Week ON5

Elmer V Villanueva

23 March 2020

SET YOUR WORKING DIRECTORY!

```r
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON5")
```

Announcements and clarifications

- Students who find new errors in the learning materials, assignment or assignment solutions should bring these to my attention via email. Genuine errors will be corrected and extra marks will be awarded. The following students have earned extra marks

Student	Items Identified
Yijia Jiang	5
Xinwen Hu	1
Jing Wang	1
Yuxuan Wu	1

- As I mentioned, the assessment of the potential violation of assumptions has an element of subjectivity. Thus, it is up to you to justify your assessments. You cannot simply state that "the assumption of linearity was met" without explaining why you think so. For the most part, reasonable explanations will garner reasonable marks.

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 one independent variable through the use of simple linear regression models. In the following sessions, we will consider the situation of one dependent variable and more than one independent variable. These models are called *multiple linear regression* models and 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.

Multiple linear regression models with two predictors

Multiple linear regression analysis is one of the most widely used of all statistical methods. Let us start our discussion with the case of the *simplest* of the multiple linear regression models – the situation of two predictor 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 $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$. This model is called a first-order model with two predictor variables, 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.

In the case of simple linear regression in which the model $Y = \beta_0 + \beta_1 X$ produces a striaght line, the multiple linear regression model with two predictors $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ can be visualised as producing a plane. Thus, frequently the regression function in multiple regression is called a *regression surface* or a *response surface*.

The vertical rule established previously between 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. Hence, the vertical distance from y_i to the response plane represents the error term $e_i = y_i - \hat{y}_i$.

Example 1A

Jansen and Keller (A-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
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

ID	AGE	EDU	CDA	ID	AGE	EDU	CDA
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				

Let's enter the data.

```r
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
```

Visualising the data

IMPORTANT: It is *inadvisable* to visualise the data using a three-dimensional scatterplot. Naive analysts think that 3D plots look "cool". This is not the case. They are known to decrease the accuracy and speed at which a reader interprets the information. At times, they can be quite misleading. In this module, you are *NOT* allowed to use 3D scatterplots.

IMPORTANT: If you choose to present your work using a 3D scatterplot, I suggest that you use `rgl` or `plotly`, which produce interactive plots that allow the user to move the graph. Both these functions are *NOT* discussed here. Being interactive, you cannot include them in static reports such as those using conventional PDF formats.

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

```
## Loading required package: rgl
```

```
## Warning: package 'rgl' was built under R version 3.6.3
```

```
library("rgl")
par(mar=c(0,0,0,0))
plot3d(
  x=ATTENTION$AGE, y=ATTENTION$EDU, z=ATTENTION$CDA,
  type = 'p',
  radius = .1,
  xlab="Age (years)", ylab="Educational Attainment (years)", zlab="CDA Score")
```

