Week ON9

Elmer V Villanueva

20 April 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON9")
```

Announcements

- The two coursework assessments and the final paper have been released. The deadlines are also stated. You are free to review the assessment files, but you will NOT be able to submit until the deadline.
- All the assessments are to be completed INDIVIDUALLY. No collusion is allowed. This means that you cannot discuss this assignment with other students, nor can you share your approach or code.

Reading

Read and understand Vittinghoff et al., Chapter 4.

Review of previous learning

Last week, we discussed the four model assumptions the multiple linear regression model. assumptions about (1) form, (2) errors, (3) predictors and (4) observations. We discussed the techniques used to assess the first three sets of assumptions. This week, we will complete the session on multiple linear regression by learning how to assess the assumptions about the observations.

Similar to last week's lecture, we will make use of Jansen and Keller's data on attention in elderly subjects [1]. You should review the previous week's lecture notes if you need to re-familiarise yourself with these data.

Lete us enter the data and derive the linear model.

```
ID <- c(1:71)
AGE <- c(72, 68, 65, 85, 84, 90, 79, 74, 69,
         87, 84, 79, 71, 76, 73, 86, 69, 66,
         65, 71, 80, 81, 66, 76, 70, 76, 67,
         72, 68, 102, 67, 66, 75, 91, 74, 90,
         79, 87, 71, 81, 66, 81, 80, 82, 65,
         73, 85, 83, 83, 76, 77, 83, 79, 69,
         66, 75, 77, 78, 83, 85, 76, 75, 70,
         79, 75, 94, 76, 84, 79, 78, 79)
EDU <- c(20, 12, 13, 14, 13, 15, 12, 10, 12,
         15, 12, 12, 12, 14, 14, 12, 17, 11,
         16, 14, 18, 11, 14, 17, 12, 12, 12,
         20, 18, 12, 12, 14, 18, 13, 15, 15,
```

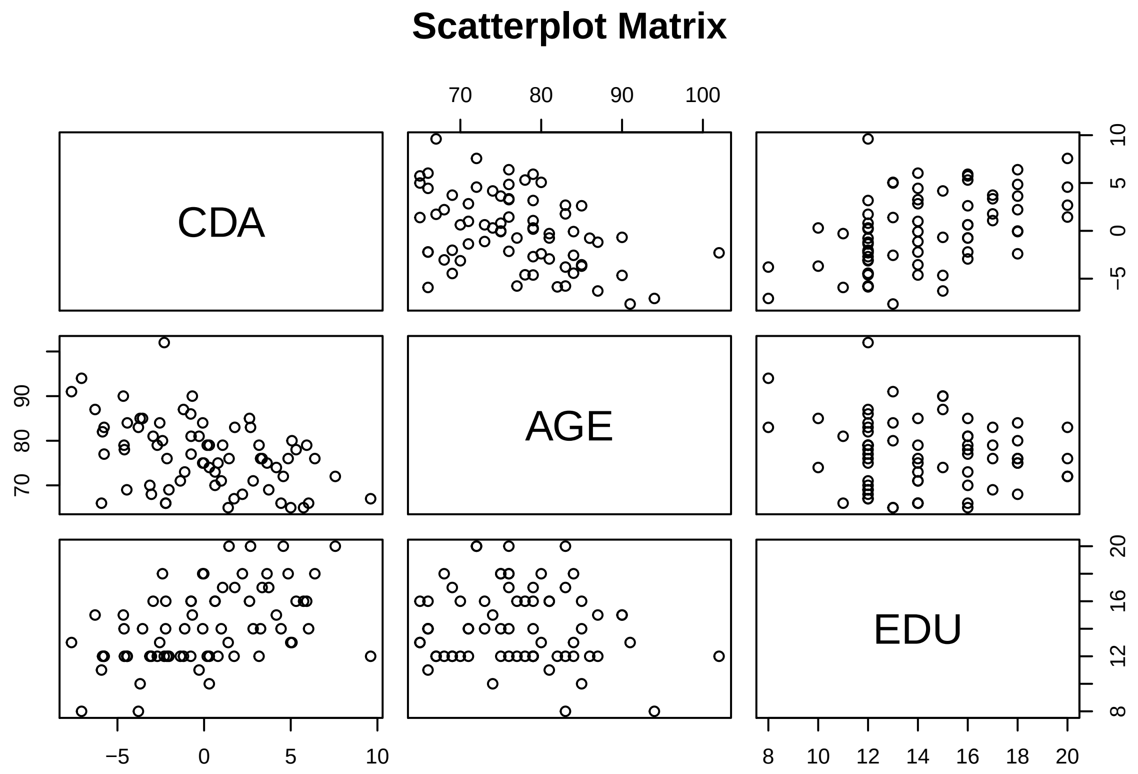
```r
         12, 12, 14, 16, 16, 16, 13, 12, 13,
         16, 16, 17, 8, 20, 12, 12, 14, 12,
         14, 12, 16, 12, 20, 10, 18, 14, 16,
         16, 18, 8, 18, 18, 17, 16, 12)
CDA <- c(4.57, -3.04, 1.39, -3.55, -2.56, -4.66, -2.70, 0.30, -4.46,
         -6.29, -4.43, 0.18, -1.37, 3.26, -1.12, -0.77, 3.73, -5.92,
         5.74, 2.83, -2.40, -0.29, 4.44, 3.35, -3.13, -2.14, 9.61,
         7.57, 2.21, -2.30, 1.73, 6.03, -0.02, -7.65, 4.17, -0.68,
         3.17, -1.19, 0.99, -2.94, -2.21, -0.75, 5.07, -5.86, 5.00,
         0.63, 2.62, 1.77, -3.79, 1.44, -5.77, -5.77, -4.62, -2.03,
         -2.22, 0.80, -0.75, -4.60, 2.68, -3.69, 4.85, -0.08, 0.63,
         5.92, 3.63, -7.07, 6.39, -0.08, 1.07, 5.31, 0.30)
ATTENTION <- data.frame(ID, AGE, EDU, CDA)
str(ATTENTION)
```

