DiZhen_1717719

dizhen

4/26/2020

Set directory

```
setwd('D:/git/DPH112-xjtlu/week09')
```

Data

```
KBI <- c(28,68,59,91,70,38,46,57,89,48,74,78,43,76,72,61,63,77,85,31,79,92,76,91,78,103,99,73,88,64,52,7

ADL <- c(39,52,89,57,28,34,42,52,88,90,38,83,30,45,47,90,63,34,76,26,68,85,22,82,80,80,81,30,27,72,46,63

MEM <- c(4,33,17,31,35,3,16,6,41,24,22,41,9,33,36,17,14,35,33,13,34,28,12,57,51,20,20,7,27,9,15,52,26,57

COG <- c(18,9,3,7,19,25,17,26,13,3,13,11,24,14,18,0,16,22,23,18,26,10,16,3,3,18,1,17,27,0,22,13,18,0,19

mydata <- data.frame(KBI,ADL,MEM,COG)
str(mydata)
```

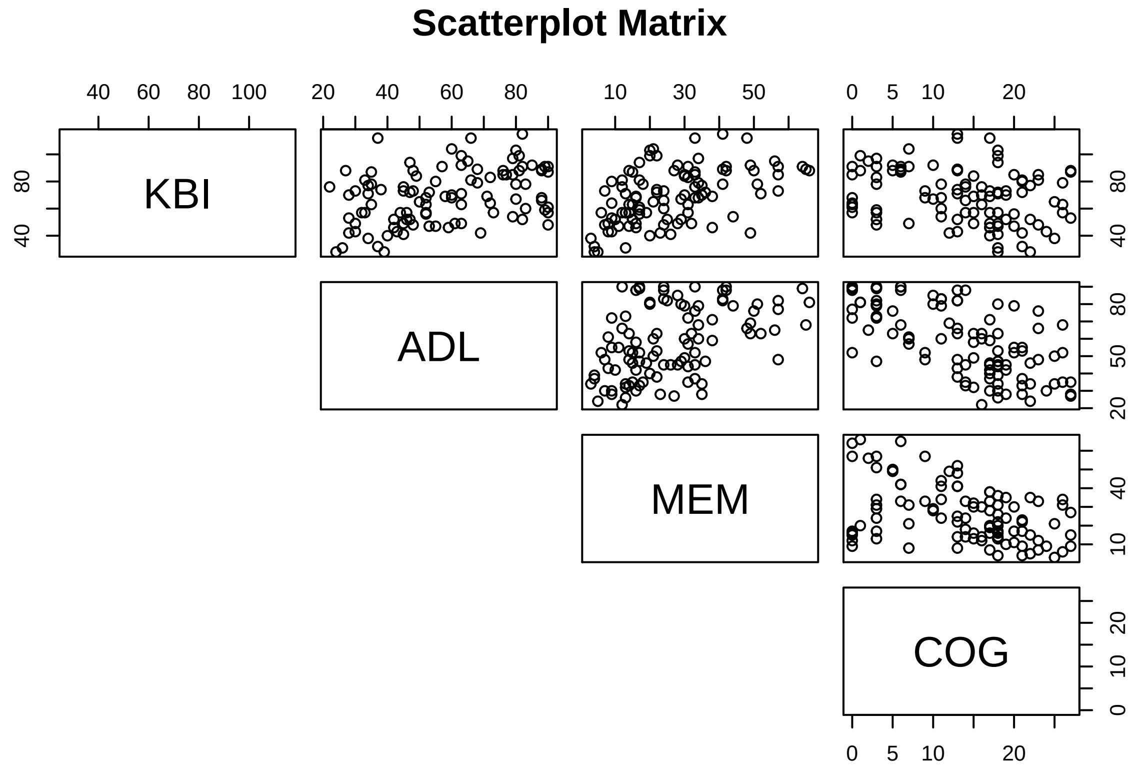
```
## 'data.frame':    100 obs. of  4 variables:
##  $ KBI: num  28 68 59 91 70 38 46 57 89 48 ...
##  $ ADL: num  39 52 89 57 28 34 42 52 88 90 ...
##  $ MEM: num  4 33 17 31 35 3 16 6 41 24 ...
##  $ COG: num  18 9 3 7 19 25 17 26 13 3 ...
```

```
head(mydata)
```

```
##   KBI ADL MEM COG
## 1  28  39   4  18
## 2  68  52  33   9
## 3  59  89  17   3
## 4  91  57  31   7
## 5  70  28  35  19
## 6  38  34   3  25
```

