Week ON8

Elmer V Villanueva

13 April 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON8")
```

Announcements

- You would hav received communications from the University regarding the teaching arrangements for the rest of the semester. We have been instructed to continue online delivery for the remainder of the semester. There will be no changes to the assessment regime that I sent to you, since all our assessments are coursework submissions.
- I will continue as your instructor for the remainder of the semester.
- I will release the two coursework assessments and the final report next week. You will not be able to submit your work before the due dates. However, you will be able to read and consider the material. Remember, all assessments are to be *COMPLETED INDIVIDUALLY*.
- The running count for students forwarding errors is as follows:

Student	Items Identified
Yijia Jiang	6
Xinwen Hu	1
Jing Wang	1
Yuxuan Wu	1

Reading

Read and understand Vittinghoff et al., Chapter 4.

Review of previous learning

In the past few weeks, we have focused on the situation of one dependent variable and more than one independent variables through the use of multiple linear regression models that take the form

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

where p is the number of predictor or independent variables.

When there are two predictor variables, the model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and is estimated as $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$, where y_i denotes the response in the ith trial and x_{i1} and x_{i2} are the values of the two predictor variables in the ith trial. The estimated parameters are b_0, b_1, b_2 and the error term is e_i. The

vertical rule established previously now becomes the vertical distance between y_i and the response plane. This represents the difference between y_i and the mean $\mu_{y_i|x_{i1},x_{i2}}$ of the probability distribution of y for the given (x_{i1}, x_{i2}) combination.

Visualising the data

Before analysing the data using multiple regression techniques, you must construct plots of the relationships among the variables. This is accomplished by making separate plots of each pair of variables, (X_1, X_2), (X_1, Y), and (X_2, Y). An easy way to do this is to construct a scatterplot matrix.

Obtaining the multiple regression equation

Unbiased estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the multiple linear regression model are obtained by the method of least squares. The output from R is very similar to that produced from a simple linear regression model.

Interpretation of regression coefficients

Intercept

The intercept is the value of the dependent variable when both independent variables are set at zero. Thus, if $x_1 = 0$ and $x_2 = 0$, then the regression equation reduces to $y = b_0$. When $x_1 = 0$ and $x_2 = 0$ are out of scope, then the intercept has no practical meaning.

Slope coefficients

The slope coefficient is interpreted as the change in the dependent variable for every unit increase in the specific independent variable while holding all other independent variables constant.

Example 1A

Let us use the example introduced in the previous session.Recall that Jansen and Keller [1] used age and education level to predict the capacity to direct attention (CDA) in elderly subjects. CDA refers to neural inhibitory mechanisms that focus the mind on what is meaningful while blocking out distractions. The study collected information on 71 community-dwelling older women with normal mental status. The CDA measurement was calculated from results on standard visual and auditory measures requiring the inhibition of competing and distracting stimuli. In this study, CDA scores ranged from -7:65 to 9.61 with higher scores corresponding with better attentional functioning. The measurements on CDA, age in years, and education level (years of schooling) for 71 subjects are shown in Table 1. We wish to obtain the sample multiple regression equation of CDA on AGE and EDU.

Table 1. CDA scores, ages, and education levels for 71 community-dwelling older women.

ID	AGE	EDU	CDA	ID	AGE	EDU	CDA
1	72	20	4.57	37	79	12	3.17
2	68	12	-3.04	38	87	12	-1.19
3	65	13	1.39	39	71	14	0.99
4	85	14	-3.55	40	81	16	-2.94
5	84	13	-2.56	41	66	16	-2.21
6	90	15	-4.66	42	81	16	-0.75
7	79	12	-2.70	43	80	13	5.07
8	74	10	0.30	44	82	12	-5.86
9	69	12	-4.46	45	65	13	5.00
10	87	15	-6.29	46	73	16	0.63
11	84	12	-4.43	47	85	16	2.62

ID	AGE	EDU	CDA	ID	AGE	EDU	CDA
12	79	12	0.18	48	83	17	1.77
13	71	12	-1.37	49	83	8	-3.79
14	76	14	3.26	50	76	20	1.44
15	73	14	-1.12	51	77	12	-5.77
16	86	12	-0.77	52	83	12	-5.77
17	69	17	3.73	53	79	14	-4.62
18	66	11	-5.92	54	69	12	-2.03
19	65	16	5.74	55	66	14	-2.22
20	71	14	2.83	56	75	12	0.80
21	80	18	-2.40	57	77	16	-0.75
22	81	11	-0.29	58	78	12	-4.60
23	66	14	4.44	59	83	20	2.68
24	76	17	3.35	60	85	10	-3.69
25	70	12	-3.13	61	76	18	4.85
26	76	12	-2.14	62	75	14	-0.08
27	67	12	9.61	63	70	16	0.63
28	72	20	7.57	64	79	16	5.92
29	68	18	2.21	65	75	18	3.63
30	102	12	-2.30	66	94	8	-7.07
31	67	12	1.73	67	76	18	6.39
32	66	14	6.03	68	84	18	-0.08
33	75	18	-0.02	69	79	17	1.07
34	91	13	-7.65	70	78	16	5.31
35	74	15	4.17	71	79	12	0.30
36	90	15	-0.68				

