

Solutions:1a)

> str(y)

num [1:120] 3 5 3 1 2 1 2 1 0 2 ...

> #MEan of the sample

> print("mean of th esample is:")

[1] "mean of th esample is:"

> mean(y)

[1] 1.816667

Sample size : 120

Mean of the sample is 1.816667

1b)

> llh <- function(x) - sum(log(dpois(y, lambda=exp(x))))

> print("the vaue of theta is:")

[1] "the vaue of theta is:"

> optim(0, llh)$par

[1] 0.5970703

The value of is 0.5970703

1c)

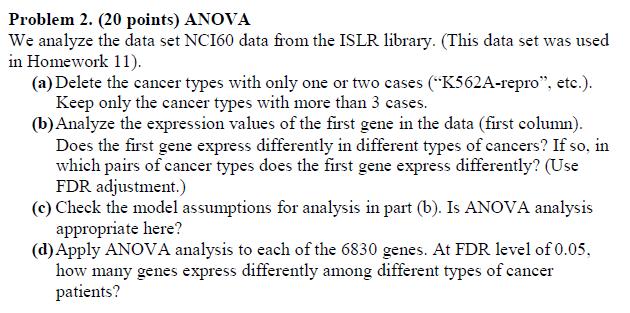
> quantile(boot.xbar,c(0.025,0.975))

2.5% 97.5%

0.4435547 0.7378906

We can thus conclude that the theta value lies between (0.4435547, 0.7378906)

Thus we reject the NULL HYPOTHESIS



Solutions:

2a)

library(ISLR)

> nci.data<- NCI60$data

> nci.labs<- NCI60$labs

> data <- NULL

> z=1

> for (i in 1:64){

+ if (sum(nci.labs==(nci.labs[i])) <= 3){

+ data[z] <- i

+ z=z+1

+ }

+ }

> new.data <- nci.data[-data,]

> new.labs <- nci.labs[-data]

2b)

> anova(lm(gene ~ new.labs))

Analysis of Variance Table

Response: gene

Df Sum Sq Mean Sq F value Pr(>F)

new.labs 7 2.8931 0.41331 2.3272 0.03928 \*

Residuals 49 8.7021 0.17759

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The p value obtained is 0.03928 which is less than 0.05, we can say that thre is no evidence that the gene is expressed differently, thus we reject the NULL hypothesis

> pairwise.t.test(gene, new.labs, p.adjust.method = 'fdr')

Pairwise comparisons using t tests with pooled SD

data: gene and new.labs

BREAST CNS COLON LEUKEMIA MELANOMA NSCLC OVARIAN

CNS 0.34 - - - - - -

COLON 0.35 0.10 - - - - -

LEUKEMIA 0.42 0.11 0.93 - - - -

MELANOMA 0.93 0.34 0.34 0.34 - - -

NSCLC 0.34 0.93 0.10 0.10 0.34 - -

OVARIAN 0.34 0.93 0.10 0.12 0.37 0.99 -

RENAL 0.93 0.34 0.34 0.35 0.93 0.34 0.34

P value adjustment method: fdr

As all the p-value are >0.05 its difficult to distinguish between the types of cancers.

2c)

> shapiro.test(residuals(lm(gene ~ new.labs)))

Shapiro-Wilk normality test

data: residuals(lm(gene ~ new.labs))

W = 0.9795, p-value = 0.4414

We fail to reject NULL hypothesis as the p-value is 0.4414 which is less than 0.05.

> bptest(lm(gene ~ new.labs), studentize = FALSE)

Breusch-Pagan test

data: lm(gene ~ new.labs)

BP = 8.8392, df = 7, p-value = 0.2644

We cant reject the NULL hypothesis as p-value is 0.2644 which is less than 0.05.

