Lab 1: A review of linear models

The purpose of this lab is to help you review basic statistical methods in linear models and understanding the implementation of these methods in R.

In general, we need to perform the following steps in most data analyses.

1. Understand the scientific background and problems of interests;
2. Understand how the data were collected and what data are available.
3. Put the scientific questions into statistical terms;
4. Perform some initial data analysis;
5. Refine your model and run model diagnostics;
6. Statistical inference or analysis based on the model you chose;
7. Interpret the results to answer the problems of interests.

We will use the breast cancer data sets collected by Richardson et al. (2006) as an example to show the above steps for building a statistical model.

Most phenotypic traits (size, weight, shape, lifespan, fecundity) in plants and animals are affected both by genetic and environmental factors. An important task for plant and animal breeders is to know how much of the phenotypic variability of the trait is due to genetic variance, and how much is due to environmental factors. Moreover, plant and animal breeders would like to find those loci of the genes which can be used to explain the genetic variation. In the paper of Richardson et al. (2006), they found that the basal-like human breast cancer is likely related to the X chromosomal abnormalities. In addition, Richardson et al. (2006) found that the basal-like breast cancer is also related to the overexpression of a small subset of X Chromosomal genes.

In this lab, we will consider the microarray gene expression data as quantitative traits and find the genetic factors that are associated with the gene overexpression. The genetic variations of human genome are mostly due to DNA polymorphisms, which include copy number variation (insertion or deletion of section of DNA) and single nucleotide polymorphism (SNP). SNP occurs when a single nucleotide differs between members of species. The frequency of SNPs is greater than any other types of polymorphism. Therefore, the objective of our analysis is to find the SNPs that lead to the overexpression of genes in X chromosome.

The tissue samples of the experiments conducted by Richardson et al. (2006) were obtained from the Harvard breast SPORE blood and tissue repository. Thus, this study is an observational study not an experimental design. No randomization was performed in the study. Because this is an observation study, we are not able to conclude any causal effects but instead we can obtain some association results. Microarray gene expression data sets and SNP data sets are both available as supplemental data sets in the paper of Richardson et al. (2006). The gene expression data sets were obtained using the Affymetrix Human Genome U133 Plus 2.0 Arrays and the SNPs data were obtained using Affymetrix GeneChip 10k arrays. Both data sets can be downloaded from the NCBI GEO database.

According to the microarray data analyses results in the paper of Richardson et al. (2006), the gene "CSF2RA" in X chromosome was overexpressed in BLC and BRCA1-associated tumors relative to non-BLC and normal breast samples. Therefore, in this lab, we will focus on the expression of this gene. We would like to find out SNPs that are associated with the overexpression of this gene. However, there are many SNPs in the data set. For simplification, in this lab, let us focus on SNPs whose physical position is close to the gene "CSF2RA". For our analyses, please choose the two nearest SNPs.

Based on above discussion, our objective is to find the association between gene expression and the nearest two SNPs. A very convenient tool for studying the association between variables is regression analysis. We should try to describe the relationship between gene expression and the SNPs using a regression model. It is quite clear that gene expression of "CSF2RA" is the response vector and the SNPs are predictors.

Before we establish the regression model, we should perform some initial data analysis. To this end, we need first obtain the data set we need. On the class page website, I have extracted the microarray data set and the SNPs data set for X chromosome. The file "ChromoXmicroarray.txt" contains all the gene expression data from Chromosome X. The file "ChromoXsnp.txt" contains all the SNPs data from Chromosome X. Based on these two files, please do the following:

Q1: Read these two data sets into R. Make sure that the column names and row names are in your data sets.

Answer: To read the data sets into R. You should first save the files into a desired folder. Then use the R command read.table() to read them into R console. Specifically, you could enter the following commands:

```
setwd("..") ## use the directory where your files were saved
chromoX<-read.table(file="ChromoXmicroarray.txt",header=TRUE)
ordered.ChromoXsnp<-read.table(file="ChromoXsnp.txt",header=TRUE)
```

Note that you should include "header=TRUE" so that the column names in the original files are also included into your R data set.

Q2: Find the gene expression of gene "CSF2RA" using the gene IDENTIFIER. The numbers with column names starting with "GSM" are gene expression values.

