Andrew Rothman STATS 202 – Homework #2 October 12th, 2018
anr248@stanford.edu SUID: 06325118

Problem #1:

Exercise 2 from section 10.7 of ITSL textbook

Suppose that we have four observations, for which we compute a dissimilarity matrix, given by:

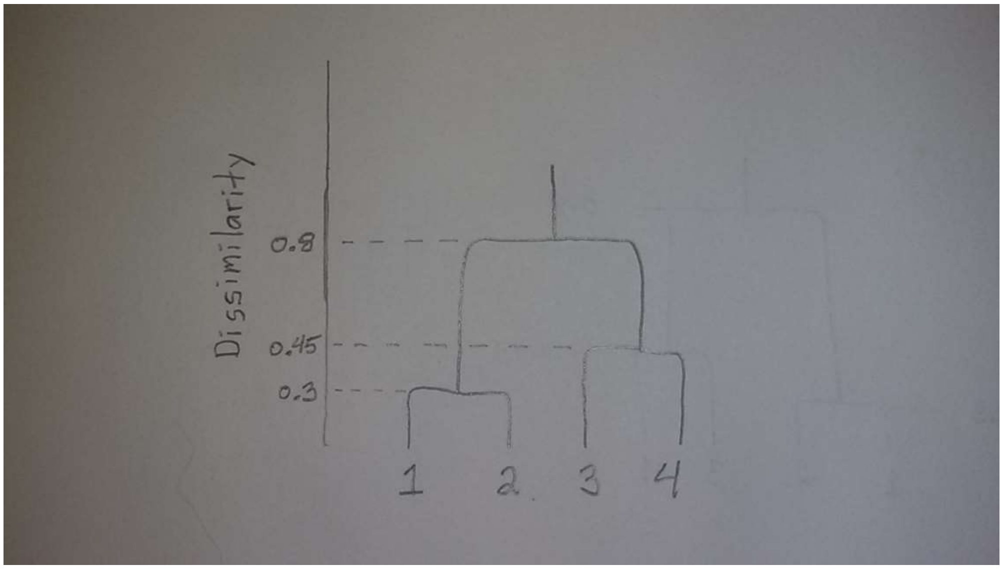
$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations 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.

Answer:

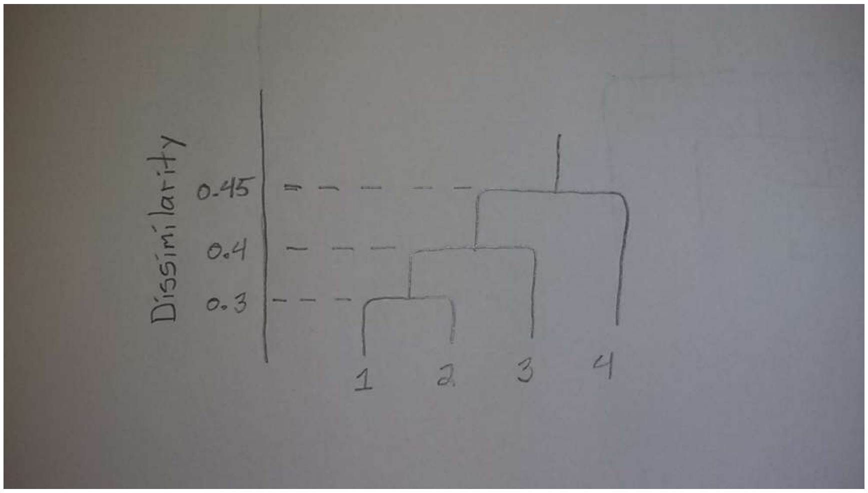
With "complete" linkage, the dissimilarity between two distinct clusters is specified as the largest of all pairwise dissimilarities between individual observations in those two distinct clusters. The dendrogram is shown below:



(b) Repeat (a), this time using single linkage clustering.

Answer:

With "single" linkage, the dissimilarity between two distinct clusters is specified as the smallest of all pairwise dissimilarities between individual observations in those two distinct clusters. The dendrogram is shown below:



(c) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster?

Answer:

As we can see from looking at the dendrogram in part (a), the observations (1,2) would be in one cluster, with observations (3,4) being in the other cluster.

(d) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?

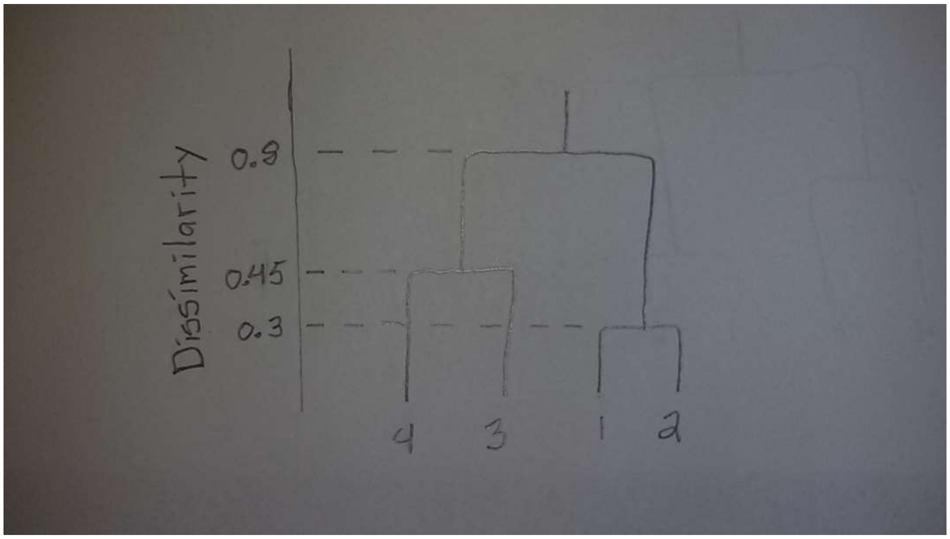
Answer:

As we can see from looking at the dendrogram in part (b), the observations (1,2,3) would be in one cluster, with observation (4) being in the other cluster.

