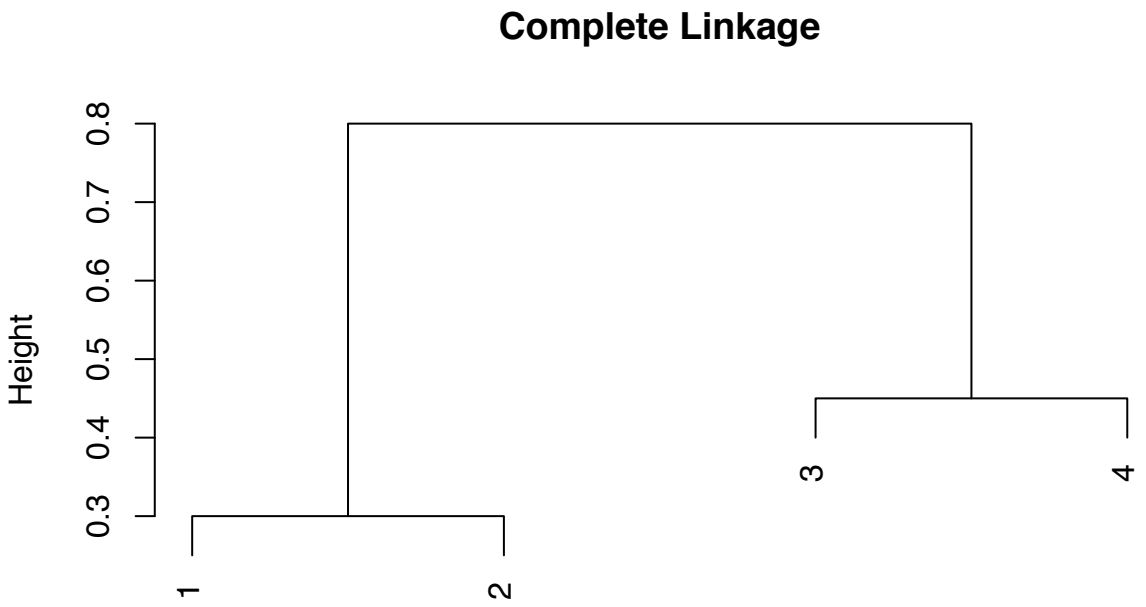
STATS 202 PS2

10/10/18

1) 10.7.2

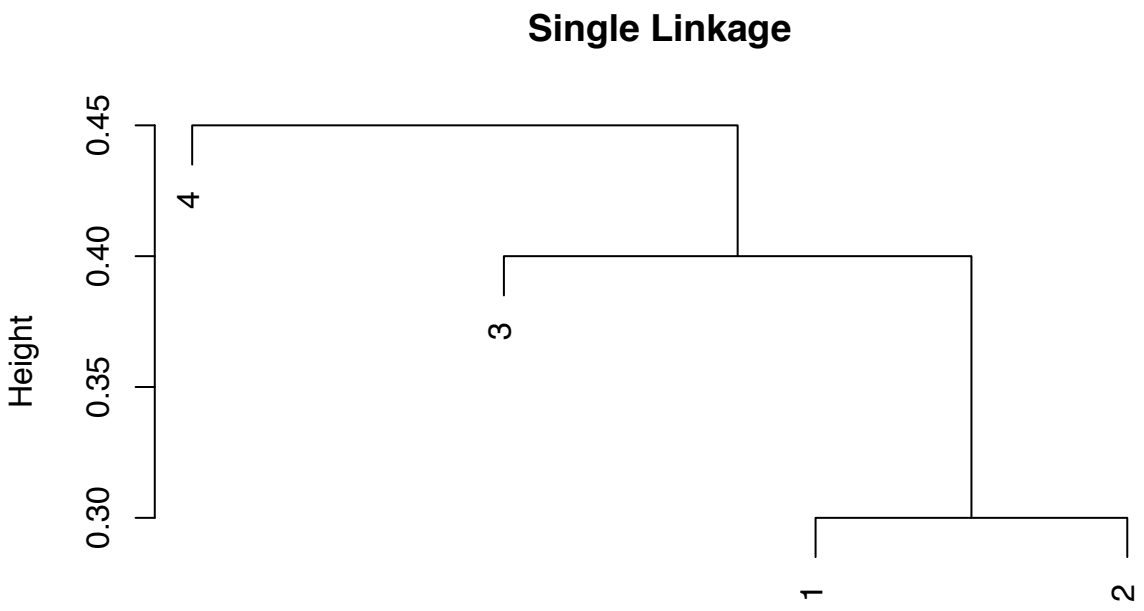
1a)

```
X = matrix(c(0, 0.3, 0.4, 0.7,
0.3, 0, 0.5, 0.8,
0.4, 0.5, 0, 0.45,
0.7, 0.8, 0.45, 0), nrow = 4)

hc.complete = hclust(as.dist(X), method="complete")

plot(hc.complete, main="Complete Linkage", xlab="", sub="")
```