```
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,
         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)
```

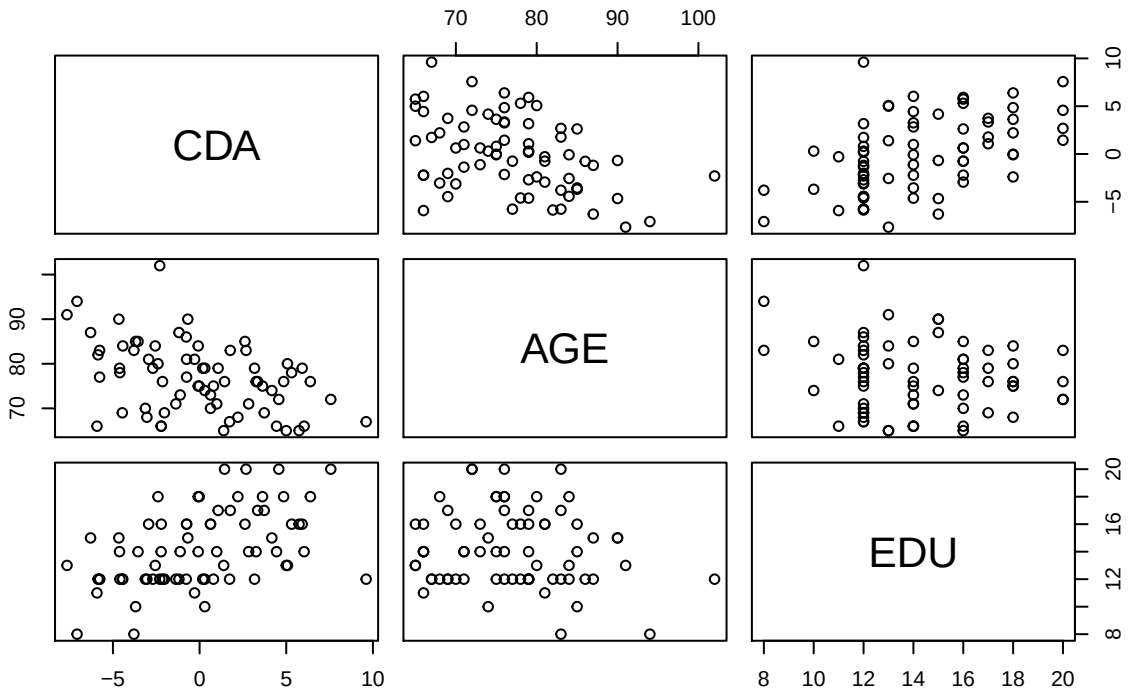
```
ATTENTION <- data.frame(ID, AGE, EDU, CDA)
str(ATTENTION)
```

```
## '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 ...
```

```
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
```

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