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same

Answer:

In the dendrogram below, I have repositioned the leaves at the bottom of the plot. Though as we can see, the resulting dendrogram provides the same results and interpretation as the dendrogram in part (a)



Problem #2:

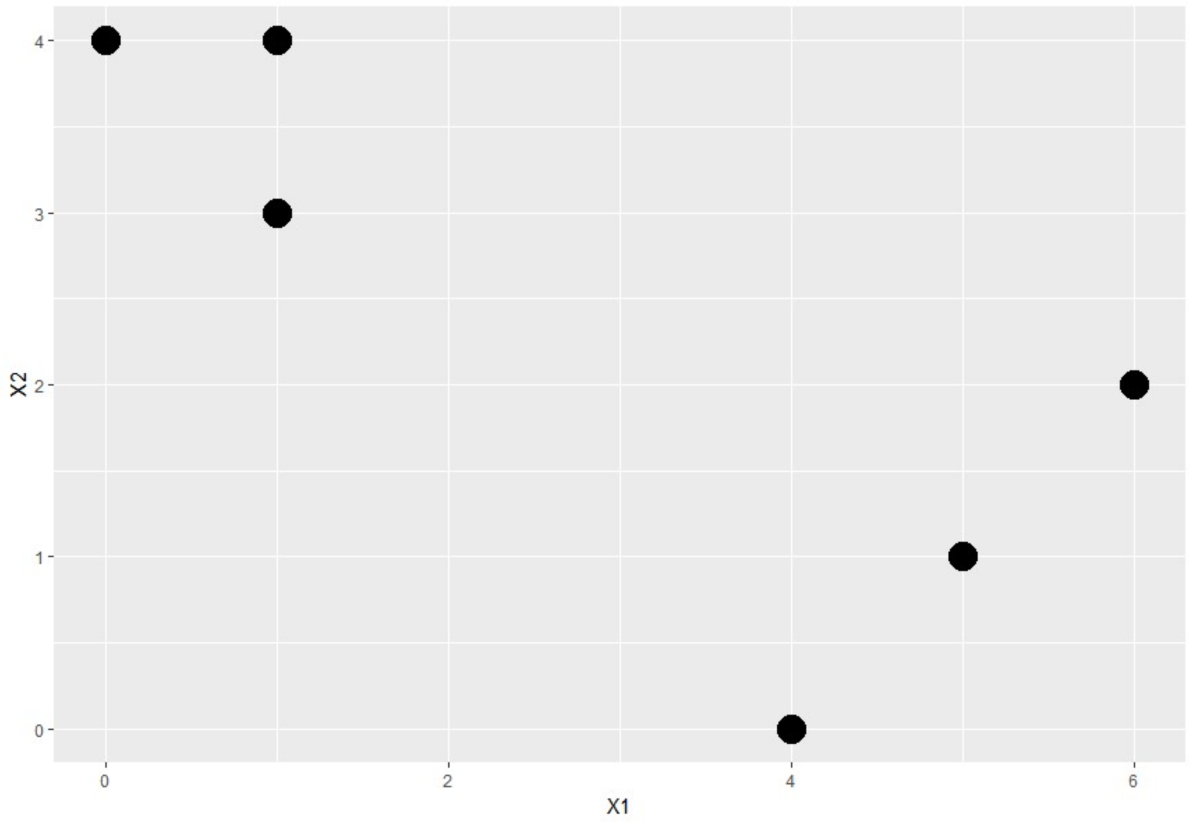
Exercise 3 from section 10.7 of ITSL textbook

In this problem, you will perform K-means clustering manually, with K=2, on a small example with n=6 observations and p=2 features. The observations are as follows:

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the observations.

Answer:



(b) Randomly assign a cluster label to each observation. You can use the “sample()” command in “R” to do this. Report the cluster labels for each observation.

Answer:

Obs.	X_1	X_2	Cluster Number
1	1	4	1
2	1	3	2
3	0	4	2
4	5	1	2
5	6	2	1
6	4	0	2

(c) Compute the centroid for each cluster.

Answer:

Cluster Number	X_1 coordinate of centroid	X_2 coordinate of centroid
1	3.5	3
2	2.5	2

(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

Answer:

Obs.	X_1	X_2	Cluster Number	Updated Cluster Number
1	1	4	1	2
2	1	3	2	2
3	0	4	2	2
4	5	1	2	1
5	6	2	1	1
6	4	0	2	2

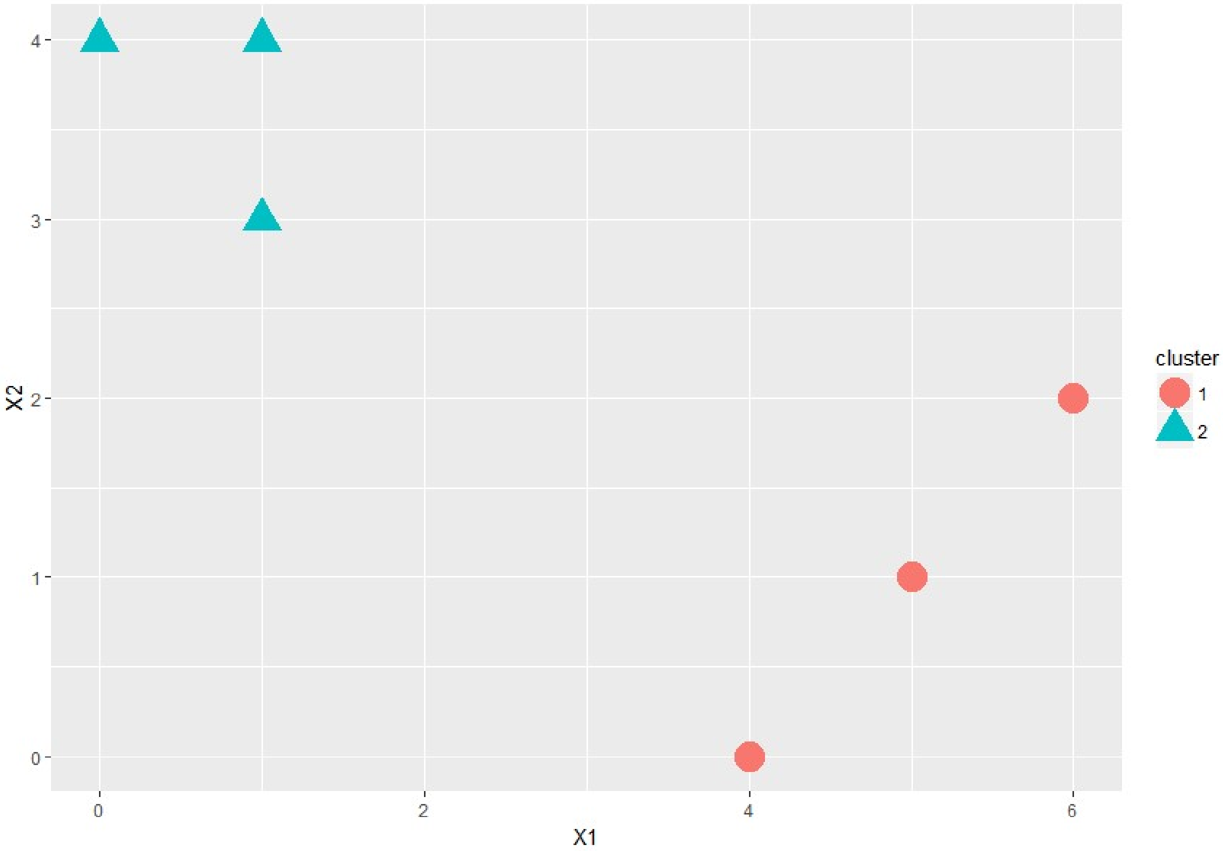
(e) Repeat (c) and (d) until the answers obtained stop changing.

