_{ii}}$$

$$= \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$