Problem 1

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?

(b) From the linear model fits, find the mean gene expression value among B3 patients.

(c) Which group's mean gene expression value is different from that of group B?

(d) Use the pairwise comparisons at FDR=0.05 to find which group means are different. What is your conclusion?

(e) 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 the question in each part directly. Relevant R outputs should be displayed to support your conclusion. Please submit your R commands separately, and label clearly which part the commands correspond to.

A)

Rscript:

```
library(ALL);data(ALL)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)["109_at",]
anova(lm(y ~ ALLB1234$BT))
```

Answer:

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
ALLB1234$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

Here p value is 0.01082 which is less than 0.05 hence we reject the null hypothesis. We can conclude that disease stages affect the mean gene expression value

B)
Rscript:
```
library(ALL);data(ALL)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)["109_at",]
summary (lm(y ~ ALLB1234$BT-1)) $coefficients
```

Answer:

	Estimate	Std. Error	t value	Pr(>|t|)
ALLB1234$BTB	6.810214	0.17384922	39.17311	2.476561e-58
ALLB1234$BTB1	6.579513	0.08918277	73.77561	2.724832e-82
ALLB1234$BTB2	6.475025	0.06478978	99.93899	5.187618e-94
ALLB1234$BTB3	**6.685333**	0.08105762	82.47631	1.382963e-86
ALLB1234$BTB4	6.914171	0.11221919	61.61309	2.192836e-75

Mean gene expression value among B3 patients is 6.68533.

C)
Rscript:
```
library(ALL);data(ALL)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)["109_at",]
pairwise.t.test(y,ALLB1234$BT)
```

Answer:
Pairwise comparisons using t tests with pooled SD

data: y and ALLB1234$BT

	B	B1	B2	B3

B1 1.00 - - -
B2 0.52 1.00 - -
B3 1.00 1.00 0.37 -
B4 1.00 0.20 0.01 0.61

P value adjustment method: holm

All the group when compared to B have p values above 0.05 which means we fail to reject null hypothesis and thus none of the group has mean gene expression value different from that of group B

D)
Rscript:

```
library(ALL);data(ALL)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)["109_at",]
pairwise.t.test(y,ALLB1234$BT,p.adjust.method='fdr')
```

Answer:
Pairwise comparisons using t tests with pooled SD

data: y and ALLB1234$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

By the above pairwise t test we can conclude that only p value of B2 and B4 is less than 0.05 so we will reject the null hypothesis in that case, which means that B2 and B4 group means are different.

Rest all the groups have same mean gene expression.

E)
Rscript:

```
install.packages("lmtest")
library(ALL);data(ALL)
library(lmtest)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)["109_at",]
shapiro.test(residuals(lm(y ~ ALLB1234$BT)))
bptest(lm(y ~ ALLB1234$BT), studentize = FALSE)
```

Answer:

```
Shapiro-Wilk normality test

data:  residuals(lm(y ~ ALLB1234$BT))
W = 0.9784, p-value = 0.1177
```

Here p value is 0.1177 which is more than 0.05 hence we fail to reject null hypothesis and can say that data is normally distributed.

```
        Breusch-Pagan test

data:  lm(y ~ ALLB1234$BT)
BP = 1.1702, df = 4, p-value = 0.883
```

Here p value is 0.883 which is more than 0.05 hence we fail to reject null hypothesis and can say that they have equal variances.

From the above two diagnostic tests we can conclude that assumptions hold true and we don't need robust tests.

Problem 2
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?

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

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

A)
Rscript:

```
library(ALL);data(ALL)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)[,]
kruskal<-apply(y,1,function(x) kruskal.test(x ~ ALLB1234$BT)$p.value)
kruskal
y.fdr<-p.adjust(kruskal,method = "fdr")
sum(y.fdr<0.05)
```

Answer:
423 genes are expressed differently.

B)
Rscript:

```
library(ALL);data(ALL)
ALLB1234 <- ALL[,ALL$BT%in%c("B","B1","B2","B3","B4")]
y <- exprs(ALLB1234)[,]
kruskal<-apply(y,1,function(x) kruskal.test(x ~ ALLB1234$BT)$p.value)
kruskal
y.fdr<-p.adjust(kruskal,method = "fdr")
sum(y.fdr<0.05)
ordered<-order(y.fdr,decreasing=F)
print(exprs(ALLB1234)[ordered[1:5],0])
```

Answer:
Top five genes with smallest p values are:
1389_at
38555_at
40268_at
1866_g_at
40155_at

Problem 3

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?

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

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.

A)

Rscript:

```
library(ALL);data(ALL)
ALLBm <- ALL[,which(ALL$BT%in%c("B1","B2","B3","B4")&
ALL$sex%in%c("F","M"))]
y <- exprs(ALLBm)["38555_at",]
Bcell<-ALLBm$BT
gender<-ALLBm$sex
anova(lm(y~Bcell*gender))
```

Answer:

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Bcell	3	24.436	8.1453	19.1179	1.818e-09	***
gender	1	0.032	0.0319	0.0748	0.7851	
Bcell:gender	3	0.230	0.0768	0.1803	0.9095	
Residuals	81	34.511	0.4261			

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

As we can see from above ANOVA analysis, p value of B stages is 1.818e-09 which is less than 0.05, Hence we reject the null hypothesis and conclude that it does affect the gene expression.

