Charles Valentine
Final
12/13/2016

Problem 1

a) See code for implementation.

OUTPUT

```
Sample size: 120
Sample mean: 1.817
```

b) The maximum likelihood value for the R routine `optimize` is found:

OUTPUT

```
MLE of θ̂: 1.817
```

c) First the bootstrap 95% confidence interval for the maximum likelihood value for $\hat{\theta}$ was found using 1000 iterations. This interval does not include 1 so it is unlikely $\theta = 1$. To test this we use a t-test using our bootstrapped samples of $\hat{\theta}$ which, according to the central limit theorem, should follow a normal distribution. A two-sided t-test to was used to test the hypothesis that $\theta = 1$. A p-value was found to be ~ 0 which is far less than our testing level of 0.05. We reject the hypothesis that $\theta = 1$ and accept the alternate hypothesis that $\theta \neq 1$.

OUTPUT

```
95 percent confidence interval of θ̂ from bootstrap:
 ( 1.292 , 2.517 )

H₀: θ̂ = 1
p-value: 0
```

Problem 2

a) See code for implementation. Effectively:

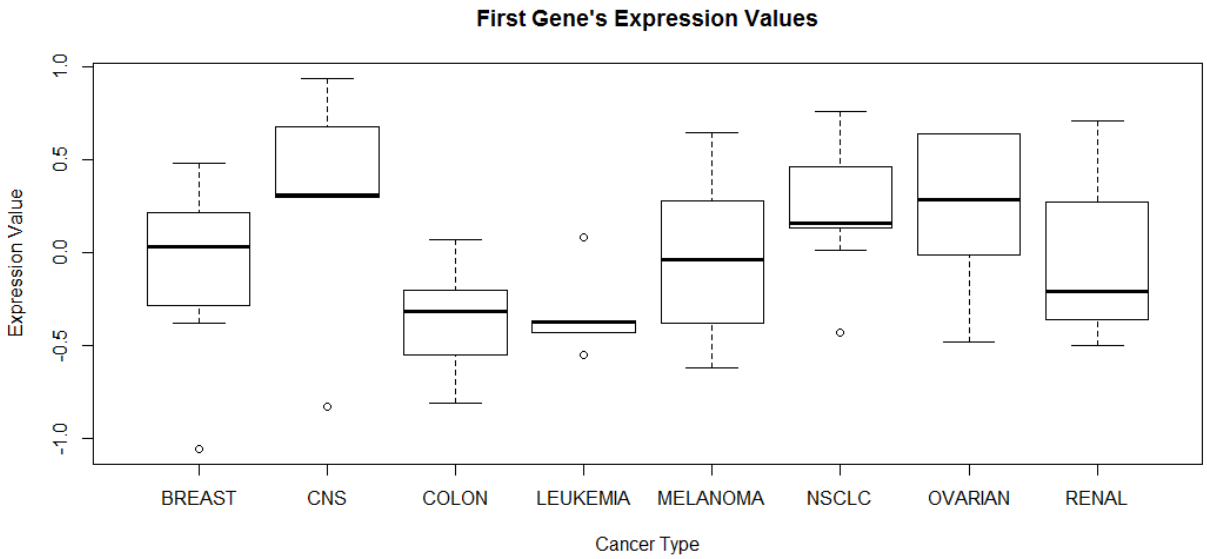
```
library(ISLR)

ncidata = NCI60$data
ncilabs = NCI60$labs

counts = table(ncilabs)
ix = ncilabs %in% names(counts[counts > 3])

data = ncidata[ix,]
labs = factor(ncilabs[ix])
```

b) An ANOVA was used on the gene in the first column of the selected `NCI60$data` data to assess whether this gene expresses differently in different types of cancer. A p-value of 0.039 was calculated when testing the null hypothesis that the gene does not express differently in different types of cancer. This p-value is below 0.05 so we choose to accept the alternate hypothesis that this gene does express differently in different types of cancer. A pairwise t-test with FDR adjustment was performed for all selected cancer types. Pairwise comparisons are shown below. Any cancer pair with a p-value of 0.1 expresses this gene differently

OUTPUT

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
labs       7   2.89   0.413    2.33  0.039 *
Residuals 49   8.70   0.178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Pairwise comparisons using t tests with pooled SD

data:  y and labs

          BREAST CNS COLON LEUKEMIA MELANOMA NSCLC OVARIAN
CNS       0.3    -   -     -        -        -     -
COLON     0.3    0.1 -     -        -        -     -
LEUKEMIA  0.4    0.1 0.9   -        -        -     -
MELANOMA  0.9    0.3 0.3   0.3      -        -     -
NSCLC     0.3    0.9 0.1   0.1      0.3      -     -
OVARIAN   0.3    0.9 0.1   0.1      0.4      1.0   -
RENAL     0.9    0.3 0.3   0.3      0.9      0.3   0.3