Answer: To find the which gene expression corresponds to the gene "CSF2RA", you could use the command `which(chromoX$IDENTIFIER=="CSF2RA")`. Since there are several rows in the data set belongs to the gene "CSF2RA", we could pick one of them or summarize them into one number for each individual. For example you could use the following code to find the gene expression of "CSF2RA".

```
CSF2RA<-chromoX[which(chromoX$IDENTIFIER=="CSF2RA")[1],]
```

Q3: Find the SNPs that are nearest to the gene "CSF2RA". The columns with column names staring with "GSM" are SNPs values.

Answer: Based on Q2, we found that the physical location of the gene on Chromosome X is between 1387701 and 1428827. The closest positions in the SNPs data start from 4065309. Note that the SNPs data were ordered by the physical locations. Therefore, the closest two SNPs are the first two SNPs in the SNP data set. Please use the following code to choose 2 nearest SNPs.

```
nearest2SNPs<-ordered.ChromoXsnp[c(1:2),]
```

Q4: Are there missing values? Any non-response? Are they categorical data or numerical continuous data or numerical counts data?

Answer: In the gene expression data set, the data are continuous and no missing values exist in the gene expression.

In the SNPs data set, all the values are categorical data. For example, “1” for SNP_A-1507407 corresponds to “CC”, “2” for “CT” and 3 for “TT”. For SNP_A-1517238, “1” represents “AA”, “2” represents “AT” and “3” for “TT”. In all the SNPs data, “4” is used for representing missing values. In the SNPs data set, there exist some missing values. For example, the value of SNP_A-1507407 for GSM85243 is missing.

Q5: Data cleaning and organizing: remove the samples with missing values and combine the SNPs data set and microarray data set together according to their subject IDs. Note that the column names are sample IDs not the subject IDs. To find their subject IDs, I also downloaded two text files (“MicroarraysampleIDs.txt” and “SNPsampleIDs.txt”) which provide the correspondence between the subject IDs and the sample IDs. In both text files, the first column is the sample ID and the second column is the subject ID. Read these two data sets into R, and replace the sample IDs in the microarray data sets and SNPs data sets with their corresponding subject IDs.

Answer: The details about how to clean the data set and how to organize the data set have been shown in the lab. Please see below for the corresponding R code.

The first part replaces the sample ID by the subject ID for SNPs data. Also removed the missing values in the SNPs data

```
SNPs2X<-ordered.ChromoXsnp[c(1:2),c(1:78)]
readsnpIDs<-read.table(file="SNPsampleIDs.txt")
colnames(SNPs2X)<-readsnpIDs[,2]
nomissingSNPs<-SNPs2X[,(SNPs2X[1,]!="4")&(SNPs2X[2,]!="4")]
```

This part replaces the sample ID by the subject ID for microarray data set.

```
readmicroIDs<-read.table(file="MicroarraysampleIDs.txt")
CSF2RAdata<-CSF2RA[,c(3:49)]
ordered.CSF2RAdata<-CSF2RAdata[,order(colnames(CSF2RAdata))]
```

```
colnames(ordered.CSF2RAdata)<-readmicroIDs[,2]
```

This part finds the common subjects who have both SNPs data and microarray data. Then combine SNPs and microarray data set together.

```
finalSNPs<-
nomissingSNPs[,colnames(nomissingSNPs)%in%colnames(ordered.CSF2RAdata)]
finalMicroarray<-
ordered.CSF2RAdata[,colnames(ordered.CSF2RAdata)%in%colnames(nomissing
SNPs)]
finalSNPs0<-finalSNPs[,order(colnames(finalSNPs))]
finalMicroarray0<-finalMicroarray[,order(colnames(finalMicroarray))]
finaldata<-rbind(finalMicroarray0,finalSNPs0)
```

The final data set contains three rows. The first row is the microarray data, the second row and the second and third row is the corresponding SNPs data.

Now, let us perform some initial data analysis based on the data we obtained.

Q6: Please give summary statistics for the microarray data and the SNPs data.