Visualization

```
pairs(~KBI + ADL + MEM + COG, data = mydata,
      lower.panel = NULL,
      main = "Scatterplot Matrix")
```



Question 1

```
mydata.LM1 <- lm(KBI ~ ADL + MEM + COG, data = mydata)
summary(mydata.LM1)
```

```
##
## Call:
## lm(formula = KBI ~ ADL + MEM + COG, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.037 -10.535  -1.503   9.213  43.151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.4908    10.1030   4.008 0.000121 ***
## ADL           0.2162     0.1168   1.851 0.067273 .
## MEM           0.5547     0.1300   4.267 4.65e-05 ***
## COG           0.1210     0.3003   0.403 0.687978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.26 on 96 degrees of freedom
## Multiple R-squared:  0.282,  Adjusted R-squared:  0.2596
## F-statistic: 12.57 on 3 and 96 DF,  p-value: 5.315e-07
```

The multiple regression equition is:

$KBI = 40.5 + 0.2 \times ADL + 0.6 \times MEM + 0.1 \times COG$

The mean KBI when ADL, MEM and COG are zero is 40.5. For every extra unit of ADL, the expected KBI increases by 0.2 points, holding other variables constant. For every extra unit of MEM, the mean KBI increases by 0.6 points, holding other variables constant. For every extra unit of COG, the mean KBI increases by 0.1 points, holding other variables constant.

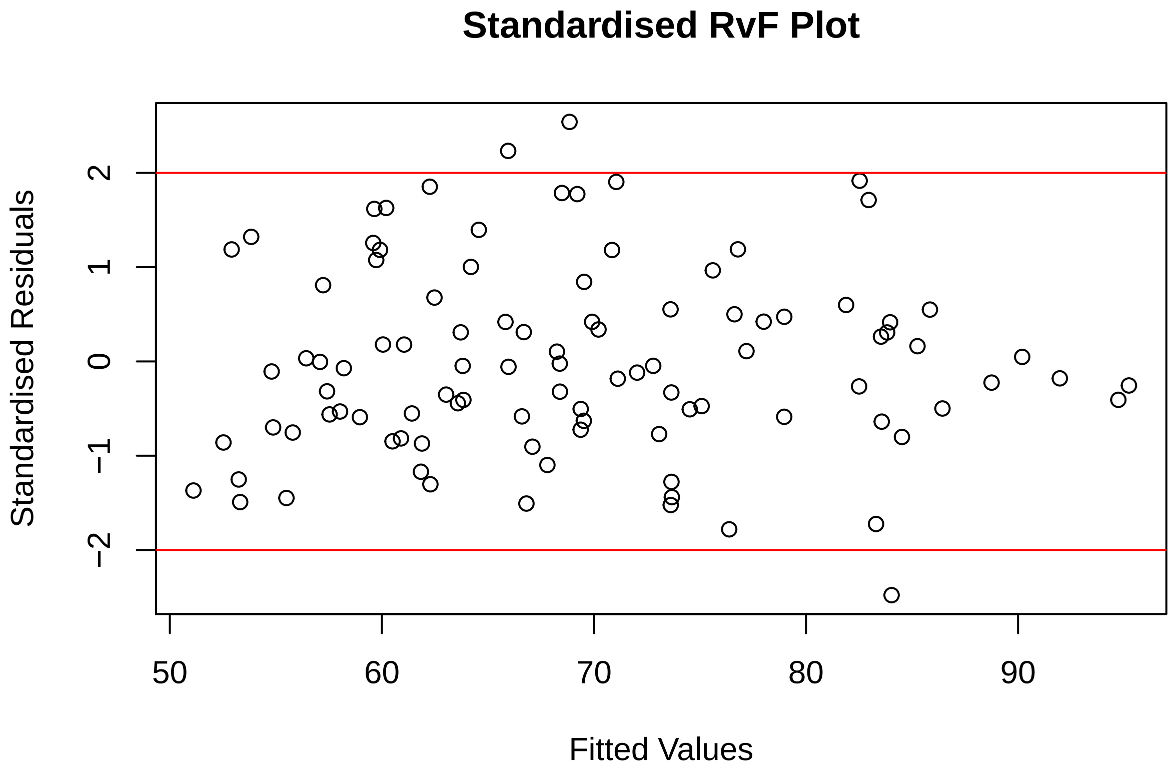
Question 2

Evaluate the presence of outliers in the dependent variable using standardised residuals.

