Lab 5 – Spring 2021

01^{st} March 2021

cavicchia@ese.eur.nl

Use Wine dataset.
These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The variables are the following ones: Class, Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline.
Here the aim is to predict the type of the wine. A further variable has been created. It is called *Class Logit*. It assumes value 0 if it belongs to the first two types of wine, 1 otherwise. Consider as independent variables the following ones: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Color intensity and Proline.

1. Descriptive Statistics: Use Task > Summary Statistics to describe the set of variables by y=default. Here, report the mean, median and standard deviation of the variables. Comment the main results focusing on: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Color intensity and Proline. Provide the main summaries statistics. By looking at the mean values in the two groups, could you draw some conclusions? Which variables do you expect to have more discriminative power?

2. Logit model. Use Task > Binary Logistic and use the window to select the dependent variables and the set of explanatory variables; to specify the model; to specify plots.
   Fit the following model: model1: Y=Class Logit; Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Color intensity and Proline, as predictors. Considering the output for the model1

   (a) Which is the value of the Chi-square test? Is it possible accept the null hypothesis that all regression coefficient are null?

   (b) Are there insignificant variable at α-level=0.01?

   (c) Which is the value of AUC (Area under the ROC) ? How is the goodness of fit of the model?

   (d) Use *Class Logit* as response variable and Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Color intensity and Proline, as predictors. Conduct a model selection (using a stepwise procedure and fixing the significance level at 0.05). What are the significant predictors in the final model?.

(e) What is the meaning of the Malic Acid coefficient and its corresponding exponential estimate?

(f) What is the meaning of Color Intensity and its corresponding exponential estimate?

(g) Which is the value of AUC of the final model? Is it better compared to the area computed in 2(c)? If yes, justify your answer.

(h) Which of the explanatory variables have a greater effect on the dependent variable?

(i) Describe the characteristics of the first two types of wine (Class Logit=0) based on the output your final model (answer 2 (d)).