实验7、R语言中线性回归分析和方差分析

年级： 15级　　专业：生信　　学号：1513401013　　姓名：郑磊

编号　　一　　二　　三　　四　　总分　　评阅人

得分

软硬件平台：

1. 硬件平台：（硬件配置）i5，2.9HZ处理器，16G内存，64位操作系统

2. 系统平台：（操作系统及其版本号）Windows10 企业版

3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL地址）R3.4.1 ，Rstudio

一、目的要求：

1、理解线性回归分析,并掌握R语言中线性回归分析函数；

2、理解单因素方差分析,并掌握R语言中单因素方差分析。

二、实验内容：

1、不同基因表达水平之间的线性回归分析

该环节需要大家提前准备好一个基因表达谱数据，如果没有，则有授课教师提供（gds4794）。

1.1、加载数据

加载GEOquery包，读取基因表达谱数据（gds4794）。

```
setwd("D:/RFile/实验七")
library(GEOquery)
gds4794 <- getGEO(filename='GDS4794.soft.gz')
```

1.2、提取数据表

从该基因表达谱数据中提取基因表达数据表。

```
data<-Table(gds4794)

rownames(data)<-data[,1]

row.name = rownames(data)
```

1.3、随机抽样

从该数据表中随机抽取一行数据，记录对应的探针ID和基因名称。

```
n = 1

set.seed(1)

sam.row.name = sample(row.name,n,replace=F)

sam.row.name

a <- unlist(data[sam.row.name,3:67])

gene_name_a <- as.character(data[sam.row.name,2])

gene_name_a
```

1.5、不同基因表达水平线性回归分析

遍历整个基因表达谱数据表，利用R语言中的线性回归分析函数，分析其他所有基因表达水平，与1.3步随机抽取的基因表达水平之间的线性回归关系；记录斜率、截距、R2以及F检验的p.value；同时为所有p.value和相关系数值（cor）关联探针ID或基因名称。

```
xb = NULL

xk = NULL

xr = NULL

xp = NULL

for(i in 1:nrow(data)){
```

```
    if(data[i,1] != sam.row.name){
      b <- unlist(data[i,3:67])
      lm.sol <- lm(b~1+a)
      suma <- summary(lm.sol)
      xb <- c(xb,lm.sol$coefficients[1])
      xk <- c(xk,lm.sol$coefficients[2])
      xr <- c(xr,suma$r.squared)
      pv <- 1-pf(suma$fstatistic[1],suma$fstatistic[2],suma$fstatistic[3])
      xp <- c(xp,pv)
    }
  }
  names(xb)<-data[-which(data$ID_REF==sam.row.name),1]
  names(xk)<-data[-which(data$ID_REF==sam.row.name),1]
  names(xr)<-data[-which(data$ID_REF==sam.row.name),1]
  names(xp)<-data[-which(data$ID_REF==sam.row.name),1]
```

1.6、高相关性基因筛选

设定p.value（至少小于0.05）和相关系数R2的筛选阈值（至少大于0.25）；对1.5步计算结果进行筛选，保留符合条件的基因；对符合条件的p.value和相关系数r所关联的基因名称进行交集运算；查看交集运算结果中是否存在非法基因信息，如果有去除它；筛选高相关性基因的斜率和截距数据。

```
  p_value = 0.01
  r_cutoff = 0.65
```

```
xp2 <- xp[xp<p_value]

xr2 <- xr[xr>r_cutoff]

genes <- intersect(names(xp2),names(xr2))

length(genes)
```

2、不同基因表达水平之间的单因素方差分析

把与所选基因表达水平相关性最高的那个基因表达数据提取出来；

两个基因的表达水平进行线性回归分析；

```
maxgene = unlist(data[which(data$ID_REF=="1556761_at"),3:67])

re_lm.sol = lm(maxgene~1+a)

summary(re_lm.sol)
```

绘制评价回归分析结果中的四张图片；

```
png(file = "plot.png")

par(mfrow=c(2,2))

plot(re_lm.sol)

dev.off()
```

