**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 9**

Prepare a short report with relevant output, your comments, and answers to the questions (this does not need to be exhaustive or polished, but should contain enough to show that you completed all tasks and analyses).

Submit the report at the end of the lab session.

Consider the dataset *swiss.txt*.

This contains standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. The variables are:

*Fertility*: common standardized fertility measure

*Agriculture:* % of males involved in agriculture as occupation

*Examination:* % draftees receiving highest mark on army examination

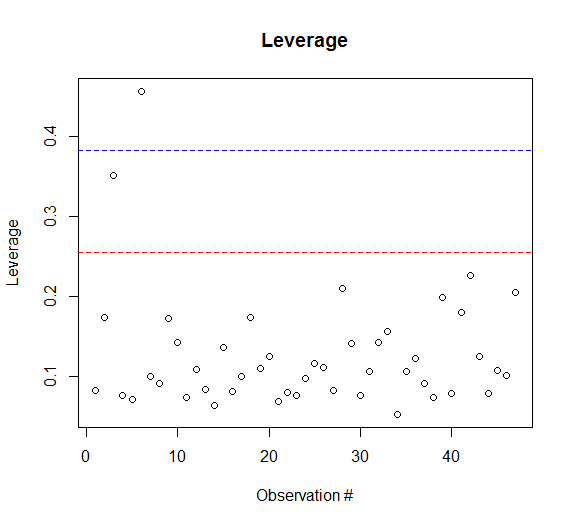
*Education:* % education beyond primary school for draftees

*Catholic:* %  ‘catholic’ (as opposed to ‘protestant’)

*Infant.Mortality:* % live births who live less than 1 year

Fit a multiple linear regression model in order to predict Fertility based on all the other 5 variables.

* Compute the leverage hi (elements on the diagonal of the hat matrix H) and draw a plot of the leverage versus the observation number. Add the horizontal lines corresponding to the thresholds 2\*p/n and 3\*p/n.



> h

1 2 3 4 5 6 7

0.08319314 0.17335920 0.35120780 0.07675938 0.07219003 0.45583631 0.09981583

8 9 10 11 12 13 14

0.09177212 0.17219053 0.14308241 0.07366702 0.10921430 0.08392561 0.06410496

15 16 17 18 19 20 21

0.13676942 0.08125700 0.10036242 0.17368296 0.10990923 0.12551221 0.06853488

22 23 24 25 26 27 28

0.07994004 0.07683116 0.09800616 0.11595449 0.11137547 0.08277061 0.21067014

29 30 31 32 33 34 35

0.14145774 0.07706245 0.10631200 0.14246210 0.15681744 0.05328210 0.10682273

36 37 38 39 40 41 42

0.12258494 0.09204746 0.07425778 0.19833238 0.07961648 0.18059095 0.22629672

43 44 45 46 47

0.12569646 0.07905125 0.10834150 0.10175031 0.20532241

**The above chart shows the leverage hi for each point.**

* + Do you observe any point with large leverage (i.e. leverage greater than 2\*p/n or 3\*p/n)?

**We can observe that two points have large leverage:**

**One is greater than 2\*p/n, the other is greater than 3\*p/n.**

* + To which provinces do those observations correspond? (hint: use function *which*)

> which(h>2\*p/n)

3 6

3 6

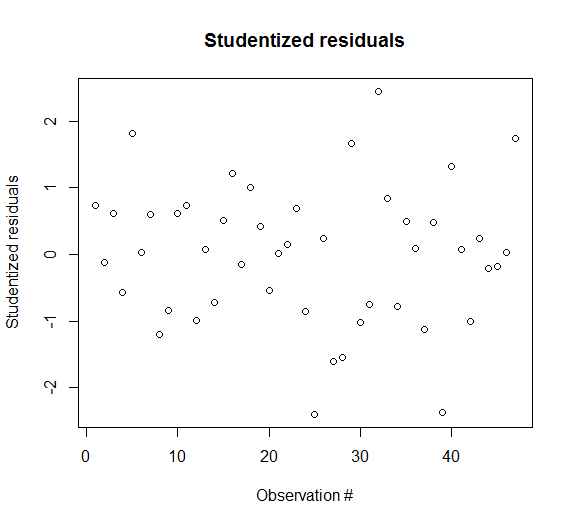
> which(h>3\*p/n)

6

6

**By the above information, we can conclude that point 3 is greater than 2\*p/n but smaller than 3\*p/n; point 6 is greater than 3\*p/n**

* Draw a plot of the **studentized deleted residuals** versus the observation number.



* + Do the points with large leverage in the previous questions also show large studentized residuals?

