Sardana_Module8-HW

Prachi Sardana

2023-03-12

Question 1 a) The mean expression values are different across all disease stages.

```
library("ALL")
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
data("ALL")

ALLB123 <- ALL[,ALL$BT%in%c("B1","B2","B3","B4")]
y <- exprs(ALLB123)["109_at",]
# one way anova
anova(lm(y ~ ALLB123$BT))
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## ALLB123$BT  3  1.9142 0.63808  4.2195 0.007817 **
## Residuals  86 13.0050 0.15122
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1 b) To find mean gene expression value among B3 patients from the linear model fits.

```
library("ALL")
data("ALL")
# Mean expression value among B3 patients
ALLB123 <- ALL[,ALL$BT =="B3"]
mean_B3 <- lm(exprs(ALLB123)["109_at",]~1)
summary(mean_B3) # summary table

##
## Call:
## lm(formula = exprs(ALLB123)["109_at", ] ~ 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9126 -0.2735  0.0931  0.2722  0.7153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.68533    0.09066   73.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4348 on 22 degrees of freedom
```

c) Use the pairwise comparisons at FDR=0.05 to find which group means are different. Show the output of your code. What is your conclusion?

```
library("ALL")
data("ALL")
ALLB123 <- ALL[,ALL$BT%in%c("B1","B2","B3")]
y <- exprs(ALLB123)["109_at",]
# pairwise t test
# method = fdr
p_values <- pairwise.t.test(y,ALLB123$BT,p.adjust.method='fdr')
p_values

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and ALLB123$BT
##
##    B1   B2
## B2 0.39 -
## B3 0.39 0.15
##
## P value adjustment method: fdr
```

d) Anova model assumption with Shapiro wilk diagnostic test

```
library("ALL")
data("ALL")
ALLB123 <- ALL[,ALL$BT%in%c("B1","B2","B3")]
y <- exprs(ALLB123)["109_at",]
# Shapiro wilk test for normality
shapiro.test(residuals(lm(y ~ ALLB123$BT)))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm(y ~ ALLB123$BT))
## W = 0.97592, p-value = 0.146
```

Question 2 a) Apply the nonparametric Kruskal-Wallis tests for every gene on the B-cell ALL patients in stage B, B1, B2, B3, B4 from the ALL data

```
library(ALL);data(ALL)
ALLB_1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB_1234)
# Kruskal test using apply function
kruskal.test <- apply(y, 1, function(x) kruskal.test(x ~
ALLB_1234$BT)$p.value)

#(a): used fdr adjustment at 0.05 level
p.fdr <-p.adjust(kruskal.test, method="fdr")
fdr<-sum(p.fdr < 0.05)
cat("\nThe genes are expressed different in some of the groups are:",fdr)

##
## The genes are expressed different in some of the groups are: 423

#(b) To find the probe names for the top five genes with smallest p-values.
top_5_genes<-names(sort(p.fdr)[1:5])
cat("\nProbe names with top 5 genes with smallest p values:",top_5_genes)

##
## Probe names with top 5 genes with smallest p values: 1389_at 38555_at
40268_at 1866_g_at 40155_at
```

Question 3

a) To conduct appropriate ANOVA analysis considering two factors in disease stages and gender of the patient # There is a significant difference in both the factors.# both factor doesn't affect the gene expression as the p values aren't closer. There is a statistical difference between the gender types and B cell stages.

```
library("ALL")
library("lmtest")

## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

data("ALL")

ALLB <- ALL[,which(ALL$BT%in%c("B1","B2","B3","B4"))]
y <- exprs(ALLB)["38555_at",]
# (a): anova test
anova_test <- anova(lm(y~ALLB$sex+ALLB$BT))
anova_test

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## ALLB$sex   1  0.366  0.3665  0.8861    0.3492
## ALLB$BT    3 24.101  8.0338 19.4248 1.174e-09 ***
## Residuals 84 34.741  0.4136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(lm(y~ALLB$BT+ ALLB$sex))

##
## Call:
## lm(formula = y ~ ALLB$BT + ALLB$sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0392 -0.4896 -0.0535  0.4686  1.6918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.94064    0.16954  40.937  < 2e-16 ***
## ALLB$BTB2   -0.79014    0.18361  -4.303 4.52e-05 ***
## ALLB$BTB3   -1.42256    0.20071  -7.088 3.92e-10 ***
## ALLB$BTB4   -1.34133    0.23714  -5.656 2.09e-07 ***
## ALLB$sexM   -0.04005    0.14428  -0.278    0.782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6431 on 84 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4132, Adjusted R-squared:  0.3853
## F-statistic: 14.79 on 4 and 84 DF,  p-value: 3.459e-09
```

```r
# The mean expression values differ while considering two factors.Hence, statistically different

#b) used shapiro and bp diagnostic tests

# Shapiro-Wilk test for normality of residuals
shapiro.test(residuals(lm(y ~ ALLB$BT + ALLB$sex)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm(y ~ ALLB$BT + ALLB$sex))
## W = 0.97097, p-value = 0.04335
```

```r
# Breusch and Pagan test
bptest(lm(y ~ ALLB$BT+ALLB$sex), studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  lm(y ~ ALLB$BT + ALLB$sex)
## BP = 4.5761, df = 4, p-value = 0.3336
```

```r
# using shapiro test p value is 0.04 since the alpha is less than 0.05. used bp diagnostic test p value = 0.33 . Since the p value is different ,anova model assumptions are violated.
```