1b)

```
hc.simple = hclust(as.dist(X), method="single")

plot(hc.simple, main="Single Linkage", xlab="", sub="")
```



1c)

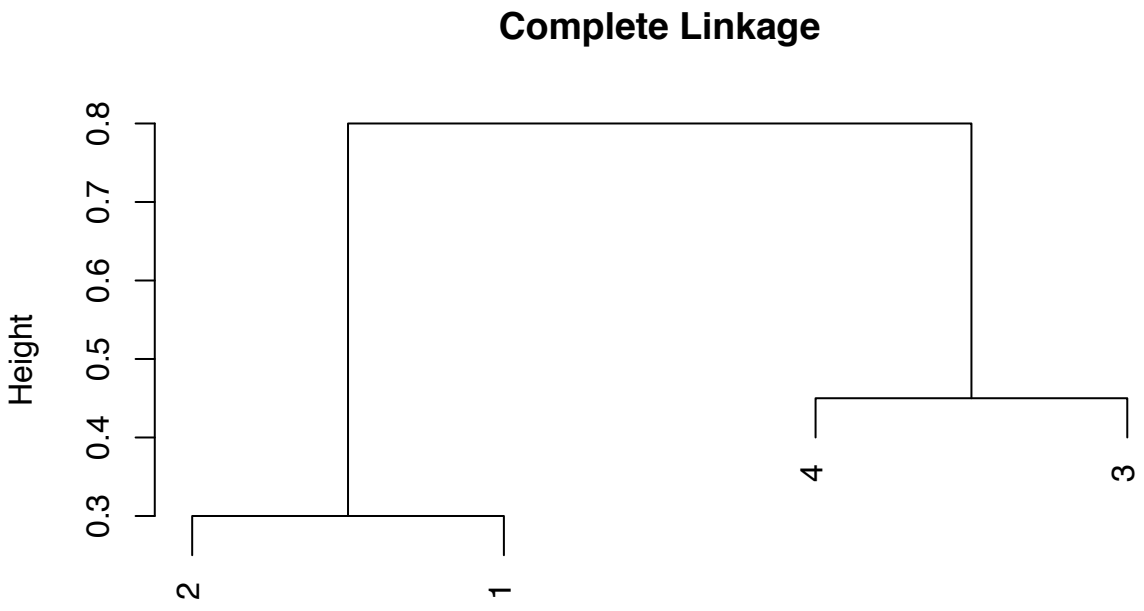
Two clusters from 1(a): (1, 2), (3, 4)

1d)

Two clusters from 1(b): (1, 2, 3), (4)

1e)

