HW1

Problem 1: Explain whether each scenario is a classification or regression prob- lem, and indicate whether we are most interested in inference or pre- diction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Regression, inference, n = 500, p = 4

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each prod- uct we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Classification Prediction n = 20, p = 14,

(c) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % hange in the German market.

Problem 2

Complete Exercise 3 from section 2.4 of the textbook (p. 52).

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training er- ror, test error, and Bayes (or irreducible) error curves, on a sin- gle plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should representthe amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

(b) Explain why each of the five curves has the shape displayed in part (a).

i. bias is inversely related to flexibility as higher flexibility creates a closer fit

ii. variance - increases monotonically because increases in flexibility yield overfitting

iii. training error - decreases monotonically because increases in flexibility yield a closer fit

iv. test error - concave up curve because increase in flexibility yields a closer fit before it overfits

v. Bayes (irreducible) error - defines the lower limit, the test error is bounded below by the irreducible error due to variance in the error (epsilon) in the output values (0 <= value). When the training error is lower than the irreducible error, overfitting has taken place. The Bayes error rate is defined for classification problems and is determined by the ratio of data points which lie at the 'wrong' side of the decision boundary, (0 <= value < 1).

Problem 3

Complete Exercise 7 from section 2.4 of the textbook (p. 53).

```
library(fields)
```

```
## Warning: package 'fields' was built under R version 3.2.5
## Loading required package: spam
## Warning: package 'spam' was built under R version 3.2.5
## Loading required package: grid
```

```
## Spam version 1.4-0 (2016-08-29) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'
```

```
## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

```
## Loading required package: maps
```

```
## Warning: package 'maps' was built under R version 3.2.5
```

```
library(SpatialTools)
```

```
## # This research was partially supported under NSF Grant ATM-0534173
```

```
Obs <- seq(1,6,length=6)
X1 <- c(0,2,0,0,-1,1)
X2 <- c(3,0,1,1,0,1)
X3 <- c(0,0,3,2,1,1)
Y <- c('Red', 'Red', 'Red', 'Green', 'Green', 'Red')
df <- data.frame(Obs, X1, X2, X3, Y)
(df)
```

```
##   Obs X1 X2 X3     Y
## 1   1  0  3  0   Red
## 2   2  2  0  0   Red
## 3   3  0  1  3   Red
## 4   4  0  1  2 Green
## 5   5 -1  0  1 Green
## 6   6  1  1  1   Red
```

```
coords_mat <- data.matrix(df[2:4])
coords_mat
```

```
##      X1 X2 X3
## [1,]  0  3  0
## [2,]  2  0  0
## [3,]  0  1  3
## [4,]  0  1  2
## [5,] -1  0  1
## [6,]  1  1  1
```

a)

```
dist <- NULL
for(i in 1:nrow(coords_mat)) {
    dist[i] <- dist(rbind(coords_mat[i,], c(0,0,0)))
}
df$dist <- dist

