1. (Ex. 2 Pg. 52)

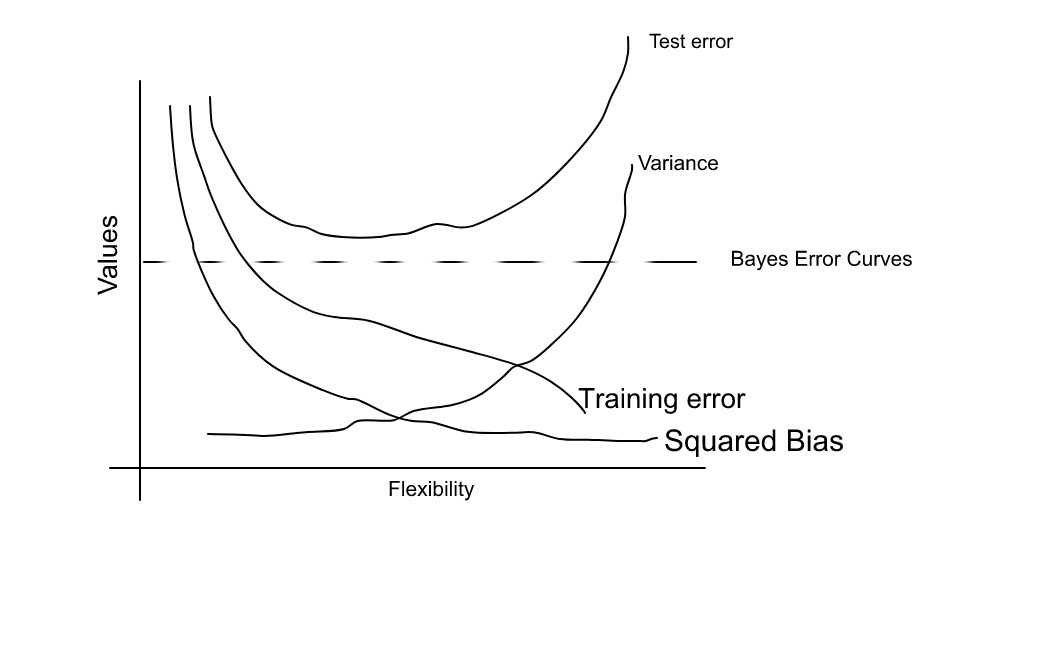
a) Regression and inference, n = 500, p = 3

b) Classification and prediction, n = 20, p = 13

c) Regression and prediction, n = 52, p = 3

2. (Ex. 3 Pg. 52)

a)



b) Training error – As flexibility increase, the curve fits more closely to data

Test error – The U shape is due to small training error will yield large test error, which results overfitting data

Squared bias – Variance increases and bias decreases, which results a decreasing curve

Variance – As flexibility increases, curve is away from the data points

Bayes Error Curves – Test is always above irreducible error

3. (Ex 7 Pg. 53)

a) By distance formula,

|  |  |
| --- | --- |
| Obs. | Distance |
| 1 | 3 |
| 2 | 2 |
| 3 | 3.162 |
| 4 | 2.236 |
| 5 | 1.732 |
| 6 | 1.414 |

b) For K = 1, closest observation to test point is 5. Probability that Y is Red given X = x0, but X can only be x5, so probability that Y is Red is 0. Similar reasoning for green, therefore probability is 1.

c) For K = 3, the closest 3 observations are 2, 5, 6. Using the formula from textbook 2.12, we can calculate P(Y = Red | X = x0) =

P(Y = Green | X = x0) =

Therefore, our prediction is Red.

d) As K becomes larger, the boundary becomes linear. Then the number of KNN is small.

4. (Ex 1 Pg 413)

a)

When each observation is assigned to closest centroid, the above is

Given , then the simplified form is

Now the right side of the equation

Given , then the simplified form is

Since index i is not in the second term, you can remove the summation and simplify

b) According to the algorithm, at each iteration when we assign an observation to the closest centroid. As we approach p, we reach a local optimum, where the centroid of i and i’ are closest for kth cluster. This is also considered as minimizing the summation of squared Euclidean distance between the two observations i, i’.

5. (Ex 2 Pg 413)

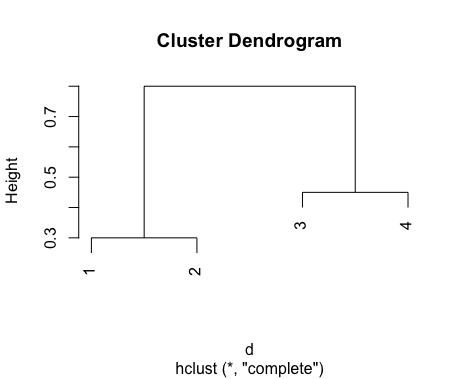
a) Given 4 observations with 6 pairwise dissimilarities, for i = 4, the most similar pair of clusters are observations 1 and 2. When we fuse the observations, we get cluster (1,2) with height 0.3. Then the dissimilarity matrix becomes

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1,2) | 3 | 4 |
| (1,2) |  | 0.5 | 0.8 |
| 3 | 0.5 |  | 0.45 |
| 4 | 0.8 | 0.45 |  |

where matrix index i = (1,2) and j = 3 is the maximum inter cluster dissimilarity(complete linkdage), therefore the max of distance between observation 1 and 3 and max of distance between observation of 2 and 3.