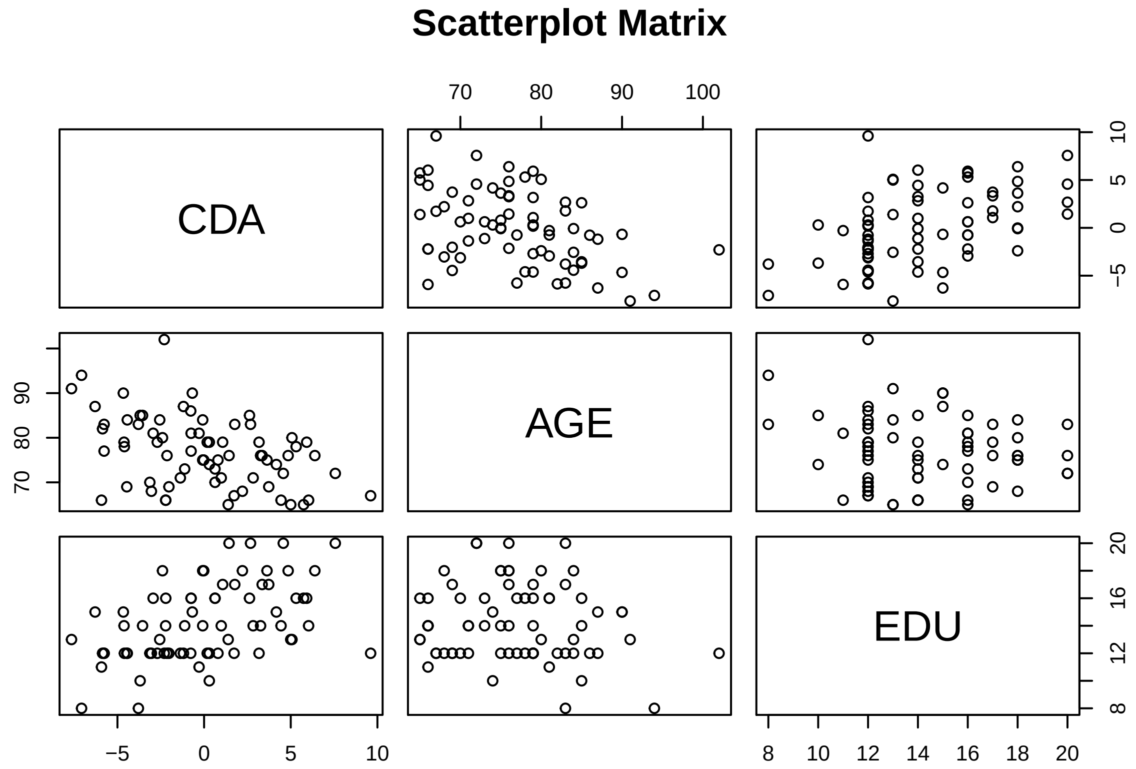
IMPORTANT: `ggplot2` does not support 3D graphics rendering. However, extensions to `ggplot2` such as `gg3d` can accommodate this chart type. This and other extensions are *NOT* discussed here.

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.