df
```

```
##   Obs X1 X2 X3     Y     dist
## 1   1  0  3  0   Red 3.000000
```

```
## 2   2  2  0  0   Red 2.000000
## 3   3  0  1  3   Red 3.162278
## 4   4  0  1  2 Green 2.236068
## 5   5 -1  0  1 Green 1.414214
## 6   6  1  1  1   Red 1.732051
```

b) If the only datapoint we care about is the one nearest neighbor, then the prediction will be Green (Obs 5)

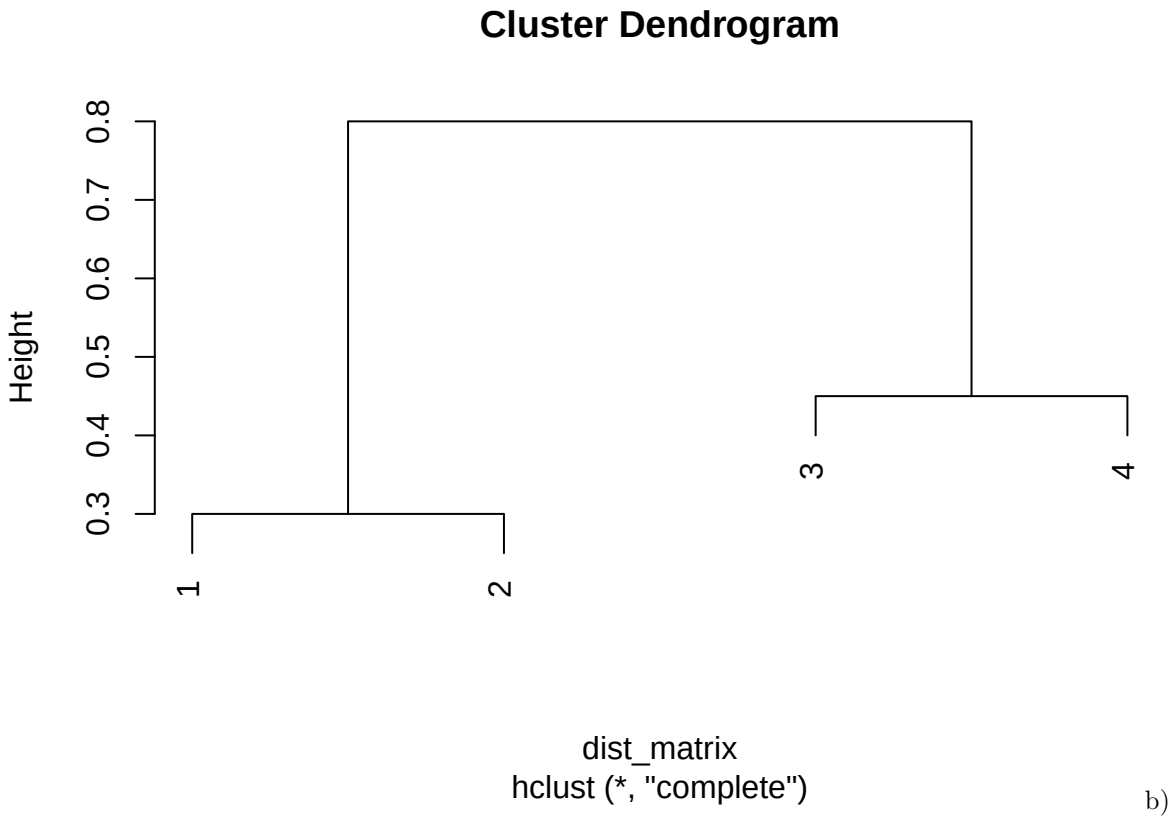
c) Obs 2, 5, 6 will be the closest 3 neighbors for [0,0,0], which corresponds with a Y of Red, Green and Red respectively; thus the prediction would be red.

d)

Problem 4: Exercise 1 (p. 413) a) K-Means Clustering: Prove equation 10.12 b)

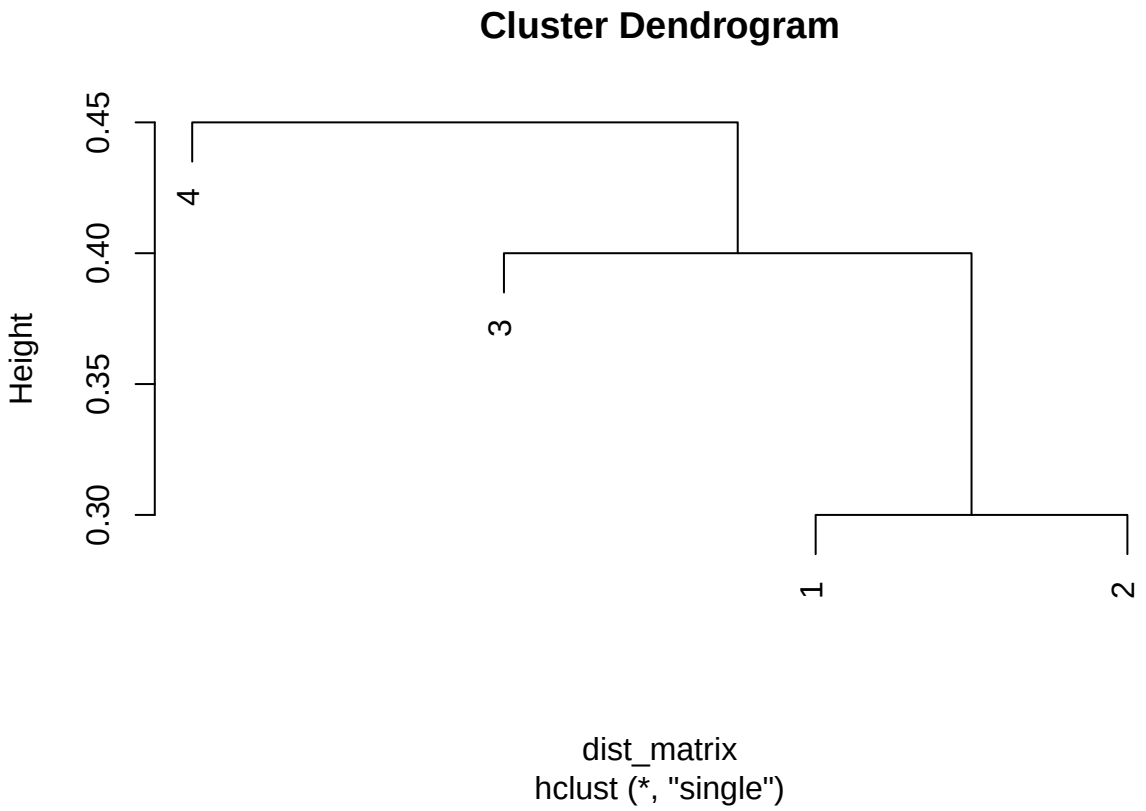
Problem 5: Exercise 2 (p. 413) For given Dissimilarity matrix, a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observa- tions using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