2d)

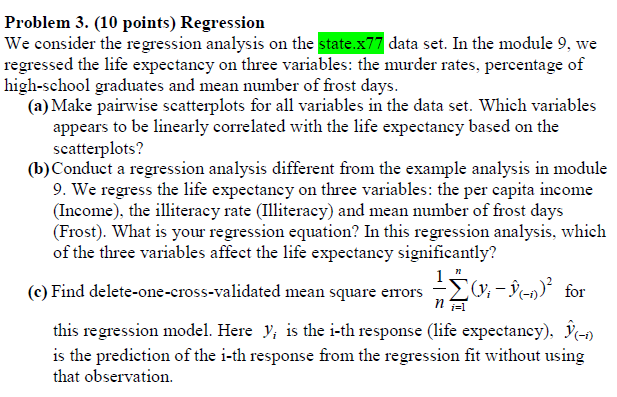
> anova <- apply(new.data, 2, function(x) anova(lm(x ~ new.labs))$Pr[1])

> p.fdr <- p.adjust(p=anova, method="fdr")

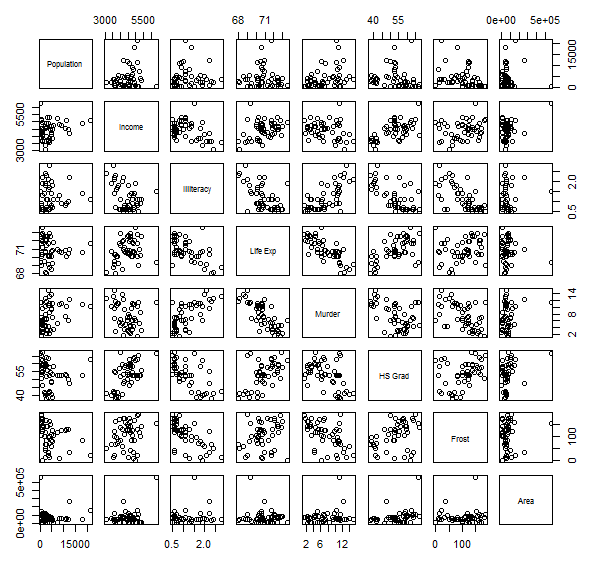
> sum(p.fdr<0.05)

[1] 2808

2808 genes were expressed differently



Solutions: 3a)



Murder, HS Grad, Frost and illiteracy appear to be correlated with life expectancy

3b)

The equation for regression analysis is

Life.Exp = 72.5+ 0.0002income – 1.56 illiteracy – 0.006 frost

As per this equation illiteracy plays a more significant role in life expectancy.

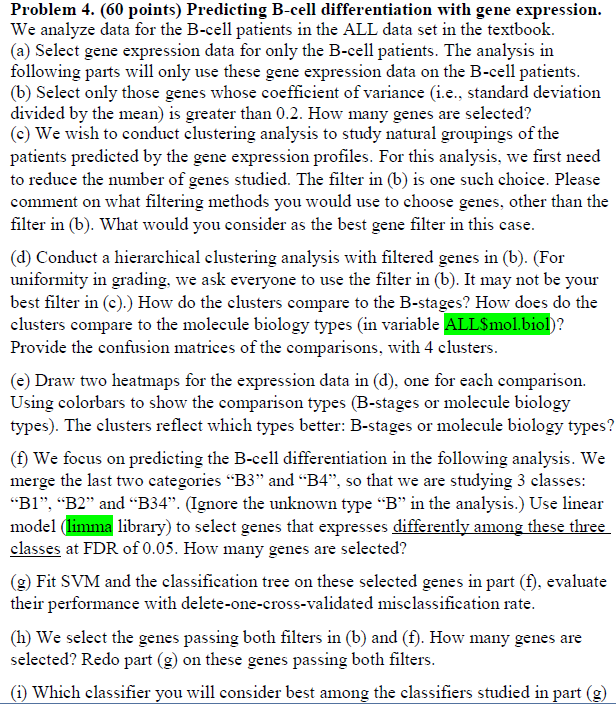
3c)

> error <- mean(error)

> print(error)

[1] 1.345143

The delete one cross validated error is 1.345143



Solutions: 4a)

> ALL.B <- ALL[,which(ALL$BT %in% c("B","B1","B2","B3","B4"))]

> B.data <- exprs(ALL.B)

> str(B.data)

num [1:12625, 1:95] 7.6 5.05 3.9 5.9 5.93 ...