Basic scatterplot matrix in R base

Example 1B

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

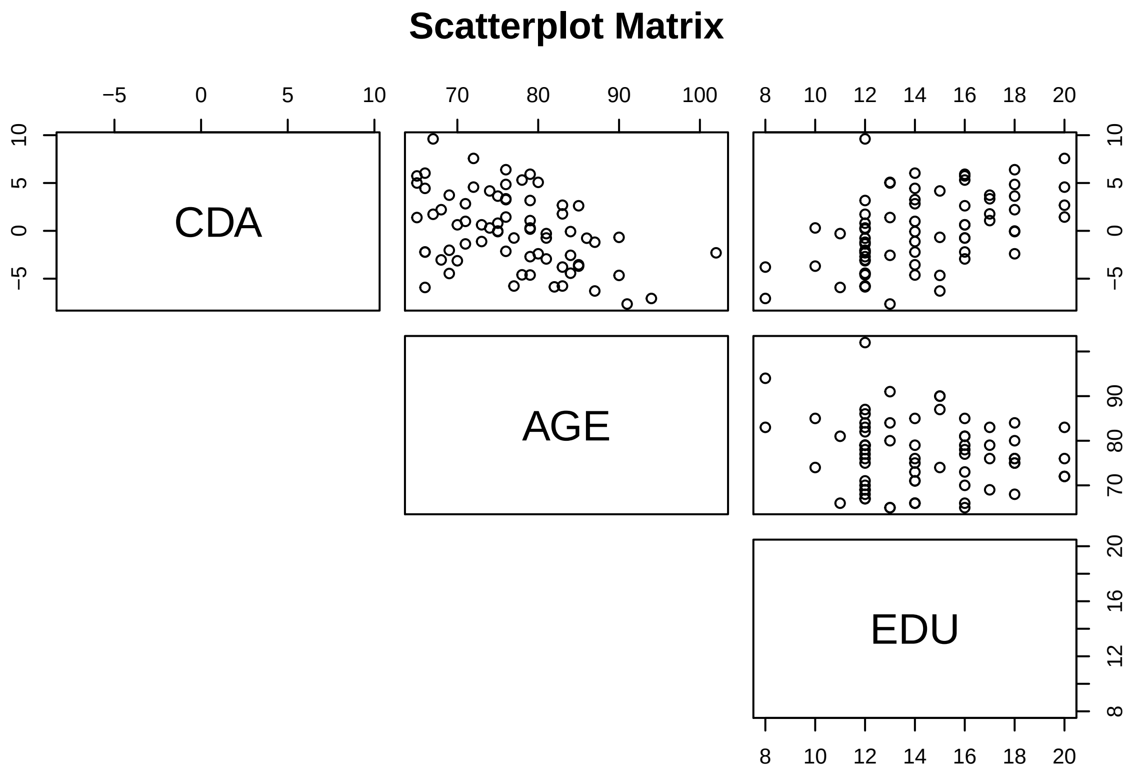


IMPORTANT: Note the function notation!

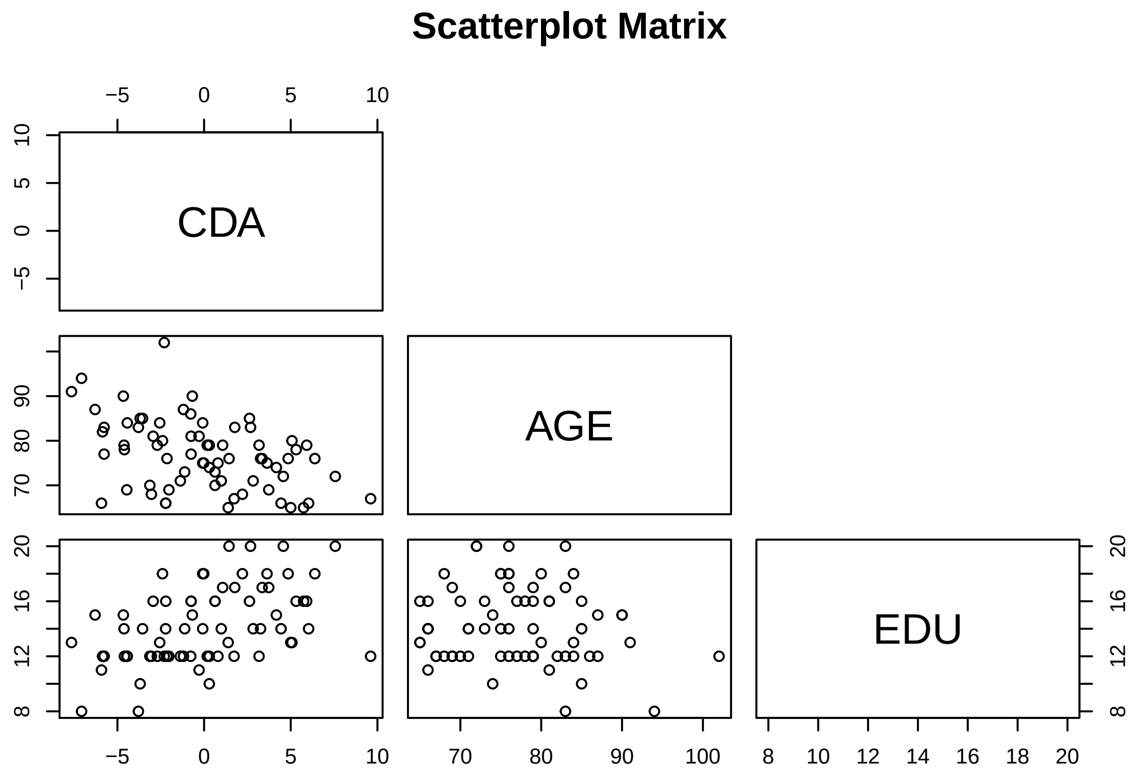
The scatterplot in the first row and the second column shows the relationship between CDA on the vertical axis and AGE on the horizontal axis. The scatterplot in the first row and the third column shows the relationship between CDA on the vertical axis and EDU on the horizontal axis. The scatterplot in the second row and the first column is a flipped version of the CDA vs AGE scatterplot, this time with the axes reversed.

