Week ON13

Elmer V Villanueva

18 May 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON13")
```

Announcements

- The submission deadline of the first coursework assessment has passed. Not all students submitted. I will be accepting submissions until Friday 22 May. After that date, no submissions will be entertained.
- The remaining coursework assessment and the final paper have deadlines as follows

Assessment	Due Date	Days Till Deadline
coursework 2	30 May	12
Final Paper	17 June	30

- 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.
- The running count for students forwarding errors is as follows:

Student	Items Identified
Yijia Jiang	6
Jing Wang	3
Xinwen Hu	1
Yuxuan Wu	1

Reading

Read and understand Vittinghoff et al., Chapter 5.

Review of previous learning

We started this module understanting the simple linear regression model, which took the form

$Y = \beta_0 + \beta_1 X$.

Then, we expanded this model by including more intependent variables, resulting in the multiple linear regression model. This model took the form

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.

We studied the properties of this model, including its assumptions. We learned to interpret its output. We also focused on tests to ensure that the model we produced was valid.

While the linear regression model was very useful in explaining and predicting data, we learned that it did not work well for all types of dependent variables. That is, the linear regression model was inappropriate when the dependent variable was continuously scaled.

In Week 11, we expanded the developed regression models to accept binary categorical dependent variables. This type of model was called the logistic regression model and took the form

$f(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, where

$$f(Y) = logit(Y) = ln(Odds(Y)) = ln\left(\frac{Pr(Y)}{1 - Pr(Y)}\right)$$

We stated that by preseving the right-hand side of the regression model, many of the techiques we learned from much simpler models can carry over.

Finally, we stated that, under the logistic regression model, the estimates regression coefficients may be interpreted as chanes in the logit of Y. However, due to the unfriendliness (or difficulty) of the undestanding of the logit by lay persons, we may interpret the exponentiated regression estimates as odds ratios, instead.

Example 1A

we will make use of Jansen and Keller's data on attention in elderly subjects [1]. We have been using these data in linear regression models. You should review previous lecture notes if you need to re-familiarise yourself with these data.

Lete us enter the data and derive the linear regression model.

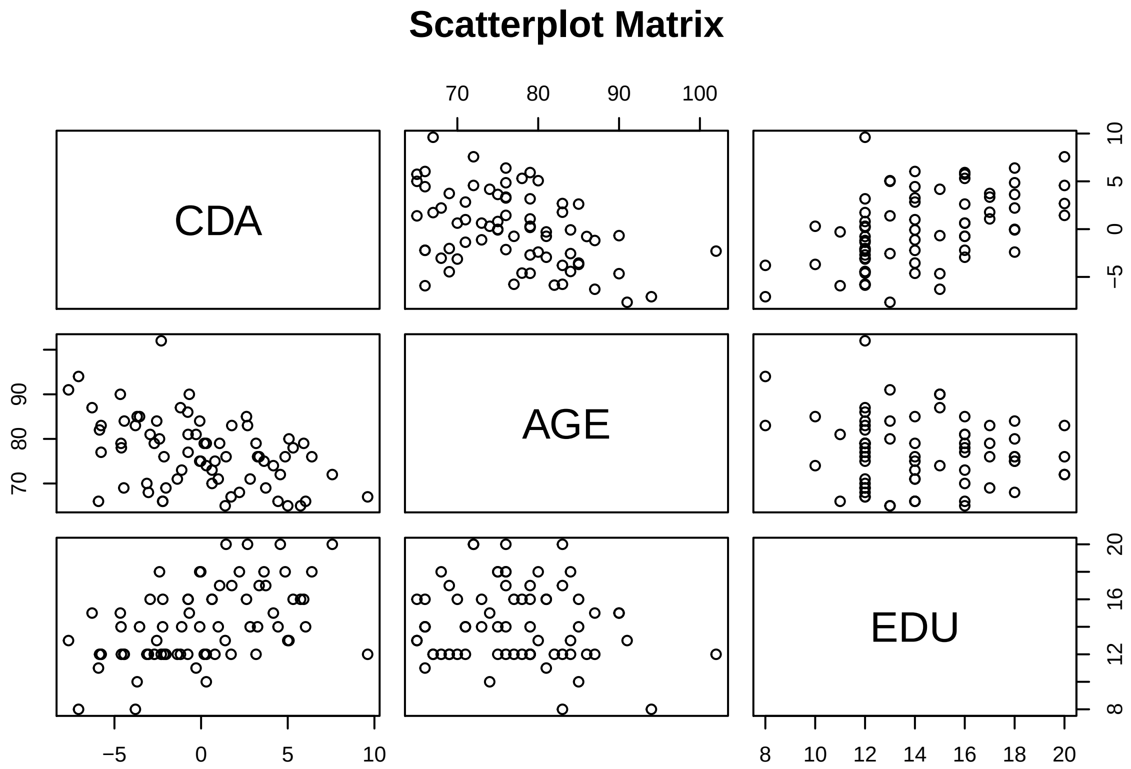
```
ID <- c(1:71)
AGE <- c(72, 68, 65, 85, 84, 90, 79, 74, 69,
         87, 84, 79, 71, 76, 73, 86, 69, 66,
         65, 71, 80, 81, 66, 76, 70, 76, 67,
         72, 68, 102, 67, 66, 75, 91, 74, 90,
         79, 87, 71, 81, 66, 81, 80, 82, 65,
         73, 85, 83, 83, 76, 77, 83, 79, 69,
         66, 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,
         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 model is $CDA = 5.494 - 0.184AGE + 0.611EDU$. This model estimates the CDA score given a subject's age and educational attainment.

