Exercise

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.R")
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.

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`:

```
result.snp12 = glm (affection ~ rs1112, family=binomial("logit"), data=dbp)
```

Print the results of the regression analysis with the following command:

```
print (result.snp12)
```

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).

The results object is part of some special R classes, namely `lm` and `glm` (same names as the functions). Membership in these classes causes R to use dedicated, specialized functions for printing, analyzing and other tasks with such objects:

```
print ( class  (result.snp12) )
print ( summary(result.snp12) )
```

The coefficients table lists the estimated values for the regression coefficients β as well as their standard errors. It further contains the *P*-values as obtained from a Wald test.

To carry out a likelihood-ratio test (LRT), first calculate the χ^2 statistic and subsequently obtain the corresponding *P*-value. Note that we have a χ^2 distribution with *two* degrees of freedom, since we test two dummy variables simultaneously against the null model:

```
dev.geno = anova (result.snp12, test="Chi")
lrt.pvalue = pchisq(dev.geno[dim(dev.geno)[1],"Deviance"],
             df=2, ncp=0, FALSE)
print ( lrt.pvalue )
```

We can also access parts of the results object with indexes. For example, we can extract the regression coefficients and calculate the odds ratios for the genotypes (reminder from the lecture: $OR=e^{\beta}$) as well as their confidence intervals:

```
print ( summary(result.snp12)$coefficients )
snp.beta = summary(result.snp12)$coefficients[2:3,1]
print ( snp.beta )
print ( exp(snp.beta) )

ci = confint (result.snp12)
print (ci)
print ( exp(ci) )
```

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):

```
snp.data = dbp[,c("affection", "rs1112")]
summary(snp.data)

snp.data[,"rs1112"] <- as.numeric(snp.data[,"rs1112"]) - 1
summary(snp.data)
```

Run the logistic regression analysis again, this time assuming an allelic model:

```
result.all = glm (affection ~ rs1112, family=binomial("logit"),
                  data=snp.data)
dev.all    = anova (result.all, test="Chi")
summary(result.all)
print(dev.all)
```

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.

First, create an excerpt from the full data set. For all subsequent analyses, we will consider an allelic (multiplicative) model for the markers:

```
snp.data = dbp[,c("affection", "trait","sex", "age", "rs1112", "rs1117")]
summary(snp.data)

snp.data[,"rs1112"] <- as.numeric(snp.data[,"rs1112"]) - 1
snp.data[,"rs1117"] <- as.numeric(snp.data[,"rs1117"]) - 1
```

Adjustment for the effects of covariates

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:

```
result.adj = glm (affection ~ sex + rs1112      , family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
```

Age is also often suspected to influence the trait of interest. Therefore, re-run the analysis with an adjusting for sample age:

```
result.adj = glm (affection ~ age + rs1112      , family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
```

Finally, adjust for both covariates, sex and age, simultaneously in the regression analysis:

```
result.adj = glm (affection ~ sex + age + rs1112, family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
```

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.

Run a logistic regression analysis for each of the two SNPs `rs1112` and `rs1117`, while adjusting for the effect of the other:

```
result.adj = glm (affection ~ rs1117 + rs1112, family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
anova (result.adj, test="Chi")

result.adj = glm (affection ~ rs1112 + rs1117, family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
anova (result.adj, test="Chi")
```

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! +

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:

```
result.adj = lm (trait ~ rs1112      , data=snp.data)
summary(result.adj)

result.adj = lm (trait ~ sex + rs1112, data=snp.data)
summary(result.adj)
```

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:

```
result.inter = glm (affection ~ sex * rs1112, family=binomial("logit"),
                    data=snp.data)
summary(result.inter)

result.inter = glm (affection ~ age * rs1112, family=binomial("logit"),
                    data=snp.data)
summary(result.inter)
```

Gene-gene (GxG) interaction

Now test markers `rs1112` and `rs1117` for significant statistical interaction:

```
result.inter = glm (affection ~ rs1112 * rs1117, family=binomial("logit"),
                    data=snp.data)
summary(result.inter)
```

