Practice 3

In this practice you will go through a case study using logistic regression model for binary response. You can refer to `prac3.R` script file for the R commands to be used.

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima` in `faraway` package.

1. Create a factor version of the `test` results and use this to produce an interleaved histogram to show how the distribution of `insulin` differs between those testing positive and negative. Do you notice anything unbelievable about the plot?

2. Replace the zero values of `insulin` with the missing value code `NA`. Recreate the interleaved histogram plot and comment on the distribution.

3. Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame?

4. Refit the model but now without the `insulin` and `triceps` predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.

5. Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?

6. Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.

7. Using the last fitted model of the previous question, what is the difference in the log-odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Then calculate the associated odds ratio value, and give a 95% confidence interval for this odds ratio.

8. Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the logistic regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

There is always some freedom in deciding which specifications of the method in use, in what order to apply them, and how to interpret the results. So there may not be one clear right answer and good analysts may come up with different models.

It is always a good idea to record your data analysis results and turn them into a technical or research report.