P value adjustment method: fdr
```



c) The model assumptions for ANOVA are normality and homoscedasticity (equal variances). We can test for normality using the Shapiro-Wilks normality test on the residuals of the linear fit used in the ANOVA. We can test equal variances by using the Breusch and Pagan test also on the residuals of the linear fit. Since both tests give p-values greater than 0.05 we accept that our data is normal and homoscedastic and that ANOVA analysis is appropriate.

OUTPUT

```
        Shapiro-Wilk normality test

data:  residuals(lm(y ~ labs))
W = 0.98, p-value = 0.4


        Breusch-Pagan test

data:  lm(y ~ labs)
BP = 8.8, df = 7, p-value = 0.3
```

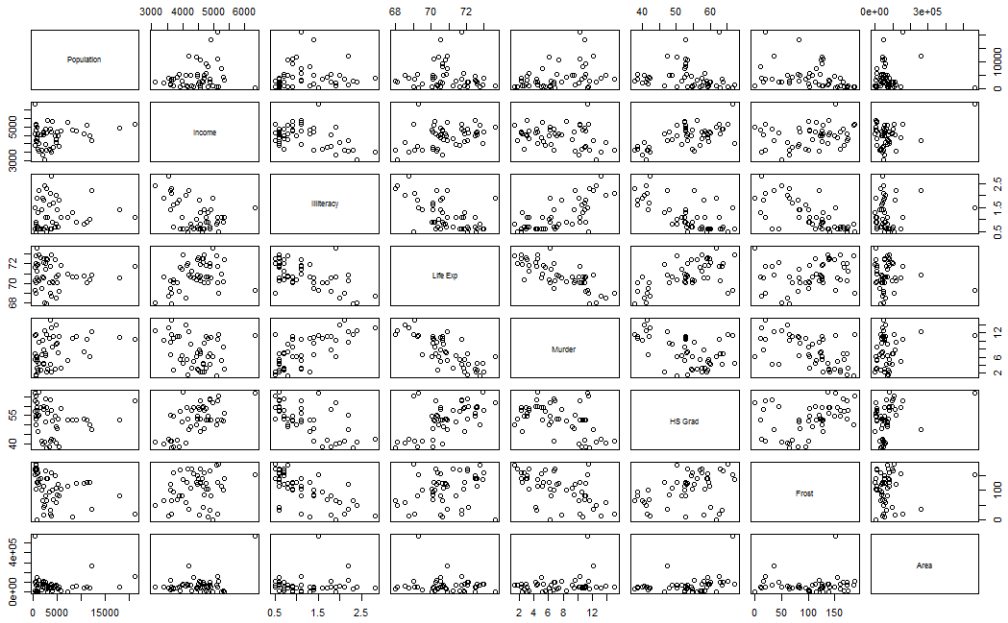
d) See code for implementation.

OUTPUT

```
The number of genes that express differently among different cancer types is: 2808
```

Problem 3

a) A scatterplot of all pairwise comparisons between population, income, illiteracy, life expectancy, murder, high school graduation, frost, and area variables is given. Income, murder, and high school graduation all appear to correlate linearly with life expectancy.



b) The regression equation `LifeExp ~ Income + Illiteracy + Frost` was used to regress life expectancy on income, illiteracy, and frost days. Illiteracy is the only variable that affect life expectancy in a significant way (p-value 0.00013).

OUTPUT

```
Call:
lm(formula = LifeExp ~ Income + Illiteracy + Frost, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5214 -0.7359 -0.0398  0.8048  3.1364

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.526008   1.658023   43.74  < 2e-16 ***
Income       0.000182   0.000282    0.64  0.52286
Illiteracy  -1.560554   0.374515   -4.17  0.00013 ***
Frost       -0.006015   0.004055   -1.48  0.14483
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.09 on 46 degrees of freedom
Multiple R-squared:  0.384,       Adjusted R-squared:  0.344
F-statistic: 9.57 on 3 and 46 DF,  p-value: 5.04e-05
```

c) The mean square errors of the delete-one-cross-validated linear regression model are outputted below. It was not clear in the assignment if this was required. If the mean was requested that is also provided.

OUTPUT