For i = 3, 0.45 is most similar which is observation 3 and 4. When we fuse the observations we get cluster (3,4) with height 0.45 with dissimilarity matrix

|  |  |  |
| --- | --- | --- |
|  | (1,2) | (3,4) |
| (1,2) |  | 0.8 |
| (3,4) | 0.8 |  |

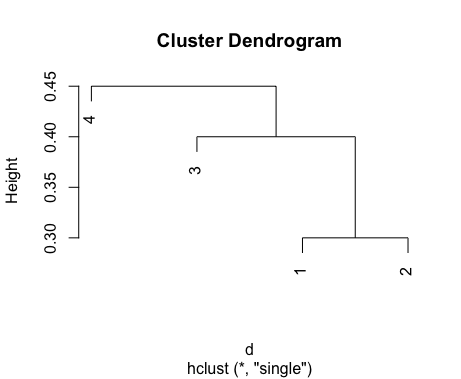


b) For i = 4, most similar pair is observation 1 and 2. Fuse these together for a new dissimilarity matrix. However, this time we will take the minimal inter cluster dissimilarity

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1,2) | 3 | 4 |
| (1,2) |  | 0.4 | 0.7 |
| 3 | 0.4 |  | 0.45 |
| 4 | 0.7 | 0.45 |  |

For i = 3, most similar pair is observation (1,2) and 3. The new dissimilarity matrix is

|  |  |  |
| --- | --- | --- |
|  | (1,2,3) | 4 |
| (1,2,3) |  | 0.45 |
| 4 | 0.45 |  |



c) (1,2) and (3,4) would be the two separate clusters

d) (4) and ((1,2),3) would be the two separate clusters

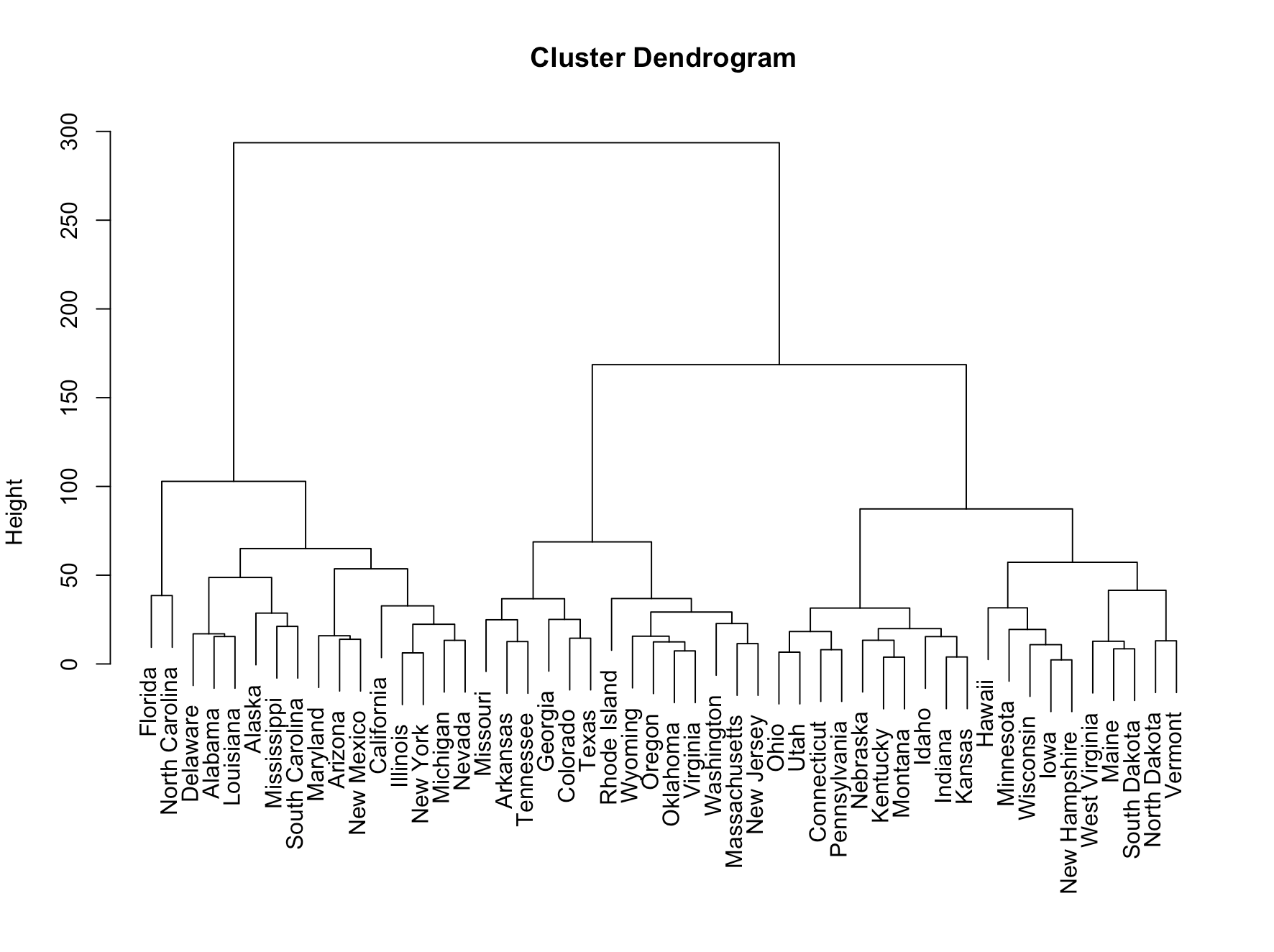
6. (Ex 4 Pg 414)

