BIOSTATISTICS HONOURS
Assignment 1: Logistic Regression
Dobutamine and Stress Echocardiography

Hand Out Date: 19th March 2024
Hand In Date: 9th April 2024

The risk of experiencing a cardiac event is typically tested by "stress echocardiography" where the patient's heart rate is raised through exercise, and then various measurements of the heart are taken. However this test can be problematic when used with older people that often cannot take the stress induced by the hard exercise.

The data for this assignment is from a study into a drug called "dobutamine" which is used as an alternative means of putting the heart under stress. The primary objective was to evaluate if the stress echocardiography test was still effective at predicting a cardiac event when the stress on the heart was induced by dobutamine rather than by exercise. A secondary objective was to explore what other measurements were useful in predicting a cardiac event.

The variables involved were:

- whether or not a cardiac event was experienced the following 12 months (0=no, 1=yes): *event*
- age of patient in years: *age*
- baseline cardiac ejection fraction (a measure of the heart's pumping efficiency): *baseef*
- ejection fraction on dobutamine: *dobef*
- stress echocardiography test was positive (0=yes, 1=no): *posse*
- wall motion anamoly on echocardiogram (0=yes, 1=no): *restwma*

Assignment Brief

1. Data

   (a) Draw a random sample of size 350 using your Assignment number as a seed (your number is listed in the spreadsheet called "AssgnSeeds"). This will be your "training data" that you fit models to. The remaining data should be stored in a "validation" data set for later use.

   The R code to set the seed used for random number generation and to draw the samples:

```
set.seed(YOURSEED#)
sample.index <- sample(1:558, size = 350, replace = F)
data <- Assign2Data[sample.index,]
val.data <- Assign2Data[-sample.index,]
```

(b) Please be careful with identifying the correct subjects based on row numbers that R prints out. The random sampling will result in these row numbers not matching the sequential row positions. Also, you need to be careful about row numbers in any vectors that are extracted from the model object as these will not align with the position in the data.

The following syntax (where XXX is the printed row name, mydata is a dataframe, and myvector is a vector of values) is useful to find the correct position of a row name:

```
which(rownames(mydata)==XXX)
which(names(myvector)==XXX)
```

2. Exploration (no formal hypothesis tests needed i.e. no p-values)

   (a) Construct a table of appropriate descriptive statistics and interpret.

   (b) Explore and interpret 1) the relationships between the outcome variable *event* and the other variables ; 2) the relationship between *dobef* and *baseef*

3. Logistic regression
   The primary objective is to determine how effective the stress echocardiography test is in predicting a cardiac event, and at the same time to build a model that can predict the outcome as well as possible. Treat age as a potential confounder. Summarise your model building procedure in a table, check the model, and interpret your final model.

4. Derive a classification scheme
   Use your final model to devise a classification scheme and then evaluate how the scheme performs by applying it to the validation data (i.e. the 208 "out-of-sample" data points: all the data that you did not use to develop the model and the classification scheme). Be sure to report the sensitivity and specificity from your scheme as well as the likelihood ratio measures.

5. Layout
   Sections to include are:
   Introduction: brief introduction to the problem and the data.
   Statistical methods - do not repeat your notes or text book. Just a brief explanation of what technique is used and why.
   Data Exploration
   - Univariate analyses
   - Bivariate analyses

   Logistic Regression
   - Model building
   - Model diagnostics
   - Model interpretation

   Classification

Conclusions

Please note the following:

- An electronic version of the R script file must be provided in addition to the final report.
- In addition to content, the layout and presentation of your report are important. Do NOT cut-and-paste R output into the report. Rather present the output in neatly constructed tables. All graphs and tables are to be appropriately labelled and must be referred to in the text.
- Your project will NOT be marked if you do not include a plagiarism declaration, or if you do not email me your script file.
- Although there is no formal page limit for this assignment, you should bear in mind that more does not mean better!