绘制表达水平的散点图和回归方程；

```
png(file = "plot2.png")

plot(a,maxgene,lwd=2,main="plot2")

y_mean=mean(maxgene)

abline(h=y_mean,col="blue")

x_mean=mean(a)

abline(v=x_mean,col="purple")
```

abline(re_lm.sol,col="red")

dev.off()

单因素方差分析。

aov(a~maxgene)

summary(aov(a~maxgene))

三、实验结果：

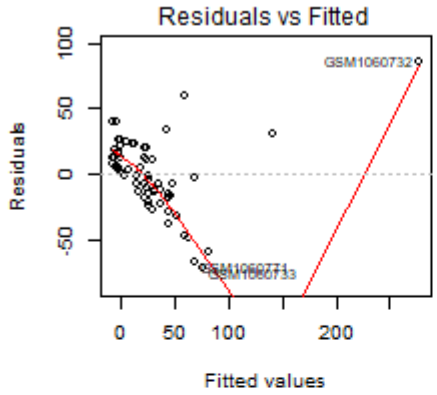
1.3

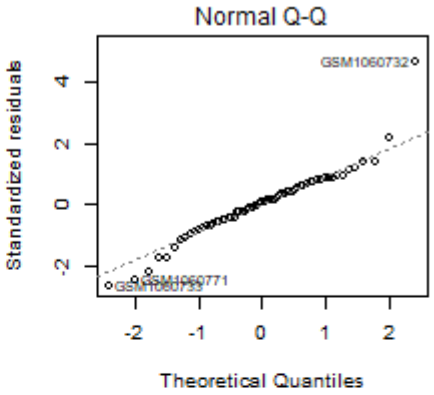
```
> sam.row.name
[1] "205069_s_at"
> gene_name_a
[1] "ARHGAP26"
```

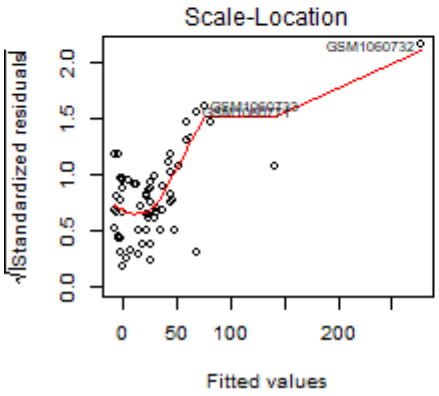
1.6

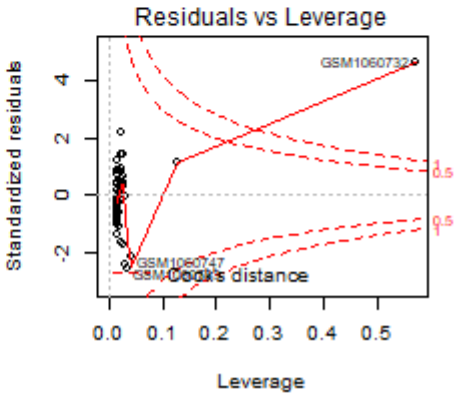
探针ID	基因名称	斜率	截距	R2	F检验pvalue
1556761_at	AI057305	2.627491	-10.1503	0.6894553	0
210252_s_at	MADD	8.684072	10.91999	0.6807879	0
230928_at	H15173	2.260044	230928_at	0.6586559	2.220446e-16