Answer:

Obs.	X_1	X_2	Cluster Number
1	1	4	2
2	1	3	2
3	0	4	2
4	5	1	1
5	6	2	1
6	4	0	1

(f) In your plot from (a), color the observations according to the cluster labels obtained.

Answer:



R-code:

```
################
## Problem #2 ##
################
library(ggplot2)
library(dplyr)
set.seed(10817645)

obs <- c(1:6)
X1 <- c(1,1,0,5,6,4)
X2 <- c(4,3,4,1,2,0)
df <- data.frame(obs, X1, X2)
rm(obs, X1, X2)

n <- dim(df)[1]
p <- (dim(df)[2])-1
K=2

## part A
ggplot(df, aes(x=X1, y=X2)) + geom_point(size=7)

## part B
df$cluster <- sample(1:K, n, replace=T)
df

## part C
cluster <- c(1:K)
df_cent <- data.frame(cluster)
rm(cluster)
df_cent$X1_cord <- NA
df_cent$X2_cord <- NA

for(i in 1:K){
  df_sub <- filter(df, cluster==i)
  df_cent$X1_cord[i] <- mean(df_sub$X1)
  df_cent$X2_cord[i] <- mean(df_sub$X2)
  rm(df_sub)
}
df_cent

## part D
df$cluster_update <- NA
for(i in 1:n){
  df_cent_sub <- df_cent
  df_cent_sub$distance <- NA

  for(j in 1:dim(df_cent_sub)[1]){
    df_cent_sub$distance[j] <- sqrt(((df$X1[i] - df_cent_sub$X1_cord[j])^2) + ((df$X2[i] -
df_cent_sub$X2_cord[j])^2))
  }
  df$cluster_update[i] <- df_cent_sub$cluster[which.min(df_cent_sub$distance)[1]]
  rm(df_cent_sub)
}
df
```

```
## part E
set.seed(10817645)

cluster_update <- function(df, K, n){
  cluster <- c(1:K)
  df_cent <- data.frame(cluster)
  rm(cluster)
  df_cent$X1_cord <- NA
  df_cent$X2_cord <- NA
  for(i in 1:K){
    df_sub <- filter(df, cluster==i)
    df_cent$X1_cord[i] <- mean(df_sub$X1)
    df_cent$X2_cord[i] <- mean(df_sub$X2)
    rm(df_sub)
  }
  df$cluster_update <- NA
  for(i in 1:n){
    df_cent_sub <- df_cent
    df_cent_sub$distance <- NA
    for(j in 1:dim(df_cent_sub)[1]){
      df_cent_sub$distance[j] <- sqrt(((df$X1[i] - df_cent_sub$X1_cord[j])^2) + ((df$X2[i] -
df_cent_sub$X2_cord[j])^2))
    }
    df$cluster_update[i] <- df_cent_sub$cluster[which.min(df_cent_sub$distance)[1]]
    rm(df_cent_sub)
  }
  return(df)
}

K_means <- function(df, K, n){
  df$cluster <- sample(1:K, n, replace=T)
  converged <- 0
  while(converged==0){
    df <- cluster_update(df, K, n)
    if(all(df$cluster==df$cluster_update)){
      converged <- 1
    }
    df$cluster <- df$cluster_update
    df <- dplyr::select(df, -cluster_update)
  }
  return(df)
}

obs <- c(1:6)
X1 <- c(1,1,0,5,6,4)
X2 <- c(4,3,4,1,2,0)
df <- data.frame(obs, X1, X2)
rm(obs, X1, X2)

n <- dim(df)[1]
p <- (dim(df)[2])-1
K=2

df <- K_means(df, K, n)
df

## Part F
df$cluster <- as.factor(df$cluster)
ggplot(df, aes(x=X1, y=X2, shape=cluster, color=cluster)) + geom_point(size=7)
```

Problem #3:

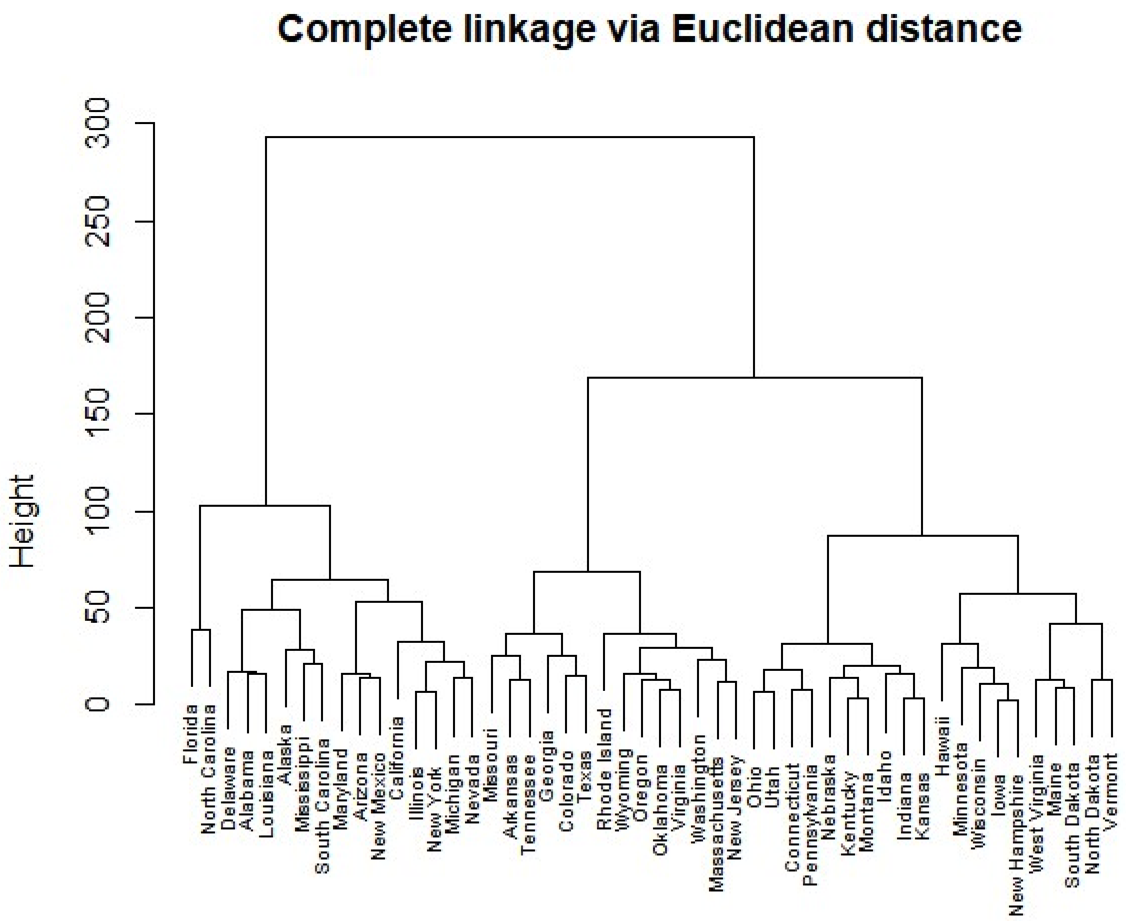
Exercise 9 from section 10.7 of ITSL textbook

