STAT537: Statistics for Research I: HW#9

Due on Nov. 8, 2016

Dr. Schmidhammer TR 11:10am – 12:25pm

Wenqiang Feng

Contents

Problem 1

Exercise 13.69

Solution. (a) **Fit the full model with all predictors.** According to the fitted results with all predictors,the model is:

$$\begin{aligned} Y &= -26.36485 + 11.76733 \cdot \text{RUN} - 7.02414 \cdot \text{SMOKE} - 0.02090 \cdot \text{HEIGHT} \\ &+ 0.05721 \cdot \text{WEIGHT} + 13.55492 \cdot \text{PHYS1} + 7.89397 \cdot \text{PHYS2}. \end{aligned}$$

The fitted results with all predictors are as follow:

```
Call:
glm(formula = PULSE ~ factor(RUN) + factor(SMOKE) + HEIGHT +
    WEIGHT + factor(PHYS1) + factor(PHYS2), data = rawdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.135   -3.875   -1.205    4.951   13.520

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -26.36485   36.53646  -0.722 0.477810
factor(RUN)1    11.76733    2.67552   4.398 0.000209 ***
factor(SMOKE)1  -7.02414    2.70636  -2.595 0.016175 *
HEIGHT          -0.02090    0.59180  -0.035 0.972128
WEIGHT           0.05721    0.08294   0.690 0.497239
factor(PHYS1)1  13.55492    4.21419   3.216 0.003825 **
factor(PHYS2)1   7.89397    3.94366   2.002 0.057250 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 46.47036)
```

(b) **Plot the studentized residuals and Cook's D and determine if any influential observations exist.** The plot of the studentized residuals and Cook's D can be found in Figure.1. And from the influence plot we can conclude that observation 13 and 18 are influential observations.

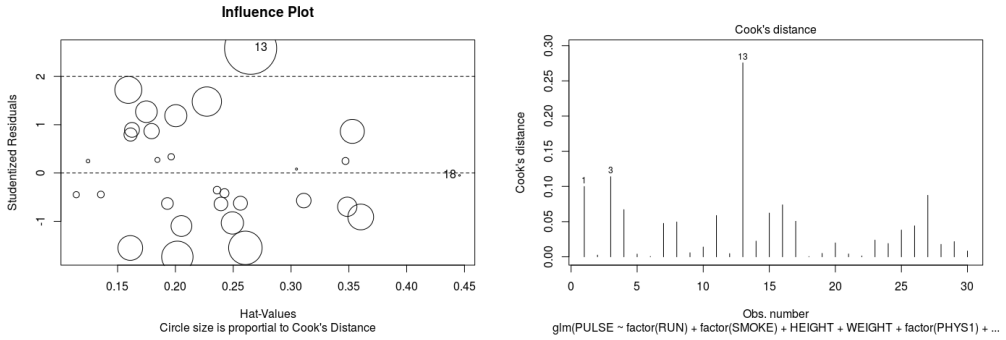

Figure 1: Studentized residuals and Cook's D plot.

```
       StudRes       Hat        CookD
13  2.58285866 0.2650344 0.2756900147
18 -0.04871369 0.4456617 0.0002849002
```

(c) **Use Stepwise Regression to select variables important to the model.** The following are the results of the backward Stepwise Regression. the results indicate that the the variables PHYS2, SMOKE, PHYS1 and RUN are the selected important variables.

```
> model_step =step(fit1,direction = "backward")
Start:  AIC=208.33
PULSE ~ factor(RUN) + factor(SMOKE) + HEIGHT + WEIGHT + factor(PHYS1) +
    factor(PHYS2)

                Df Deviance    AIC
- HEIGHT         1   1068.9 206.33
- WEIGHT         1   1090.9 206.94
<none>               1068.8 208.33
- factor(PHYS2)  1   1255.0 211.15
- factor(SMOKE)  1   1381.8 214.04
- factor(PHYS1)  1   1549.6 217.47
- factor(RUN)    1   1967.7 224.64

Step:  AIC=206.33
PULSE ~ factor(RUN) + factor(SMOKE) + WEIGHT + factor(PHYS1) +
    factor(PHYS2)

                Df Deviance    AIC
- WEIGHT         1   1100.6 205.21
<none>               1068.9 206.33
- factor(PHYS2)  1   1260.5 209.28
- factor(SMOKE)  1   1381.9 212.04
- factor(PHYS1)  1   1549.6 215.47
- factor(RUN)    1   1972.4 222.71

Step:  AIC=205.21
PULSE ~ factor(RUN) + factor(SMOKE) + factor(PHYS1) + factor(PHYS2)

                Df Deviance    AIC
<none>               1100.6 205.21
- factor(PHYS2)  1   1274.0 207.60
- factor(SMOKE)  1   1408.5 210.61
- factor(PHYS1)  1   1567.3 213.81
- factor(RUN)    1   1981.8 220.85
```

(d) **Interpret the final model.** The fitting results of the final model can be found in the following, which can be formulated as

$$Y = -18.302 + 11.132 \cdot \text{RUN} - 6.963 \cdot \text{SMOKE} + 13.325 \cdot \text{PHYS1} + 7.451 \cdot \text{PHYS2}.$$

The final model indicates that

- The interception is -18.302.
- If the volunteers RUN , the pulse rate will increase 11.132 units when the other variables are fixed.
- If the volunteers SMOKE , the pulse rate will decrease 6.963 units when the other variables are fixed.
- If the volunteers do a lot of physical exercise (PHYS1), the pulse rate will increase 13.325 units when the other variables are fixed.
- If the volunteers do moderate physical exercise (PHYS2), the pulse rate will increase 7.451 units when the other variables are fixed.

```
Call:
glm(formula = PULSE ~ factor(RUN) + factor(SMOKE) + factor(PHYS1) +
    factor(PHYS2), data = rawdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-11.1862   -4.1927   -0.5269    4.6858   12.9764

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -18.302      3.649  -5.016 3.58e-05 ***
factor(RUN)1     11.132      2.488   4.474 0.000146 ***
factor(SMOKE)1   -6.963      2.633  -2.645 0.013919 *
factor(PHYS1)1   13.325      4.092   3.256 0.003238 **
factor(PHYS2)1    7.451      3.754   1.985 0.058276 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 44.02283)

    Null deviance: 2938.7  on 29  degrees of freedom
Residual deviance: 1100.6  on 25  degrees of freedom
AIC: 205.21

Number of Fisher Scoring iterations: 2
```

□

Appendix

R code for HW#9

Listing 1: Source code for problem 1

```r
# reference: http://www.stat.columbia.edu/~martin/W2024/R3.pdf
rm(list = ls())
```

```r
# set the path or enverionment
setwd("/home/feng/Dropbox/UTK_Course/Stat537/Excel/CH13")

# (b)
#install.packages("readxl") # CRAN version
library(readxl)
#install.packages("moments")
library(moments)
rawdata = read_excel("ex13-69.xls",sheet = 1)
attach(rawdata)

# (a)
fit1 = glm(PULSE~factor(RUN)+factor(SMOKE)+HEIGHT+
            WEIGHT+factor(PHYS1)+factor(PHYS2),data=rawdata)
summary(fit1)

#(b)
library(car)
fit1$residuals
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(rawdata)-length(fit1$coefficients)-2))
plot(fit1, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit1, main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance" )


# (c)
model_step =step(fit1,direction = "backward")

# (d)
fit2 = glm(PULSE~factor(RUN)+factor(SMOKE)
           +factor(PHYS1)+factor(PHYS2),data=rawdata)
summary(fit2)
```