Quitting

Quit the R session by calling the quit function:

```
q()
```

Questions

1. Please enter the *P*-values for marker rs1112 from the analyses in the table below.

	Type of Analysis	*P*-value
I.	Single marker, case-control, genotypic model	Wald test: het 1/2: hom 2/2: LRT:
	Single marker, case-control, allelic model	Wald: LRT:
II.	Single marker, case-control, adjustment for age	
	Single marker, case-control, adjustment for sex	
	Single marker, case-control, adjustment for sex & age	
	Single marker, case-control, adjustment for marker rs1117	Wald: LRT:
III.	Single marker, quantitative trait, adjustment for sex	
IV.	Interaction *P*-value with covariate sex	
	Interaction *P*-value with marker rs1117	

2. Please give the odds ratio (OR) and its 95% confidence interval for marker rs1112 in the unadjusted, genotype-based case-control analysis.

$OR_{het(1/2)}$ = ... 95% CI = ... – .....

$OR_{hom(2/2)}$ = ... 95% CI = ... – .....

3. In the combined analysis of rs1112 and rs1117 (section II., no interaction), the LRT-based *P*-values strongly depended on the order of the markers in the regression model (`affection ~ rs1117+rs1112`: p_{rs1117}=5.547e-07 / p_{rs1112}=1.193e-03; `affection ~ rs1112+rs1117`: p_{rs1112}=5.438e-09 / p_{rs1117}=0.21). Do you have an explanation?

Answers

Regression Analysis

I. Logistic regression on a single SNP genotype

```
# --- Regression + Wald --- #
result.snp12 = glm (affection ~ rs1112, family=binomial("logit"), data=dbp)
print (result.snp12)
Call:  glm(formula = affection ~ rs1112, family = binomial("logit"),
data = dbp)

Coefficients:
(Intercept)      rs11123      rs11124
    -0.4449       0.7582       1.5435

Degrees of Freedom: 599 Total (i.e. Null);  597 Residual
Null Deviance:                      831.8
Residual Deviance: 797.7        AIC: 803.7


print ( class (result.snp12) )
[1] "glm" "lm"


print ( summary(result.snp12) )
Call:
glm(formula = affection ~ rs1112, family = binomial("logit"),
    data = dbp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651  -0.9952  -0.1183   1.0476   1.3712

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4449     0.1189  -3.741 0.000183 ***
rs11123       0.7582     0.1746   4.343 1.40e-05 ***
rs11124       1.5435     0.3416   4.518 6.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 797.75  on 597  degrees of freedom
AIC: 803.75

Number of Fisher Scoring iterations: 4
```

The SNP genotype of marker rs1112 has been stored as `factor` type in R. Since there is no distance defined between categories, R has automatically defined two dummy variables, namely rs1112**3** and rs1112**4**, to code for the presence or absence of the ‘3’ and of the ‘4’ genotypes, respectively. The dummy variable assumes the value 1 if the respective genotype is present in the

particular individual and 0 otherwise. If both '3' and '4' are absent, then the baseline genotype '2' is assumed to be present. Using a Walt test, each of the dummy variables has been tested separately for significant association with the phenotype (affection status). Both variables are significantly associated with the phenotype.

```
# --- Likelihood-ratio test --- #
dev.geno = anova (result.snp12, test="Chi")
lrt.pvalue = pchisq(dev.geno[dim(dev.geno)[1],"Deviance"],
            df=2, ncp=0, FALSE)
print ( lrt.pvalue )
[1] 4.077856e-08
```