Consider the “USArrests” data. We will now perform hierarchical clustering on the states.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

Answer:

Complete linkage via Euclidean distance to cluster the states was performed in R. The dendrogram is shown below.



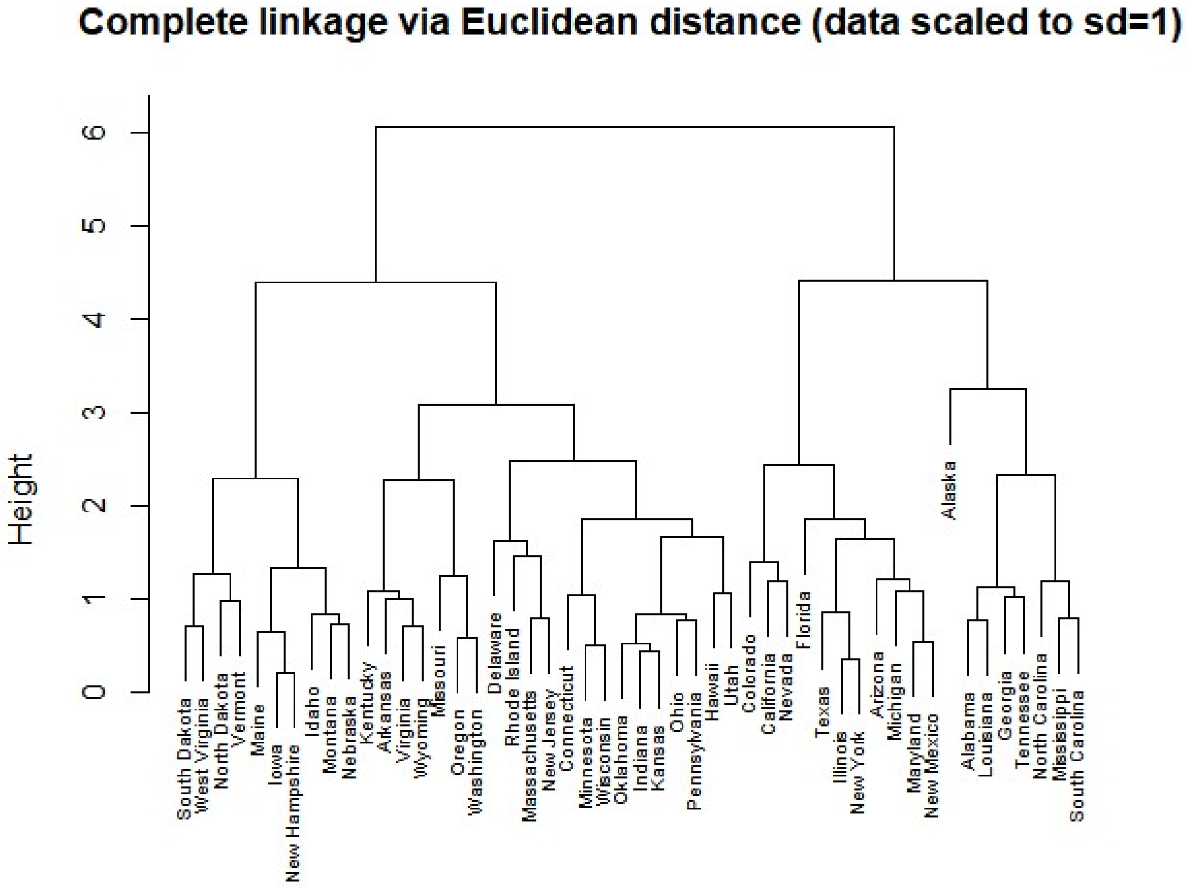
(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

Answer:

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

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, *after scaling the variables to have standard deviation one*.

Answer:



(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for you answer.

Answer:
Comparing the dendrogram in parts (a) and (c) of this problem, we can see that scaling has a significant impact on the resulting hierarchical clustering using complete linkage and Euclidean distance. The two dendrograms are very different; if we were to cut the tree to produce three unique clusters, several of the states would be in differing clusters between the two dendrograms.

As to whether variables should be scaled before the inter-observation dissimilarities are computed, I believe that's a bit of a subjective question, and a question that is case-specific. In clustering problems in general, if one is supplementing the analysis and methods design with significant expert subject-matter knowledge, and there is a justified reason to keep the variables unscaled, then that may be the preferred choice. Additionally, if all of the variables are measured in the same "units", then leaving them unscaled may be suitable. However, if less subject-matter knowledge is known, and there is little known about the functional form of the relationship between the variables used as inputs for the clustering methodology, it may be best to first scale the variables.