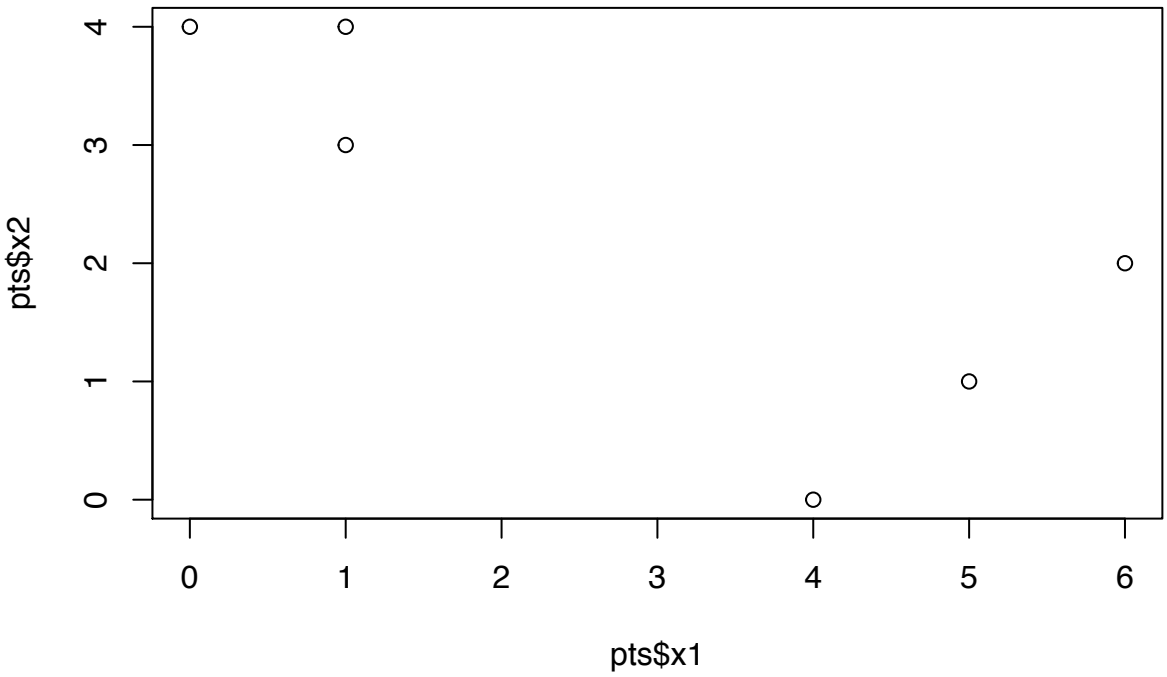
```
plot(hc.complete, main="Complete Linkage", xlab="", sub="", labels=c(2,1,4,3))
```



2) 10.7.3

2a)

```
set.seed(202)

pts = data.frame(
   obs = c(1:6),
   x1 = c(1,1,0,5,6,4),
   x2 = c(4,3,4,1,2,0)
)

plot(pts$x1, pts$x2)
```



2b)

```
pts$cluster = sample(c(1,2), 6, replace=TRUE)

pts
```

```
##   obs x1 x2 cluster
## 1   1  1  4       1
## 2   2  1  3       2
## 3   3  0  4       1
## 4   4  5  1       1
## 5   5  6  2       1
## 6   6  4  0       2
```

2c)

```
c1 = c(mean(pts[pts$cluster == 1, 2]), mean(pts[pts$cluster == 1, 3]))
c2 = c(mean(pts[pts$cluster == 2, 2]), mean(pts[pts$cluster == 2, 3]))

c1
```

```
## [1] 3.00 2.75
```

```
c2
```

```
## [1] 2.5 1.5
```

2d)

```
pts$d1 = ((pts$x1 - c1[1]) ^ 2) + ((pts$x2 - c1[2]) ^ 2)
pts$d2 = ((pts$x1 - c2[1]) ^ 2) + ((pts$x2 - c2[2]) ^ 2)

pts$cluster[(pts$d1 < pts$d2)] = 1
pts$cluster[(pts$d1 >= pts$d2)] = 2

pts
```

```
##   obs x1 x2 cluster      d1   d2
## 1   1  1  4       1  5.5625  8.5
## 2   2  1  3       1  4.0625  4.5
## 3   3  0  4       1 10.5625 12.5
## 4   4  5  1       2  7.0625  6.5
## 5   5  6  2       1  9.5625 12.5
## 6   6  4  0       2  8.5625  4.5
```

2e)