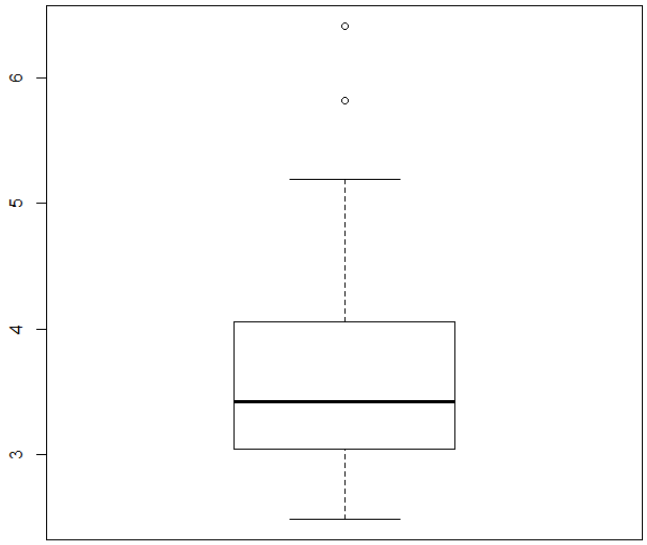
Answer:

The summary statistics for the microarray data is

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.485	3.045	3.422	3.681	4.059	6.414

We can see that median is less than mean, which might indicate that the distribution is right skewed.

If a boxplot of above quantiles is given below



Since the SNPs data are discrete, we can summarize the data using their frequency. For SNP 1, the frequencies are

1	2	3
4	4	21

For SNP 2, the frequencies are

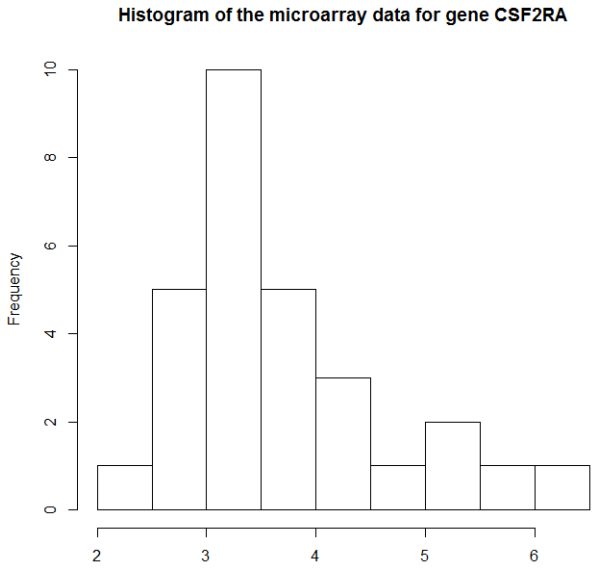
1	2	3
27	1	1

If we combine these two SNPs together, the frequency table is

	SNP 2		
SNP1	1	2	3
1	3	0	1
2	3	1	0
3	21	0	0

We observe that many cells are empty and most of data come from the cell (3,1).

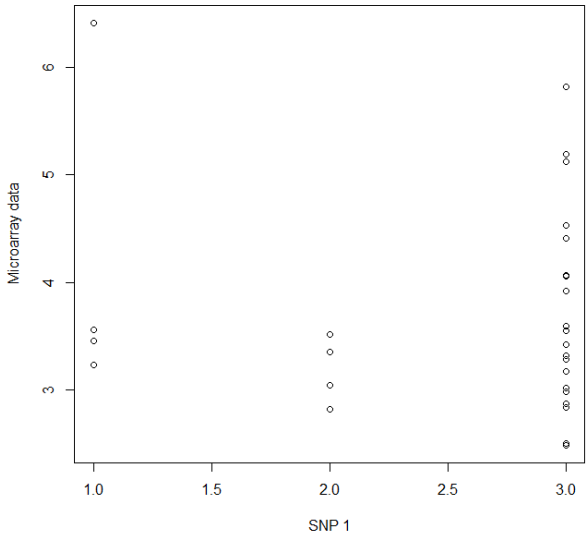
Q7: Give a histogram of the microarray data set. Describe the shape of the histogram.