There are two observations greater than two standard deviations larger than the mean and there are one observations less than two standard deviations below the mean. There are three outliers in the dependent variable in total.

```r
mydata$FITTED <- predict(mydata.LM1, type = "response")
mydata$RESID <- resid(mydata.LM1)

mydata$RSTAND <- rstandard(mydata.LM1)
plot(RSTAND ~ FITTED, data = mydata,
     ylab = "Standardised Residuals",
     xlab = "Fitted Values",
     main = "Standardised RvF Plot")
abline(h = c(-2, 2), col = "red")
```



```
mydata[abs(mydata$RSTAND) > 2,]
```

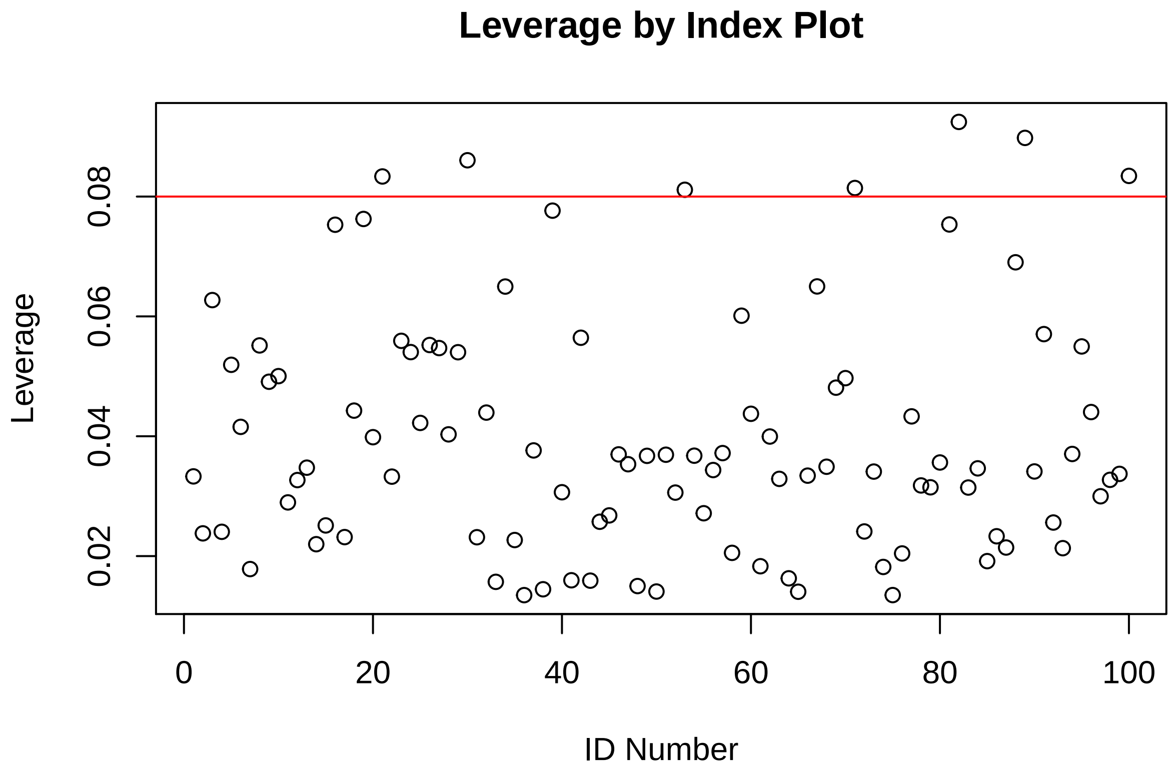
```
##    KBI ADL MEM COG   FITTED     RESID    RSTAND
## 79 112  37  33  17 68.84927  43.15073  2.540513
## 84  42  69  49  12 84.03658 -42.03658 -2.478963
## 92 104  60  21   7 65.95588  38.04412  2.233089
```

Question 3

Evaluate the presence of outliers in the independent variables using leverage values.

The plot shows 7 observations are high-leverage points.They are outliers in the independent variable.

```
mydata$HAT <- hatvalues(mydata.LM1)
HAT.CUT <- 2 * (3 + 1)/ length(KBI)
ID <- seq(1,length(KBI),by = 1)
plot(HAT ~ ID, data = mydata,
     ylab = "Leverage",
     xlab = "ID Number",
     main = "Leverage by Index Plot")
abline(h = HAT.CUT, col = "red")
```



```r
mydata[mydata$HAT > HAT.CUT,]
```

