Module 08 HW

Diana McSpadden

10/13/2020

Stat 6021: Homework Set 8

H. Diana McSpadden

UID: hdm52

Date: 10/22/2020

Attended group with: Wright, McSpadden, Nam, Alvarado, Chivaluri, Barbre, Bernhardt, Bushkar

In this question, you will revisit the swiss data set that you worked on in Homeworks 4 and 5. The data set contains information regarding a standardized fertility measure and socio-economic indictors 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.

```
library(ggplot2)

attach(swiss)
#?swiss

swissModel <- lm(Fertility ~ Education + Catholic + Infant.Mortality)
```

Question 1

(a) Are there any observations that are outlying in the response variable?

Be sure to show your work and explain how you arrived at your answer.

**** Work on q1a****

First, I will use R to get the externally studentized residuals:

```
##residuals
#regResiduals <- swissModel$residuals # regular residuals

##studentized residuals
#studentResiduals <- rstandard(swissModel) # rstandard give studentized resid
uals == ri

##externally studentized residuals
extStudentResiduals <- rstudent(swissModel) # externally studentized
```

Next, I will calculate the critical value to use for comparison. Externally studentized residuals follow a t distribution with n - p - 1 degrees freedom, so that is the critical value I will use.

```
n <- nrow(swiss)
p <- 4
##critical value using Bonferroni procedure
criticalValue <- qt(1-(0.05/(2*n)), n-p-1)
criticalValue
```

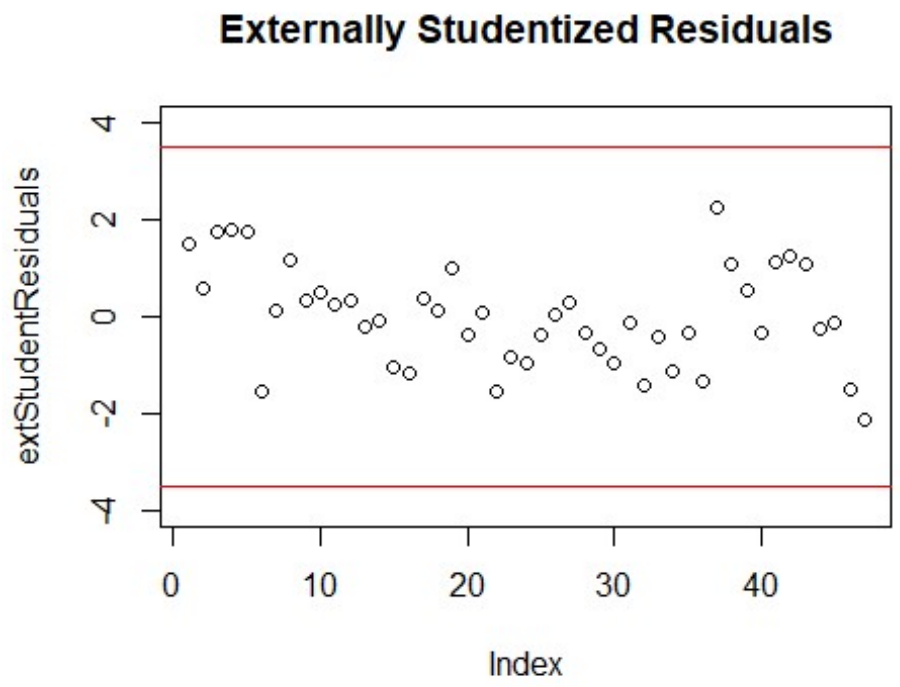
```
## [1] 3.516461
```

Next, I will sort the externally studentized residuals and plot with the +/- critical value also plotted. This will show if any of the absolute values of externally studentized residuals are greater than the critical value.

```
sort(extStudentResiduals)
```

```
##          47          22           6          46          32          36
## -2.12565903 -1.53143864 -1.53112557 -1.48760356 -1.40449006 -1.32506771
##          16          34          15          24          30          23
## -1.19047678 -1.14710495 -1.05530766 -0.97516163 -0.95608183 -0.82725155
##          29          33          20          25          35          40
## -0.65212871 -0.44132091 -0.39639684 -0.39340945 -0.35817316 -0.33803518
##          28          44          13          45          31          14
## -0.33455347 -0.25138105 -0.19397095 -0.14834340 -0.11644471 -0.07004982
##          26          21           7          18          11          27
##  0.03395282  0.07846824  0.12749104  0.12968550  0.24340591  0.27639251
##          12           9          17          10          39           2
##  0.31655920  0.33167701  0.38546504  0.51281448  0.53665268  0.58850210
##          19          43          38          41           8          42
##  1.01354330  1.07015829  1.07562852  1.13116821  1.16036081  1.25986063
##           1           5           3           4          37
##  1.51413731  1.72950210  1.75386263  1.78317342  2.25437539
```

```
plot.new()
plot(extStudentResiduals,main="Externally Studentized Residuals", ylim=c(-4,4
))
abline(h=criticalValue, col="red")
abline(h=-criticalValue, col="red")
```




```
extStudentResiduals[abs(extStudentResiduals)>criticalValue]
```

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

Answer Question 1a: No, it does not appear that there are any outlying responses by comparing the critical value: 3.52 to each ....

