**Lab3\_Practical exercises**

1. Using the chdagesex.csv dataset (available on Learn):

• Fit a logistic regression model to test the association between age and coronary heart disease (model M1). Compute the odds ratio for age (answer: 1.06) and the 95% confidence interval (answer: 1.04-1.09).

* Test the association of CHD with age after adjusting for sex (model M2): produce odds ratios and confidence intervals for both covariates (answer: age: 1.06, 1.04-1.09, sex: 2.62, 1.34-5.32).
* Perform a likelihood ratio test comparing model M1 to model M2 and confirm that the addition of sex to the model is significant at α = 0.05 by computing a *p*-value (answer: 0.0047).
* Investigate the effect of sex after stratifying for age ≤ 50 and age > 50. Report odds ratios and confidence intervals for the two strata (answer: age ≤ 50: 0.62, 0.12-2.62, age > 50: 3.9, 1.81-8.88).
* Create dataframe agesex containing two columns: AGE with values in the sequence from 1 to 100, and SEX created as follows: set.seed(1)

SEX <- factor(rbinom(100, 1, 0.5), labels=c("F", "M"))  
Predict the probabilities of CHD for the agesex data according to model M2 and plot them using a different colour according to sex.

* Plot the ROC curves for the two models in the same graph (hint: use option add=TRUE for the second curve) and report their AUCs (answer: 0.734, 0.76).
* Write function glm.cv(formula, data, folds) that given a model formula (an expression of the type outcome ~ predictors), a dataframe containing outcome and predictors, and a set of cross-validation folds produced by createFolds(), fits a logistic regression model in each of the folds and returns a list of fitted models. After setting the random seed to 1, generate a set of 10 cross-validation folds and use glm.cv() to cross-validate model M1 and model M2.
* Write function predict.cv(regr.cv, data, outcome, folds) where regr.cv is a list of fitted models produced by glm.cv(), data is a dataframe of covariates, outcome is the vector of observed outcomes and folds is the set of cross-validation folds: the function should use the model fitted on the training set of each fold to predict the outcome of the corresponding test set. The function should return a list of dataframes, each containing observed and predicted outcome for the test observations.
* Use predict.cv() to make predictions for both model M1 and model M2. Using these predictions, compute AUCs for all folds and report the mean cross-validated AUCs. (answer: 0.756, 0.754).

2. Using the hemophilia.csv dataset (available on Learn):

* + Using as.integer(), convert the “group” variable to a 0-1 integer variable in a sensible way so to be used as outcome variable of a classification model. Use logistic regression to model the probability of being a carrier of hemophilia A using the two “AHF” variables as predictors and retrieve the predicted probabilities.
  + Create a scatter plot of AHFactivity (x axis) vs AHFantigen (y axis) using red for cases and green for controls. Add a line corresponding to the decision boundary obtained from the fitted logistic regression coefficients, with intercept and slope defined as follows:

A close up of a logo

Description automatically generated

* + Using a classification threshold θ = 0.5, count the number of misclassified observations (answer: 9). For this threshold, derive sensitivity and specificity (answer: 0.91, 0.83).
  + Write function sens.spec(y.obs, y.pred, threshold) that computes sensitivity and specificity from a 0-1 vector of observed outcomes, a vector of probabilities and a threshold value, and returns the two quantities as a vector.
  + Use function sens.spec() to compute sensitivity and specificity of the model fitted above for at least 10 equally-spaced values of θ spread in the interval (0,1) and plot the points in the ROC space.

3. Consider the following summary statistics from a study of smoking in students:

A screenshot of a cell phone

Description automatically generated

* Compute the odds ratio of smoking in students according to the exposure to smoking in parents directly from the values in the table (answer: 1.58).
* Create a synthetic dataset with the same characteristics as those in the table and fit a logistic regression model to it. Check that the odds ratio for exposure to smoking in parents matches what you computed before and report the 95% confidence interval (answer: 1.34, 1.88), and the Wald test *p*-value (answer: 1.71e-07).
* By using the deviance, test the goodness-of-fit of the model by deriving a *p*-value (answer: 6.8e-08).
* From the regression coefficients compute the probability of smoking for a student whose parents do not smoke (answer: 0.139) and for a student whose parents smoke (answer: 0.203).
* Determine an appropriate value of θ and compute the corresponding sensitivity and specificity of the model (answer: 0.813, 0.267).