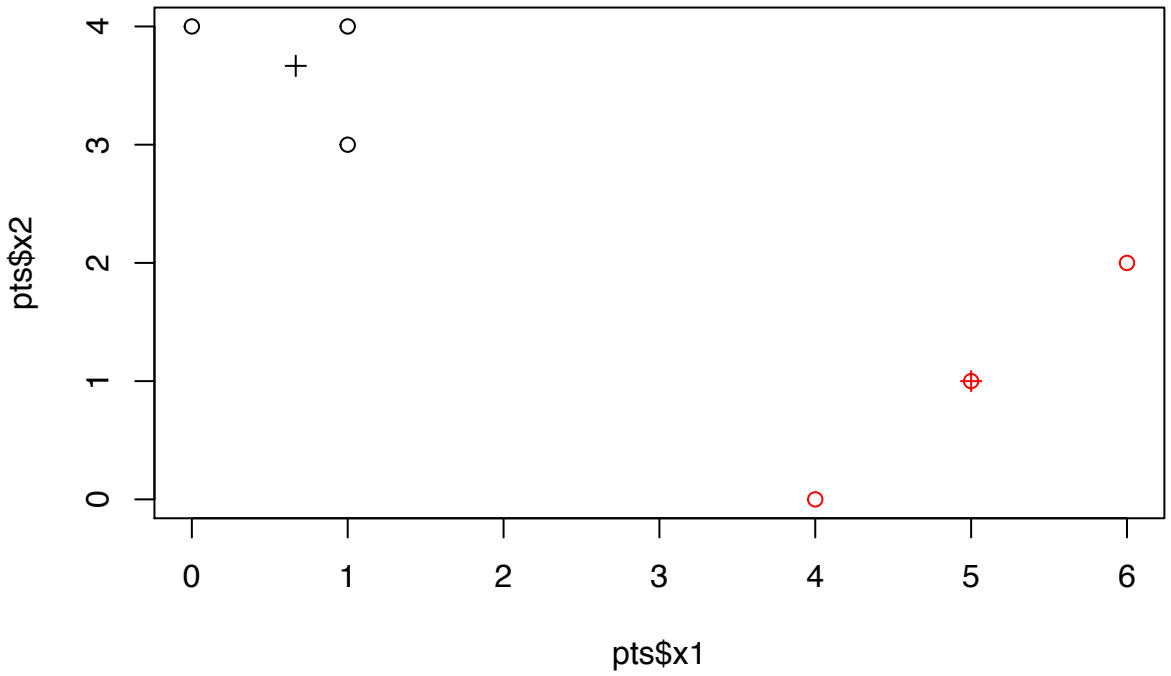
It converges after another iteration:

```
c1 = c(mean(pts[pts$cluster == 1, 2]), mean(pts[pts$cluster == 1, 3]))
c2 = c(mean(pts[pts$cluster == 2, 2]), mean(pts[pts$cluster == 2, 3]))

pts$d1 = ((pts$x1 - c1[1]) ^ 2) + ((pts$x2 - c1[2]) ^ 2)
pts$d2 = ((pts$x1 - c2[1]) ^ 2) + ((pts$x2 - c2[2]) ^ 2)

pts$cluster[(pts$d1 < pts$d2)] = 1
pts$cluster[(pts$d1 >= pts$d2)] = 2

pts
```

```
##   obs x1 x2 cluster      d1   d2
## 1   1  1  4       1  1.5625 24.5
## 2   2  1  3       1  1.0625 18.5
## 3   3  0  4       1  4.5625 32.5
## 4   4  5  1       2 14.0625  0.5
## 5   5  6  2       2 17.5625  4.5
## 6   6  4  0       2 14.5625  0.5
```

2f)