(b) Are there any observations that have high leverage?

Be sure to show your work and explain how you arrived at your answer.

Work on q1b

First, I will get the leverage values from that hat matrix. The diagonal of the **hat matrix** are the standardized measures of the distance of the i-th observation from the center of the x space.

Large hat diagonals reveal observations that are potentially influential.

```
lev <- lm.influence(swissModel)$hat
```

It is obvious that the average size of hat diagonal is p/n.

Thus, if a diagonal is **greater than 2p/n** then we consider it remote enough to be a **leverage point**.

```
comparisonValue <- 2*p/n
comparisonValue
```

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

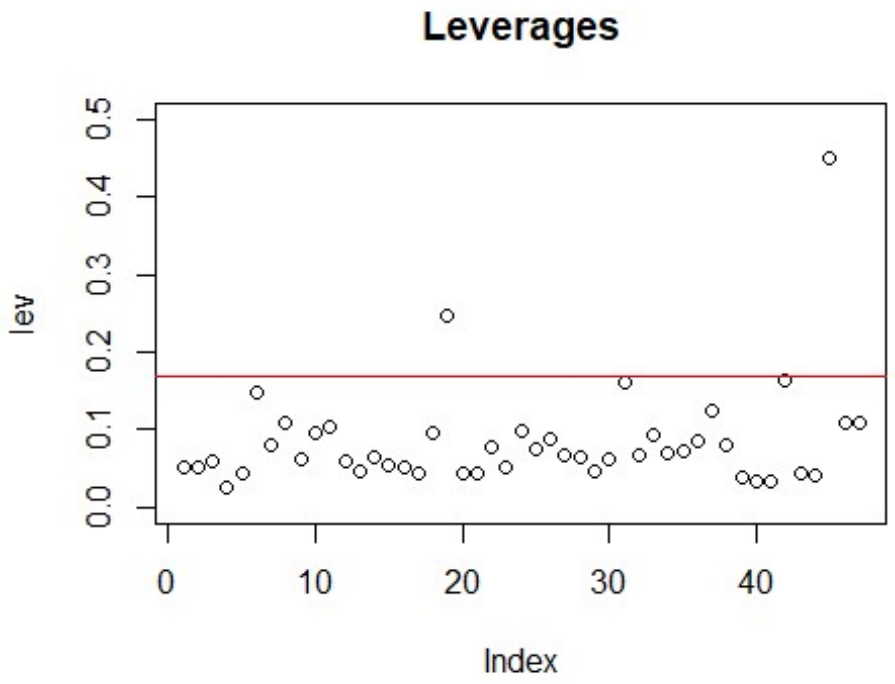
```
sort(lev)
```

```
##          4         40         41         39         44         20
5
## 0.02661704 0.03286437 0.03382830 0.03871182 0.04088238 0.04284048 0.042852
77
##         17         43         21         29         13         16
1
## 0.04415246 0.04468945 0.04503096 0.04544170 0.04770628 0.05112156 0.051930
09
##         23          2         15         12          3          9
30
## 0.05242355 0.05297992 0.05493888 0.05924589 0.06010649 0.06242374 0.063157
48
##         28         14         32         27         34         35
25
## 0.06499491 0.06538567 0.06840751 0.06842411 0.06973133 0.07141262 0.075418
99
##         22          7         38         36         26         33
18
## 0.07812621 0.07933908 0.07965007 0.08480751 0.08869935 0.09287236 0.095910
96
##         10         24         11          8         46         47
37
## 0.09730273 0.09772627 0.10276388 0.10872122 0.10900165 0.10906673 0.125789
18
##          6         31         42         19         45
## 0.14876552 0.16245603 0.16503569 0.24610559 0.45013921
```

I notice there are two high leverage points: **19 and 45**

For fun, I will plot the leverages with a plotted line to identify the high leverage "level" and the points that exceed that level.

```
plot(lev, main="Leverages", ylim=c(0,0.5))
abline(h=2*p/n, col="red")
```



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

Work on q1c

First, calculate DFFITS to see how drastically the fitted value changes with and without the presence of a particular observation.