Now, let us revise the problem a little

Example 1B

The higher the CDA score, the better a subject's capacity to direct attention. Conversely, the lower the CDA score, the worse the subject's attention capacity. Let us define subjects with CDA scores at or less than the first quartile as being of "low attention" and subjects with CDA scores above the first quartile as being of "normal attention".

The first quartile cut-off is

```
quantile(CDA, probs = 0.25)
```

```
##   25%
## -2.82
```

What, then, is the proper regression approach to analyse these dichotomisesd data?

First, let us produce our data.

```
CDA.BIN <- CDA < quantile(CDA, probs = 0.25)
CDA.BIN2 <- factor(CDA.BIN, labels = c("Normal", "Low"))
ATTENTION2 <- data.frame(ID, AGE, EDU, CDA.BIN2)
str(ATTENTION2)
```

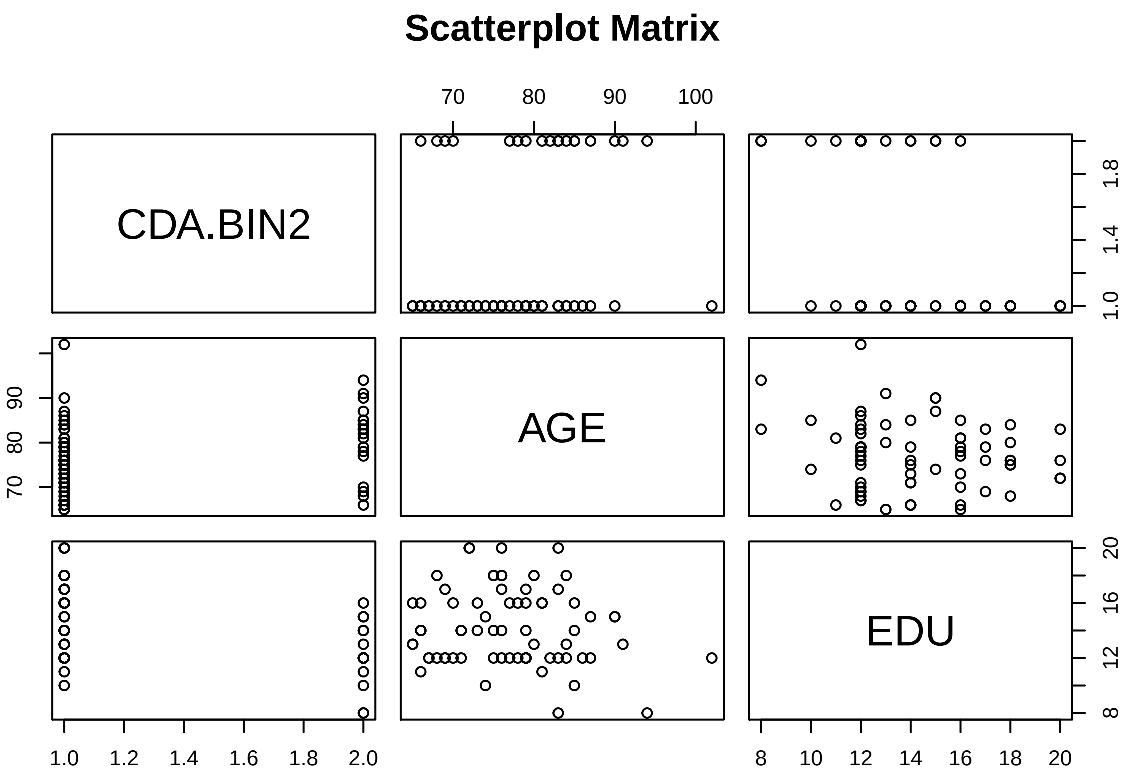
```
## 'data.frame':    71 obs. of  4 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ AGE     : num  72 68 65 85 84 90 79 74 69 87 ...
##  $ EDU     : num  20 12 13 14 13 15 12 10 12 15 ...
##  $ CDA.BIN2: Factor w/ 2 levels "Normal","Low": 1 2 1 2 1 2 1 1 2 2 ...
```

```
head(ATTENTION2)
```

```
##   ID AGE EDU CDA.BIN2
## 1  1  72  20   Normal
## 2  2  68  12      Low
## 3  3  65  13   Normal
## 4  4  85  14      Low
## 5  5  84  13   Normal
## 6  6  90  15      Low
```

Let us visualise the data.

```
pairs(~CDA.BIN2 + AGE + EDU, data = ATTENTION2,
      main = "Scatterplot Matrix")
```



The logistic regression model is

```
ATTENTION.LM2 <- glm(CDA.BIN2 ~ AGE + EDU, data = ATTENTION2,
                     family = binomial(link = 'logit'))
summary(ATTENTION.LM2)
```

```
##
## Call:
## glm(formula = CDA.BIN2 ~ AGE + EDU, family = binomial(link = "logit"),
##     data = ATTENTION2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7643  -0.6937  -0.3834   0.1155   2.0647
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.42492    3.76817  -0.113  0.91022