Often, we are not interested in the effect of a particular genotype, e.g. '3' or '4', but in the overall significance of a marker. To test this, we have to compare the null model (without both dummy variables) against the alternative model (with both dummy variables) using a likelihood-ratio test. Because the two models differ in two parameters, we have to compare the deviance against a χ^2 distribution with two parameters.

```
# --- OR + CI --- #
print ( summary(result.snp12)$coefficients )
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -0.4449068  0.1189351 -3.740754 1.834691e-04
rs11123      0.7582015  0.1745740  4.343154 1.404519e-05
rs11124      1.5435191  0.3416277  4.518132 6.238747e-06

# Coefficients (betas) for the both dummy variables #
snp.beta = summary(result.snp12)$coefficients[2:3,1]
print ( snp.beta )
  rs11123   rs11124
0.7582015 1.5435191

# Odds ratios (OR) for both dummy variables [OR=exp(beta)] #
print ( exp(snp.beta) )
rs11123  rs11124
2.134434 4.681034

# 95% confidence interval for betas #
ci = confint (result.snp12)
print (ci)
                2.5 %     97.5 %
(Intercept) -0.6802726 -0.2135169
rs11123      0.4176220  1.1023701
rs11124      0.8984800  2.2475097

# 95% confidence intervals for OR #
print ( exp(ci) )
               2.5 %    97.5 %
(Intercept) 0.5064789 0.8077385
rs11123     1.5183466 3.0112947
rs11124     2.4558674 9.4641382

# --- Allelic model --- #
snp.data = dbp[,c("affection", "rs1112")]
summary(snp.data)
affection rs1112
 0:300     2:297
```

```
 1:300     3:251
           4: 52

snp.data[,"rs1112"] <- as.numeric(snp.data[,"rs1112"]) - 1
summary(snp.data)
affection     rs1112
 0:300     Min.   :0.0000
 1:300     1st Qu.:0.0000
           Median :1.0000
           Mean   :0.5917
           3rd Qu.:1.0000
           Max.   :2.0000
```

Because we have coded the marker genotype as `numeric`, another summary than for `factor` data is provided.

```
# --- Allelic model for allele 2 of marker rs1112 --- #
result.all = glm (affection ~ rs1112, family=binomial("logit"),
                  data=snp.data)
dev.all    = anova (result.all, test="Chi")
summary(result.all)
Call:
glm(formula = affection ~ rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6582  -0.9944  -0.1154   1.0456   1.3722

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4470     0.1142  -3.913 9.10e-05 ***
rs1112        0.7652     0.1356   5.642 1.68e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 797.75  on 598  degrees of freedom
AIC: 801.75

Number of Fisher Scoring iterations: 4

print(dev.all)
Analysis of Deviance Table

Model: binomial, link: logit
Response: affection
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      599     831.78
rs1112   1    34.03       598     797.75 5.438e-09
```

Due to the numeric coding of the genotype, there is now an interpretable distance between 0, 1 and 2 copies of the second allele. This allele count can readily enter the regression model, without any need for creating dummy variables. With large-enough sample sizes, the P-values from the Wald test and from the likelihood-ratio test are very similar.

II. Adjustment for the effects of covariates

```
# --- Data conversion for all subsequent analyses --- #
snp.data = dbp[,c("affection", "trait","sex", "age", "rs1112", "rs1117")]
summary(snp.data)
affection     trait        sex          age        rs1112  rs1117
 0:300     Min.   : 60.50   1:329   Min.   :18.00   2:297   2:396
 1:300     1st Qu.: 77.44   2:271   1st Qu.:38.00   3:251   3:190
           Median : 82.00           Median :55.00   4: 52   4: 14
           Mean   : 81.85           Mean   :55.49
           3rd Qu.: 86.09           3rd Qu.:74.00
           Max.   :101.49           Max.   :90.00

snp.data[,"rs1112"] <- as.numeric(snp.data[,"rs1112"]) - 1
snp.data[,"rs1117"] <- as.numeric(snp.data[,"rs1117"]) - 1
```

For the subsequent analysis, we convert the marker genotypes to numeric coding, i.e. we will consider an allele-based model.

Adjustment for the effects of covariates

```
# Adjustment for sex #
result.adj = glm (affection ~ sex + rs1112      , family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
Call:
glm(formula = affection ~ sex + rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.82645  -1.12415  -0.09007   1.21323   1.57462

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.08386    0.13730  -0.611    0.541
sex2        -0.81412    0.17253  -4.719 2.37e-06 ***
rs1112       0.77139    0.13840   5.574 2.49e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 774.98  on 597  degrees of freedom
AIC: 780.98
```

```
Number of Fisher Scoring iterations: 4
```