```
## 'data.frame':    71 obs. of  4 variables:
##  $ ID : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ AGE: num  72 68 65 85 84 90 79 74 69 87 ...
##  $ EDU: num  20 12 13 14 13 15 12 10 12 15 ...
##  $ CDA: num  4.57 -3.04 1.39 -3.55 -2.56 -4.66 -2.7 0.3 -4.46 -6.29 ...
```

```r
head(ATTENTION)
```

```
##   ID AGE EDU   CDA
## 1  1  72  20  4.57
## 2  2  68  12 -3.04
## 3  3  65  13  1.39
## 4  4  85  14 -3.55
## 5  5  84  13 -2.56
## 6  6  90  15 -4.66
```

```r
pairs(~CDA + AGE + EDU, data = ATTENTION,
      main = "Scatterplot Matrix")
```



```r
ATTENTION.LM1 <- lm(CDA ~ AGE + EDU, data = ATTENTION)
summary(ATTENTION.LM1)
```

```
##
## Call:
## lm(formula = CDA ~ AGE + EDU, data = ATTENTION)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9804 -2.2125 -0.0761  2.2824  9.1230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.49407    4.44297   1.237 0.220498
## AGE         -0.18412    0.04851  -3.795 0.000316 ***
## EDU          0.61078    0.13565   4.503  2.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.134 on 68 degrees of freedom
## Multiple R-squared:  0.3706, Adjusted R-squared:  0.3521
## F-statistic: 20.02 on 2 and 68 DF,  p-value: 1.454e-07
```

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

The estimated regression model is $CDA = 5.494 - 0.184AGE + 0.611EDU$.

Assumptions about the Observations

When we run a regression we are assuming that all observations are equally reliable and have an approximately equal role in determining the regression result. We need to ensure that the fit of the model is not overly determined by one or few observations.