- attr(\*, "dimnames")=List of 2

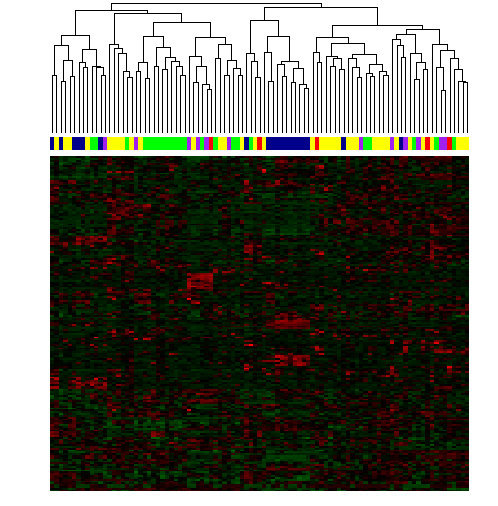
..$ : chr [1:12625] "1000\_at" "1001\_at" "1002\_f\_at" "1003\_s\_at" ...

..$ : chr [1:95] "01005" "01010" "03002" "04006" ...

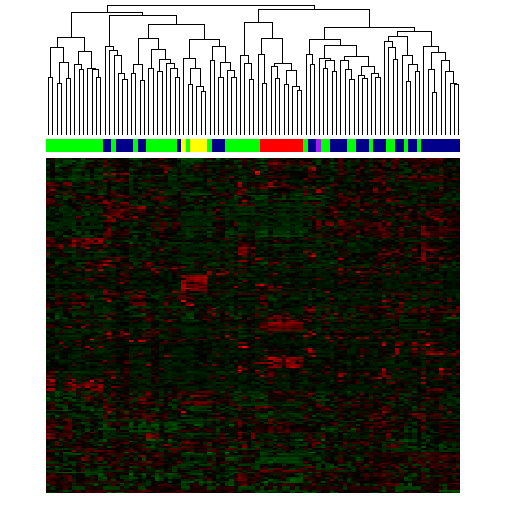
4b)

Selecting the genes whose coefficience of variance is greater than 0.2 , we get 184 genes

4e)



Most of the B1 is in the first cluster. Most of the B2 (yellow) is in the second cluster. Most of the B3 (green) is in the first cluster.



Most of the green(NEG) is in the first cluster. Entire yellow(E2A/PBXI) is in the first cluster. Entire red(ALL/AF4) and (p15/p16) are in the second cluster. I think the clusters reflect molecular biology types better.

4c) Shapiro–Wilks test has the power to detect even small deviations from normality and score them as significant. Thus we use Shapiro-Wilks test.

4d)

> table(ALL$BT[1:95,drop=T],cluster)

cluster

1 2 3 4

B 3 1 1 0

B1 2 0 11 6

B2 20 10 2 4

B3 5 15 1 2

B4 6 5 0 1

> table(ALL$mol.bio[1:95],cluster)

cluster

1 2 3 4

ALL1/AF4 0 0 10 0

BCR/ABL 25 12 0 0

E2A/PBX1 0 5 0 0

NEG 10 14 5 13

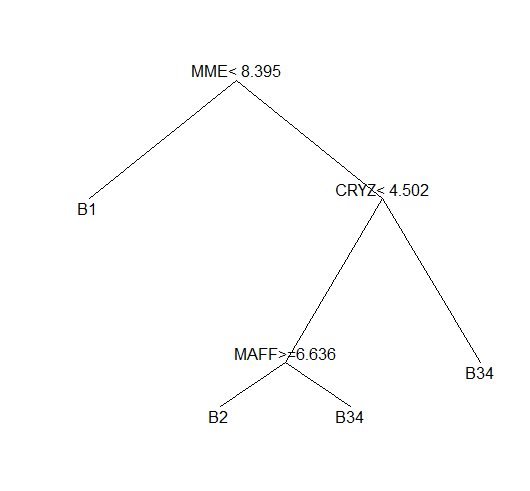
NUP-98 0 0 0 0

p15/p16 1 0 0 0

4f)

Using linear model I could select 1169 genes.

4g)



For SVM

Delete one cross validated misclassification rate is 0.3

> mcr.cv<-mean(mcr.raw)

> mcr.cv

[1] 0.3

For classification tree Delete one cross validated misclassification rate is 0.102

> tree.cv <- mean(tree.raw)

> tree.cv

[1] 0.102

4h)

Passing both the filters in (b) and (f) we can get 55 genes.

4i)

I think the genes in (h) is better as we are filtering down using more rigorous filtering methods

Problem 5

5a) Code in r script