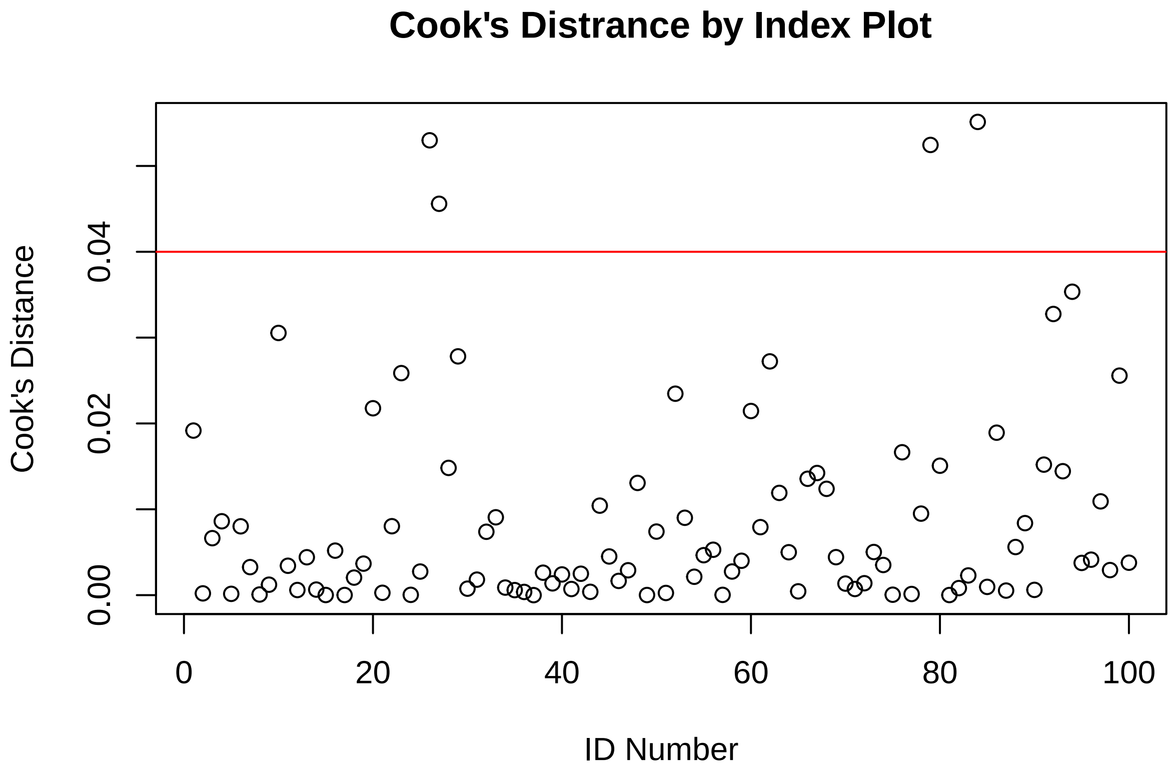
```
##     KBI ADL MEM COG   FITTED      RESID     RSTAND        HAT
## 21   79  68  34  26 77.19408   1.805925  0.1092915 0.08335487
## 30   64  72   9   0 61.04743   2.952570  0.1789481 0.08605394
## 53   73  48  57   9 83.57120 -10.571198 -0.6389755 0.08112930
## 71   89  68  65   6 91.96908  -2.969081 -0.1794948 0.08142400
## 82   63  52  15   0 60.05181   2.948193  0.1793117 0.09245325
## 89   57  90  12   0 66.60258  -9.602580 -0.5831815 0.08978768
## 100  88  81  66   1 94.72923  -6.729226 -0.4072611 0.08344422
```

Question 4

Evaluate the presence of influential observations using Cook's distance.

The plot shows there are 4 observations with high Cook's distance values. Two of them (ID79 and ID84) are detected as having high standardised residuals.

```r
mydata$COOK <- cooks.distance((mydata.LM1))
COOK.CUT <- 4/length(ID)
plot(COOK ~ ID, data = mydata,
     ylab = "Cook's Distance",
     xlab = "ID Number",
     main = "Cook's Distrance by Index Plot")
abline(h = COOK.CUT, col = "red")
```



```
mydata[mydata$COOK > COOK.CUT,]
```

```
##    KBI ADL MEM COG   FITTED     RESID    RSTAND        HAT       COOK
## 26 103  80  20  18 71.05536  31.94464  1.904236 0.05522324 0.05298758
## 27  99  81  20   1 69.21517  29.78483  1.775010 0.05471437 0.04559108
## 79 112  37  33  17 68.84927  43.15073  2.540513 0.03148128 0.05244777
## 84  42  69  49  12 84.03658 -42.03658 -2.478963 0.03464047 0.05512833
```

THE END