## AGE          0.07102    0.04037   1.759  0.07850 .
## EDU         -0.45833    0.15868  -2.888  0.00387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 80.396  on 70  degrees of freedom
## Residual deviance: 62.434  on 68  degrees of freedom
## AIC: 68.434
##
## Number of Fisher Scoring iterations: 5
```

The estimated reqression equation is $logit(CDA.BIN2) = -0.425 + 0.071AGE - 0.458EDU$.

Now this is a little strange. The above result states that as age increases, the outcome variable increases, after controlling for educational attainment. Why, then, does the linear regression model suggest the opposite? The same goes for educational attainment – in the linear regression model, the regression coefficient is positive, but in the logistic regression model, it is negative. Why are the two models inconsistent?

There is actually no consistency here. The problem arises because `R` is forced to choose a reference category when it analyses qualitative data. By default, `R` chooses the first level of a factor. In this case, the first factor is "Normal", so R has chosen that as the reference category. The interpretation of the model is based on this reference category. Thus, the logit of being in the "Low" category increases with age after adjusting for educational attainment. This is similar to the result of the linear regression model that states that the CDA score decreases with age after controlling for educational attainment.

The estimated regression coefficients are interpreted as follows.

- The odds ratio for age is $e^{0.071} = 1.073$. For every year of life lived, the odds of a "Low" CDA compared to a "Normal" CDA increases by $(1.073 - 1) \times 100\% = 7.3\%$ afer holding educational attainment constant.
- The odds ratio for educational attainment is $e^{-0.458} = 0.633$. For every year of education completed, the odds of a "Low" CDA compared to a "Normal" CDA changes by $(0.633 - 1) \times 100\% = -36.7\%$ after adjusting for age.