Compare DFFITS to 2 * sqrt(p/n).

```
DFFITS <- dffits(swissModel)
DFFITS[abs(DFFITS)>2*sqrt(p/n)] # how drastically fitted value changes with a
nd without the presence of the observation

##          6         37         47
## -0.6400846  0.8551451 -0.7437332
```

Three points significantly change the fitted values of the model if they are not (or are) included in the creation of the model: points 6, 37, and 47.

Second, calculate DFBETAS to see how the significantly the coefficients will change with and without observations.

DFBETAS is compared to 2 / sqrt(n)

```
#DFBETAS -
DFBETAS <- dfbetas(swissModel) # measures how estimated coefficients change w
```

```
ithout presence of the observation
DFBETAS[abs(DFBETAS)>2/sqrt(n)]
```

```
##  [1]  0.4908561  0.5129647  0.5583841 -0.2955973  0.4503229 -0.4388348
##  [7] -0.6455617  0.3154725 -0.2949364 -0.3099066  0.5420013 -0.5135508
## [13] -0.5111900 -0.5581878
```

```
# how to find the data point
DFBETAS # can find that value with eyeballing
```

```
##     (Intercept)     Education     Catholic Infant.Mortality
## 1  -0.148417854  1.368007e-02 -0.201907799      0.211495818
## 2  -0.051391914 -4.251943e-05  0.081764597      0.053389912
## 3   0.054337113 -1.187854e-01  0.315472521     -0.044363266
## 4   0.033031440 -1.161608e-01 -0.069343924      0.032866664
## 5  -0.049009914  8.295364e-02 -0.225043619      0.105740766
## 6   0.490856120  1.897387e-02 -0.190320047     -0.513550845
## 7  -0.019502528 -2.894163e-03  0.019685734      0.020385595
## 8  -0.264745405  1.155477e-03  0.190548903      0.269250491
## 9  -0.006061838 -1.004128e-02  0.064219968      0.005657703
## 10 -0.113464500  3.989431e-02  0.080480396      0.109289535
## 11 -0.048218675 -7.961994e-03  0.040939039      0.050177581
## 12  0.061282889 -4.497493e-03 -0.028450418     -0.051254555
## 13 -0.015600029  1.685151e-02  0.028328224      0.004984045
## 14  0.008708059 -5.739414e-04  0.011165532     -0.011959604
## 15 -0.115342325  1.271807e-01  0.153601466      0.052472984
## 16  0.007930851  1.753286e-01  0.111787700     -0.078778289
## 17  0.011747765 -2.572471e-02 -0.056844158      0.008626132
## 18 -0.008098990  3.426208e-02 -0.009624175      0.006836464
## 19  0.512964671  9.997667e-02 -0.052578519     -0.511189993
## 20 -0.010474172  2.022508e-02  0.058213966     -0.009443586
## 21  0.009733236 -3.277344e-03 -0.009317609     -0.006429376
## 22  0.111160547  2.125187e-01  0.272353591     -0.221897900
## 23 -0.151021208  8.217882e-03  0.055106425      0.128019328
## 24 -0.278407198  1.169649e-01  0.108573100      0.229247300
## 25 -0.003194789  6.988139e-02  0.069968406     -0.026951391
## 26 -0.005656149 -1.767865e-03 -0.006014528      0.007706287
## 27  0.042651506 -4.326839e-02 -0.040346480     -0.024796420
## 28 -0.071144128  1.594110e-02  0.032081977      0.058539208
## 29  0.039175320 -7.812019e-02  0.048673173     -0.048092476
## 30  0.093305361  5.358667e-02  0.153133610     -0.147305298
## 31 -0.037554745  1.637670e-02 -0.029724563      0.037494928
## 32 -0.066699571  7.051737e-02 -0.294936437      0.069412058
## 33 -0.063095234  5.388727e-02 -0.096277272      0.060172073
## 34 -0.078165884  5.989151e-02 -0.241949760      0.080755079
## 35 -0.014069718  3.506907e-02 -0.069702555      0.011033574
## 36 -0.184703436  1.351203e-02 -0.309906594      0.203590956
## 37  0.558384110 -2.620255e-01  0.542001320     -0.558187838
## 38  0.115116731  5.888086e-02  0.248688451     -0.140792559
## 39 -0.000529050 -3.711456e-04 -0.070944967      0.022224709
```

```
## 40  0.002441136  3.900297e-03  0.035691117     -0.015444311
## 41  0.074259731  1.420696e-02 -0.109759815     -0.038820844
## 42 -0.295597310  4.503229e-01 -0.093722986      0.270623733
## 43  0.036919051 -8.755421e-02 -0.153601586      0.021608016
## 44 -0.015013586  2.054061e-02  0.031505633      0.002109957
## 45  0.009901006 -1.294500e-01 -0.022351708      0.010452812
## 46 -0.055649664 -4.388348e-01 -0.137697822      0.117387597
## 47  0.043882710 -6.455617e-01 -0.240469967      0.050528230
```

Observations With High DFBETAS: 6, 19, 37, 42

Third, calculate COOKS to determine how fitted values change for all values, not just the removed observation, when an observation is removed.

