# Regression Analysis

In this exercise, a number of logistic and linear regression analyses will be carried out to test for the association of a number of SNPs with an affection status and with a quantitative trait, respectively. This includes the use of different tests, the calculation of odds ratios (OR), and the consideration of different genetic models. Further objectives are the adjustment for the effects of co­variates and the testing of a SNP for association given the effect of another SNP. Finally, we will consider gene-gene and gene-environmental interaction as well as model selection.

The data set is in R format and has been stored in the file **dbp.R**.

Since the syntax for many of the commands is highly repetitive and in order to save time, please use the copy & paste functionality of your text editor and subsequently make the neces­sa­ry changes to the copied text.

Please also answer the questions at the end of the exercise.

### Data set import

Start R and change the working directory as requested. Load the data set for the exercise and get an overview which objects have been loaded into the R working memory:

load(“dbp.Rdata”)

ls()

dbp[1:5,]

### I. Logistic regression on a single SNP genotype

Logistic regression models are implemented through the glm function in R. This function re­quires a model formulation. This includes a specification of what is regressed on what (e.g. affection ~ rs1112), the error family, the link function, and the data set to be used.

1. Run a logistic regression analysis of the affection status regressed on the genotype of marker rs1112, using the data in the data frame dbp. Assign the results from the regression analysis to the new object result.snp12:
2. Print the results of the regression analysis with the following command:

The marker variable rs1112 is of data type factor (nominal). Thus, we have considered a gene­ral *genotypic* model. R has therefore created two dummy variables, named rs1112**3** and rs1112**4**, which separately describe the effects of the genotypes coded as **3** (hetero­zygous 1/2) and **4** (homozygous 2/2), respectively. The effects of these two genotypes are com­pared to the baseline genotype 2 (homozygous 1/1).

1. To carry out a likelihood-ratio test (LRT), first calculate the χ2 statistic and subsequently obtain the corresponding *P*-value. Note that we have a χ2 distribution with *two* degrees of freedom, since we test two dummy variables simultaneously against the null model:
2. Extract the regression coefficients and calculate the odds ratios for the genotypes (reminder from the lecture: OR=e) as well as their confidence intervals:
3. So far, the marker data are of type factor (nominal) and we have considered a general genotypic model. For an allelic (multiplicative) model, the data type has to be changed to numeric. This way, the genotype is recoded from nominal 2/3/4 (for 11/12/22) to numeric 0/1/2 (for the number of copies of the “2” allele with each sample):
4. Run the logistic regression analysis again, this time assuming an allelic model.

### II. Adjustment for the effects of covariates and of other SNPs

Analyses can be confounded by external factors. If such factors are known and measured, regression analysis allows for adjusting for their effect by simply incorporating them into the statistical model.

1. First, create an excerpt from the full data set with "affection", "trait","sex", "age", "rs1112", "rs1117". For all subsequent analyses, we will consider an allelic (multiplicative) model for the markers
2. Does sex have an effect on the affection status and is the effect of the SNP independent of such a potential influence? To answer this question, re-run the regression analysis for SNP rs1112, this time with an adjustment for sex:
3. Age is also often suspected to influence the trait of interest. Therefore, re-run the analysis with an adjusting for sample age:
4. Finally, adjust for both covariates, sex and age, simultaneously in the regression analysis:

Adjustment for the effects of other SNPs: For many diseases and phenotypes, there are already established genetic factors. In many genetic epidemiological studies, one would therefore like to assess if some newly found association is independent of such established ones. This is equivalent to adjusting for the effect of the already established SNP.

1. Run a logistic regression analysis for each of the two SNPs rs1112 and rs1117, while adjusting for the effect of the other:
2. Note that the *P*-values from a Wald test do not differ for the different orders of markers, but that the *P*-values from a likelihood-ratio test (obtained from the anova function) do! Why?

### III. Analysis of quantitative instead of dichotomized trait

Dichotomization of quantitative trait values can result in a power loss, because information is discarded. In our example data set, the original trait value (diastolic blood pressure) had been dichotomized to case-control status: All individuals with a value greater than a certain threshold were defined as having high blood pressure (“cases”), whereas the others were considered to be controls with normal blood pressure.

The column trait in the data frame dbp contains the original quantitative trait values. Run two linear regression analyses, one without and one with adjust for the effect of sex.

### IV. Gene-environment (GxE) and gene-gene (GxG) interaction

Interaction between factors (genetic and non-genetic) can also be tested. The model then additionally includes the product term of the two factors. In R, this is achieved by using the \* operator in the model formulation, for example affection ~ sex \* snp, which is equivalent to affection ~ sex + snp + sex:snp. The variables sex and snp denote the main effect terms, while sex:snp denotes the interaction term.

It is important to note, however, that statistical interaction does not necessarily imply biological interaction, such as epistasis or synergy. *Statistical interaction only denotes the deviation from linearity within the regression model!*

Gene-environment (GxE) interaction

Test SNP rs1112 for significant interaction with each of the two covariates sex and age:

Gene-gene (GxG) interaction

Now test markers rs1112 and rs1117 for significant statistical interaction:

### Quitting

Quit the R session by calling the quit function:

q()