```
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
```

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

The mean CDA score when age is zero and education is zero is 5.494. For every extra year of life lived, the expected CDA score decreases by 0.184 points, holding education constant. The mean CDA score increases by 0.611 points for every year of education completed after adjusting for age.

In this session we will learn to test the model that we just developed. As described previously, the testing phase – termed *model diagnostics* – is important because models may not always be appropriate.

Model Assumptions

We begin by stating the assumptions of the multiple linear regression model. We divide these assumptions into four types: assumptions about (1) form, (2) errors, (3) predictors and (4) observations. We will discuss each in turn.

IMPORTANT: Plot all the things!

- Plot the residuals against each of the predictors
- Plot the residuals against the fitted values (the so-called RvF plot we learned last time)
- Plot the residuals against a Gaussian curve (the qq plot)

Assumptions about the Form of the Model

This is the linearity assumption that we met previously. In the simple linear regression case, it was easy for us to test this suing a scatterplot of the residuals versus the fitted values. This is the case here, too.

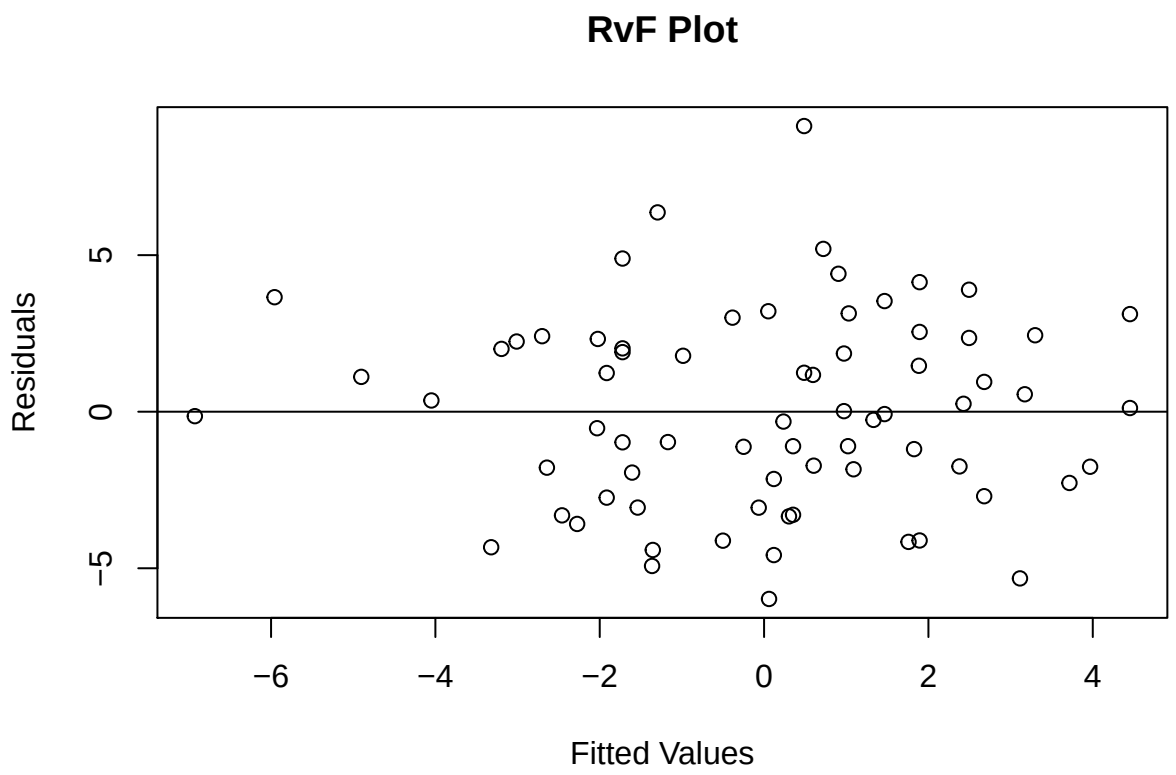
In addition, we can check the residuals against each of the x_p values. We are looking for evidence of non-linear patterns.

Example 1B

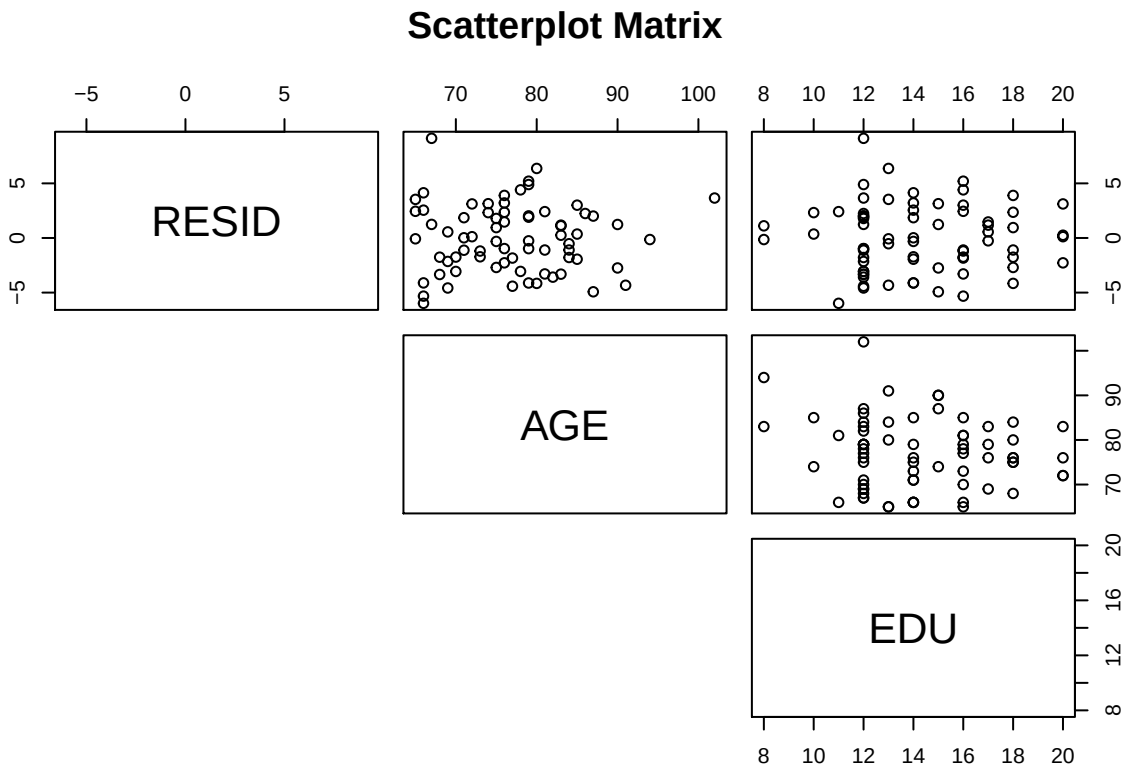
Test the linearity assumption of the CDA data.