**In this plot, the 3rd and the 6th point do not have studentized residuals which are greater than 2. Thus, the 2 points with large leverage in the previous questions do not show large studentized residuals.**

* + Which are the three most extreme provinces in terms of studentized residuals, among the 47 in the dataset? (hint: use function *which*)

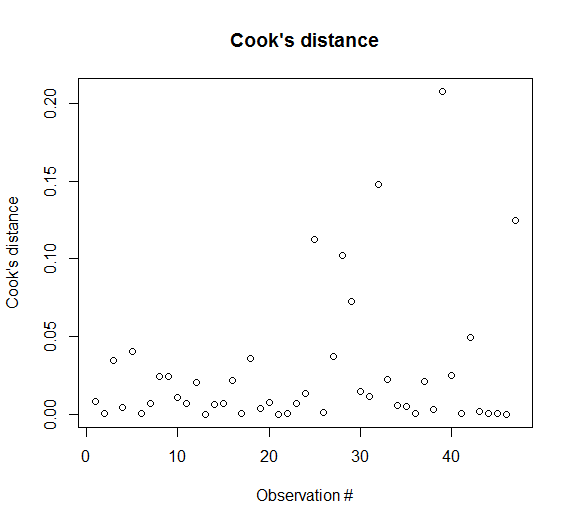
> which(abs(r)>2)

25 32 39

25 32 39

**These three most extreme points are the 25th, 32th , 39th point.**

* Compute Cook’s distance and draw a plot of the Cook’s distance versus the observation number.



> Di

1 2 3 4 5 6

8.249015e-03 4.749927e-04 3.436545e-02 4.466169e-03 4.050222e-02 2.046583e-04

7 8 9 10 11 12

6.830752e-03 2.414395e-02 2.425355e-02 1.066647e-02 7.148244e-03 2.026133e-02

13 14 15 16 17 18

8.221766e-05 5.916059e-03 7.171732e-03 2.176441e-02 4.535869e-04 3.576983e-02

19 20 21 22 23 24

3.723219e-03 7.259171e-03 2.870831e-06 3.471875e-04 6.754670e-03 1.330057e-02

25 26 27 28 29 30

1.123648e-01 1.251196e-03 3.727352e-02 1.024707e-01 7.265696e-02 1.449967e-02

31 32 33 34 35 36

1.118028e-02 1.476237e-01 2.233176e-02 5.704113e-03 5.086126e-03 1.799632e-04

37 38 39 40 41 42

2.134955e-02 3.121719e-03 2.077343e-01 2.460093e-02 2.374726e-04 4.936628e-02

43 44 45 46 47

1.517361e-03 6.562694e-04 7.091540e-04 1.656458e-05 1.249500e-01

**The above information shows the Cook’s distance for each point.**

* + Do the points with large leverage in the previous question also show large Cook’s distance? What about the points that were extreme with respect to studentized residuals?

**The points with large leverage in the previous question do not show large Cook’s distance.**

**Points that were extreme with respect to studentized residuals do show large Cook’s distance.**

* + Do you conclude that the points with large leverage in the previous question influence the model more than the other points?

**No, because points with large leverage in the previous question show neither large studentized residuals nor large Cook’s distance.**

* + Do you conclude that the points with extreme studentized residuals in the previous question influence the model more than the other points?

**Yes, because the extreme studentized residuals in the previous question also show large Cook’s distance.**

**R code:**

setwd("//udrive.win.psu.edu/Users/j/q/jql5883/Desktop/math462")

getwd()

swiss=read.csv("swiss.txt", header=T, sep="")

head(swiss)

attach(swiss)

model=lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality)

residuals=model$residuals

sigma\_hat=summary(model)$sigma

X1=model.matrix(model)

H=X1%\*%solve(t(X1)%\*%X1)%\*%t(X1)

h=diag(H)

r=residuals/(sigma\_hat\*sqrt(1-h))

p=6

n=nrow(swiss)

sum(h)

thresh2=2\*p/n

thresh3=3\*p/n

plot(h,xlab='Observation #',ylab='Leverage',main='Leverage')

abline(h=thresh2,lty=2,col="red")>abline(h=thresh3,lty=2,col="blue")

which(h>2\*p/n)

which(h>3\*p/n)

t=r\*sqrt((n-p-1)/(n-p-r^2))

plot(t,xlab='Observation #',ylab='Studentized residuals',main='Studentized residuals')

which(abs(r)>2)

Di=(1/p)\*r^2\*h/(1-h)

Di

plot(Di,xlab='Observation #',ylab='Cook\'s distance',main='Cook\'s distance')