```
dist_matrix <- as.dist(matrix(c(0, 0.3, 0.4, 0.7,
                      0.3, 0, 0.5, 0.8,
                      0.4, 0.5, 0.0, 0.45,
                      0.7, 0.8, 0.45, 0.0), nrow=4))
plot(hclust(dist_matrix, method="complete"))
```



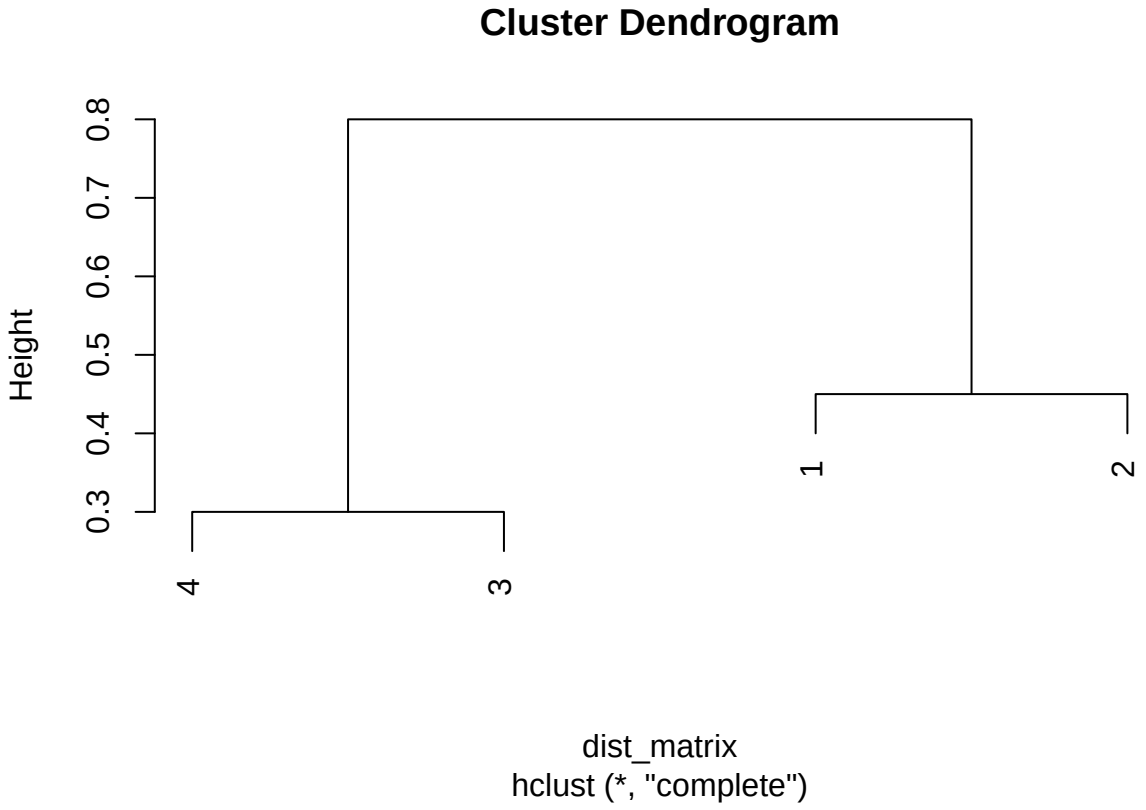
b)

```
plot(hclust(dist_matrix, method="single"))
```



c) Cluster1: Observations 1 and 2; Cluster2: Observations 3 and 4
d) Cluster1: Observations 1,2, and 3; Cluster2: Observation 4
e) The following dendrogram swaps positions of the two clusters without changing the meaning