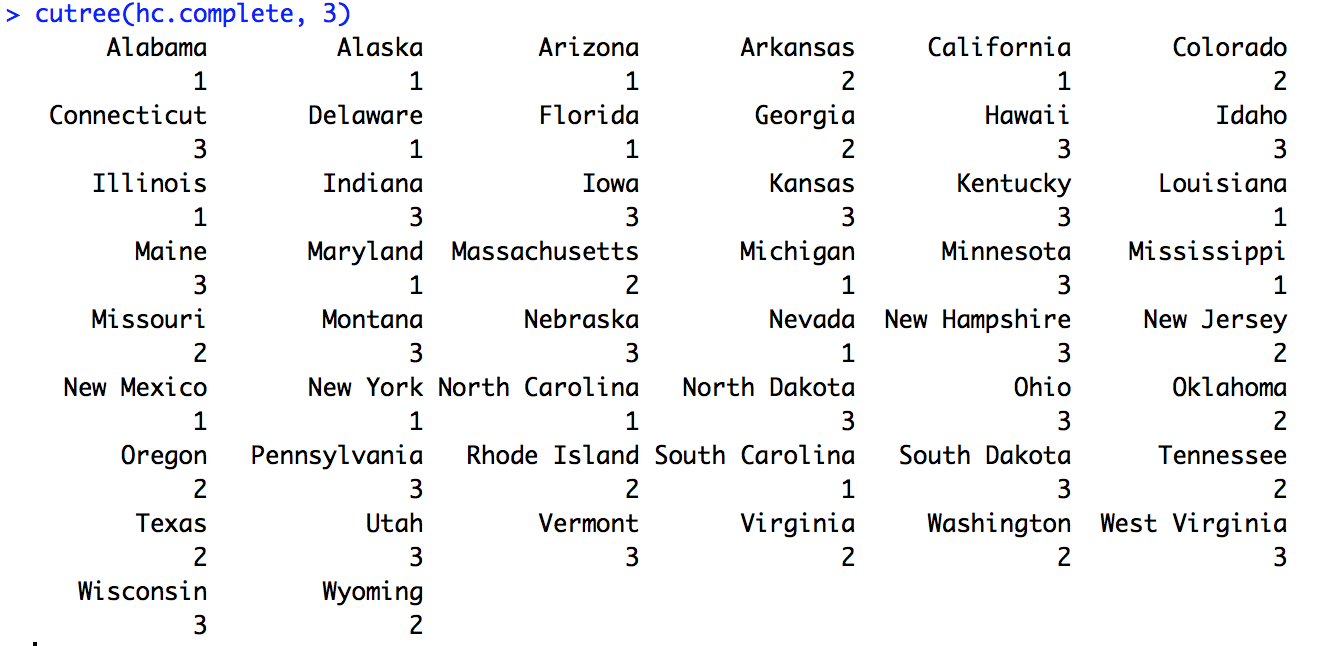
a) There is not enough information, because you do not know the distance between each observation. Therefore you can come up with your own counterexample for each case. An example is if all observations differ by 2 with each other. Then the complete and single linkage would be 2.

