

Homework 1

Nan Cao

January 31, 2016

```
library(car)
```

```
#####Q1
```

```
# import of dataset
```

```
setwd("C:/Users/nan66/documents/msu/stt864/LAB1") MSID<-read.table(file="MicroarraysampleIDs.txt")
SNPID<-read.table(file="SNPsampleIDs.txt") CXM<-read.table(file="ChromoXmicroarray.txt",head=T)
CXSNP<-read.table(file="ChromoXsnp.txt",head=T)
```

```
# Q2
```

```
# CSF2RA
```

```
CSF2RA<-CXM[which(CXM$IDENTIFIER=="CSF2RA")[1,],] CSF2RAdata<-CSF2RA[,c(3:49)]
CSF2RAdata ordered.CSF2RAdata<-CSF2RAdata[,order(colnames(CSF2RAdata))] colnames(ordered.CSF2RAdata)<-
MSID[,2]
```

```
group<-c(rep(1,27),rep(2,20)) ttests<-function(x,group) { x<-as.numeric(x) pvalue<-t.test(x ~ group)$p.value;
return(pvalue) } pvalues_CFS2RA<-ttests(CSF2RA[,c(3:49)],group) pvalues_CFS2RA
```

```
#Q3 Q4 Q5
```

```
# the nearest Q
```

```
nearestSNPs<-CXSNP[c(1:2),] which(CXM$IDENTIFIER=="CSF2RA") SNPsX<-CXSNP[1,c(1:78)] SNPsX
colnames(SNPsX)<-SNPID[,2] nomissingSNPs<-SNPsX[,SNPsX!="4"] finalSNPs<-nomissingSNPs[,colnames(nomissingSNPs)]
finalMicroarray<-ordered.CSF2RAdata[,colnames(ordered.CSF2RAdata)%in%colnames(nomissingSNPs)]
finalSNPs0<-finalSNPs[,order(colnames(finalSNPs))] finalMicroarray0<-finalMicroarray[,order(colnames(finalMicroarray))]
finaldataX<-rbind(finalMicroarray0,finalSNPs0)
```

```
#the neatest two SNPs
```

```
nearest2SNPs<-CXSNP[c(1:2),] SNPs2X<-CXSNP[c(1:2),c(1:78)] SNPs2X colnames(SNPs2X)<-
SNPID[,2] nomissingSNPs2X<-SNPs2X[, (SNPs2X[1,]!="4") & (SNPs2X[2,]!="4")] finalSNPs2X<-
nomissingSNPs2X[,colnames(nomissingSNPs2X)%in%colnames(ordered.CSF2RAdata)] finalMicroarray2X<-
ordered.CSF2RAdata[,colnames(ordered.CSF2RAdata)%in%colnames(nomissingSNPs2X)] finalSNPs02X<-
finalSNPs2X[,order(colnames(finalSNPs2X))] finalMicroarray02X<-finalMicroarray2X[,order(colnames(finalMicroarray2X))]
finaldata2X<-rbind(finalMicroarray02X,finalSNPs02X)
```

```
# summary statistics
```

```
summary(finaldata2X)
```

```
#Q7
```

```
hist(as.matrix(CSF2RA[,c(4:49)]),main="histogram of CSF2RA")
```

#Q8

```
MA<-as.numeric(finalMicroarray02X[1,]) snp1<-as.numeric(finalSNPs02X[1,]) snp2<-as.numeric(finalSNPs02X[2,])
scatterplot(snp1|MA) scatterplot(snp2|MA) boxplot(MA~snp1) boxplot(MA~snp2)
```

#Q9

MIC

Q11

Two-way ANOVA models

```
snp1.factor<-as.factor(snp1) snp2.factor<-as.factor(snp2) y<-MA lm2anova<-lm(y~snp1.factor*snp2.factor)
lm2anova summary(lm2anova) anova(lm2anova)
```

#QQnorm & interaction plot

```
qqnorm(lm2anova$res) interaction.plot(snp1.factor,snp2.factor,y)
```

Q12

```
x0<-as.numeric(finalSNPs0)-1 y0<-as.numeric(finalMicroarray0)
```

```
xvec4AA<-c(1,1,0) xvec4aa<-c(1,0,-1) xvec4Aa<-c(1,1,-1) Xmat<-matrix(0,length(finalSNPs0),3) for (i in
1:length(finalSNPs0)) { Xmat[i,<-xvec4AA (finalSNPs0[i]=="1")+xvec4Aa(finalSNPs0[i]=="2")+xvec4aa*(finalSNPs0[i]=="3")
}
```

```
betahat<-solve(t(Xmat)%%Xmat)%%t(Xmat)%%y0 lm3<-lm(y0~Xmat[,2]+Xmat[,3]) lm3
```

Check residuals

```
predy<-Xmat%%solve(t(Xmat)%%Xmat)%%t(Xmat)%%y0 epsilonhat<-y0-predy plot(predy, epsilonhat)
```

Confidence interval for b1-b2 or hypothesis testing for

H0: b1=b2 vs H1: b1≠b2

```
n<-length(y0) rankx<-qr(Xmat)$rank SSE<-sum(epsilonhat^2) MSE<-SSE/(n-rankx) sigmahat<-
sqrt(MSE) cvec<-c(0,1,-1) cbeta<-cvec%%solve(t(Xmat)%%Xmat)%%t(Xmat)%%y0 varcbeta<-
t(cvec)%%solve(t(Xmat)%%Xmat)%%cvec low95cbeta<-cbeta-qt(0.975,n-rankx)sigmahatsqrt(varcbeta)
upp95cbeta<-cbeta+qt(0.975,n-rankx)sigmahatsqrt(varcbeta) CI95cbeta<-c(low95cbeta,upp95cbeta) Tn<-
cbeta/(sigmahatsqrt(varcbeta)) pvalue<-2*(1-pt(abs(Tn),n-rankx)) CI95cbeta pvalue cov(c(y,snp1,snp2))
```

Q13

Check normal distribution

```
qqnorm(y)
```

#Q14

```
anova(lm2anova)
```

Q15

```
cvec2<-c(0,1,1) cbeta2<-cvec2%%solve(t(Xmat)%%Xmat)%%t(Xmat)%%y0 varcbeta2<-t(cvec2)%%solve(t(Xmat)%%Xmat)%%
low95cbeta2<-cbeta2-qt(0.975,n-rankx)sigmahat2sqrt(varcbeta2) upp95cbeta2<-cbeta2+qt(0.975,n-
rankx)sigmahat2sqrt(varcbeta2) CI95cbeta2<-c(low95cbeta2,upp95cbeta2) Tn2<-cbeta2/(sigmahat2sqrt(varcbeta2))
pvalue2<-2*(1-pt(abs(Tn),n-rankx)) CI95cbeta2 pvalue2
```