Math7340 HW8

Chengbo Gu

Problem 1 (40 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?

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

```
## Analysis of Variance Table
##
## Response: y
##                Df  Sum Sq Mean Sq F value  Pr(>F)
## ALLBgroups$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
```

Since the p-value for BT is 0.0108235 which is less than 0.05, we reject the null hypothesis and conclude that the disease stages do affect the mean gene expression value.

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

```
summary( lm( y ~ ALLBgroups$BT ) )
```

```
##
## Call:
## lm(formula = y ~ ALLBgroups$BT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09026 -0.27845  0.03999  0.26618  0.71532
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.8102     0.1738  39.173   <2e-16 ***
## ALLBgroups$BTB1  -0.2307     0.1954  -1.181   0.2408
## ALLBgroups$BTB2  -0.3352     0.1855  -1.807   0.0742 .
## ALLBgroups$BTB3  -0.1249     0.1918  -0.651   0.5167
## ALLBgroups$BTB4   0.1040     0.2069   0.502   0.6166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3887 on 90 degrees of freedom
## Multiple R-squared:  0.134,  Adjusted R-squared:  0.09556
```

```
## F-statistic: 3.483 on 4 and 90 DF,  p-value: 0.01082
```

The mean gene expression value among B3 patients is $6.8102 - 0.1249 = 6.6853$.

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

```
pairwise.t.test(y, ALLBgroups$BT)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and ALLBgroups$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 p-values of pairs related to group B is greater than 0.05. So there is no group that has different mean gene expression 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?

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

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and ALLBgroups$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
```

The p-value of group B2 and group B4 is 0.01 which is less than 0.05. So group B2 and group B4 have different group means.

(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.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm(y ~ ALLBgroups$BT))
## W = 0.97839, p-value = 0.1177
```

```
bptest( lm( y ~ ALLBgroups$BT ), studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  lm(y ~ ALLBgroups$BT)
## BP = 1.1702, df = 4, p-value = 0.883
```

The p-values of Shapiro-Wilk normality test and Breusch-Pagan test are greater than 0.05. So both the normality and the homoscedasticity assumptions are met. There is no need to apply robust ANOVA tests here.

Problem 2 (20 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 differently in some of the groups?

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

```
## [1] 423
```

There are 423 genes are expressed differently in some of the groups.

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

```
rownames(exprs(ALLBgroups))[order(p.fdr)][1:5]
```

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

The probe names for the top five genes with smallest p-values are "1389_at", "38555_at", "40268_at", "1866_g_at", "40155_at".

Problem 3 (20 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?

```r
rm(list=ls())
data(ALL)
ALLBs <- ALL[,which(ALL$BT %in% c("B1","B2","B3","B4"), ALL$sex %in% c("F", "M"))]
y <- exprs(ALLBs)["38555_at",]
Bcell <- ALLBs$BT
sex <- ALLBs$sex
anova( lm( y ~ Bcell*sex ) )
```

```
## 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 ***
## sex        1  0.032  0.0319  0.0748    0.7851
## Bcell:sex  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
```

The p-value of Bcell is 1.818e-09. So disease stage (Bcell/BT) affects the gene expression values.

Since the p-value of Bcell*sex is 0.9095 which is greater than 0.05, we accept the null hypothesis. That is, there is no interaction between the two factors.

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

```r
shapiro.test( residuals( lm( y ~ Bcell*sex )))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm(y ~ Bcell * sex))
## W = 0.96926, p-value = 0.03291
```

```r
bptest(lm(y ~ Bcell*sex), studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  lm(y ~ Bcell * sex)
## BP = 6.7635, df = 7, p-value = 0.4539
```

The normality assumption is violated because the p-value of Shapiro-Wilk normality test is 0.03291 which is less than 0.05.

Problem 4 (20 points)

We wish to conduct a permutation test for ANOVA on $(y_1, ..., y_N)$, with the group indentifiers 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.

```
rm(list=ls())

permutation.test <- function(probe, groups) {
  data(ALL)
  ALLB123 <- ALL[,ALL$BT %in% groups]
  data<- exprs(ALLB123)[probe,]
  group<-ALLB123$BT[,drop=T]
  n.group <- length(groups)
  n <- length(data)

  estimatedMean <- sum( by(data, group, mean) ) / n.group
  T.obs <- sum( (by(data, group, mean) - estimatedMean) ^ 2) / (n.group - 1)

  n.perm <- 2000
  T.perm <- rep(NA, n.perm)
  for(i in 1:n.perm) {
    data.perm <- sample(data, n, replace=F) #permute data
    estimatedMean <- sum( by(data.perm, group, mean) ) / n.group
    T.perm[i] <- sum( (by(data.perm, group, mean) - estimatedMean) ^ 2) /
                 (n.group - 1) #Permuted statistic
  }
  mean(T.perm >= T.obs) #p-value

}
```

(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.

```
p.value <- permutation.test("1242_at", c("B1", "B2", "B3"))
p.value
```

```
## [1] 0.525
```

The p-value here is 0.525 which is greater than 0.05. Thus, we accept the null hypothesis of ANOVA test that the group means are equal (probe: "1242_at", group: B1, B2, B3).