```
ATTENTION$FITTED <- predict(ATTENTION.LM1, type = "response")
ATTENTION$RESID <- resid(ATTENTION.LM1)
plot(RESID ~ FITTED, data = ATTENTION,
     main = "RvF Plot",
     ylab = "Residuals",
```

```
     xlab = "Fitted Values")
abline(h = 0)
```




```
pairs(~RESID + AGE + EDU, data = ATTENTION,
      lower.panel = NULL,
      main = "Scatterplot Matrix")
```



In the RvF plot, we do not notice any systematic non-linear deviation. In the scatterplot matrix, we are only interested in the first row (that is, (1) RESID versus AGE and (2) RESID versus EDU). Note that neither shows any systematic non-linear deviations. *Conclusion: the assumption of linearity is not violated.*

If the graphs show evidence of the violation of this assumption, you should try transforming the data.

Assumptions about the Errors

The errors are assumed to be independent and identically dsitributed normal random variables with a mean of zero and a common variance σ^2. This implies four criteria:

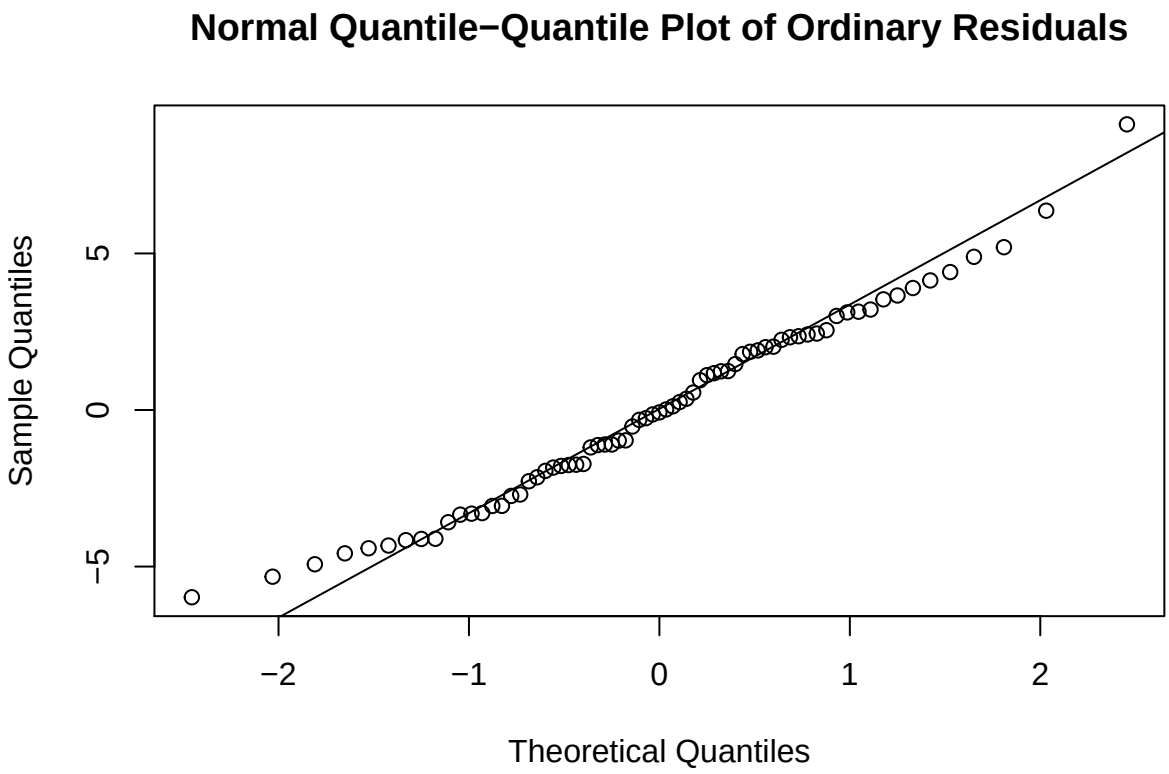
- Errors are normally distributed. This is the normality assumption.
- Errors have a mean of zero.
- Errors have an unknown but constant variance. This is the homoskedasticity assumption.
- Errors are independent of each other. This is the assumption of independence.

The normality assumption is tested by producing a quantile-quantile plot of the residuals.

Example 1C

Next the assumption of normality of the CDA data.