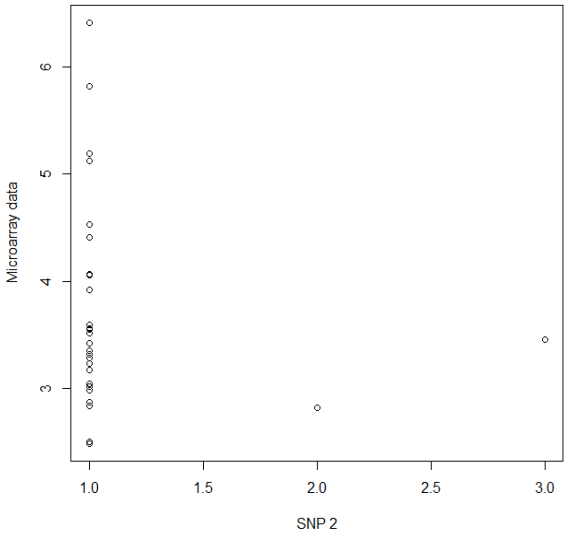


Answer: The histogram of the microarray data set is given above. The histogram is unimodal with center around 3.5. The histogram is also not symmetric and a little right skew.

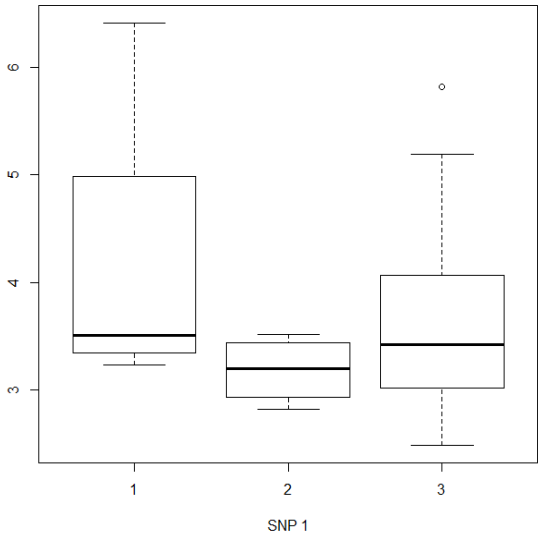
Q8: To check the relationship between SNPs and microarray data, draw a scatter plot of each SNP and the microarray data. If we consider the SNPs data as categorical data set, we could draw boxplots of microarray data according to the values of each SNP.

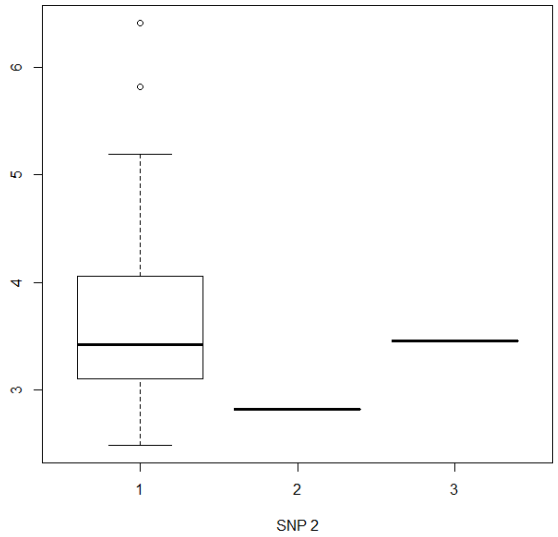
Answer: See below for the scatter plots of SNP 1 vs. microarray data, and SNP 2 vs. microarray data.





As we can see from the above scatter plots, the data are very imbalance. Specifically, for SNP 1, most microarray data are corresponding to SNP value 3. For SNP 2, most data have SNP value 1. The box plots of microarray data versus the SNPs data are given below. It shows that the means of microarray data do not change significantly at different level of SNP.





Since the response variable is a continuous random variable, we could fit a linear regression model using SNPs as predictors and microarray data as response vector. To begin with, let us consider the SNPs data as categorical data and fit a two-way ANOVA model.

Q9: Please describe the two-way ANOVA model in statistical formulation for this data set.