An observation is *influential* if its deletion, singly or in combination with other observations, causes substantial changes in the fitted model (estimated coefficients, fitten values, t-tests, etc.). In general, deletion of any observation will cause changes in the fit. However, we are interested in detecting those points whose deletion causes large changes. That is, these points exercise undue influence.

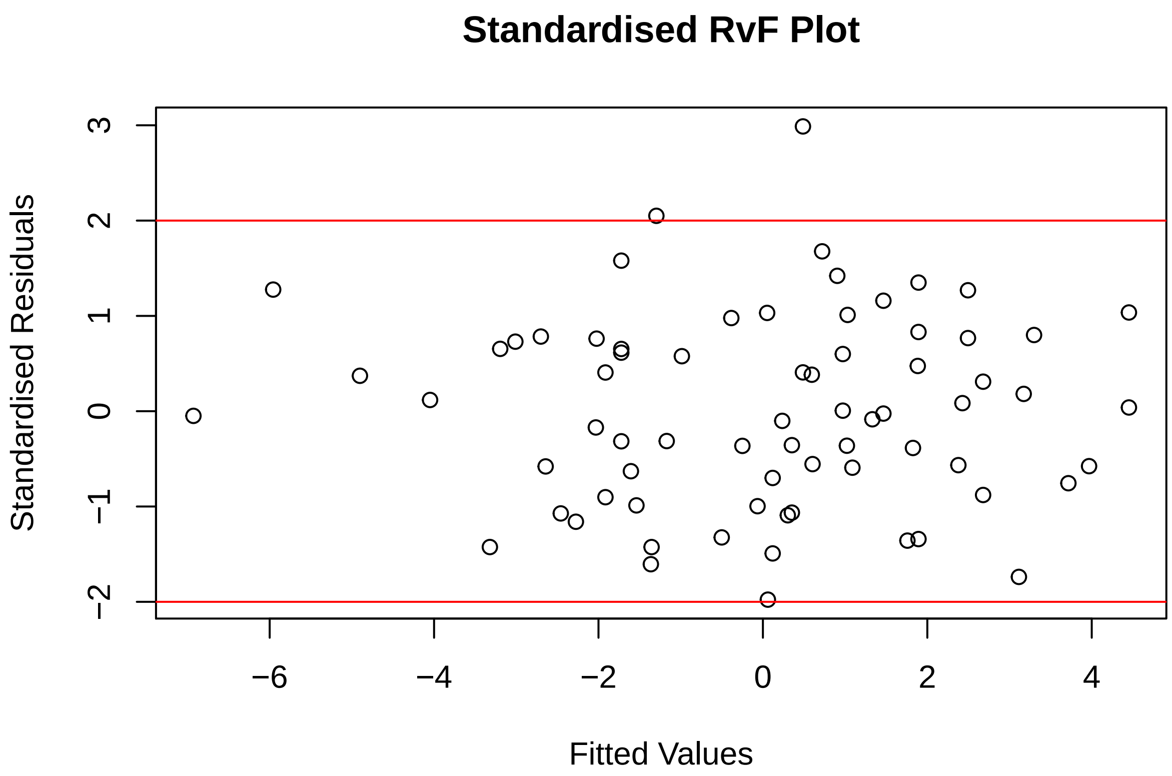
Outliers in the Dependent Variable

Observations with large standardised residuals are outliers in the dependent variable because they lie far from the fitted equation in the vertical direction. Points with standardised residuals larger than 2 are called *outliers*.

Example 1A

Detect the presence of outliers in the dependent variable.

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



Here we can see that there are two observations greater than two stsandard deviations larger than the mean and there are no observations less than two standard deviations below the mean. The two obserevations are

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

```
##    ID AGE EDU  CDA     FITTED    RESID   RSTAND
## 27 27  67  12 9.61  0.4870423 9.122958 2.987872
## 43 43  80  13 5.07 -1.2958039 6.365804 2.050035
```

Let us run the model with and without the two outliers.

```
summary(ATTENTION.LM1)
```

```
##
## Call:
## lm(formula = CDA ~ AGE + EDU, data = ATTENTION)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9804 -2.2125 -0.0761  2.2824  9.1230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.49407    4.44297   1.237 0.220498
## AGE         -0.18412    0.04851  -3.795 0.000316 ***
## EDU          0.61078    0.13565   4.503  2.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.134 on 68 degrees of freedom
## Multiple R-squared:  0.3706, Adjusted R-squared:  0.3521
## F-statistic: 20.02 on 2 and 68 DF,  p-value: 1.454e-07
```

```
ATTENTION.LM2 <- lm(CDA ~ AGE + EDU, data = subset(ATTENTION, abs(RSTAND)<2))
summary(ATTENTION.LM2)
```

