Univariate Data and Modelling – Exercises

Session 8 – Logistic Regression

Exercise 1

Import the dataset called BLOOD.DAT into RStudio. It contains data from a case-control study assessing several plasma risk factors for breast cancer.. The dataset contains following variables:

ID	Identification Code
matchid	Matched ID
case	Case (1) / Control (0)
curpmh	Current PMH use – yes(1) / no(0)
ageblood	Age at blood drawn
estradol	Estradiol concentration
estrone	Estrone concentration (missing value: 999)
testost	Testosterone concentration (missing value: 999)
prolactn	Prolactine concentration (missing value: 99.99)

1. Use logistic regression to assess the association between testosterone and breast cancer risk.
    a. Calculate and interpret the Odds Ratio.
    b. Look at the predicted values and visualize them.
    c. Create the classification table with different cut-off values.
    d. Calculate the "deviance" of the model and interpret.
    e. Create the ROC curve, the AUC and interpret.
2. Repeat exercise 1 but now with age at blood sampling as predictor.
3. Repeat exercise 1 but now with testosterone and age at blood sampling as predictors. Is this multiple logistic regression model better than the simple one?