In this problem, I have little subject matter expertise regarding US prisons and the legal justice system. Additionally, the variables used as inputs to the hierarchical clustering are not all measured in the same units. In this case, it may best to first scale the variables. Again though, this question is a bit subjective…

R-code:

```
################
## Problem #3 ##
################
library(dplyr)

df <- USArrests
head(df)

## Part A
hc_complete <- hclust(dist(df), method='complete')
plot(hc_complete, main="Complete linkage via Euclidean distance", cex=0.6)

## Part B
cutree(hc_complete , 3)

## Part C
df_scaled <- df
df_scaled$Murder_scaled <- df_scaled$Murder / sd(df_scaled$Murder)
df_scaled$Assault_scaled <- df_scaled$Assault / sd(df_scaled$Assault)
df_scaled$UrbanPop_scaled <- df_scaled$UrbanPop / sd(df_scaled$UrbanPop)
df_scaled$Rape_scaled <- df_scaled$Rape / sd(df_scaled$Rape)
df_scaled <- select(df_scaled, Murder_scaled, Assault_scaled, UrbanPop_scaled, Rape_scaled)
head(df_scaled)

sd(df_scaled$Murder_scaled)
sd(df_scaled$Assault_scaled)
sd(df_scaled$UrbanPop_scaled)
sd(df_scaled$Rape_scaled)

hc_complete_scaled <- hclust(dist(df_scaled), method='complete')
plot(hc_complete_scaled, main="Complete linkage via Euclidean distance (data scaled to sd=1)",
cex=0.6)
```

Problem #4:

Exercise 10 from section 10.7 of ITSL textbook

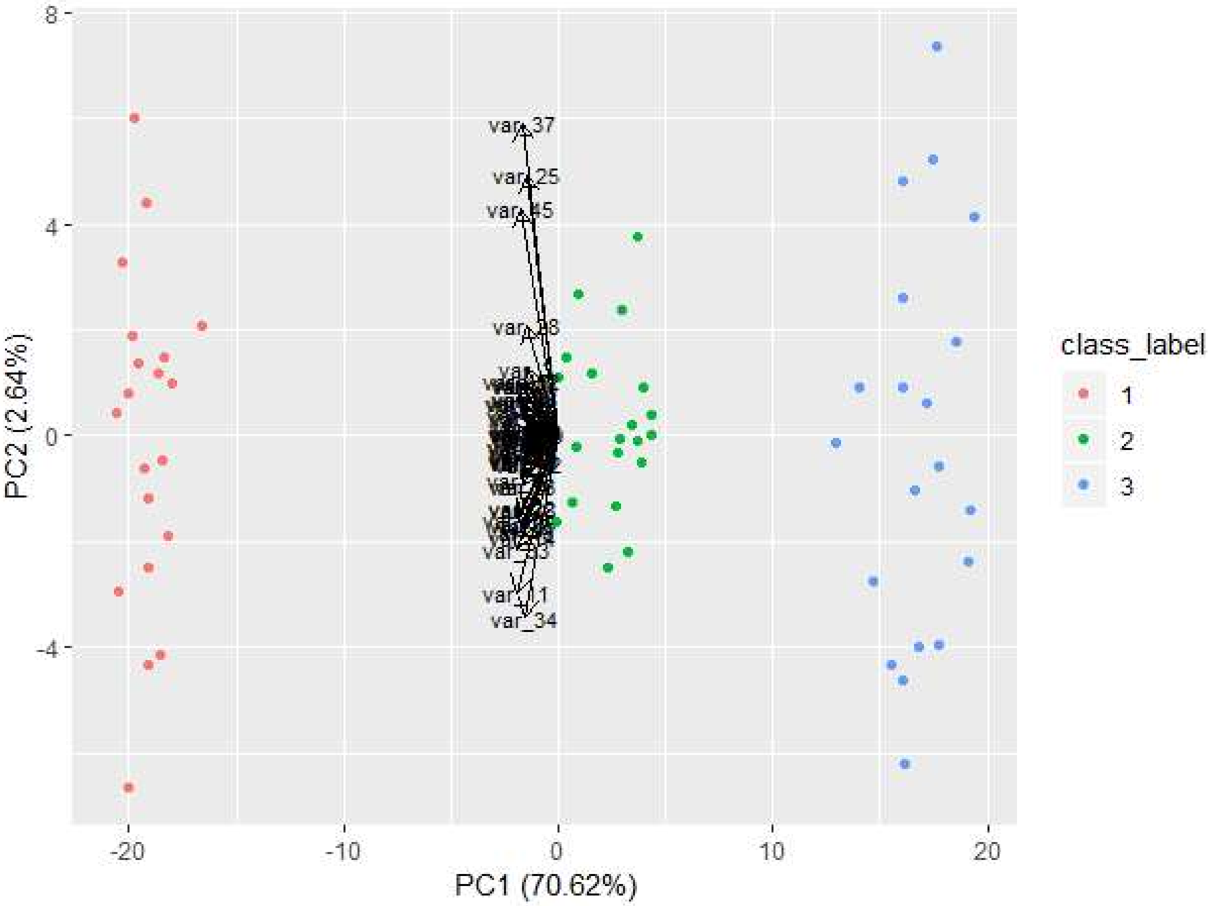
In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.

Answer:
The simulated data set was created in R. Please see the supplemented R code at the end of this problem.

(b) Perform PCA on the 60 observations and plot the first two principal component vectors. Use a different color to indicate the observations in each of the three classes.

Answer:



(c) Perform K-means clustering of the observations with K=3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

Answer:

K-means clustering with K=3 was performed using R. The resulting crosstab (K-means cluster assignment vs. actual class label) is shown in the table below. As shown in the table below, K-means perfectly separated the data according to their actual three class labels:

- All 20 observations with actual class label “1” were clustered into K-means cluster “2”
- All 20 observations with actual class label “2” were clustered into K-means cluster “3”
- All 20 observations with actual class label “3” were clustered into K-means cluster “1”

… So the K-means clustering with K=3 performed extremely well.

		Actual Class Label		
		Class 1	*Class 2*	*Class 3*
K-means Cluster	*K-means cluster 1*	0	0	20
	K-means cluster 2	20	0	0
	K-means cluster 3	0	20	0

(d) Perform K-means clustering with K=2. Describe your results.

Answer:

K-means clustering with K=2 was performed using R. The resulting crosstab (K-means cluster assignment vs. actual class label) is shown in the table below. As shown in the table below, the 40 observations with actual class labels “2” or “3” were clustered together, with the remaining 20 observations in class label “1” being clustered alone.

- All 20 observations with actual class label “1” were clustered into K-means cluster “2”
- All 20 observations with actual class label “2” were clustered into K-means cluster “1”
- All 20 observations with actual class label “3” were clustered into K-means cluster “1”

		Actual Class Label		
		Class 1	*Class 2*	*Class 3*
K-means Cluster	*K-means cluster 1*	0	20	20
	K-means cluster 2	20	0	0

(e) Now perform K-means clustering with K=4, and describe your results.

Answer:

K-means clustering with K=4 was performed using R. The resulting crosstab (K-means cluster assignment vs. actual class label) is shown in the table below. As shown in the table below, all 20 observations with class label "1" were clustered together (cluster 4), and all 20 observations with class label "2" were clustered together (cluster 3). The 20 observations with class label "3" however were split among K-means clusters 1 and 2.