```
##
## Call:
## lm(formula = CDA ~ AGE + EDU, data = subset(ATTENTION, abs(RSTAND) <
##     2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3222 -2.3861 -0.0039  2.3463  5.2749
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.77070    4.13020   0.671 0.504663
## AGE         -0.16318    0.04485  -3.639 0.000539 ***
## EDU          0.67287    0.12453   5.403 9.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.854 on 66 degrees of freedom
## Multiple R-squared:  0.4285, Adjusted R-squared:  0.4112
## F-statistic: 24.74 on 2 and 66 DF,  p-value: 9.589e-09
```

The results are compared in Table 1.

**Table 1. Comparison regression estimates between the two fitted models produced in Example 1A.

Variable	All Data	No Outliers	% Change
b_0	5.494	2.771	-49.56
b_1	-0.184	-0.163	-11.41
b_2	0.611	0.673	10.15

Note how the intercept estimate is reduced by about 50% if the two outliers are removed.

IMPORTANT: DELETING observations simply because they are outliers SHOULD NOT BE DONE. Instead, the purpose of diagnostic tests is to identify these observations so that you can confirm that these are indeed accurate. DO NOT DELETE OBSERVATIONS simply because they do not behave as your model expects.

IMPORTANT: Why did I delete the observations when I ran the second model? I did that to demonstrate the meaning of the term “influential” points. In practice, I would not have been so casual with my actions.

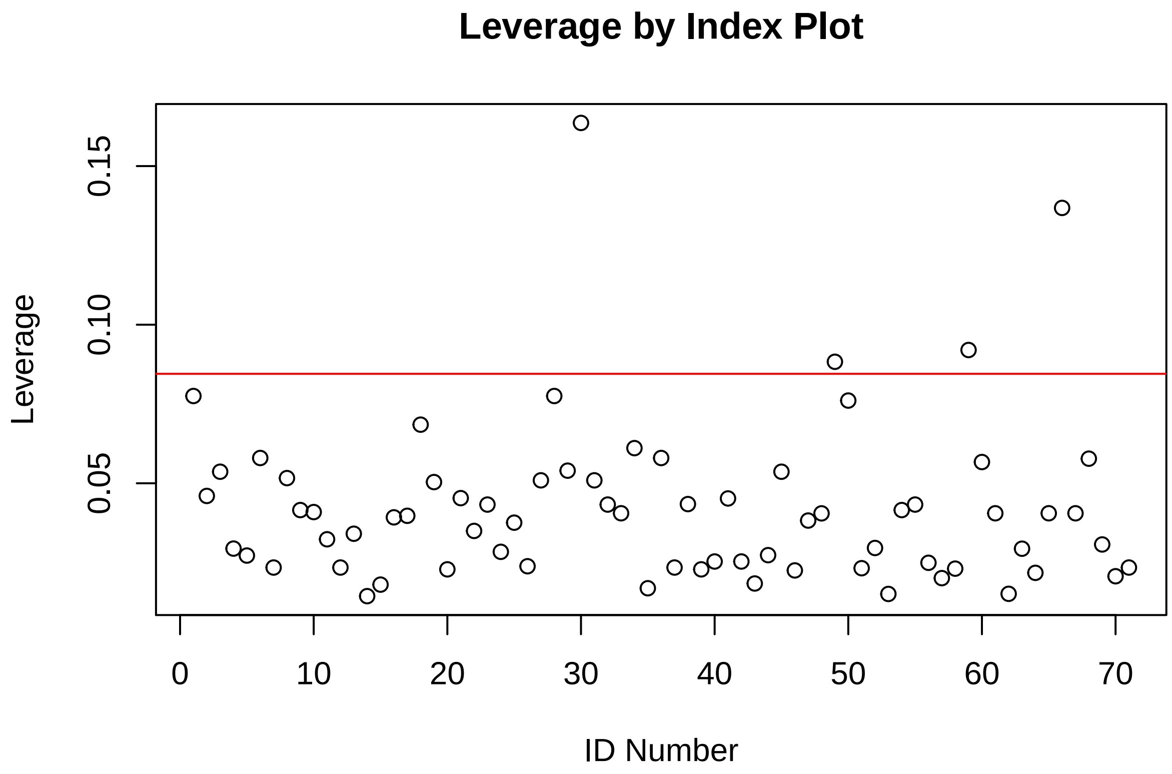
Outliers in the Independent Variable

Outliers can also occur in the independent variable space (that is, the horizontal axis). The diagonal of the *projection matrix* (a concept NOT tackled here) can be used to measure outliers in the predictor variables. Because the prjection matrix is also called the *hat matrix*, we call these values *hat* values. Another word for them is *leverage*. Outliers that are outliers in the horizontal direction are called high-leverage points. Hat values greater than $2(p+1)/n$, where p is the number of independent variables, is the cutoff.

Example 1B

Detect high-leverage values.

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



The plot shows that four observations are high-leverage points.