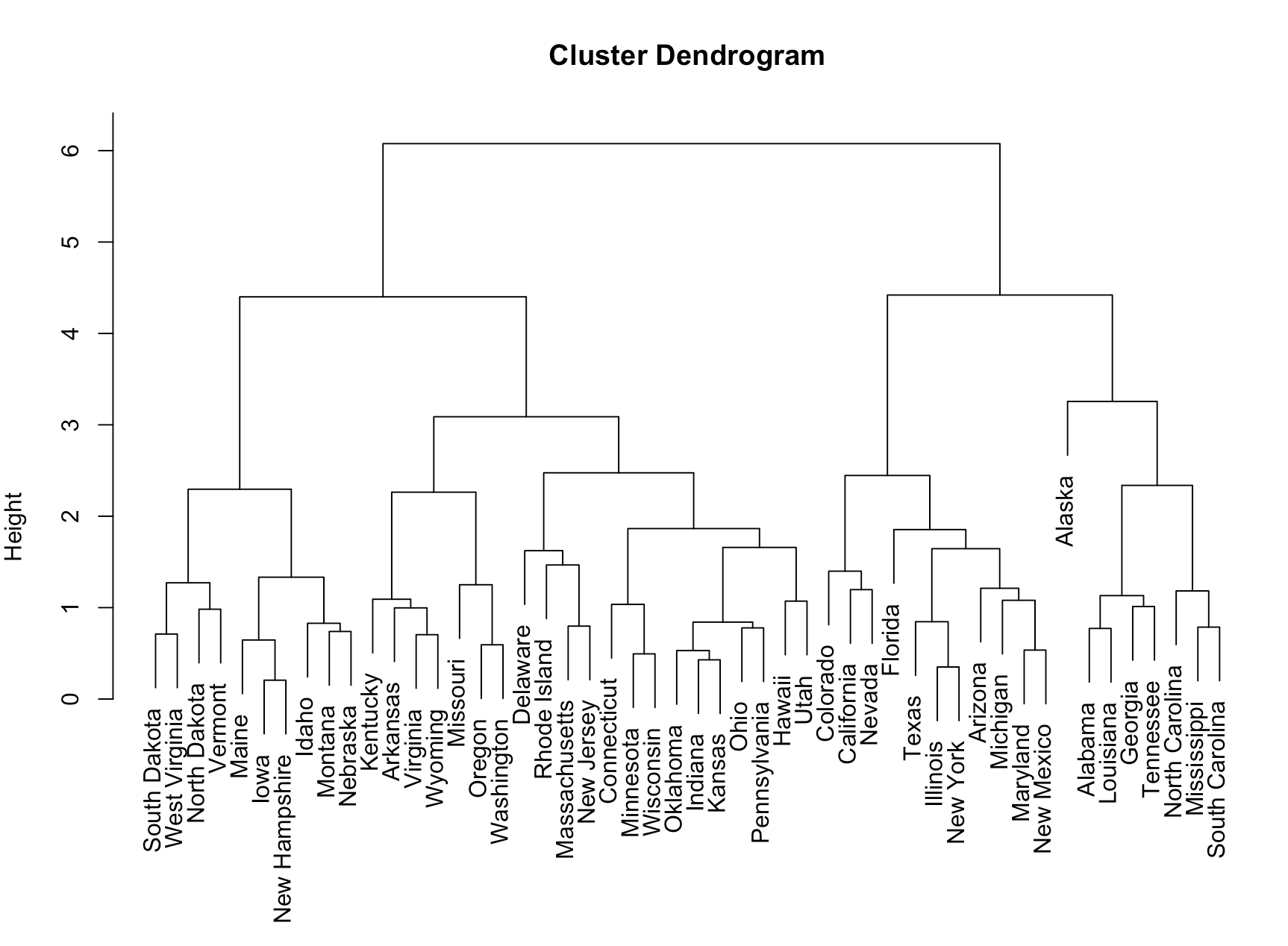
b) They would fuse at the same height. If the distance between observation 5 and 6 is 1, they would fuse at a height of 1 for single and complete linkage because the max and min distance between the two observations is the same value.

7. (Ex 9 Pg 416)

a)



b) 

c) 

d) Trees of scaled data versus not scaled data are pretty similar. They should be scaled before hand, because the data may have different units.

8. (Ex 4 Pg 120)

a) There is not enough information. RSS is defined as e12+ … + en2 where ei = (regression equation with beta hat symbol). You do not know the difference of the RSS between cubic and linear, so there’s not enough information.