```
plot(hclust(dist_matrix, method="complete"), labels=c(4,3,1,2))
```



Problem 6: Exercise 4 (p. 414) Suppose that for a particular data set, we perform hierarchical clus- tering using single linkage and using complete linkage. We obtain two dendrograms. (a) At a certain point on the single linkage dendrogram, the clus- ters {1,2,3} and {4,5} fuse. On the complete linkage dendro- gram, the clusters {1, 2, 3} and {4, 5} also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

This question requires more information to answer and is dependent on both the organization of information as well as the dissimilarity measure (euclidian distance, correlation etc). Complete linkage joins on maximal intercluster dissimilarity, while single linknage joins on minimal intercluster dissimilarity; were these to be equal, then the two clusters in question would fuse at the same height. Otherwise, a dendrogram formed with complete linkage would fuse them at a greater height than a dendrogram formed with single linkage.

(b) At a certain point on the single linkage dendrogram, the clusters {5} and {6} fuse. On the complete linkage dendrogram, the clus- ters {5} and {6} also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell? They would fuse at the same height because the choice of complete vs single linkage operates on observations between two clusters instead of the clustering of two individual observations that are not yet clustered.

Problem 7: Exercise 9 (p. 416)

```
data("USArrests")
names(USArrests)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

```
dim(USArrests)
```

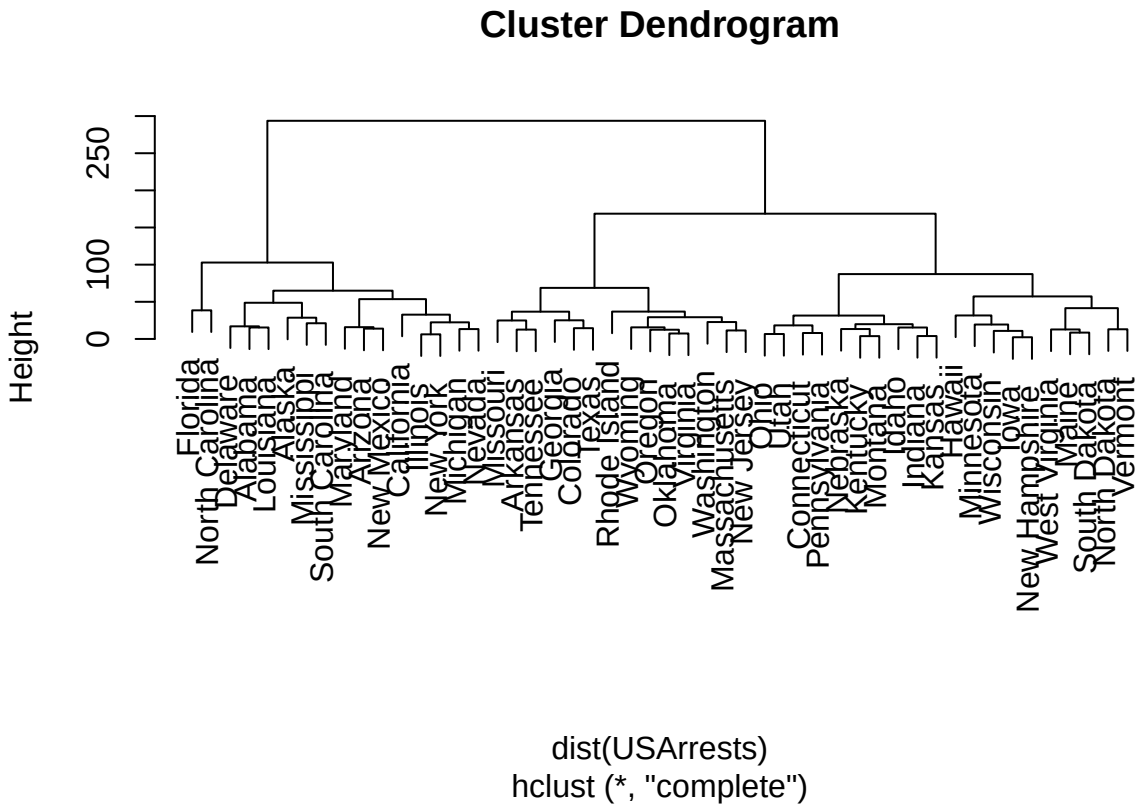
```
## [1] 50  4
```

```
class(USArrests)
```

```
## [1] "data.frame"
```

a)

```
cluster_USArrests <- hclust(dist(USArrests), method="complete")
plot(cluster_USArrests)
```