```
[1.097, 1.068, 1.109, 1.092, 1.109, 1.098, 1.057, 1.073, 1.100, 1.076, 0.876,
 1.108, 1.081, 1.101, 1.095, 1.095, 1.110, 1.110, 1.097, 1.075, 1.097, 1.102,
 1.063, 1.088, 1.099, 1.093, 1.087, 0.967, 1.109, 1.109, 1.069, 1.109, 1.098,
 1.044, 1.103, 1.106, 1.109, 1.102, 1.071, 1.074, 1.101, 1.109, 1.074, 1.066,
 1.106, 1.101, 1.101, 1.091, 1.082, 1.083]

The average MSE is 1.085
```

Problem 4

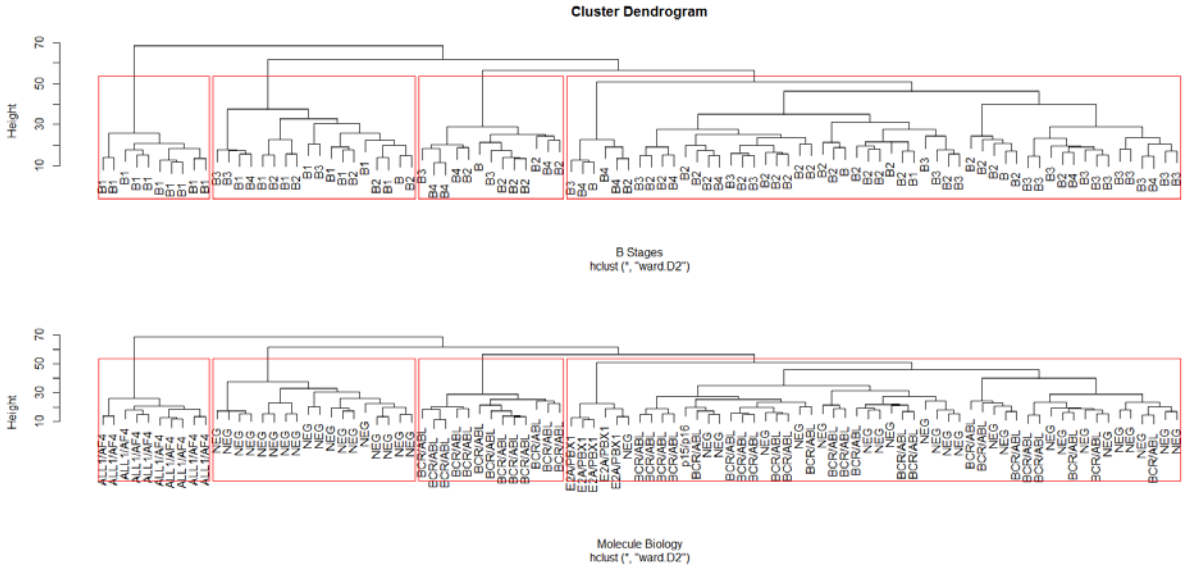
a) See code for implementation.
b) The following genes were analyzed for only B-cell patients.

OUTPUT