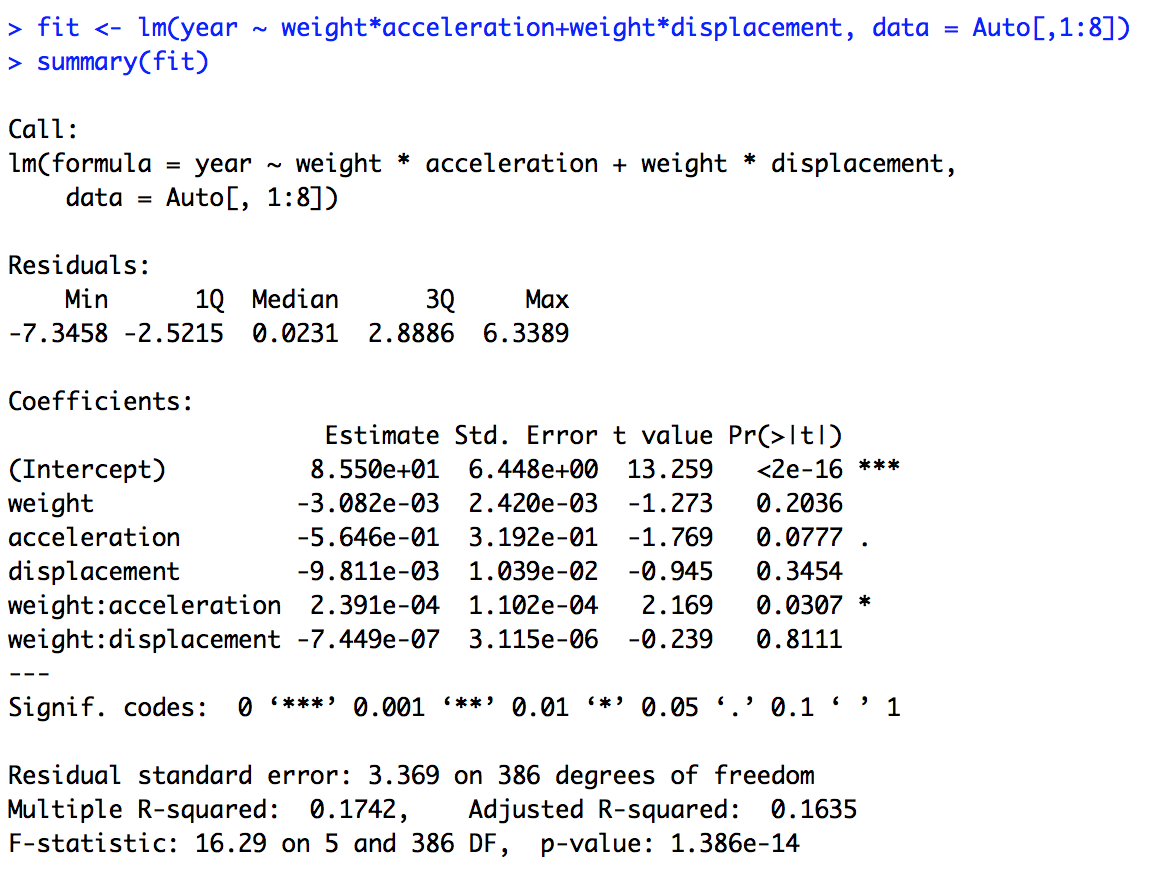
b) Since our test data relies on training data, there is still not enough information.

c) There’s not enough information, because we do not know how far it is from linear. If it was a bit far, then we would expect training RSS for linear to be lower than cubic.

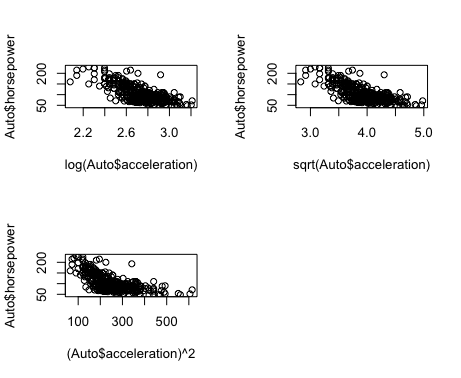
d) Since our test data relies on training data, there is not enough information.

9. (Ex 9e, 9f Pg 122)

e) Weight and displacement were not as statistically significant as weight and acceleration.



f)