```
qqnorm(ATTENTION$RESID,
       main = "Normal Quantile-Quantile Plot of Ordinary Residuals")
qqline(ATTENTION$RESID)
```



The graph shows some minor deviation in the left tail, but this only affects less than 10 observations. Overall, the points hug the qq line quite closely. *Conclusion: the assumption of normality is not violated.*

If the graphs show evidence of the violation of this assumption, you should try transforming the data.

Testing the assumption of homoskedasticity also involves the examination of the RvF plot and the residduals versus each fo the predictors. We produced those plots in Example 1B above. This time, we are looking for wedge-shaped patterns or trumpteting.

Example 1D

Test the assumption of homoskedasticity. The graphs in Example 1B above do not show any wedge-shaped patterns. *Conclusion: the assumption of homoskedasticity is not violated.*

As before, evidence about independence is available from the design of the data.

Assumptions about the Predictors

There are three assumptions about the predictors. However, the first two are almost never validated; they are taken as given. The last assumption needs to be examined because its violation can lead to severe problems in the model.

The first assumption is that the independent variables are nonrandom. That is, they are assumed to be fixed and selected in advance. This assumption is satisfied when the experimenter can set the values of the independent variables at predetermined levels. Of course, in nonexperimental or observational studies, this assumption cannot be satisfied.

The second assumption is that the independent variables are measured without error. Any errors in measurement will affect the residual variance, the multiple correlation coefficient and the estimates of the

regression coefficient.

Note that these two assumptions are assumptions that we held in the case of the simple linear regression model. They simply carry over to the multiple linear regression model. As stated earlier, in this module, we accept these assumptions are being fulfilled.

The third assumption is that the independent variables are linearly independent of each other. That is, there is no linear equation that relates one independent variable to another. This assumption is needed to guarantee the uniqueness of the least squares solution used to derive estimates of the regression equation. If this assumption is violated, then there are no unique solutions to the normal equations. The violation of this assumption is called the problem of *collinearity*.

Example 2

To see the effect of the collinearity problem, consider the sample observations shown in Table 2.

Table 2. Sample observations demonstrating collinearity.

Y	X1	X2
23	2	6
83	8	9
63	6	8
103	10	10

Student A was asked to fit the multiple linear regression model

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Student A returned with the function

$\hat{y}_A = -87 + x_1 + 18x_2$

This equation fit the data perfectly.

Student B was asked to fit the same model and returned with a different function

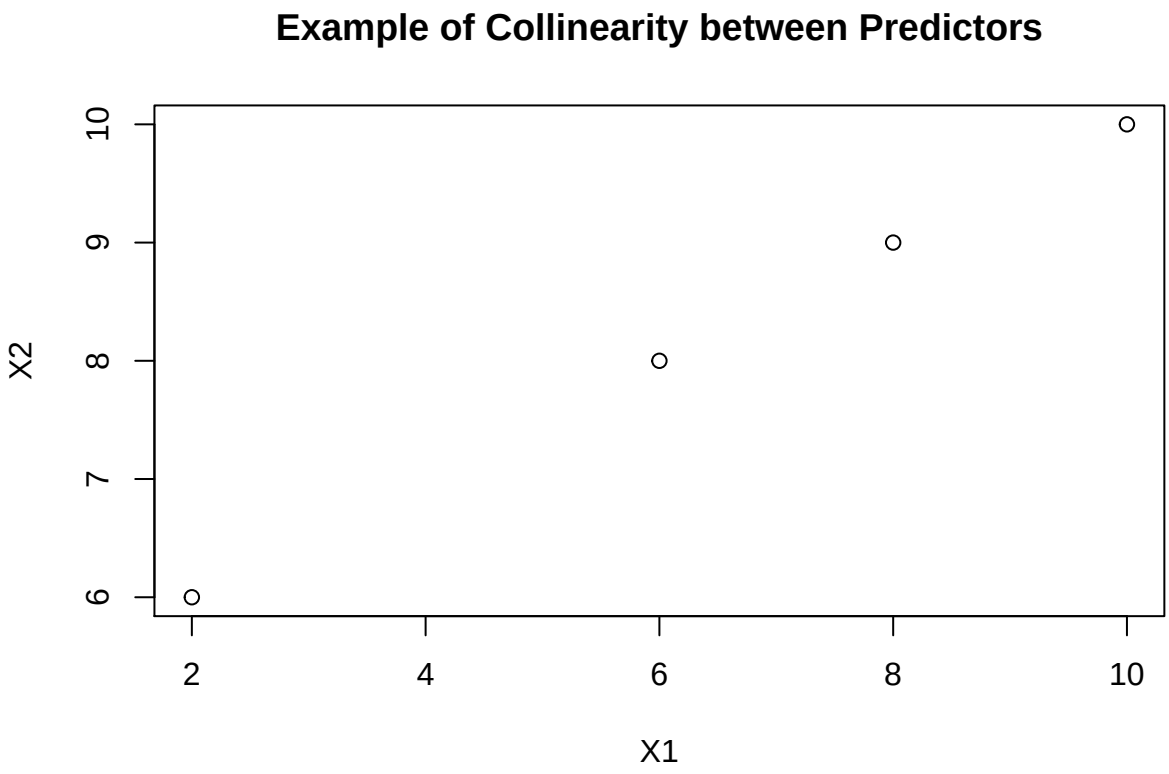
$\hat{y}_B = -7 + 9x_1 + 2x_2$

This function *ALSO* fit the data perfectly!

Indeed, it can be shown that there are infinitely many functions that will fit the data perfectly. The reason for this is that the independent variables X_1 and X_2 are perfectly related according to the relation $X_2 = 5 + 0.5X_1$. The effects of this are twofold. First, the perfect relationship between the independent variables did not inhibit our ability to obtain a good fit. In fact, we produced equations that fit the data perfectly. Second, Since there are many functions that produce the same good fit, we cannot interpret any one set of regression coefficients.

How is this situation detected? Well, relatively easy, it appears.