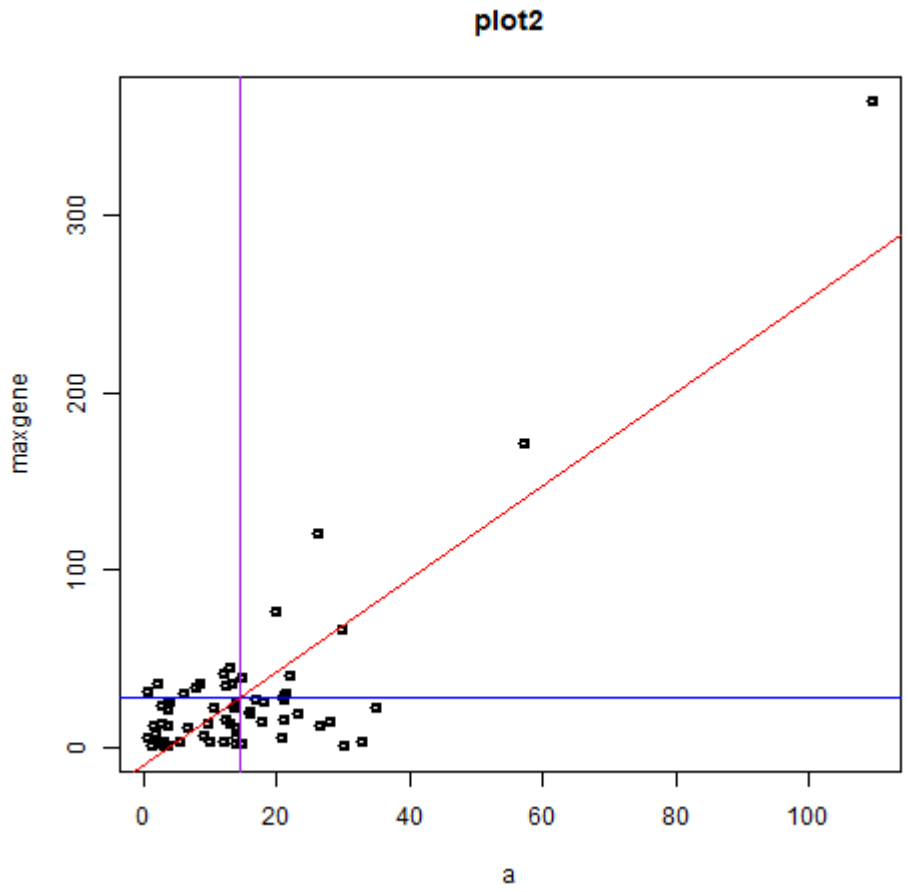
2.


Residuals vs Fitted
Residuals
Fitted values
GSM1060732
GSM1060771
GSM1060733


Normal Q-Q
Standardized residuals
Theoretical Quantiles
GSM1060732
GSM1060771
GSM1060733


Scale-Location
√|Standardized residuals|
Fitted values
GSM1060732


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
GSM1060732
GSM1060747




```
> aov(a~maxgene)
Call:
   aov(formula = a ~ maxgene)

Terms:
                  maxgene Residuals
Sum of Squares  11167.438  5030.043
Deg. of Freedom         1        63

Residual standard error: 8.935432
Estimated effects may be unbalanced
> summary(aov(a~maxgene))
            Df Sum Sq Mean Sq F value Pr(>F)
maxgene      1  11167   11167   139.9 <2e-16 ***
Residuals   63   5030      80
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

四、讨论：

1.7

在生物体中一个性状往往是由多个基因共同调控的，甚至会影响到其他性状或代谢过程，还有一些反馈调节也可能会影响到基因的表达。还有一种可能，这些基因可能调控着同一条代谢通路。

在DAVID网站处理结果

ID	Gene Name	Species	GOTERM_BP_DIRECT	GOTERM_CC_DIRECT	GOTERM_MF_DIRECT
ELAVL3	ELAV like RNA binding protein 3(ELAVL3)	Homo sapiens	GO:0007399~nervous system development,GO:0030154~cell differentiation,		GO:0000166~nucleotide binding,GO:0003676~nucleic acid binding,GO:0003723~RNA binding,GO:0017091~AU-rich element binding,
STH	saitohin(STH)	Homo sapiens		GO:0005634~nucleus,GO:0005737~cytoplasm,	
SLC25A41	solute carrier family 25 member 41(SLC25A41)	Homo sapiens	GO:0006412~translation,GO:0055085~transmembrane transport,	GO:0005743~mitochondrial inner membrane,GO:0016021~integral component of membrane,	GO:0003735~structural constituent of ribosome,