Since the lower triangle and upper triangles are mirror images, it is sometimes useful to only present one or the other.

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



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



The results suggest that CDA is negatively correlated with AGE and positively correlated with EDU. Also, AGE and EDU show a small negative relationship.

Scatterplot matrices in the `car` package

Example 1C

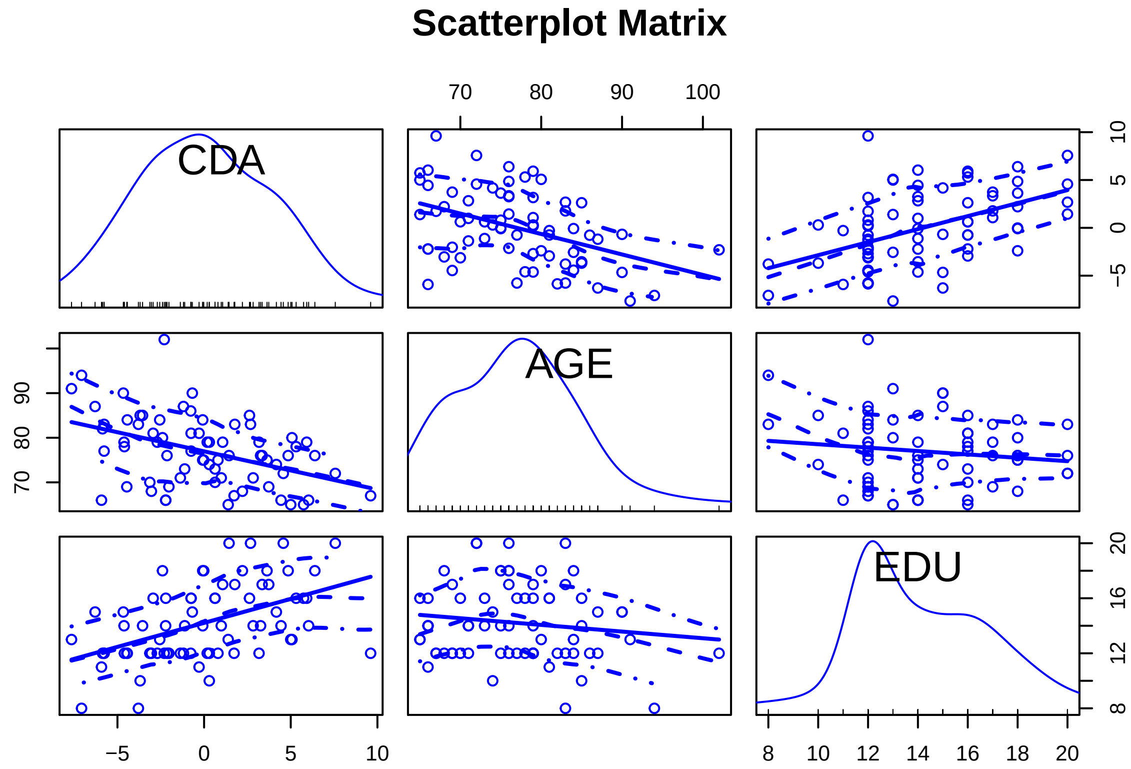
The `car` package produces scatterplot matrices with more information. However, you need to be careful that you are not overloaded with the complexity of the figure.

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

```
## Loading required package: car

## Loading required package: carData
```

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



I leave to you the interpretation of all the information presented here.

