Computer exercise 3

Association analysis using logistic regression

The aim of this exercise is to learn how to perform logistic regression models in R, and to apply such models for association analysis of genetic markers. Here, we mainly focus on the interpretation of parameters and models.

The main R-functions you'll need

Enter and save data

To enter data into R from a file named dataset.txt, and save it in an object called ds, type `ds<-read.table("dataset.txt",header=TRUE)`. The argument `header=TRUE` is necessary if the first row of the file contains names of variables. You can get the size of the dataset ds with `dim(ds)`, and `names(ds)` will show all variables that ds contains. To look at a part of it, type e.g. `ds[1:10,1:3]` and you will see the first 10 observations (=rows) of the first 3 variables (=columns).

Logistic regression in R

As an example, say that the data set ds contains a binary response variable Y, and 3 explanatory variables X1, X2, and X3. To check how many 1:s (=cases) Y contains, you can use for instance `sum(ds$Y)`. To investigate the relationship between Y and X1, using the model $logit(\pi) = \alpha + \beta X1$, we fit the model with the glm()-function: `mod1<-glm(Y ~ X1,family=binomial, data=ds)`

Here the model fit was saved in a glm-object called `mod1` (the name is your choice). A glm object has many components. The most essential information is printed by calling `summary(mod1)`.

To fit a model that includes several covariates, such as $logit(\pi) = \alpha + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$, type `mod3<-glm(Y ~ X1+X2+X3,family=binomial, data=ds)`.

The function `anova()` can be used to compare two or more nested models. For instance `anova(mod1,mod3,test="Chisq")` will perform a LR-test of the hypotheses H0: $\beta_2 = 0$ and $\beta_3 = 0$, against H1: $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$, in the model $logit(\pi) = \alpha + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$.

Assignment

The dataset you will analyse is available at the course home page in the file
http://www.math.chalmers.se/Stat/Grundutb/CTH/tms121/1011/diabetes.txt.
It consists of a number of type 1 diabetes patients, and a number of controls (all are unrelated). The variable Y is the case - control classification, M1, ... , M5 are allele-counts for 5 different SNPs (counting the number of '1'-alleles for each marker). The genotypes from which M1, .. ,M5 have been calculated are also included.

The DQDR-variables contains the genotypes of a gene known to have a strong association with type 1 diabetes. (It is actually the haplotypes across 3 genes situated close to each other in the HLA-complex on chromosome 6.) It is multi-allelic, and the genotypes have been recoded into a number of dummy-variables D1, D6, D7, etc., where Di=1 if the subject has at least one 'i'-allele, and =0 otherwise. D99 is a grouping of all rare alleles.

1. Import the dataset into R and check out what it contains. Make sure you understand what the different variables represent. Check how many cases and controls there are.

   Fit a 'baseline'-model, that is a model with only an intercept-parameter $logit(\pi) = \alpha$. Use `mod.b<-glm(Y ~ 1,family=binomial, data=ds)`.

2. Modeling the 5 SNPs:

   (a) Fit a logistic model $logit(\pi) = \alpha + \beta_1 \text{Mi}$, for each of the 5 SNPs.
   Does any of them show significant association with the disease?

   (b) Include all 5 SNPs as covariates in the same model. What happens with the SNP-associations? Any changes from the estimated associations in the separate models?
   Also, look at the estimated standard errors. When some standard errors blow up in this way, it is likely a sign of strong dependency (colinearities) between some of the covariates in the model. The corresponding p-values can then not be trusted.

3. Fit a model for the DQDR-association. Use D1 as reference category (which means you leave that one out in the model statement). Look at the overall association for the DQDR-gene, using the anova()-function and comparing with the baseline-model.

4. Include the best SNP (the strongest associated one) into the DQDR-model. Is it still significantly associated? Does it change the overall significance of the DQDR-model? Interpretation?

Hand in

Write a short summary of the results and conclusions from the analysis you have performed. Please hand in no later than Friday October 8.