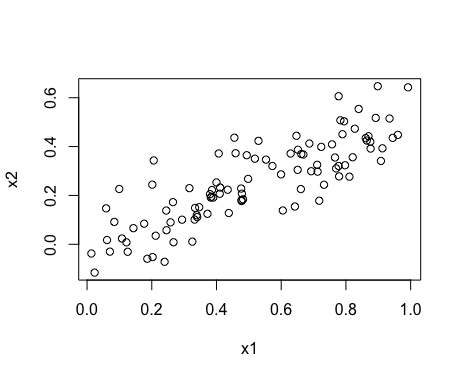


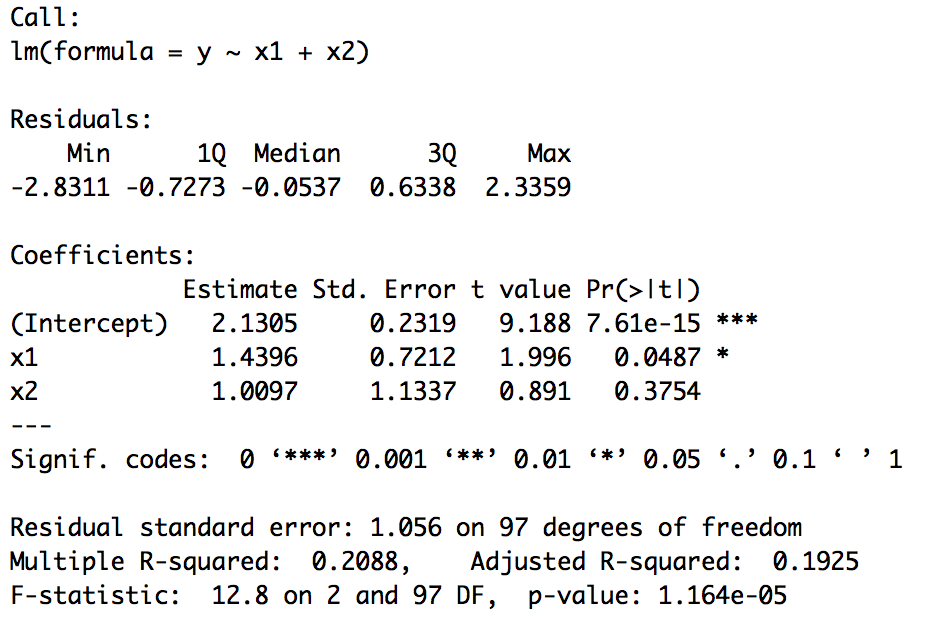
10. (Ex 14 Pg 125)

a) Form of linear model: y = 2 + 2x1 + 0.3x2 + error, where error is a N(0,1) (standard Normal distribution) random variable. Regression coefficients are given by model 2, 2, 0.3 respectively.

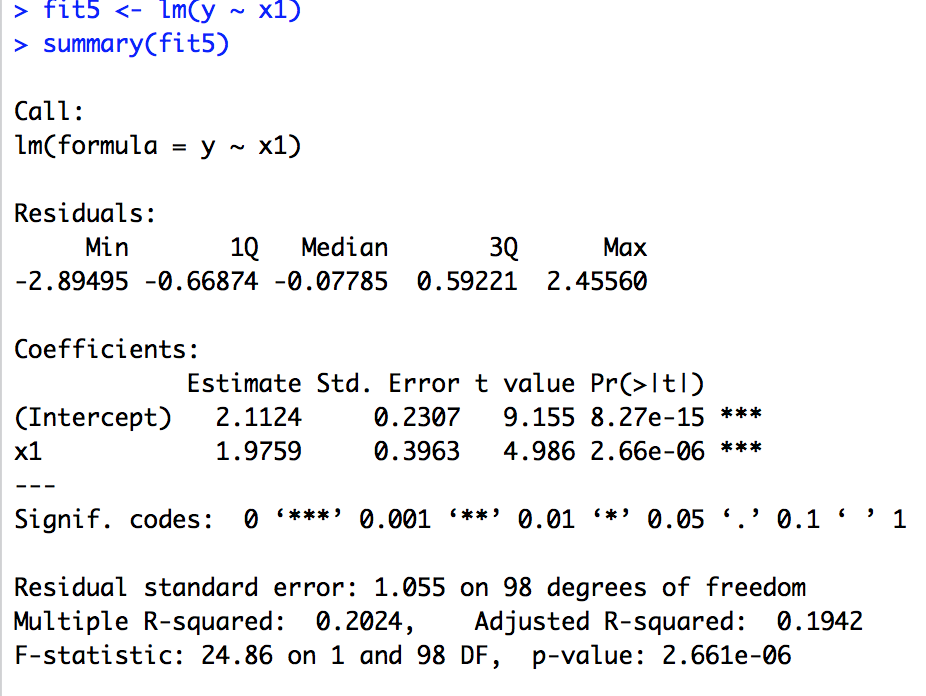
b) Correlation between x1 and x2 is 0.8351212 by using cor(x1,x2).