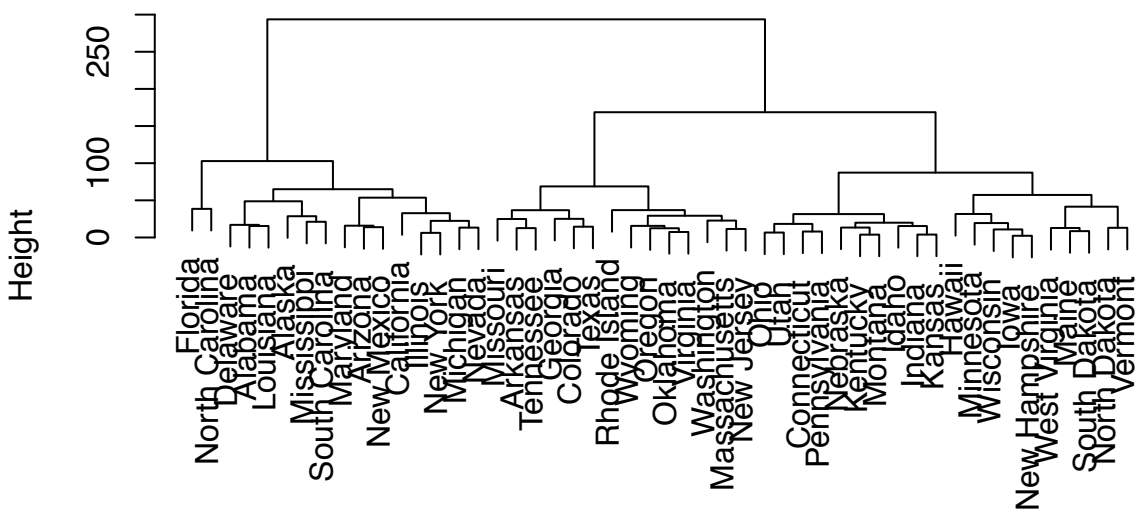
```
plot(pts$x1, pts$x2, col=pts$cluster)

c1 = c(mean(pts[pts$cluster == 1, 2]), mean(pts[pts$cluster == 1, 3]))
c2 = c(mean(pts[pts$cluster == 2, 2]), mean(pts[pts$cluster == 2, 3]))

points(c1[1], c1[2], col=1, pch = 3)
points(c2[1], c2[2], col=2, pch = 3)
```



3) 10.7.9

3a)

```
hc.arrests.complete = hclust(dist(USArrests), method="complete")

plot(hc.arrests.complete, main="Complete Linkage", xlab="", sub="")
```



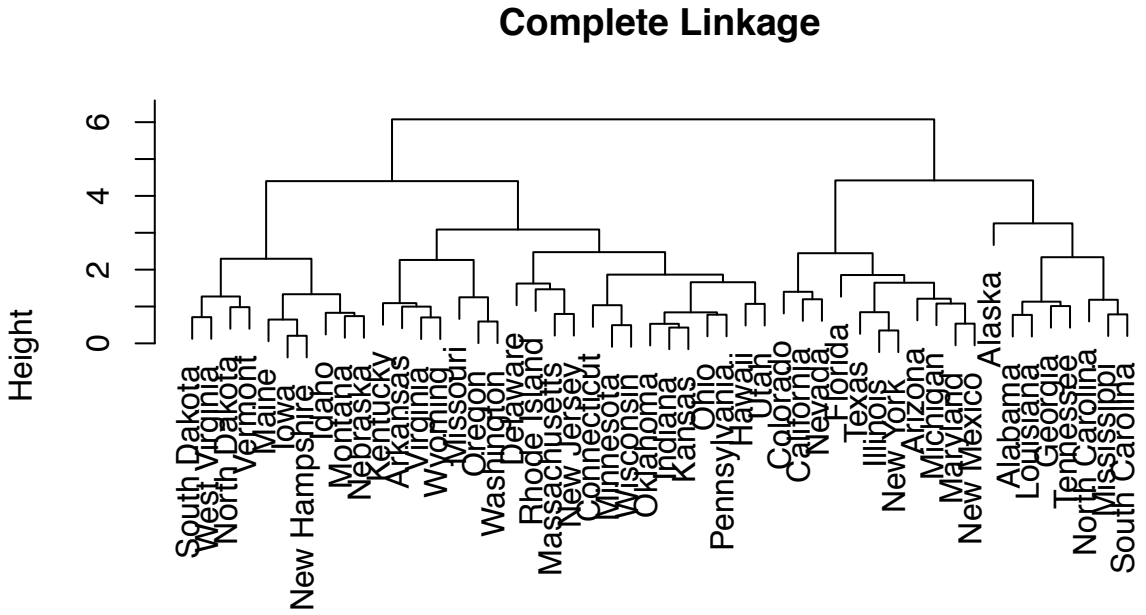
3b)

```
cutree(hc.arrests.complete, 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
```