COOKS is compared to an F distribution with p, and n-p degress of freedom.

```
# COOKS considered how fitted values changes for all values (not just the obs
ervation) when data point is removed.
COOKS<-cooks.distance(swissModel)
COOKS[COOKS>qf(0.5,p,n-p)] # f distribution

## named numeric(0)
```

There are no observations that need to be considered based on COOK'S Distance.

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

DFFITs checks if the fitted value for the removed observation predictor value(s) significantly changes if the observation is removed.

Cook's distance checks the amount of change for all fitted values if an observation is removed from all the data used to create the model.

Question 2 (No R Required)

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

… Also included were residuals and diagonals from hat matrix …

(a) Calculate the externally studentized residual, ti,

for observation 6. Will this be considered outlying in the response?

Work on q2a

Needed Formulas:

S(i)^2 = (((n-p) * MSE) - (e-i^2/ (1 - hii))) / (n - p - 1)

with that, calculate ti

ti = e-i / sqrt((S(i)^2 * (1-hii)))

```r
# Knowns
n <- 19
p <- 2
MSE <- 40.13^2
ei <- 120.829070
```

```
hii <- 0.23960510

Si2 <- (((n-p) * MSE) - (ei^2/ (1 - hii))) / (n - p - 1)
print(paste("Sihat^2: ", Si2))

## [1] "Sihat^2:  511.061209996762"

ti <- ei / sqrt((Si2 * (1-hii)))
print(paste("ti: ", ti))

## [1] "ti:  6.12936347001077"
```

Externally standardized residuals follow a t distribution with n - p - 1 degrees freedom.

Use t (alpha/2n), n - p - 1 as a comparison.

```
criticalValue <- qt(1-(0.05/(2*n)), n-p-1)
print(paste("critical value for comparison: ", criticalValue))

## [1] "critical value for comparison:  3.55624214386903"
```

Question 2a Answer: 6.13 > 3.56, so yes, this observation has an outlying response value.

(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?

Question 2b Answer

```
leverageComparison <- (2 * p) / n
leverageComparison

## [1] 0.2105263
```

The leverage, from the hat matrix diagonals, is 0.23960510. The leverage value for comparison, (2 * p) / n is 0.21. Since **0.24 > 0.21** this observation does have high leverage on our model.

(c) Calculate the DFFITS for observation 6.

Briefly describe the role of leverages in DFFITS.

Work on Question 2c

Formula for DFFITSi: DFFITSi = (sqrt(hii / 1 - hii) * ti)

```
#hii
#ti
DFFITSi6 <- (sqrt(hii / (1 - hii)) * ti)

print(paste("DFFITSi for observation 6: ", DFFITSi6))
```

```
## [1] "DFFITSi for observation 6:  3.44067622811447"
```

DFFITs uses the leverage for the observation with the (hii / (1 - hii)) multiplier of the formula. hii approaching 1 (with high leverage) causes hii / (1 - hii) to increase. This (hii / (1 - hii)) moderates the effect of the externally studentized residuals which calculate whether the observation is an outlier response. A point can be an outlier response, but not have high leverage, thus not skew the model towards the point. DFFITs uses both a leverage measure and the measure of outlying to identify high leverage points.

(d) Calculate Cook's distance for observation 6.

Work on Question 2d

Formulas Needed:

Di = (ri^2 / p) * *(hii / (1 - hii))

where ri = e-i / (sqrt(MSE * (1 - hii)))

We compare Cook's distance to an F distribution: F alpha, p, n-p

```
#ei
#MSE
ri = ei / (sqrt(MSE * (1 - hii)))
#ri

Di = (ri^2 / p) * (hii / (1 - hii))

criticalValueCooks = qf(0.5,p,n-p)

print(paste("Cook's Distance for Observation 6: ", Di))

## [1] "Cook's Distance for Observation 6:  1.87841789748727"

print(paste("Cook's Distance F Comparison: ", criticalValueCooks))

## [1] "Cook's Distance F Comparison:  0.722193265970915"
```

Question 3 (No R Required)