- All 20 observations with actual class label "1" were clustered into K-means cluster "4"
- All 20 observations with actual class label "2" were clustered into K-means cluster "3"
- Of the 20 observations with actual class label "3", 7 observations were clustered into K-means cluster "2", with the remaining 13 observations clustered into K-means cluster "1"

		Actual Class Label		
		Class 1	*Class 2*	*Class 3*
K-means Cluster	*K-means cluster 1*	0	0	13
	K-means cluster 2	0	0	7
	K-means cluster 3	0	20	0
	K-means cluster 4	20	0	0

(f) Now perform K-means clustering with K=3 on the first two principal component score vectors, rather than on the raw data. Comment on the results.

Answer:

K-means clustering with K=3 was performed on just the first two principal components using R. The resulting crosstab (K-means cluster assignment vs. actual class label) is shown in the table below. As shown in the table below, similar to part (c) of this problem, K-means perfectly separated the data according to their actual three class labels:

- All 20 observations with actual class label "1" were clustered into K-means cluster "2"
- All 20 observations with actual class label "2" were clustered into K-means cluster "1"
- All 20 observations with actual class label "3" were clustered into K-means cluster "3"

		Actual Class Label		
		Class 1	*Class 2*	*Class 3*
K-means Cluster	*K-means cluster 1*	0	20	0
	K-means cluster 2	20	0	0
	K-means cluster 3	0	0	20

(g) Using the "scale()" function, perform K-means clustering with K=3 on the data *after scaling each variable to have standard deviation one*. How do these results compare to those obtained in (b)? Explain.

Answer:

K-means clustering with K=3 was performed on the "scaled" data using R. The resulting crosstab (K-means cluster assignment vs. actual class label) is shown in the table below. As shown in the table below, similar to part (c) of this problem, K-means perfectly separated the data according to their actual three class labels:

- All 20 observations with actual class label "1" were clustered into K-means cluster "2"
- All 20 observations with actual class label "2" were clustered into K-means cluster "3"
- All 20 observations with actual class label "3" were clustered into K-means cluster "1"

		Actual Class Label		
		Class 1	*Class 2*	*Class 3*
K-means Cluster	*K-means cluster 1*	0	0	20
	K-means cluster 2	20	0	0
	K-means cluster 3	0	20	0

We can see the three respective clusters correspond perfectly to the three class labels in part (b) of this problem. Note that I do recognize that in my dataset, parts (c), (f), and (g) of this exercise all produced essentially equivalent results. This equivalency of results is because in my particular simulated dataset, the 60 observations are extremely well separated based on their actual class labels. I incorporated this extreme separation deliberately in part (a) of this problem *(perhaps a bit too well…)*. If I were to change the conditions of the simulated set to where the class labels are less deliberately separated, this would likely result in *(at least slightly)* different output and conclusions from the K-means clustering algorithm for parts (c), (f), and (g) of this problem.

R-code:

```
################
## Problem #4 ##
################
set.seed(1234)
library(dplyr)
library(ggfortify)
library(psych)

## Part A
## initialize array
data <- array(NA, dim=c(60,50))
column_labels <- array(NA, dim=c(50,1))

class_label <- array(NA, dim=c(60,1))
class_label[1:20,1] <- 1
class_label[21:40,1] <- 2
class_label[41:60,1] <- 3

for(i in 1:50){
  column_labels[i,1] <- paste("var_",i,sep='')
  mean <- runif(1,-1,1)
  sd = runif(1,0.5,2)

  data[1:20, i] <- rnorm(20, mean+2, sd)
  data[21:40, i] <- rnorm(20, mean-1, sd)
  data[41:60, i] <- rnorm(20, mean-3, sd)
}
colnames(data) <- column_labels
data_df <- data.frame(data)
data_df$class_label <- as.factor(class_label)
head(data_df)
describe(data_df)

## Part B
autoplot(prcomp(select(data_df, -class_label), scale=FALSE), scale=0, data=data_df, colour =
'class_label', loadings = TRUE, loadings.colour = 'black', loadings.label = TRUE,
loadings.label.colour = 'black', loadings.label.size = 3)

## Part C
Kmeans_k3 <- kmeans(select(data_df, -class_label), 3, nstart=20)
data_df$k3 <- Kmeans_k3$cluster
table(data_df$k3, data_df$class_label)

## Part D
Kmeans_k2 <- kmeans(select(data_df, -class_label), 2, nstart=20)
data_df$k2 <- Kmeans_k2$cluster
table(data_df$k2, data_df$class_label)

## Part E
Kmeans_k4 <- kmeans(select(data_df, -class_label), 4, nstart=20)
data_df$k4 <- Kmeans_k4$cluster
table(data_df$k4, data_df$class_label)

## Part F
first_2_PCs <- prcomp(data)$rotation[,1:2]
Kmeans_k3_2PC <- kmeans((data %*% first_2_PCs), 3, nstart=20)
data_df$k3_2PC <- Kmeans_k3_2PC$cluster
table(data_df$k3_2PC, data_df$class_label)

## Part G
data_scaled <- data.frame(scale(data, center = TRUE, scale = TRUE))
Kmeans_k3_scaled <- kmeans(data_scaled, 3, nstart=20)
data_df$k3_scaled <- Kmeans_k3_scaled$cluster
table(data_df$k3_scaled, data_df$class_label)
```

Problem #5:

Exercise 3 from section 3.7 of ITSL textbook

Suppose we have a data set with five predictors:

- $X_1 = GPA$
- $X_2 = IQ$
- $X_3 = Gender$ (1 for Female and 0 for Male)
- $X_4 = Interaction\ between\ GPA\ and\ IQ$
- $X_5 = Interaction\ between\ GPA\ and\ Gender.$

The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get:

- $\hat{\beta}_0 = 50$
- $\hat{\beta}_1 = 20$
- $\hat{\beta}_2 = 0.07$
- $\hat{\beta}_3 = 35$
- $\hat{\beta}_4 = 0.01$
- $\hat{\beta}_5 = -10$

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Answer:

The linear regression model for response $\hat{Y}$ *(starting salary after graduation)* is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5$$

$$\hat{Y} = 50 + (20 * GPA) + (0.07 * IQ) + (35 * Female) + (0.01 * GPA * IQ) - (10 * GPA * Female)$$

As we can see from the model specified above, the estimated regression coefficient in-front of the variable “female” is positive 35. However the interaction term between “GPA” and “female” is -10. This means for females specifically, there is a negative additive effect on their predicted outcome of starting salary as their GPA increases. Therefore based on the estimated model above, males are predicted to earn more on average than females provided their GPA is sufficiently high.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Answer:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5$

