Logistic regression 1

We will now examine how logistic regression can be used to investigate the risk of developing acute graft versus host disease after a bone marrow transplant. We will use the data set bmt, which should already be available in your work space.

The purpose of the exercise is to learn to work with the R functions for logistic regression (`glm`, `coef`, `confint`, `anova`, etc.).

When reading in the SPSS file, the file path between the quotes should correspond the location where the file is stored on your computer. R will give you a warning (as is frequently the case when importing SPSS files). In these exercises you can ignore this type of warning. Here (in the shiny app), we have already loaded the bmt data.

Question 1

a) Define a new variable sexmatch with value 1 if the receiver and donor have the same sex and value 0 if not. In R you can do this using:

b. Make a contingency table of the outcome agvhd against sexmatch using the `table` function. Add the proportions of healthy and diseased with the `prop.table` function. Use `chisq.test` to perform a chi-square test.

Is there a significant relation between these variables?

Question 2

Compute the odds ratio of sexmatch = 1 versus sexmatch = 0. Use R as a calculator to calculate the odds ratio from the contingency table.

Question 3

Now we will look at the relation using logistic regression.

In R we can estimate a logistic regression model using the command `glm()`. The first parameter indicates the model formula. On the left-hand side of the ~ sign the dependent variable is specified (here `agvhd`); on the right hand side the independent/explanatory variables (here the single variable sexmatch). The parameter 'family' indicates the distribution we use. The default model for a binomial distribution is the logistic regression model. The last parameter specifies the data set.

Question 4

Use the `summary` command on the returned object to obtain the estimated coefficients, standard errors and p-values from a Wald test. The odds ratios can be obtained by using `exp(coef(glm1))`. The function `confint` is used to obtain confidence intervals, again the `exp` function transforms them to the odds ratio scale. Compare the OR with the result you obtained when you calculated it by hand.

Question 5

Using `drop1(glm1, test='LRT')` we also obtain the p-values from a likelihood ratio (LR) test. Compare the results between the LR test and the Wald test.

We will now look at the effect of a continous variable on the odds of `agvhd`.

Question 6

Specify a model using `agedon` as independent variable. What is the estimated odds ratio of this continuous variable and the confidence interval? What is the meaning of the odds ratio? Call the estimated model `glmAge`.

Question 7

We are now going to plot the relation between the age of the donor and the estimated probabilities. We can use the `predict` function for this. The first argument is the estimated `glm`. When we use `type = 'response'` we obtain predictions on the scale of the response (the 0-1 scale of probabilities, this in contrast to the scale of the linear predictor that is obtained by specifying `type='link'`).

Question 8

In the model above we assume a linear relation between the log odds for agedon and the outcome. To test this assumption we can add the square of agedon to the model. Add the quadratic term to the model below (hint: use I(agedon^2)). Call the model glmAge2. Is the quadratic term statistically significant?

Question 9

Estimate a model using `agedon`, `agedon` squared, `agerec`, `sexmatch` and `diag`. What are the odds ratios?

Question 10

Let's visualize the risk of acute graft versus host disease for a 30 year old male with acute nonlymphoblastic leukemia with a male and a female donor.