Northeastern University
College of Science

# Module 8 Homework

## Problem 1 (50 points)

On the ALL data set, consider the ANOVA on the gene with the probe “109_at” expression values on B-cell patients in 5 groups: B, B1, B2, B3 and B4.

(a) Conduct the one-way ANOVA. Do the disease stages affect the mean gene expression value?

**Answer)**

```
data(ALL,package="ALL");library(ALL)

library(lmtest)

ALLB12345 <- ALL[,ALL$BT %in% c("B","B1","B2","B3","B4")]

y<-exprs(ALLB12345)["109_at",]

anova(lm(y ~ ALLB12345$BT))
```

**Output)**

```
Analysis of Variance Table

Response: y
             Df  Sum Sq Mean Sq F value  Pr(>F)  
ALLB12345$BT  4  2.1053 0.52632  3.4829 0.01082 *
Residuals    90 13.6006 0.15112                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From Anova table, p-value=0.01082 is very small and we reject the null hypothesis. hence we conclude that the 109_at gene expression is related to the disease stages for B-cells:B,B1,B2,B3,B4.

(b) From the linear model fits, find the mean gene expression value among B3 patients. Make sure you show the summary table in your submission.

**Answer)**

```
ALLB3 <- ALL[,ALL$BT =="B3"]
mean <- lm(exprs(ALLB3)["109_at",]~1)
summary(mean)
```

**Output)**

```
Call:
lm(formula = exprs(ALLB3)["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?

Answer)

```
ALLB12345 <- ALL[,ALL$BT %in% c("B","B1","B2","B3","B4")]
y<-exprs(ALLB12345)["109_at",]
pairwise.t.test(y, ALLB12345$BT, p.adjust.method = 'fdr')
```

**Output)**

```
> pairwise.t.test(y, ALLB12345$BT, p.adjust.method = 'fdr')

        Pairwise comparisons using t tests with pooled SD 

data:  y and ALLB12345$BT 

   B    B1   B2   B3  
B1 0.40 -    -    -   
B2 0.19 0.48 -    -   
B3 0.57 0.48 0.15 -   
B4 0.62 0.11 0.01 0.20

P value adjustment method: fdr
```

**Conslusion:**

The results indicate that only for B4/B2, the value is 0.01, which is less than 0.05.

(d) Check the ANOVA model assumptions with diagnostic tests? Do we need to apply robust ANOVA tests here? If yes, apply the appropriate tests and state your conclusion.

**Answer:**

```
ALLB12345 <- ALL[,ALL$BT %in% c("B","B1","B2","B3","B4")]
y<-exprs(ALLB12345)["109_at",]
shapiro.test(residuals(lm(y ~ ALLB12345$BT)))
bptest(lm(y~ALLB12345$BT), studentize=FALSE)
```

**Output:**

```
> shapiro.test(residuals(lm(y ~ ALLB12345$BT)))

Shapiro-Wilk normality test

data:  residuals(lm(y ~ ALLB12345$BT))
W = 0.97839, p-value = 0.1177

> bptest(lm(y~ALLB12345$BT), studentize=FALSE)

        Breusch-Pagan test

data:  lm(y ~ ALLB12345$BT)
BP = 1.1702, df = 4, p-value = 0.883
```

**Conclusion:**

For shapiro test, the p-value is 0.1177, so we don't reject null hypothesis of normally distributed residuals. Therefore, the normality assumption does hold. For Besusch-Pagan test, the p-value is 0.883, so we don't reject the null hypothesis of equal variances (homoscedasticity).

Answer the question in each part directly. Relevant R outputs should be displayed to support your conclusion.

**Problem 2 (25 points)**

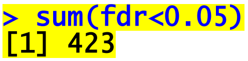
Apply the nonparametric Kruskal-Wallis tests for every gene on the B-cell ALL patients in stage B, B1, B2, B3, B4 from the ALL data. (Hint: use the apply() function.)

(a) Use FDR adjustments at 0.05 level. How many genes are expressed different in some of the groups?

**Answer:**

```
ALLB12345 <- ALL[,ALL$BT %in% c("B","B1","B2","B3","B4")]
y<-exprs(ALLB12345)
kruskal_test <- apply(y, 1, function(x) kruskal.test(x ~
ALLB12345$BT))
p.values<- sapply(kruskal_test, function(x) x$p.value)
fdr <- p.adjust(p=p.values, method ='fdr')
sum(fdr<0.05)
```

**Output:**



```
> sum(fdr<0.05)
[1] 423
```

(b) Find the probe names for the top five genes with smallest p-values.

**Answer:**

```
genes5 <- names(sort(fdr)[1:5])
genes5
```

**Output:**

```
  > genes5

[1] "1389_at"   "38555_at" "40268_at"
[4] "1866_g_at" "40155_at"
```

Please submit your R commands together with your answers to each part of the question.

**Problem 3 (25 points)**

On the ALL data set, we consider the ANOVA on the gene with the probe "38555_at" expression values on two factors. The first factor is the disease stages: B1, B2, B3 and B4 (we only take patients from those four stages). The second factor is the gender of the patient (stored in the variable ALL$sex).

(a) Conduct the appropriate ANOVA analysis. Does any of the two factors affects the gene expression values? Are there interaction between the two factors?

**Answer:**

```
ALLBs <- ALL[ALL$BT %in% c("B1","B2","B3","B4")]
y<-exprs(ALLBs)["38555_at",]
anova(lm(y~ALLBs$BT * ALL$sex))
```

**Output:**

```
  > anova(lm(y~ALLBs$BT * ALL$sex))
Analysis of Variance Table

Response: y
                  Df Sum Sq Mean Sq F value    Pr(>F)    
ALLBs$BT           9 26.060 2.89561  6.1298 6.509e-07 ***
ALL$sex            1  0.023 0.02260  0.0479    0.8273    
ALLBs$BT:ALL$sex   8  0.654 0.08170  0.1729    0.9941    
Residuals        106 50.073 0.47238                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Conclusion:**

The results show that only ALLBs$Bt/Pr(>F) value is very low, which is 6.509e-07 when compared to ALL$sex and ALLBs$BT:ALL$sex.

(b) Check the ANOVA model assumption with diagnostic tests? Are any of the assumptions violated?

**Answer:**

```
ALLBs <- ALL[ALL$BT %in% c("B1","B2","B3","B4")]
y<-exprs(ALLBs)["38555_at",]
shapiro.test(residuals(lm(y~ALLBs$BT * ALL$sex)))
bptest(lm(y~ALLBs$BT * ALL$sex), studentize = FALSE)
```

**Output:**

```
> shapiro.test(residuals(lm(y~ALLBs$BT * ALL$sex)))

Shapiro-Wilk normality test

data:  residuals(lm(y ~ ALLBs$BT * ALL$sex))
W = 0.97555, p-value = 0.02282


> bptest(lm(y~ALLBs$BT * ALL$sex), studentize = FALSE)
```

```
        Breusch-Pagan test

data:  lm(y ~ ALLBs$BT * ALL$sex)
BP = 15.091, df = 18, p-value = 0.6557
```

```
Conclusion:

> Since the p-value 0.02282 is very small, we reject the null-hypothesis of n
ormally distributed residuals. Therefore, the normality assumption does not h
old.
> From the p-value 0.6557, we don't reject the null hypothesis of equal varia
nces (homoscedasticity).
```

Please submit your R commands together with your answers to each part of the question. Relevant R outputs should be displayed to support your conclusion.