$\hat{Y} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$

$\hat{Y} = 50 + (20 * 4) + (0.07 * 110) + (35 * 1) + (0.01 * 4 * 110) - (10 * 4 * 1)$

$\hat{Y} = 50 + 80 + 7.7 + 35 + 4.4 - 40$

$\hat{Y} = 137.1$

The predicted salary of a female with IQ of 110 and a GPA of 4.0 is $137,100

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Answer:

Given the information provided in the problem statement, I would argue we cannot make this conclusion given the results presented. We do not have a sense of the precision of any of these estimates (we don't have any measure of inferential coverage; all we have are point estimates). I would also say, even if inferential coverage was provided and showed that estimate $\hat{\beta}_4$ was highly significant at an alpha threshold of $\alpha = 0.05$ for null hypothesis $H_0: \beta_4 = 0$, there is no indication in the problem statement that we as the investigators analyzing this data have the sufficient subject-matter expertise to specify how large β_4 in truth would need to be to be considered a substantively relevant effect size. I would further comment I think this problem is very poorly specified, and in its current state is subjective and is not a scientifically refutable question.

Problem #6:

Exercise 10, parts (a)-(e), from section 3.7 of ITSL textbook

This question should be answered using the "Carseats" data set.

(a) Fit a multiple regression model to predict "Sales" using "Price", "Urban", and "US".

Answer:

The estimated model that was fit with OLS was:

$$\widehat{Sales} = \hat{\beta}_0 + \hat{\beta}_1 Price + \hat{\beta}_2 Urban + \hat{\beta}_3 US$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	13.043469	< 2e-16
β_1	-0.054459	< 2e-16
β_2	-0.021916	0.936
β_3	1.200573	4.86e-06

(b) Provide an interpretation of each coefficient in the model

Answer:

The interpretation of each estimated coefficient:

$\hat{\beta}_0 = 13.043469$

- The estimated average unit sales (in thousands) among stores that charge $0 for car seats, are in a rural location, and are not in the United States is 13.043469

$\hat{\beta}_1 = -0.054459$

- The estimated average unit sales (in thousands) in stores decreases by 0.054459 unit sales (in thousands) for each $1 increase the store charges for a car seat, adjusting for both statistical area type (urban vs rural) and country location (US vs not US) of the store.

$\hat{\beta}_2 = -0.021916$

- The estimated average unit sales (in thousands) is 0.021916 unit sales (in thousands) less in stores in Urban areas compared to stores in Rural areas, adjusting for both the amount of money the store charges for a car seat and country location (US vs not US) of the store.

$\hat{\beta}_3 = 1.200573$

- The estimated average unit sales (in thousands) is 1.200573 unit sales (in thousands) more in stores in the United States compared to stores not in the United States, adjusting for both the amount of money the store charges for a car seat and statistical area type (urban vs rural) of the store.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

Answer:

$$\widehat{Sales} = \hat{\beta}_0 + \hat{\beta}_1 Price + \hat{\beta}_2 Urban + \hat{\beta}_3 US$$

Where:

- $Sales$ = Unit sales (in thousands) at each location
- $Price$ = Price in US dollars company charges for cat seats at each site
- $Urban$ = indicator variable specifying if store is located in an urban or rural area (1: urban, 0: rural)
- US = indicator variable specifying if store is located in the United States or not (1: located in the USA, 0: not located in the USA)

(d) For which of the predictors car you reject the null hypothesis $H_0: \beta_j = 0$?

Answer:

Looking at the results from part (a), the calculated p-values for estimates $\hat{\beta}_j$ for the null hypothesis $H_0: \beta_j = 0$ are all well below 0.05 except for $\hat{\beta}_2$ (p-value = 0.936) which corresponds to the "Urban" indicator variable. Therefore at an alpha threshold of 0.05, we can reject the null hypotheses for predictor variables "Price" and "US".

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

Answer:

The estimated model that was fit with OLS was:

$$\widehat{Sales} = \hat{\beta}_0 + \hat{\beta}_1 Price + \hat{\beta}_2 US$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	13.03079	< 2e-16
β_1	-0.05448	< 2e-16
β_2	1.19964	4.71e-06

R-code:

```
################
## Problem #6 ##
################
library(ISLR)

df <- Carseats
head(df)
colnames(df)
dim(df)

## Part A
summary(lm(Sales ~ Price + Urban + US, data=df))
table(df$Price, exclude=NULL)
table(df$Urban, exclude=NULL)
table(df$US, exclude=NULL)

## Part E
summary(lm(Sales ~ Price + US, data=df))
```

Problem #7:

Exercise 14 from section 3.7 of ITSL textbook

This This problem focuses on the *collinearity* problem.

(a) Perform the following commands in R:

```
> set.seed(1)
> x1=runif(100)
> x2 =0.5*x1+rnorm(100)/10
> y=2+2*x1+0.3*x2+rnorm(100)
```

The last line corresponds to creating a linear model in which *y* is a function of *x1* and *x2*. Write out the form of the linear model. What are the regression coefficients?

Answer:

The linear model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

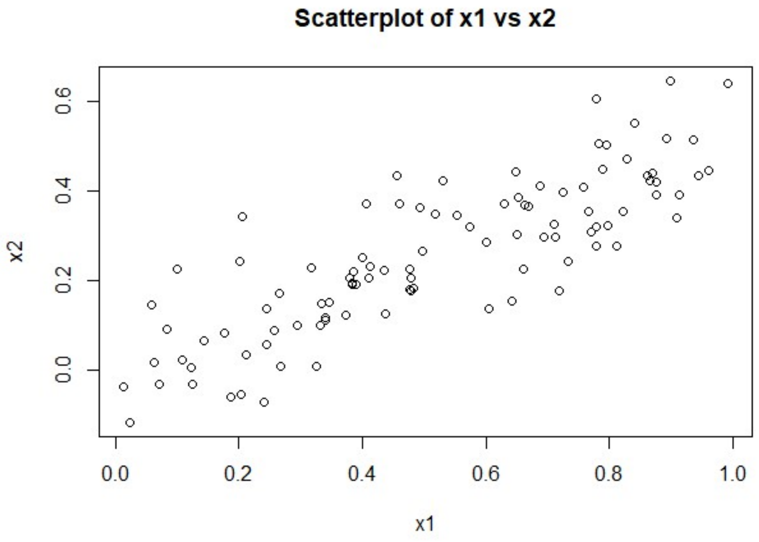
With:

- $\beta_0 = 2$
- $\beta_1 = 2$
- $\beta_2 = 0.3$
- $\epsilon \sim N(0,1)$

(b) What is the correlation between *x1* and *x2*? Create a scatterplot displaying the relationship between the variables.