The 95% confidence intervals for the estimated slope coefficients are calculated as

```
exp(confint(ATTENTION.LM2, level = 0.95))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %       97.5 %
## (Intercept) 0.0003270204 1108.5696431
## AGE         0.9956891390    1.1694482
## EDU         0.4413836635    0.8325202
```

In tabular format, I would present the information as

Variable	Odds Ratio (95% CI)	p-value
Age, years	1.073 (0.996, 1.169)	0.0785
Educational attainment, years	0.633 (0.441, 0.833)	0.0039

Model Diagnostics

As stated in previous lectures, the assessment of violations to the assumptions of a model is an important step in model building. After all, it is quite inappropriate to report the result of a model that has problems with its characteristics.

Some of the techniques that we previously learned when diagnosing linear regression models apply here. The folloiwng, in particular, are important: collinearity measured by variance inflation factors, leverage and Cook's distance. We will not discuss these further; if you are unclear, please review the corresponding lectures.

We will introduce a set of diagnostic tests particular to logistic regression. For this, we will require the `LogisticDx` package to be installed.

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

```
## Loading required package: LogisticDx
```

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

```
library("LogisticDx")
```

The function `gof()` returns various measures of goodness of fit.

```
gof(ATTENTION.LM2, plotROC = FALSE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##       chiSq df    pVal
## PrI  55.986 68 0.85080
## drI  62.434 68 0.66761
## PrG  46.192 54 0.76606
## drG  54.117 54 0.46995
## PrCT 46.192 54 0.76606
## drCT 54.117 54 0.46995
##                        val df     pVal
## HL chiSq         11.917706  8 0.154914
## mHL F             1.742657  9 0.105638
## OsRo Z                  NA NA 0.514903
## SstPgeq0.5 Z      0.158305 NA 0.874216
## SstPl0.5 Z        1.605872 NA 0.108302
## SstBoth chiSq     2.603885  2 0.272003
## SllPgeq0.5 chiSq  0.023324  1 0.878618
## SllPl0.5 chiSq    3.281584  1 0.070061 .
## SllBoth chiSq     6.990034  2 0.030348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For our purposes, we are interested in the reading for `HL chiSq`, which stands for the Hosmer-Lemeshow chi-squared test. This tests the null hypothesis that the model is a poor fit for the data. Here, the p-value is 0.1549, meaning that we are unable to reject the null hypothesis at the 5% level of significance.

Another useful package is `rms`.

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

```
## Loading required package: rms
```

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

```
## Loading required package: Hmisc
```

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

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
## 
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
## 
##     format.pval, units

## Loading required package: SparseM

## 
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
## 
##     backsolve
```

```
library("rms")
```

One the package is loaded, we use the function `lrm()` to produce some output.

```
ATTENTION.LM2.DX <- lrm(CDA.BIN2 ~ AGE + EDU)
print(ATTENTION.LM2.DX)
```

```
## Logistic Regression Model
## 
##  lrm(formula = CDA.BIN2 ~ AGE + EDU)
## 
##                        Model Likelihood     Discrimination    Rank Discrim.    
##                           Ratio Test           Indexes           Indexes       
##  Obs            71    LR chi2      17.96    R2       0.330    C       0.803    
##   Normal        53    d.f.             2    g        1.656    Dxy     0.607    
##   Low           18    Pr(> chi2) 0.0001    gr       5.236    gamma   0.609    
##  max |deriv| 7e-08                          gp       0.237    tau-a   0.233    
##                                             Brier    0.150                     
## 
##           Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept -0.4249 3.7682 -0.11  0.9102  
##  AGE        0.0710 0.0404  1.76  0.0785  
##  EDU       -0.4583 0.1587 -2.89  0.0039  
## 
```

Here, we are looking for the R2 reading. This is called Nagelkirke's R^2 or pseudo-R^2. Since R^2, or the coefficient of determination, has no real meaning for categorical variables, there was a need to create measures similar to R^2. The value may be interpreted in a similar fashion. In the present case, we can state that the model explains about 33% of the variability in CDA classification.

References

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

THE END