Compared to males, females have a significantly decreased risk ($p=2.4\times10^{-6}$; $OR=e^{-0.814}=0.44$) of becoming affected. The sex-adjusted *P*-value for marker rs1112 is highly significant ($p=2.5\times10^{-8}$), which causes an increase in risk of $e^{0.771}=2.16$ for heterozygous carriers and of $2.16^2=4.67$ for homozygous carriers of the risk allele (allele-based, i.e. multiplicative risk model!).

```
# Adjustment for age #
result.adj = glm (affection ~ age + rs1112      , family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
Call:
glm(formula = affection ~ age + rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6776  -1.0066  -0.1132   1.0550   1.3937

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.520422   0.250956  -2.074   0.0381 *
age          0.001322   0.004020   0.329   0.7423
rs1112       0.765189   0.135624   5.642 1.68e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 797.64  on 597  degrees of freedom
AIC: 803.64

Number of Fisher Scoring iterations: 4


# Adjustment for sex and age #
result.adj = glm (affection ~ sex + age + rs1112, family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
Call:
glm(formula = affection ~ sex + age + rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.84985  -1.12493  -0.08714   1.19367   1.60989

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.198133   0.263732  -0.751    0.452
sex2        -0.817603   0.172736  -4.733 2.21e-06 ***
age          0.002084   0.004105   0.508    0.612
rs1112       0.771546   0.138411   5.574 2.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 774.72  on 596  degrees of freedom
AIC: 782.72

Number of Fisher Scoring iterations: 4
```

Age has no significant impact on the phenotype, while sex does. Marker rs1112 is highly significantly associated with affection status ($p=2.5\times10^{-8}$) after adjusting for the effects of the covariates sex and age.

Adjustment for the effects of other SNPs

```
# Association analysis of rs1112, adjusted for the effects of rs1117 #
result.adj = glm (affection ~ rs1117 + rs1112, family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
Call:
glm(formula = affection ~ rs1117 + rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7636  -0.9923  -0.1518   1.1154   1.3745

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4523     0.1144  -3.955 7.66e-05 ***
rs1117        0.2853     0.2297   1.242  0.21431
rs1112        0.5999     0.1883   3.186  0.00144 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 796.21  on 597  degrees of freedom
AIC: 802.21

Number of Fisher Scoring iterations: 4

anova (result.adj, test="Chi")
Analysis of Deviance Table

Model: binomial, link: logit
Response: affection
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      599     831.78
rs1117   1    25.06       598     806.71 5.547e-07
rs1112   1    10.50       597     796.21 1.193e-03
```

The Wald test compares the model with the predictor (here: rs1117) against the model without it; marker rs1112 is included in both models. The ANOVA sequence procedure first applied a likelihood-ratio test for the null model (no predictors) against the model that only includes rs1117 ($p=5.5\times10^{-7}$) and subsequently compares the latter model against the one that includes rs1117 and rs1112 ($p=1.2\times10^{-3}$). Thus, it first assesses the contribution of rs1117 and subsequently the *additional* contribution of rs1112.

```
# Association analysis of rs1117, adjusted for the effects of rs1112 #
result.adj = glm (affection ~ rs1112 + rs1117, family=binomial("logit"),
                  data=snp.data)
summary(result.adj)
Call:
glm(formula = affection ~ rs1112 + rs1117, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7636  -0.9923  -0.1518   1.1154   1.3745

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4523     0.1144  -3.955 7.66e-05 ***
rs1112        0.5999     0.1883   3.186  0.00144 **
rs1117        0.2853     0.2297   1.242  0.21431
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 796.21  on 597  degrees of freedom
AIC: 802.21

Number of Fisher Scoring iterations: 4

anova (result.adj, test="Chi")
Analysis of Deviance Table

Model: binomial, link: logit
Response: affection
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      599     831.78
rs1112   1    34.03       598     797.75 5.438e-09
rs1117   1     1.54       597     796.21      0.21
```

The Wald test again compares the model with the predictor (here: rs1117) against the model without it; marker rs1112 is included in both models. Results are therefore the identical for the models `affection ~ rs1117 + rs1112` (see above) and `affection ~ rs1112 + rs1117`. However, the ANOVA sequence procedure (using a likelihood-ratio test) now first compares model `affection ~ constant` (null model) against `affection ~ rs1112` and only includes rs1117: `affection ~ rs1112 + rs1117`. Marker rs1112 makes a

significant contribution ($p=5.4\times10^{-9}$). On top of this, marker rs1117 does not contain any new information and does not provide a significant additional contribution (p=0.2).