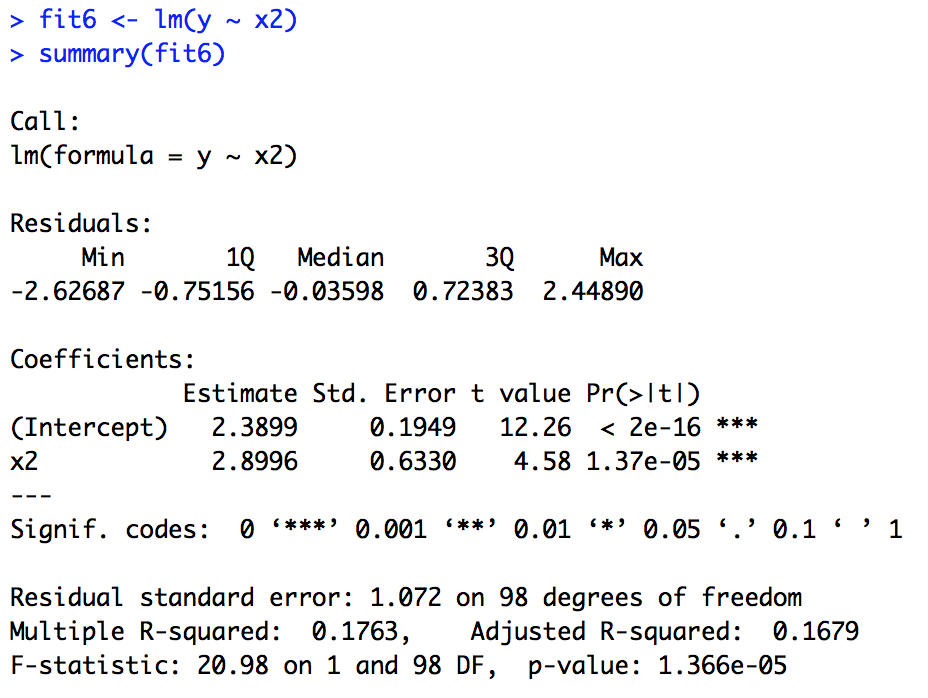
Plot(x1,x2)



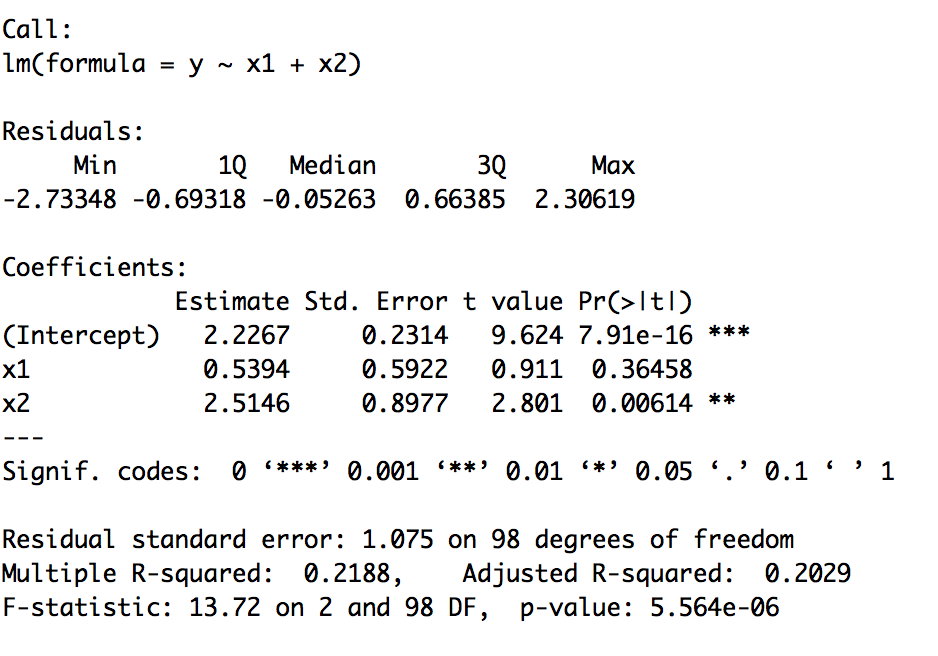
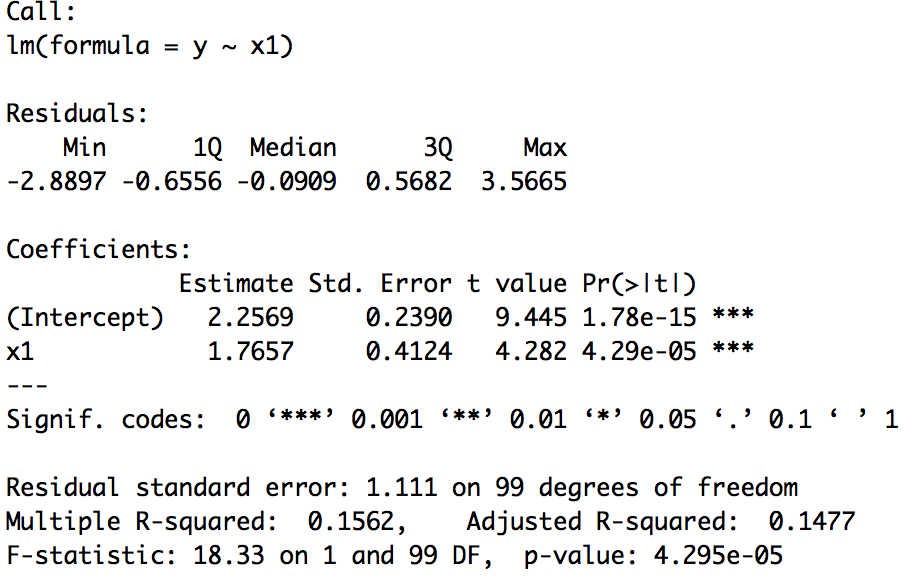
c)