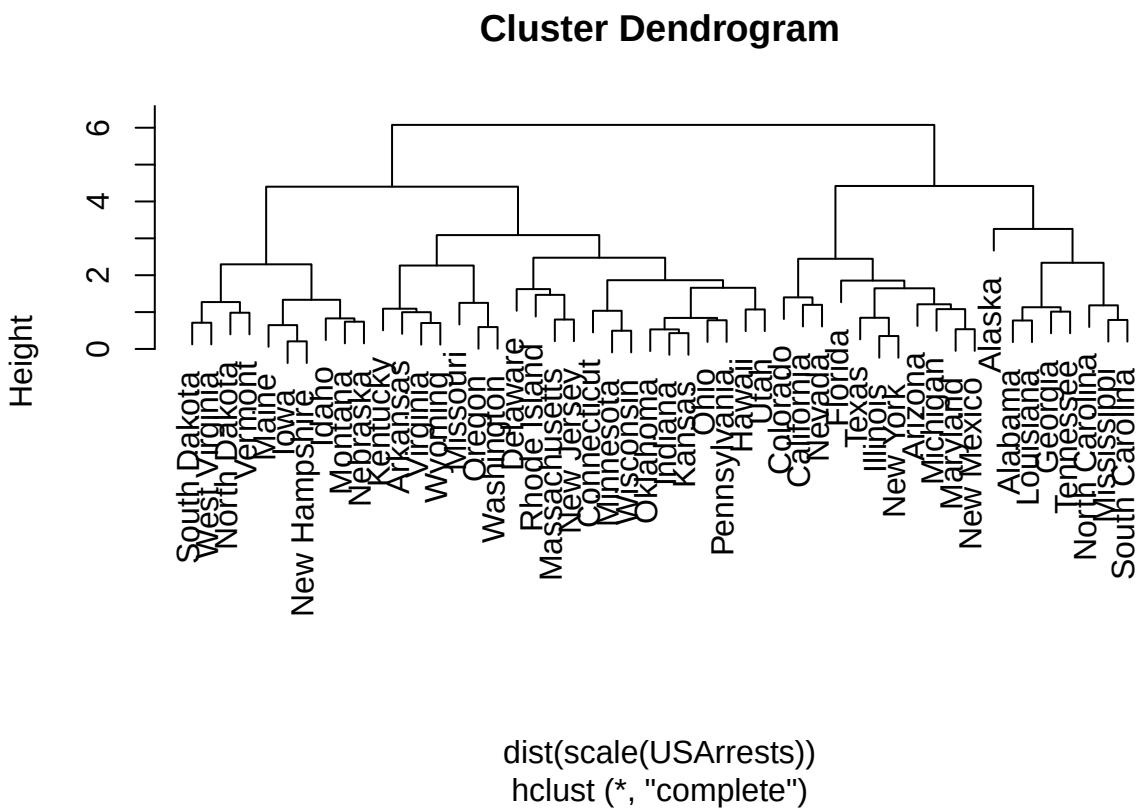
b)

```
cutree(cluster_USArrests, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

c)

```
cluster_USArrests_scaled = hclust(dist(scale(USArrests)), method="complete")
plot(cluster_USArrests_scaled)
```



d) Number of states in each cluster without scaling USArrests:

```
table(cutree(cluster_USArrests, 3))
```

```
##
##  1  2  3
## 16 14 20
```

Number of states in each cluster after scaling USArrests:

```
table(cutree(cluster_USArrests_scaled, 3))
```

```
##
##  1  2  3
##  8 11 31
```

Dendrogram with scaled USArrests:

Scaling each variable vector to standardize variance makes sense. The variables in USArrests dataset have different units with different inherent variance. Units with a larger variance has a greater effect on euclidian distance, and thus have a greater influence on how clusters are formed.

Problem 8: Exercise 4 (p. 120)

4. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. Y = 0 + 1X + 2X2 + 3X3 + .

(a) Suppose that the true relationship between X and Y is linear, i.e. Y = 0 + 1X + . Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic

regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer. Adding more variables to the least squares equations always improves the fit to the training data; thus, the RSS to training data should decrease

(b) Answer (a) using test rather than training RSS. test RSS should decrease due to the overfitting and failing to generalize overfit model to test dataset

(c) Suppose that the truerelationshipbetweenXandYisnotlinear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer. The increased flexibility from polynomial regression will lead to a better fit to training data over a linear regression.

(d) Answer (c) using test rather than training RSS. Since the true relationship is not known, there is not enough information to exactly tell whether test dataset RSS will be better with a polynomial fit;

Problem 9: Exercise 9 (p. 122). In parts (e) and (f), you need only try a few interactions and transformations.

Problem 10: Exercise 14 (p. 125)