```
There are 184 genes with a coefficient of variance greater than 0.2.
```

c) A second filter may be to reject any gene expression data that does not fit a normal distribution. One could assume that genes that have different expression levels *per* B-cell class should not necessarily follow a normal distribution.
d) Using `ward.D2` linkage and Euclidean distance we conduct a hierarchical clustering analysis with the filtered genes from Problem 4b. Below are two dendrograms with different labels, the top figure for B-cell class and the bottom figure for molecular biology type. When the tree is cut in four places it appears as if the algorithm predicts molecular biology type better. The confusion matrix for the comparisons are provided and it is clear that molecular biology type is better classified.

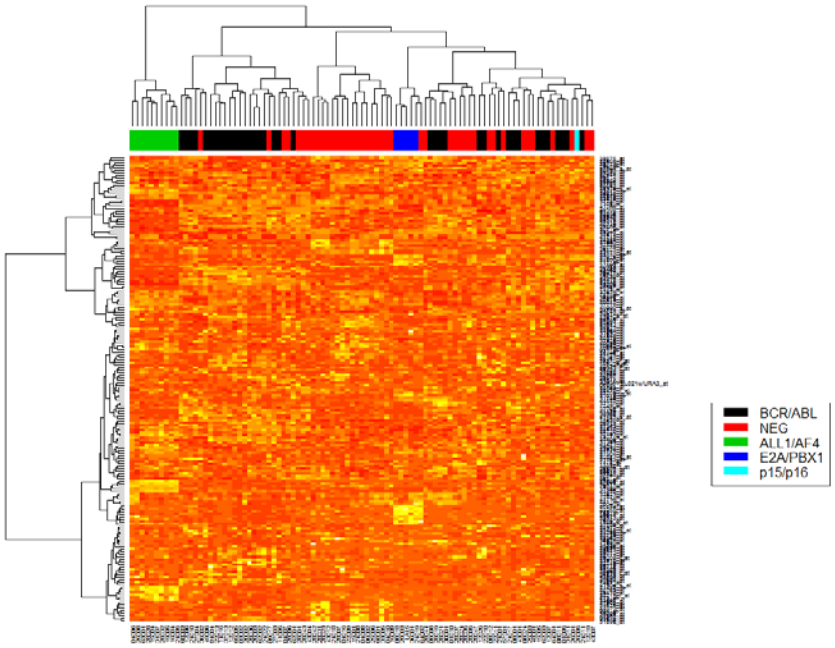
OUTPUT

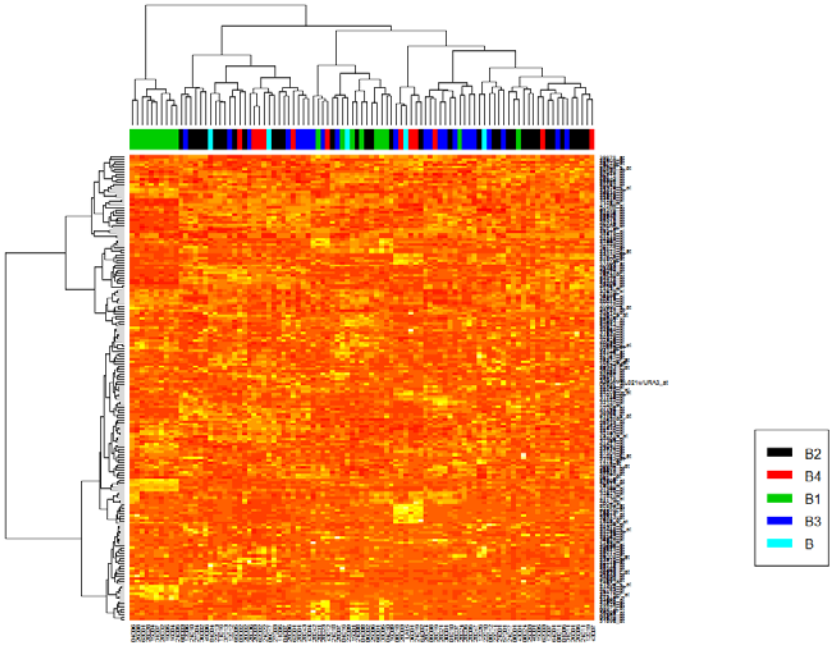



```
          pred.groups
names       1  2  3  4
   B        3  1  0  1
   B1       1  0 10  8
   B2      25  6  0  5
   B3      18  2  0  3
   B4       7  4  0  1

          pred.groups
mol.biol    1  2  3  4
  ALL1/AF4  0  0 10  0
  BCR/ABL  24 13  0  0
  E2A/PBX1  5  0  0  0
  NEG      24  0  0 18
  p15/p16   1  0  0  0
```

e) Two heatmaps were drawn of the expression data (hierarchically ordered) with a color bar for molecular biology type (top) and B-cell type (bottom). It is again our conclusion that the clustering algorithm classifies better based on molecular biology types.

OUTPUT





f) Using the `limma` library we fit a linear model to select genes that express differently among three specified classes.

OUTPUT

```
There are 1169 genes at FDR p-value < 0.05 that are significantly different among the
classes B1, B2, and B34
```

g) SVM and classification tree were run on the genes that were selected in Problem 4f and the classifiers were evaluated with a delete-one-cross-validated misclassification rate.

OUTPUT

```
The average MCR of a n-fold cross-validation for SVM classification is: 0.2

The average MCR of a n-fold cross-validation for tree classification is: 0.3333
```

h) The intersection of the genes selected for in Problem 4b and 4f were selected. This was done as: `intersect(partb.genes, partf.genes)`. SVM and classification tree were run on these genes and the classifiers were evaluated with a delete-one-cross-validated misclassification rate.

OUTPUT

```
There are 55 genes that pass the filters in 4b and 4f

The average MCR of a n-fold cross-validation for SVM classification is: 0.2444

The average MCR of a n-fold cross-validation for tree classification is: 0.1889
```

i) Having only evaluated the two classifiers in two instances it is tough to draw a comparison of their shortcomings. It is understandable that the tree classification performs well when there are fewer data points as a classification tree finds simple rules to fit a dataset and all combinations of simple rules may not be tractable to find when the measured variables increases dramatically. Therefore, a classification tree may have a higher misclassification rate when there are many measured variables. SVM appears to perform well in this scenario and has a lower maximum MCR than the maximum MCR for tree classification.