```
ATTENTION[ATTENTION$HAT > HAT.CUT,]
```

```
##    ID AGE EDU   CDA    FITTED      RESID      RSTAND        HAT
## 30 30 102  12 -2.30 -5.957332  3.6573322  1.27595208 0.16360487
## 49 49  83   8 -3.79 -4.902072  1.1120718  0.37160763 0.08830865
## 59 59  83  20  2.68  2.427271  0.2527288  0.08462341 0.09201248
## 66 66  94   8 -7.07 -6.927447 -0.1425534 -0.04895542 0.13681302
```

These points are different from the ones identified by the standardised residuals. You will note that observation 30 and 66 are very old individuals.

Cook’s Distance

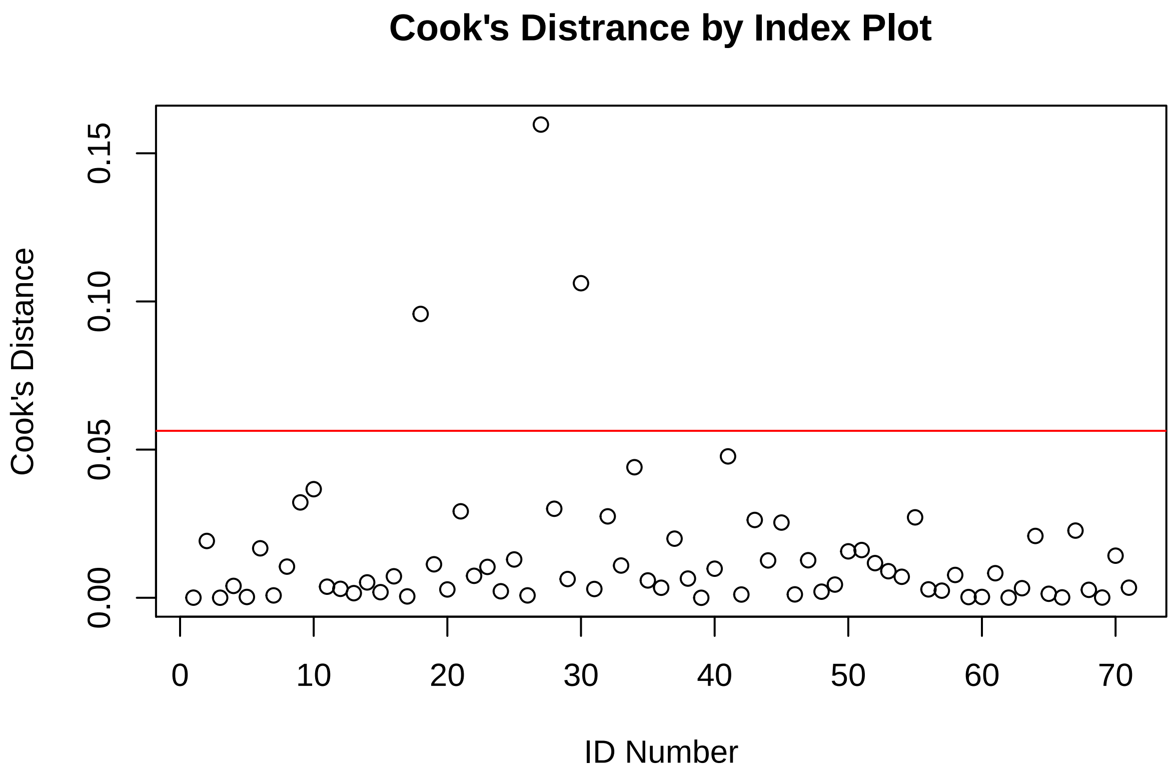
An influence measure developed in 1977 by Cook attempts to combine the effect of outliers in the dependent variable with the outliers in the independent variable in a singel measure. In addition, it measures the difference between regression coefficients obtained from the full dataset and the regression coefficeints obtained by deleting the ith observation. You can think of Cook's distance as the produce of standardised resituals and hats.

The cutoff value for Cook's distance is $4/n$.

Example 1C

Examine Cook's D.

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



The plot shows that there are three observations with high Cook's distance values.

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

```
##    ID AGE EDU   CDA      FITTED     RESID    RSTAND        HAT       COOK
## 18 18  66  11 -5.92  0.06038868 -5.980389 -1.977012 0.06848104 0.09578024
## 27 27  67  12  9.61  0.48704228  9.122958  2.987872 0.05092909 0.15968720
## 30 30 102  12 -2.30 -5.95733219  3.657332  1.275952 0.16360487 0.10615298
```

The three observations are ID18, ID27 and ID30. The last two were identified previously as having high standardised residuals or high hat values. Observation ID18 was not detected previously.

What To Do with the Outliers?

Outliers and influential observations should not routinely be deleted or automatically down-weighted because they are not necessarily bad observations. On the contrary, if they are correct, they may be the most informative points in the data. For example, they may indicate that the data did not come from a normal population or that the model is not linear.

References

1. Jansen DA, Keller ML. Cognitive function in community-dwelling elderly women. *Journal of Gerontological Nursing.* 2003;29:34-43.

THE END