Charles Valentine
Homework 8
10/30/2016

Note:
All code is in script hw08.R --- I have made use of cat and print statements to display information easily!

Problem 1

The following statements refer to gene expression data for probe “109_at” in B-cell patients in 5 groups [B, B1, B2, B3, B4] from the ALL dataset.

a. According to a one-way ANOVA the disease stages do affect gene expression values (p-value = 0.0108) and we should reject the null hypothesis that the disease stages do not affect gene expression values.

OUTPUT

```
Probe `109_at` in groups B, B1, B2, B3, and B4:
Pr(>F) = 0.0108 via ANOVA
```

b. The estimated gene expression mean for B3 patients is -0.1249 according to a linear model fit.

OUTPUT

```
B3 Estimate Mean: -0.1249
```

c. Only one group’s gene expression values are different from that of group B. That group is B2. This difference, however, is not statistically significant to an alpha level of 0.05 (p-value = 0.5191

OUTPUT

```
B2's mean gene expression is most different to
B's mean gene expression ( p-value: 0.5191 )
```

d. A pairwise t-test with FDR correction and an alpha level of 0.05 was used and only one pair of groups were significantly different. Those groups were B2 and B4 (corrected p-value = 0.0104).

OUTPUT

```
Pairwise t-test with FDR correction of B, B1, B2,
B3, and B4 mean expression values:

        B     B1     B2     B3
B1 0.4014     NA     NA     NA
B2 0.1854 0.4778     NA     NA
B3 0.5741 0.4778 0.1522     NA
B4 0.6166 0.1089 0.0104 0.2036

There are 1 p-values < 0.05
p-values:
 0.0104
```

e. Two diagnostic tests were used to check the assumptions of using a one-way ANOVA. A test for normality was taken using the Shapiro-Wilks method and a test for homoscedasticity was taken using the Breusch and Pagan test. The tests indicate the residuals of the linear fit follow a normal distribution and have equivalent variances (homoscedasticity) (p-value = 0.1177 and p-value = 0.8830 respectively).

OUTPUT

```
Ho: The residuals of the linear fit follow a normal dist. ( p-value: 0.1177 )

Ho: The residuals of the linear fit are homoscedastic ( p-value: 0.883 )
```

Problem 2

a. After applying the Kruskal-Wallis nonparametric test for every gene on the B-cell ALL patients in stage B, B1, B2, B3, and B4 from the ALL dataset (with an FDR correction and alpha level of 0.05) it was found that 423 gene expression values are different in the different stages of B-cell patients.

OUTPUT

```
There are 423 p-values less than 0.05
```

b. The top five probe names with the smallest p-values are (descending order):

OUTPUT

```
 1. 1389_at
 2. 38555_at
 3. 40268_at
 4. 1866_g_at
 5. 40155_at
```

Problem 3

a. A two-way ANOVA test was used on the ALL data set for probe “3855_at” with the factors disease stage [B1, B2, B3, and B4] and sex [male and female] of patients. The test was used with an interaction term in the linear model. The results show that there is no interaction between the two factors (p-value = 0.9095).

OUTPUT

```
                                   Df Sum Sq Mean Sq F value    Pr(>F)    
ALL.selection$BT                    3 24.436  8.1453 19.1179 1.818e-09 ***
ALL.selection$sex                   1  0.032  0.0319  0.0748    0.7851    
ALL.selection$BT:ALL.selection$sex  3  0.230  0.0768  0.1803    0.9095    
Residuals                          81 34.511  0.4261                      
```

b. Two diagnostic tests were used to check the assumptions of using a two-way ANOVA. A test for normality was taken using the Shapiro-Wilks method and a test for homoscedasticity was taken using the Breusch and Pagan test. The tests indicate the residuals of the linear fit do not follow a normal distribution but do have equivalent variances (homoscedasticity) (p-value = 0.0329 and p-value = 0.4539 respectively).

OUTPUT

```
Ho: The residuals of the linear fit follow a normal dist. ( p-value: 0.0329 )

Ho: The residuals of the linear fit are homoscedastic ( p-value: 0.4539 )
```

Problem 4

a. A permutation test was used to assess the statistic $\frac{1}{g-1}\sum_{j=1}^{g}\left(\hat{u}_j - \hat{u}\right)^2$ where $\hat{u}_j$ is the jth group sample mean and $\hat{u} = \frac{1}{g}\sum_{j=1}^{g}\hat{u}_j$ on the Ets2 repressor gene “1242_at” on the patients in stage B1, B2, and B3 from the ALL dataset. We are testing that the data in all groups come from a common distribution so permutations should not change our distribution under this null hypothesis. We assess how many statistics of permuted data are less than or equal to our statistic on the actual data. With a p-value = 0.9805 we choose to accept our null hypothesis that this statistic does not explain any differences between disease groups for this gene expression.

OUTPUT

```
P(F>=F_obs) = p-value = 0.9805
```