Northeastern University
College of Science

Module 8 - Homework

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?

(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.

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?

(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.

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?

(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.

Problem 4 (20 points)

We wish to conduct a permutation test for ANOVA on $(y_1,...,y_N)$, 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),** submit your conclusion and R outputs for **(b).**

Hint: the sample group means can be found by R command by(y, group, mean).