The explanation for this observation is that both markers, rs1112 and rs1117, are in linkage disequilibrium with each other and also with the causal genetic variant, but that marker rs1112 shows the higher allelic correlation with that variant. If marker rs1117 is first included in the regression model, then marker rs1112 can still provide some additional association information. However, if marker rs1112 is first included, then marker rs1117 has nothing to offer and will be insignificant.

III. Analysis of quantitative instead of dichotomized trait

```
# --- Single-marker analysis with *linear* model --- #
result.adj = lm (trait ~ rs1112      , data=snp.data)
summary(result.adj)
Call:
lm(formula = trait ~ rs1112, data = snp.data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5556  -3.9106   0.2194   4.0144  15.4809

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.1021     0.3301 242.680  < 2e-16 ***
rs1112        2.9535     0.3774   7.826 2.29e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.954 on 598 degrees of freedom
Multiple R-squared: 0.09291,Adjusted R-squared: 0.09139
F-statistic: 61.25 on 1 and 598 DF,  p-value: 2.292e-14
```

The genotype of marker rs1112 has been coded as `numeric`; an allele-based test has therefore been performed. In a linear regression model, this corresponds to an *additive* risk model. Marker rs1112 is highly associated with the quantitative trait (Wald test: $p=2.3\times10^{-14}$). The linear regression model is able to explain about 9% of the observed variance in the quantitative trait ($R^2=0.093$).

```
# --- Additional adjustment for sex --- #
result.adj = lm (trait ~ sex + rs1112, data=snp.data)
summary(result.adj)
Call:
lm(formula = trait ~ sex + rs1112, data = snp.data)

Residuals:
     Min       1Q   Median       3Q      Max
-20.9404  -3.6272   0.2234   3.7815  16.3480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.4542     0.3904 208.654  < 2e-16 ***
```

```
sex2         -2.8823     0.4748  -6.071 2.27e-09 ***
rs1112        2.8685     0.3668   7.820 2.41e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.784 on 597 degrees of freedom
Multiple R-squared: 0.1456,Adjusted R-squared: 0.1428
F-statistic: 50.89 on 2 and 597 DF,  p-value: < 2.2e-16
```

Marker rs1112 is highly associated with the quantitative trait after adjusting for the effect of sex (Wald test: $p=2.4\times10^{-14}$; likelihood-ratio test: $p<2.2\times10^{-16}$). The linear regression model is now able to explain more than 14% of the observed trait variance, i.e. inclusion of sex provides a better model fit to the sample data.

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

Gene-environment (GxE) interaction

```
# --- Interaction between sex and marker --- #
result.inter = glm (affection ~ sex * rs1112, family=binomial("logit"),
                    data=snp.data)
summary(result.inter)
Call:
glm(formula = affection ~ sex * rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8388  -1.1205  -0.0965   1.2176   1.5685

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09415    0.15371  -0.613 0.540174
sex2        -0.79026    0.23515  -3.361 0.000777 ***
rs1112       0.79049    0.18896   4.183 2.87e-05 ***
sex2:rs1112 -0.04141    0.27771  -0.149 0.881472
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 774.96  on 596  degrees of freedom
AIC: 782.96

Number of Fisher Scoring iterations: 4
```

While marker rs1112 is associated with affection status after adjusting for the effect of sex (Wald test: $p=2.9\times10^{-5}$), there is no evidence that the marker risk allele causes different risk *increases* in females compared to males (p=0.9).

```
# --- Interaction between age and marker --- #
result.inter = glm (affection ~ age * rs1112, family=binomial("logit"),
                    data=snp.data)
summary(result.inter)
```

```
Call:
glm(formula = affection ~ age * rs1112, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8044  -1.0479  -0.1256   1.0606   1.4655

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.764365   0.328207  -2.329  0.01986 *
age          0.005719   0.005508   1.038  0.29909
rs1112       1.193715   0.393377   3.035  0.00241 **
age:rs1112  -0.007716   0.006585  -1.172  0.24130
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 796.26  on 596  degrees of freedom
AIC: 804.26

Number of Fisher Scoring iterations: 4
```

While marker rs1112 is associated with affection status after adjusting for the effect of age (Wald test: $p=2.9\times10^{-5}$), there is no evidence for a significant association of age with affection status ($p=0.3$) nor an interaction between age and the marker ($p=0.2$).