Answer:

The Pearson correlation coefficient between *x1* and *x2* is 0.8351212. The scatterplot displaying the relationship between the two variables is below:



(c) Using this data, fit a least squares regression to predict *y* using *x1* and *x2*. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0: \beta_1 = 0$? How about the null hypothesis $H_0: \beta_2 = 0$?

Answer:

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	2.1305	7.61e-15
β_1	1.4396	0.0487
β_2	1.0097	0.3754

The interpretation of each estimated coefficient:

$\hat{\beta}_0 = 2.1305$

- The estimated average Y value when x1 and x2 are both zero is 2.1305

$\hat{\beta}_1 = 1.4396$

- The estimated average Y increases by 1.4396 units for each one unit increase in x1, adjusting for x2.

$\hat{\beta}_2 = 1.0097$

- The estimated average Y increases by 1.0097 units for each one unit increase in x2, adjusting for x1.

Looking at the parameter estimates above, they are at least all in the same direction of the actual parameter values (all positive in this case). Some of the magnitudes (in particular $\hat{\beta}_2 = 1.0097$) are off from the true values of the parameters specified in part (a) of the problem.

From the table above, the calculated p-value for the estimate $\hat{\beta}_1$ is 0.0487. Therefore at an alpha threshold of $\alpha = 0.05$ we can reject the null hypothesis $H_0: \beta_1 = 0$

From the table above, the calculated p-value for the estimate $\hat{\beta}_2$ is 0.3754. Therefore at an alpha threshold of $\alpha = 0.05$ we cannot reject the null hypothesis $H_0: \beta_2 = 0$

(d) Now fit a least squares regression to predict *y* using only *x1*. Comment on your results. Can you reject the null hypothesis $H_0 \colon \beta_1 = 0$?

Answer:
For the fit linear model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	2.1124	8.27e-15
β_1	1.9759	2.66e-06

The interpretation of each estimated coefficient:

$\hat{\beta}_0 = 2.1124$

- The estimated average Y value when x1 is zero is 2.1124

$\hat{\beta}_1 = 1.9759$

- The estimated average Y increases by 1.9759 units for each one unit increase in x1.

From the table above, the calculated p-value for the estimate $\hat{\beta}_1$ is 2.66e-06. Therefore at an alpha threshold of $\alpha = 0.05$ we can reject the null hypothesis $H_0 \colon \beta_1 = 0$

(e) Now fit a least squares regression to predict *y* using only *x2*. Comment on your results. Can you reject the null hypothesis $H_0 \colon \beta_1 = 0$?

Answer:
For the fit linear model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_2$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	2.3899	< 2e-16
β_1	2.8996	1.37e-05

The interpretation of each estimated coefficient:

$\hat{\beta}_0 = 2.3899$

- The estimated average Y value when x2 is zero is 2.3899

$\hat{\beta}_1 = 2.8996$

- The estimated average Y increases by 1.9759 units for each one unit increase in x2.

From the table above, the calculated p-value for the estimate $\hat{\beta}_1$ is 1.37e-05. Therefore at an alpha threshold of $\alpha = 0.05$ we can reject the null hypothesis $H_0\colon \beta_1 = 0$

(f) Do the results obtained in (c)-(e) contradict each other? Explain your answer.

Answer:

No, the results in (c)-(e) do not contradict each other. Because we've simulated this data, we know the true functional form of the model, we know how Y was specified in terms of it's relation to x1 and x2, and we know how x1 and x2 were specified in relation to each other (i.e. x2 was specified as a function of x1). Because x2 is a function of x1, x1 and x2 are highly correlated. This is why when we try to fit both variables in the linear model simultaneously with OLS regression (like in part (c)), our parameters estimates (particular the estimate for β_2) were not that great (compared to the true values of parameters β_j).

Both x1 and x2 explain much of the same variance inherent in the variable Y, so stability in the fitting of parameter estimates when x1 and x2 are both in the model is an issue. When we fit x1 or x2 in the model alone without the other, we find the parameter estimates to be highly statistically significant for $H_0\colon \beta_1 = 0$. This is because x1 and x2 are now not "fighting" with each other to explain the same variance in Y.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
> x1=c(x1 , 0.1)
> x2=c(x2 , 0.8)
> y=c(y,6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

Answer:

For the fit linear model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	2.2267	7.91e-16
β_1	0.5394	0.36458
β_2	2.5146	0.00614

For this model, I would say that observation “101” is not an outlier, but is a high leverage point. The R^2 fit of this model with and without this additional observation is nearly identical (R^2=0.2188 with observation; R^2=0.2088 without), providing evidence it’s probably not an outlier. However, the coefficient estimates and inferential coverage above for β_1 and β_2 are very different than in part (c), telling us this additional observation is a high leverage point.

For the fit linear model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	2.2569	1.78e-15
β_1	1.7657	4.29e-05

For this model, I would say that observation “101” is an outlier, but is not a high leverage point. The R^2 fit of this model decreases quite a bit with this additional observation (R^2=0.1562 with observation; R^2=0.2024 without), providing reasonable evidence it’s an outlier. The coefficient estimate and inferential coverage above for β_1 is not very different than in part (d), telling us this additional observation is likely not a high leverage point.

For the fit linear model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_2$$

The estimated parameters of the model are shown below:

	Estimated Coefficient	p-value
β_0	2.3451	< 2e-16
β_1	3.1190	1.25e-06

For this model, I would say that observation "101" is not an outlier or a high leverage point. The fit of this model actually increased with this additional observation (R^2=0.2122 with observation; R^2=0.1763 without), and the coefficient estimate and inferential coverage above for β_1 is not very different than in part (e). Therefore there is reasonable evidence to say, for this model, the additional observation is not an outlier or high leverage point.

R-code:

```
################
## Problem #7 ##
################

## Part A
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)

df <- data.frame(y, x1, x2)
head(df)

## Part B
cor(df$x1, df$x2)
plot(df$x1, df$x2, main="Scatterplot of x1 vs x2", xlab="x1", ylab="x2")

## Part C
summary(lm(y ~ x1 + x2, data=df))

## Part D
summary(lm(y ~ x1, data=df))

## Part E
summary(lm(y ~ x2, data=df))

## Part G
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
df_add <- data.frame(y, x1, x2)

summary(lm(y ~ x1 + x2, data=df))
summary(lm(y ~ x1 + x2, data=df_add))
summary(lm(y ~ x1, data=df))
summary(lm(y ~ x1, data=df_add))
summary(lm(y ~ x2, data=df))
summary(lm(y ~ x2, data=df_add))
```