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. This means that the sum of the squared deviations of the observed values of the dependent variable y_i from the resulting response surface is minimised. In the case of two predictor variables, the method of least squares produces sample estimates for $\beta_0, \beta_1, \ldots, \beta_p$ so that

$\sum e_i^2 = \sum(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2 = \sum(y_i - \hat{y}_i)^2$

is minimised. While it is possible to calculate the regression estimates by hand using a calculator, it is more practical to use a statistical program.

Example 1D

Let us calculate the regression estimates for the `ATTENTION` dataset. Note that the same basic function `lm()` is used.

```
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 output is very similar to that produced from a simple linear regression model. In the first chunk, the regression equation is given. You need to confirm that this equation is not erroneous. In the second chunk, summary statistics for the residuals are shown. In the third chunk, the coefficients of the regression model are provided, including their p-values. The last chunk gives information about the model. (We will ignore the fourth chunk for now and discuss it next week.)

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

IMPORTANT: You will note that data for AGE and EDU are given as whole numbers while those for CDA are given to the hundredths place. Technically, the application of the reporting rules previously means that we should present the coefficients for AGE and EDU to the tenths place. At this point, the reporting rules are difficult to apply. Thus, I will allow the presentation of results to the precision required of the *dependent variable*.

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.

Example 1E

The mean CDA score when age is zero and education is zero is 5.494.

Slope coefficients

Recall that in the case of the simple linear regression model, the slope coefficient is the value of the dependent variable when the independent variable increases by one unit.

Example 1F

```
ATTENTION.LM2 <- lm(CDA ~ AGE, data = ATTENTION)
summary(ATTENTION.LM2)
```

```
##
## Call:
## lm(formula = CDA ~ AGE, data = ATTENTION)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -8.2659 -2.4254 -0.2197  3.0502  7.4779 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 16.45267    4.20424   3.913 0.000211 ***
## AGE         -0.21374    0.05437  -3.931 0.000198 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.545 on 69 degrees of freedom
## Multiple R-squared:  0.183,  Adjusted R-squared:  0.1712 
## F-statistic: 15.46 on 1 and 69 DF,  p-value: 0.0001982
```

In the simple linear regression model above, CDA score decreases by 0.214 points for every year of life lived.

In the multiple linear regression case, the interpretation of the slope coefficients takes on an important difference. To understand this, let's use an example.

Example 1G

What is the expected CDA score of a 65-year-old woman who has 12 years of previous schooling? Using the regression estimates, we calculate the answer to be

$CDA_1 = 5.494 - 0.184 AGE + 0.611 EDU = 5.494 - 0.184 \times 65 + 0.611 \times 12 = 0.866.$

Now, what is the expected CDA score of a woman who is one year older than our previous subject (i.e., the new subject is 66 years old) but has the same educational attainment?

$CDA_2 = 5.494 - 0.184 AGE + 0.611 EDU = 5.494 - 0.184 \times 66 + 0.611 \times 12 = 0.682,.$

Finally, what is the difference in CDA scores between these two women?

$\Delta CDA = CDA_2 - CDA_1 = 0.682 - 0.866 = -0.184,.$

Thus, the slope coefficient for `AGE` is interpreted as the change in CDA score for every unit increase in AGE while holding EDU constant. The interpretation of the coefficient for age is as follows: For every extra year of life lived, the expected CDA score decreases by 0.184 points, *holding education constant*.

IMPORTANT: Note the addition of the phrase *holding education constant*. This is the important difference between the interpretation of the slope coefficient in the simple linear regression model and the multiple linear regression model. If you do not add the phrase, then you are NOT providing the multiple linear regression interpretation.

IMPORTANT: Other valid phrases include *after adjusting for education* or *while controlling for education*. For example, you can state "For every year of life lived, the mean CDA score decreases by 0.184 points, while controlling for education."

Example 1H

Interpret the slope coefficient of education.

The mean CDA score increases by 0.611 points for every year of education completed after adjusting for age.

Inferences regarding the regression estimates

To test the null hypothesis that b_p is equal to some particular value, say, b_{p0}, the t statistic may be computed as $t = (b_p - b_{p0})/s_{b_p}$. We don't usually have a ready value for the standard deviation of the coefficient, so we substitute the standard error. The degrees of freedom is $n - p - 1$. This is called the *Wald test* and its result is provided under the last two columns of the third chunk of results in the R output.

Example 1I

Test the null hypotheses that $\beta_{AGE} = 0$ at $\alpha = 1\%$.

The output gives the value of the t statistic as -3.795 and the p-value is 0.000316. The p-value is less than α providing us with evidence to reject the null hypothesis.

Confidence intervals for the regression estimates

Confidence intervals for the b_p may be constructed in the usual way by using a value from the t distribution for the reliability factor and standard errors given above. A $100(1 - \alpha)$ percent confidence interval for b_p is given by

$b_p \pm t_{1-\alpha/2,df=n-p-1} s_{b_p}$

Example 1J

Construct 99% confidence intervals for the slope coefficients of age and educational attainment.