```
COLLI <- data.frame(Y <- c(23, 83, 63, 103),
                    X1 <- c(2, 8, 6, 10),
                    X2 <- c(6, 9, 8, 10))
plot(COLLI$X1, COLLI$X2,
     ylab = "X2",
     xlab = "X1",
     main = "Example of Collinearity between Predictors")
```



Note that evidence of collinearity is detected by a simple scatterplot of the independent variables. Here, the strong linear relationship is quite apparent. This is further evidence, if you required it, that proper exploratory data analysis is very useful prior to any formal regression modeling.

What happens if we fit the data anyway?

```
COLLI.LM <- lm(Y ~ X1 + X2, data = COLLI)
summary(COLLI.LM)
```

```
## Warning in summary.lm(COLLI.LM): essentially perfect fit: summary may be
## unreliable
```

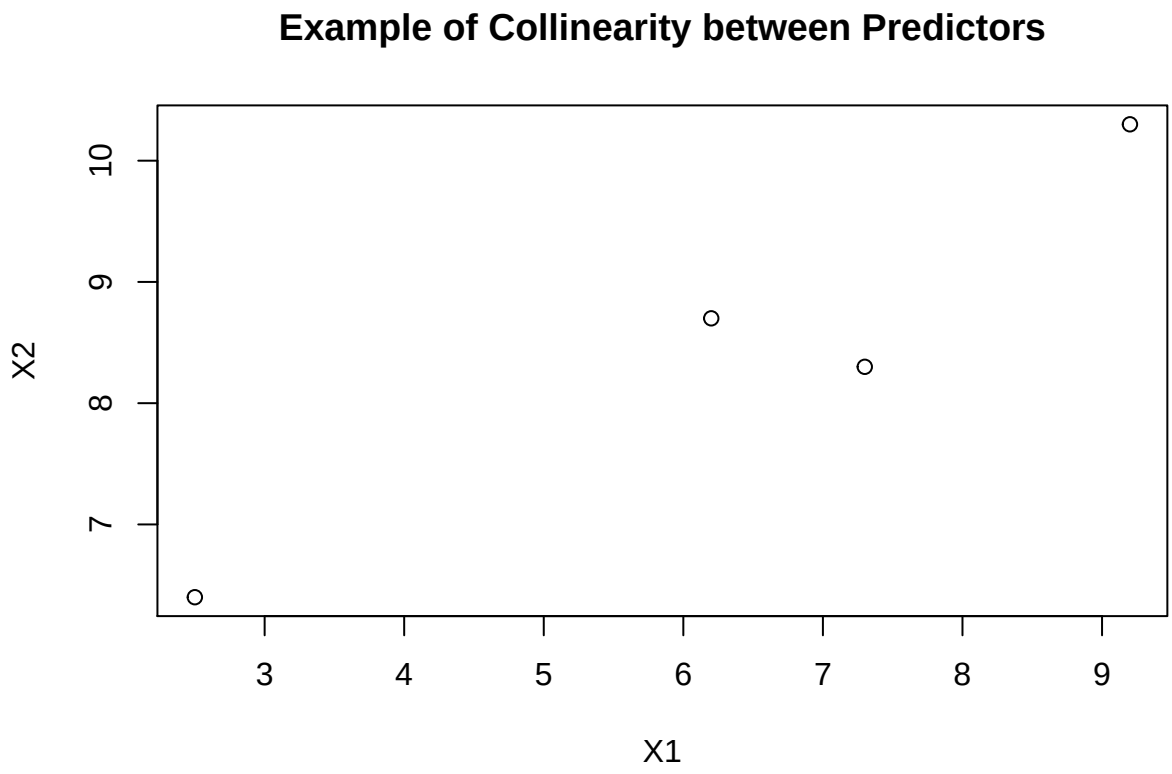
```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = COLLI)
##
## Residuals:
##          1          2          3          4
##  1.470e-15 -5.515e-15 -1.834e-16  4.228e-15
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 3.000e+00  6.065e-15 4.946e+14   <2e-16 ***
## X1          1.000e+01  8.493e-16 1.177e+16   <2e-16 ***
## X2                 NA         NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.024e-15 on 2 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.386e+32 on 1 and 2 DF,  p-value: < 2.2e-16
```

In the present case, R provides an error because of the perfect relationship between the two independent variables. If you look at the regression estimates, you'll note that R has dropped b_2.

Let us make this a little interesting by wiggling the data a little