On the other hand for gender p value is 0.7851 which is more than 0.05 so we fail to reject null hypothesis and conclude that gender does not affect gene expression.

Another thing to notice here is that p value of interaction of gender and B stage is 0.9095 which is also greater than 0.05 hence we fail to reject null hypothesis and conclude that this interaction does not affect gene expression.

B)

Rscript:

```
library(ALL);data(ALL)
ALLBm <- ALL[,which(ALL$BT%in%c("B1","B2","B3","B4")&
ALL$sex%in%c("F","M"))]
y <- exprs(ALLBm)["38555_at",]
Bcell<-ALLBm$BT
gender<-ALLBm$sex
shapiro.test(residuals(lm(y ~ Bcell*gender)))
bptest(lm(y ~ Bcell*gender), studentize = FALSE)
```

Answer:

```
Shapiro-Wilk normality test

data:  residuals(lm(y ~ Bcell * gender))
W = 0.9693, p-value = 0.03291
```

As we see the p value here is 0.03291, which is less than 0.05 so we reject the null hypothesis and conclude that data is not normally distributed. So the normality assumption is violated.

Breusch-Pagan test

data: lm(y ~ Bcell * gender)
BP = 6.7635, df = 7, p-value = 0.4539

Here the p value is 0.4539 which is more than 0.05 so we fail to reject null hypothesis and conclude that data has equal variance. Thus this assumption of homoscedasticity is not violated.

Problem 4

We wish to conduct a permutation test for ANOVA on $(y_1,\ldots,y_n)$, with the group identifiers stored in the vector 'group'. We wish to use

$$\frac{1}{g-1}\sum_{j-1}^{g}(\hat{\mu}_j-\hat{\mu})^2$$

as the test statistic. Here $\hat{\mu}_j$ is the j-th group sample mean, and $\hat{\mu}=\frac{1}{g}\sum_{j-1}^{g}\hat{\mu}_j$.

(a) Program this permutation test in R.

(b) Run this permutation test on the Ets2 repressor gene 1242_at on the patients in stage B1, B2, and B3 from the ALL data set.

Submit your R script for **(a)** and the answer and R outputs for **(b).**
Hint: the sample group means can be found by R command by(y,group,mean).

A)

Rscript:

```
ANOVApermutation<- function (gene,stages){
data(ALL,package="ALL");library(ALL)
ALLB123 <-ALL[,ALL$BT%in%stages]
data<- exprs(ALLB123)[gene,]
group<-ALLB123$BT[,drop=T]
g<-length(stages)
mewj <- summary(lm(data ~ group-1))$coefficients[1:g]
mew<- (1/g)*sum(mewj)
SSM<-NULL
for (i in 1:g){
 SSM[i]<- (mewj[i]-mew)^2
}
t.obs<- (1/(g-1))*sum(SSM)
n<-length(data)
n.perm=2000
T.perm = rep(NA, n.perm)
for(i in 1:n.perm) {
 data.perm = sample(data, n, replace=F)
 mewjperm <- summary(lm(data.perm ~ group-1))$coefficients[1:g]
 mewperm<- (1/g)*sum(mewjperm)
```

```
  SSMperm<-NULL
  for(j in 1:g){
    SSMperm[j]<- (mewjperm[j]-mewperm)^2
  }
  T.perm[i]<-(1/(g-1))*sum(SSMperm)
}
mean(T.perm>=T.obs)
}
```

b)

Rscript:

```
ANOVApermutation<- function (gene,stages){
data(ALL,package="ALL");library(ALL)
ALLB123 <-ALL[,ALL$BT%in%stages]
data<- exprs(ALLB123)[gene,]
group<-ALLB123$BT[,drop=T]
g<-length(stages)
mewj <- summary(lm(data ~ group-1))$coefficients[1:g]
mew<- (1/g)*sum(mewj)
SSM<-NULL
for (i in 1:g){
  SSM[i]<- (mewj[i]-mew)^2
}
t.obs<- (1/(g-1))*sum(SSM)
n<-length(data)
n.perm=2000
T.perm = rep(NA, n.perm)
for(i in 1:n.perm) {
  data.perm = sample(data, n, replace=F)
  mewjperm <- summary(lm(data.perm ~ group-1))$coefficients[1:g]
  mewperm<- (1/g)*sum(mewjperm)
  SSMperm<-NULL
  for(j in 1:g){
    SSMperm[j]<- (mewjperm[j]-mewperm)^2
  }
  T.perm[i]<-(1/(g-1))*sum(SSMperm)
}
mean(T.perm>=T.obs)
}
```

```
gene<-"1242_at"
stages<-c("B1","B2","B3")
ANOVApermutation(gene, stages)
```

Answer:
0.5385

here p value is 0.5385 which is greater than 0.05, We fail to reject null hypothesis and thus we conclude equal distribution of expression values.