Gene-gene (GxG) interaction

```
# --- Interaction between markers rs1112 and rs1117 --- #
result.inter = glm (affection ~ rs1112 * rs1117, family=binomial("logit"),
                    data=snp.data)
summary(result.inter)
Call:
glm(formula = affection ~ rs1112 * rs1117, family = binomial("logit"),
    data = snp.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7167  -0.9899  -0.1342   1.1126   1.3773

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.45855    0.11749  -3.903  9.5e-05 ***
rs1112         0.61285    0.19612   3.125  0.00178 **
rs1117         0.37232    0.43522   0.855  0.39228
rs1112:rs1117 -0.07464    0.31590  -0.236  0.81323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 831.78  on 599  degrees of freedom
Residual deviance: 796.16  on 596  degrees of freedom
AIC: 804.16
```

```
Number of Fisher Scoring iterations: 4
```

While marker rs1112 is associated with affection status after adjusting for the effect of marker rs1117 (Wald test: p=0.002), there is no evidence for a significant association of marker rs1117 with affection status (p=0.4) nor an interaction between both markers (p=0.8), i.e. that the genotype of rs1117 may have a modifying effect on the risk increase caused by marker rs1112.

The model selection procedure started with the full model, i.e. it contained all predictor terms that were to be considered. If interaction terms should also be considered, these would have to be included in the initial model estimation request (i.e. the `glm` call). Covariates age and rs1117 were subsequently discarded from the model since there do not significantly improve the model fit (i.e. decrease the error) or, with backward selection, did not significantly worsened the model fit (i.e. dropping did not lead to largely increased error). Model fit is measured by the AIC criterion that, in addition to the deviance, penalized larger numbers of predictors in the regression model. The final model includes sex and rs1112 as predictors.

Questions

1. Please enter the *P*-values for marker rs1112 from the analyses in the table below.

	Type of Analysis	***P*-value**
I.	Single marker, case-control, genotypic model	Wald test: het 1/2: **1.40e-05** hom 2/2: **6.24e-06** LRT: **4.077856e-08**
	Single marker, case-control, allelic model	Wald: **1.68e-08** LRT: **5.438e-09**
II.	Single marker, case-control, adjustment for age	**1.68e-08**
	Single marker, case-control, adjustment for sex	**2.49e-08**
	Single marker, case-control, adjustment for sex & age	**2.48e-08**
	Single marker, case-control, adjustment for marker rs1117	Wald: **0.00144** LRT: **1.193e-03**
III.	Single marker, quantitative trait, adjustment for sex	**2.29e-14**
IV.	Interaction *P*-value with covariate sex	**0.881472**
	Interaction *P*-value with marker rs1117	**0.81323**

2. Please give the odds ratio (OR) and its 95% confidence interval for marker rs1112 in the unadjusted, genotype-based case-control analysis.

$OR_{het(1/2)}$ = **2.134434** 95% CI = **1.5183466 – 3.0112947**

$OR_{hom(2/2)}$ = **4.681034** 95% CI = **2.4558674 – 9.4641382**

3. In the combined analysis of rs1112 and rs1117 (section II., no interaction), the LRT-based *P*-values strongly depended on the order of the markers in the regression model (`affection ~ rs1117+rs1112`: p_{rs1117}=5.547e-07 / p_{rs1112}=1.193e-03; `affection ~ rs1112+rs1117`: p_{rs1112}=5.438e-09 / p_{rs1117}=0.21). Do you have an explanation?

Both markers are correlated, i.e. they show allelic association (LD), and represent the *same* phenotypic association signal at the locus (remember that association analysis usually pursues an indirect approach). One of the SNPs, namely rs1112, is more strongly correlated with the causative variant than rs1117. The likelihood-ratio test (LRT) compares the following models: `affection~SNP1` vs. `affection~SNP1+SNP2`. If the marker rs1117 is included in the model first as `SNP1`, it already contains *some but not all* information on the phenotypic association at the locus. Inclusion of rs1112 as `SNP2` still contributes significant additional information. However, if marker rs1112 is included first as `SNP1`, it already contains *all* information available from the data set on the association of the locus. In this case, inclusion of marker rs1117 as `SNP2` cannot contribute any further information and is, thus, tested as being insignificant.