```
COLLI2 <- data.frame(Y <- c(25, 81, 65, 110),
                     X1 <- c(2.5, 7.3, 6.2, 9.2),
                     X2 <- c(6.4, 8.3, 8.7, 10.3))
plot(COLLI2$X1, COLLI2$X2,
     ylab = "X2",
     xlab = "X1",
     main = "Example of Collinearity between Predictors")
```



Note that the relationship between the predictors is still present, but not as perfect. In fact, the correlation is 95.72%.

```
cor(COLLI2$X1, COLLI2$X2)
```

```
## [1] 0.957156
```

Let's see what R does when we fit the data.

```
COLLI2.LM <- lm(Y ~ X1 + X2, data = COLLI2)
summary(COLLI2.LM)
```

```
##
```

```
## Call:
## lm(formula = Y ~ X1 + X2, data = COLLI2)
##
## Residuals:
##       1       2       3       4
##  2.2252 -0.7367 -4.5030  3.0145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -16.214     37.602  -0.431    0.741
## X1            11.678      4.174   2.798    0.219
## X2             1.530      7.344   0.208    0.869
##
## Residual standard error: 5.904 on 1 degrees of freedom
## Multiple R-squared:  0.9908, Adjusted R-squared:  0.9723
## F-statistic: 53.59 on 2 and 1 DF,  p-value: 0.09615
```

You'll note that R did not complain, despite the presence of a strong relationship between the two independent variables.

In this case, there is one other way to detect the presence of collinearity. This is to calculate the *variance inflation factor* or VIF. This is available through the `car` package.

```
if(!require(car)){install.packages("car")}
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
library(car)
vif(COLLI2.LM)
```

```
##       X1       X2
## 11.92571 11.92571
```

IMPORTANT: Any $VIF(b_p) > 10$ is a sign of collinearity. Some use a threhold of 5.

In this example, the VIFs of both b_1 and b_2 are greater than 10, indicating the presence of collienarity.

When collinearity is detected, then you must choose to drop one of the variables and refit the model.

IMPORTANT: Once a varaible is dropped and the model refit, then ALL the diagnostic tests must be performed again.

Example 1E

Test for collienarity in the CDA data.

The scatterplot matrix in Example 1B shows no strong relationship between AGE and EDU. The variance inflation factors from the model are all less than 10.

```
vif(ATTENTION.LM1)
```

```
##      AGE      EDU
## 1.018724 1.018724
```

Conclusion: there is no evidence that collinearity is a problem in these data.

Assumptions about the Observations

We will leave this topic for next week, when we will close the session on multiple linear regression. Next week will also be the time you will have enough knowledge to tackle the first coursework assessment.

References

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

THE END