Homework10

Dima Mikhaylov

11/15/2021

PART 1

```
library(datasets)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data(swiss)
head(swiss)
```

```
##              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
```

Three predictors (Education, Catholic, and Infant Mortality) will be used in the preferred model:

```
# Fit the model
mlr <- lm(Fertility ~ Education + Catholic + Infant.Mortality, data = swiss)
summary(mlr)
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality,
##     data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4781  -5.4403  -0.5143   4.1568  15.1187
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      48.67707    7.91908   6.147 2.24e-07 ***
## Education        -0.75925    0.11680  -6.501 6.83e-08 ***
## Catholic          0.09607    0.02722   3.530  0.00101 **
## Infant.Mortality  1.29615    0.38699   3.349  0.00169 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.505 on 43 degrees of freedom
## Multiple R-squared:  0.6625, Adjusted R-squared:  0.639
## F-statistic: 28.14 on 3 and 43 DF,  p-value: 3.15e-10
```

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

Using studentized as well as externally studentized residuals to find possible outliers:

```
# Critical value using Bonferroni procedure at 95% confidence level
n <- dim(swiss)[1]
p <- 4 #3 slopes and 1 for intercept
critical <- qt(1-0.05/(2*n), n-1-p)
cat("Critical value is", critical)
```

```
## Critical value is 3.516461
```

```
##Number of outliers predictors based on studentized residuals
student.res<-rstandard(mlr)
student.res[abs(student.res)>critical]
```

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

```
# Number of outliers predictors based on externally studentized residuals
ext.student.res <- rstudent(mlr)
ext.student.res[abs(ext.student.res)>critical]
```

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

Conclusion: based on standartized residuals, there are no outliers.

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

Checking for high leverage observations, below are 2 possible areas:

```
# High leverage points above the cutoff
leverage <- lm.influence(mlr)$hat
leverage[leverage > 2* p/n]
```

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

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

Checking for influential observations DFFITs, below are 3 possible areas:

```
# Influential observations in terms of y_hat_i
DFFITS <- dffits(mlr)
DFFITS[abs(DFFITS) >  2*sqrt(p/n)]
```

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

When using Cook's Distance, none detected:

```
# Influential observations in terms of least-squares
COOKS <- cooks.distance(mlr)
COOKS[COOKS>qf(0.5,p,n-p)]
```

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

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

Comment: DFFITS checks how removing an observation will change its own predicted value in terms of standard deviations. Cook's distance is a measure of how much the entire regression model changes when the high leverage observation is removed.

Part 2

Data from n = 19 bears of varying ages are used to develop an equation for estimating Weight from Neck circumference

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

Externally studentized residual: t_6 = e_6 / sqrt(S2 * (1-h_6))

```
h_6 = 0.23960510 # Leverage of the outlier
S2_6 = 22.6 # Residual standard error with outlier was removed
e_6 = 120.829070 # Residual of the outlier
t_6 = (e_6) / sqrt( S2_6 * (1-h_6)) # Esternally studentized residual
t_6
```

```
## [1] 29.14725
```

Check if it is greater than critical value. Yes, it appears highly significant.

```
n<-18 # one bear is out
p<-2 # number of params
crit<-qt(1-0.05/(2*n), n-1-p) #critical value using Bonferroni procedure
cat("Externally studentized residual greater than critical value: ", abs(t_6)>crit)
```

```
## Externally studentized residual greater than critical value:  TRUE
```

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?

Comment: yes, it is high leverage because it is greater than 2p/n cutoff.

```
h_6 = 0.23960510
p = 2 # Numer of params: 1 slope + intercepts
n = 19 # Number of observations
cutoff = (2*p)/n
cat("Leverage for #6 is greater than the cutoff:" , (h_6 > cutoff))
```

```
## Leverage for #6 is greater than the cutoff: TRUE
```

c. Calculate the DFFITS for observation 6. Briefly describe the role of leverages in DFFITS.

DFFITS for #6 is given by the following: (y_hat - y_hat_6) / sqrt(S2 * h_6). The higher is the leverage the smaller is the overall difference between predictions with and without the influencial observation, because it is “pulling” the prediction toward itself harder when leverage is high.

```
x_6 = 10.5 # Neck for bear #6

y_hat = -158.78 + 16.95 * x_6 # Predicted weight for bear #6 with bear #6
y_hat_6 = -234.60 + 20.54 * x_6 # Predicted weight for bear #6 without bear #6
DFFITS_6 = (y_hat-y_hat_6)/sqrt(S2_6*h_6)
cat("DFFITS for observation 6", DFFITS_6)
```

```
## DFFITS for observation 6 16.38353
```

d. Calculate Cook’s distance for observation 6.

Cook’s distance for #6 is given by (r_6^2 / p)*(h_6/(1-h_6)) where residual = 120.829070

```
MS_res = 40.13 # Residual standard error for all observations
r_6 = e_6 / sqrt( MS_res * (1-h_6))
D_6 = (r_6^2/p)*(h_6/(1-h_6))
cat("Cook's distance D_6 is", D_6)
```

```
## Cook's distance D_6 is 75.38091
```