```
confint(ATTENTION.LM1, level = 0.99)
```

```
##                  0.5 %     99.5 %
## (Intercept) -6.2801566 17.2683031
## AGE         -0.3126912 -0.0555588
## EDU          0.2512943  0.9702629
```

The 99% confidence interval for AGE is -0.313 to -0.056. The 99% confidence interval for EDU is 0.251 to 0.970.

The confidence intervals and p-values are interpreted in the usual fashion.

Obtaining predictions from the regression equation

Just as was the case in simple linear regression, we may, in multiple regression, interpret a $\hat{y}$ value in one of two ways. First, we may interpret $\hat{y}$ as an estimate of the *mean* of the subpopulation of Y values assumed to exist for particular combinations of X_p values. Under this interpretation $\hat{y}$ is called an *estimate*, and when it is used for this purpose, the equation is thought of as an *estimating equation.*

The second interpretation of $\hat{y}$ is that it is the value Y is most likely to assume for given values of the X_p. In this case, $\hat{y}$ is called the *predicted value* of Y, and the equation is called a *prediction equation.* In both cases, intervals may be constructed about the $\hat{y}$ value when the normality assumption holds.

When $\hat{y}$ is interpreted as an estimate of a population mean, the interval is called a *confidence interval*, and when ˆy is interpreted as a predicted value of Y, the interval is called a *prediction interval.* The calculation of confidence and prediction intervals is complicated by the estimation of the standard error. Thus, it will not be covered in this module. Nevertheless, R provides us with easy ways to produce these estimates.

The confidence interval for the mean of a subpopulation of Y values given particular values of X_p

Example 1K

Estimate the 99% confidence interval of the mean CDA score for 75-year-old women with 12 years of education.

```
ATTENTION.NEW1 <- data.frame(AGE = 75, EDU = 12)
predict(ATTENTION.LM1, ATTENTION.NEW1,
        interval = "confidence",
        level = 0.99)
```

```
##          fit       lwr       upr
## 1 -0.9859576 -2.297386 0.3254708
```

The 99% confidence interval for the mean CDA score is between -2.297 and 0.325.

The prediction interval for a particular value of Y given particular values of X_p

Example 1L

Estimate the 99% confidence interval of the CDA score for a 77-year-old woman with 10 years of education.

```r
ATTENTION.NEW2 <- data.frame(AGE = 77, EDU = 10)
predict(ATTENTION.LM1, ATTENTION.NEW2,
        interval = "prediction",
        level = 0.99)
```

```
##         fit       lwr     upr
## 1 -2.575765 -11.07562 5.92409
```

The 99% prediction interval for the CDA score is between -11.076 and 5.924.

Multiple linear regression models with more than two predictors

In the situation in which there are more than two predictors, then the knowledge we have gained previously can be applied, too.

In the next session, we will learn to evaluate the assumptions of the multiple linear regression model.

References

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

THE END