Answer: Let Y_{ijk} be the gene expression value for the k-th individual with values i and j, respectively, for SNPs 1 and 2 ($i, j = 1, 2, 3$). As we can see from Q6, only 5 cells in the frequency table are non-zeros, which are the cells (1, 1), (2, 1), (3, 1), (2, 2) and (1, 3). Therefore, for this data set, the corresponding linear regression model (two-way ANOVA) is

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{131} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{221} \\ Y_{311} \\ \vdots \\ Y_{31,21} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \alpha\beta_{11} \\ \alpha\beta_{13} \\ \alpha\beta_{21} \\ \alpha\beta_{22} \\ \alpha\beta_{31} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{113} \\ \varepsilon_{131} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{213} \\ \varepsilon_{221} \\ \varepsilon_{311} \\ \vdots \\ \varepsilon_{31,21} \end{pmatrix}$$

Note that the design matrix in the above model is of rank 5. The independent estimable linear combination of the parameters is 5. Please also note that we adapt the two-way ANOVA model to our data set.

Q10: Fit the two-way ANOVA model using R. Fit the models with or without interactions. Can we fit a model with interaction for this data set? Why or why not?

Answer: To fit a linear model without interaction, we issue the following command:

```
snp1<-as.factor(unlist(finalSNPs0[1,]))
```

```
snp2<-as.factor(unlist(finalSNPs0[2,]))
y1<-as.numeric(finalMicroarray0)
lm2anova<-lm(y1~snp1+snp2)
summary(lm2anova)
```

The output the above model fitting is

```
Call:
lm(formula = y1 ~ snp1 + snp2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.4003     0.5623   7.826 4.65e-08 ***
snp12        -1.0923     0.7952  -1.374    0.182
snp13        -0.7176     0.6011  -1.194    0.244
snp22        -0.4850     1.1245  -0.431    0.670
snp23        -0.9463     1.1245  -0.842    0.408
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.9739 on 24 degrees of freedom
Multiple R-squared:  0.1081,    Adjusted R-squared:  -0.0406
F-statistic: 0.7269 on 4 and 24 DF,  p-value: 0.5823
```

To fit a two way ANOVA model with interactions, we could use the following command:

```
lm3anova<-lm(y1~snp1*snp2)
summary(lm3anova)
```

The outputs are the same as the model without interactions.

For this data set, we are not able to estimate the interactions. The reason is due to the following. Let μ_{ij} be the mean of Y_{ijk}. According to the definition of the interactions, the interactions are $(\mu_{ij} - \mu_{i',j})$- $(\mu_{ij'} - \mu_{i',j'})$ for any $i \neq i'$ and $j \neq j'$. As we can see from the frequency table in Q6, none of the interactions is estimable due to many empty cells (the cells with 0 observation).

Q11: Do you have any ideas on improving the above two-way ANOVA model?

Answer: To improve the model, one may consider simple models such as additive models which contain few parameters. Let Y_i be the gene expression data for the ith individual. Let X_{1i} be the number of allele "T" for SNP 1 and X_{2i} be the number of allele "T" for SNP 2. Specifically, one could consider an additive model as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

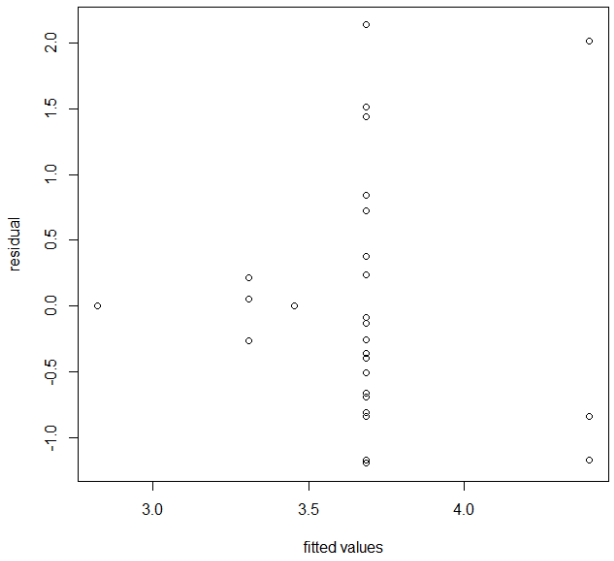
There are many other possible models. Based on the fitting of these models, one could compare the AIC, BIC or Cp values of these models to select a model which is more appropriate.