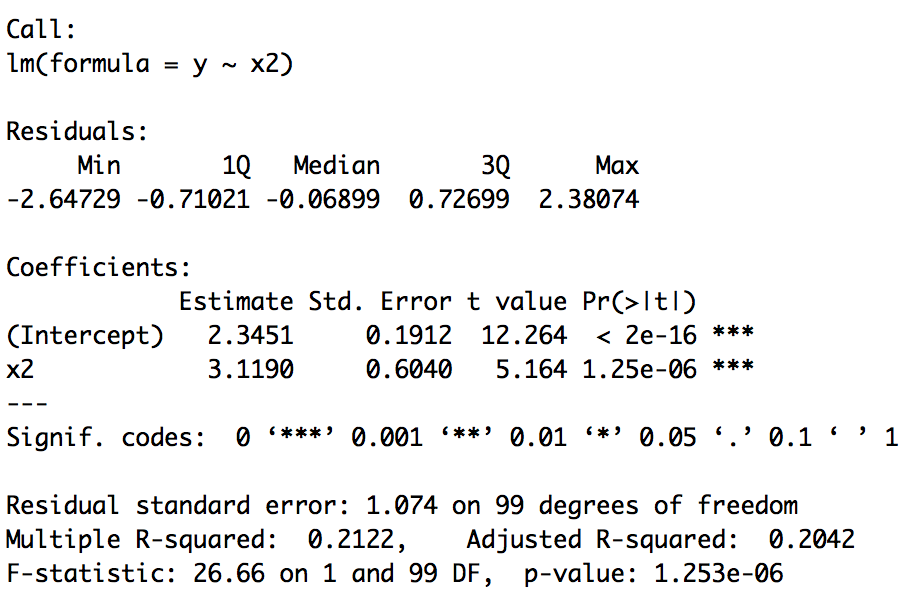
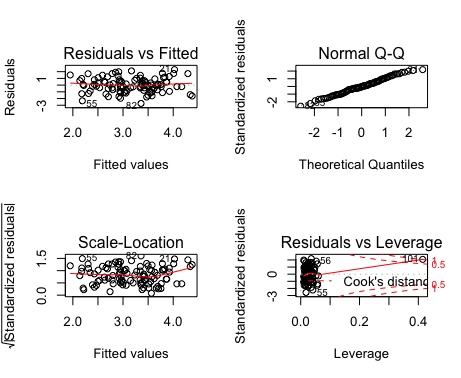
B0, B1, B2 is 2.1305, 1.4396, 1.0097 respectively. Only B0 is close to B^0.We can reject H0 for B1, because its p-value is less than 0.05, but not for B2 since 0.3754 is not less than 0.5.

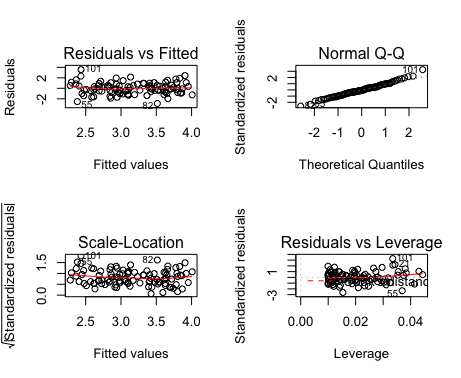
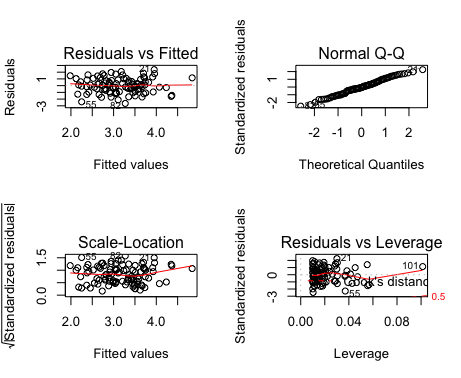
d) x1 here is different from the previous question. Since the p-value is very low, we may reject H0.

e) x2 is different from (a)’s x2. Since p-value of x2 is low, we may reject the null hypothesis.

f) The results from c to e do not contradict each other. The predictors x1 and x2 are correlated to each other, therefore they are collinear. Based on the plot from (a), we can tell that predictors that correlate each other are more bound together. If they do not correlate each other, observations are more scattered.

g)





The model with x1 and x2 as predictors, the last point is the high-leverage point. The model with x1 only, the last point is an outlier. The model with x2, the last point is the high-leverage point.