Model diagnostic is an important step for checking the assumptions of the fitted model and searching for an improvement of the fitted model.

Q12: To check if the model is sufficient, you might plot the fitted value versus the residual of the two-way ANOVA model. What is your conclusion for this data set?

Answer: The plot of the fitted value versus the residual is given below:

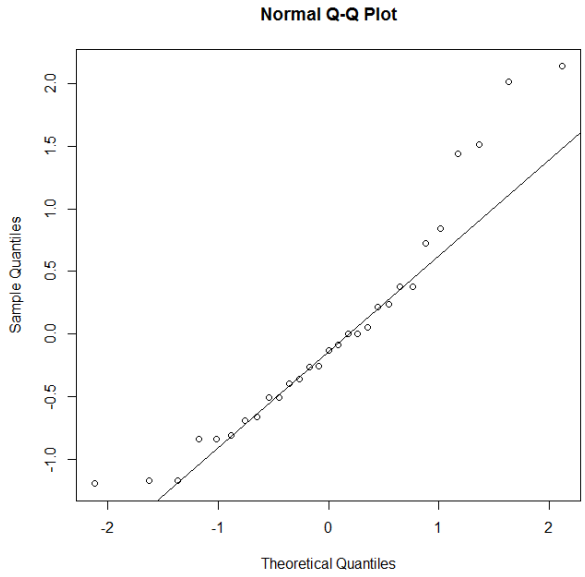
```
plot(lm2anova$fitted.value,lm2anova$residuals,xlab="fitted values",
ylab="residual")
```



Based on the residual plot, we observed that the mean of the residuals is almost symmetric around 0. But it seems that the variations of the residual are very different at different level of fitted values. This might indicate that we should use a linear model with heteroskedastic variances. The above phenomena might be also due to the extreme unbalance of this data set because most data points are within the cell (3, 1).

Q13: To check the normality assumption, you might use a Quantile-Quantile plot.

Answer: The Q-Q plot of the residual is given below:



Based on the above QQ plot, we can observe see that normality assumption is reasonable since most points are around the line. But there exist some deviations from normality on the right tail of the data as we observed in the histogram. A skewed distribution such as Gamma distribution might be more suitable for this data set.

Now assume that the two-way ANOVA model without interaction is sufficient for the data set. Answer the following questions using statistical inference methods.

Q14: Are these two SNPs associated with the gene expression of the gene "CSF2RA"?

Answer: Consider the following two-way ANOVA model without interactions:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

To see if these two SNPs are associated with the gene expression, it is equivalent to test if all the α_i, β_j s are zero or not. This could be done using an F-statistic. Based on the R out, the F-statistic value is 0.7269 and the corresponding p value is 0.5823. These results imply that these two SNPs are not significantly associated with the gene expression of interest.

Q15: For SNP 1, please check if the effects of two alleles on the gene expression are the same or not.

Answer: Based on the two-way ANOVA model without interaction, the different effects of the two alleles of SNP 1 could be inferred from the differences between α_1 and α_2 or α_2 and α_3. So we could construct a confidence interval for α_1-α_2 or α_2-α_3. First, we compute the variance covariance matrix of the estimated coefficients as following:

> vcov(lm2anova)

	(Intercept)	snp12	snp13	snp22	snp23
(Inter)	3.161335e-01	-0.3161335	-3.161335e-01	-2.287924e-17	-3.161335e-01
snp12	-3.161335e-01	0.6322671	3.161335e-01	-3.161335e-01	3.161335e-01
snp13	-3.161335e-01	0.3161335	3.612955e-01	2.627225e-17	3.161335e-01
snp22	-2.287924e-17	-0.3161335	2.627225e-17	1.264534e+00	2.026377e-17
snp23	-3.161335e-01	0.3161335	3.161335e-01	2.026377e-17	1.264534e+00

Using the model fitting results in Q10 and the above variance covariance matrix, a 95% confidence interval for α_2-α_3 is

(-1.0923+0.7176-1.96*0.3612955, -1.0923+0.7176+1.96*0.3612955),

which gives a confidence interval (-1.0828392, 0.3334392). The confidence interval includes 0. This might imply that the two alleles of the SNP 1 have some effects on the gene expression.