3c)

```
arrests.sc = scale(USArrests)
hc.arrests.sc.complete = hclust(dist(arrests.sc), method="complete")
plot(hc.arrests.sc.complete, main="Complete Linkage", xlab="", sub="")
```



3d)

Scaling the variables has the effect of balancing the clustering branches and reducing the height of the tree; it is needed because the variables are measured in different units and have different variances. The variance for Assault is much higher than that of the other variables.

```
apply(USArrests , 2, var)
```

```
##     Murder    Assault   UrbanPop       Rape
##   18.97047 6945.16571  209.51878   87.72916
```

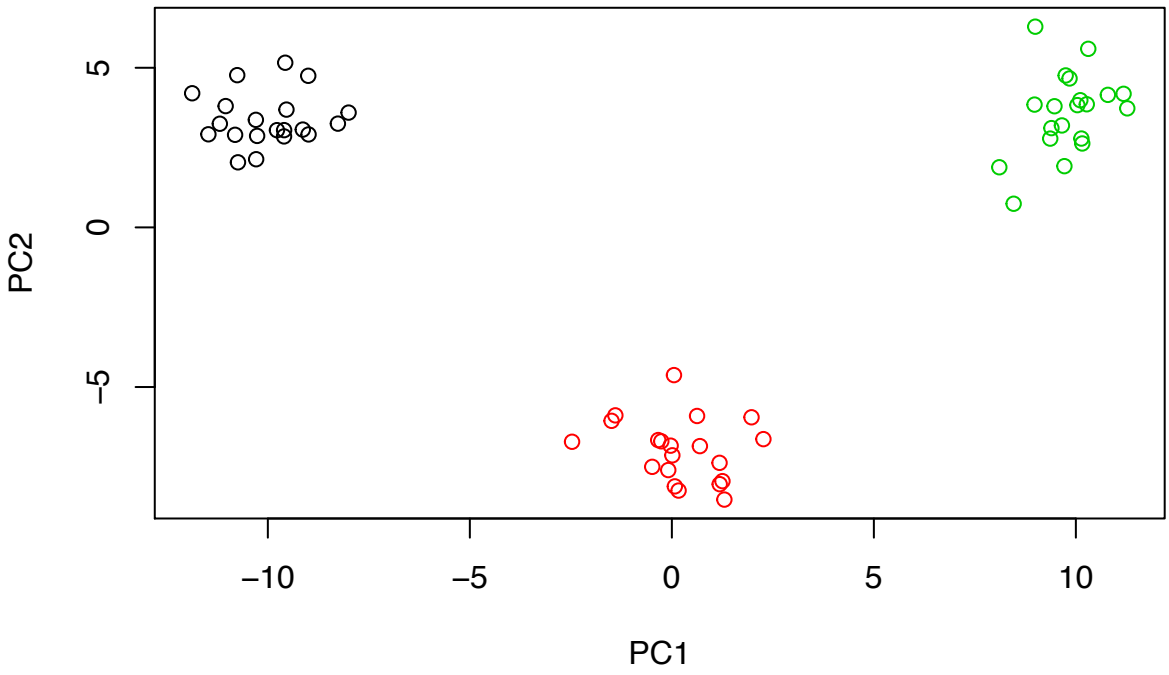
4) 10.7.10

4a)

```
X = matrix(rnorm(60*50), nrow = 60)
X[1:20] = X[1:20] + 10
X[21:40, 2] = X[21:40, 2] + 10
X[41:60, 1] = X[41:60, 1] - 10
```

4b) Perform PCA:

```
pr.out = prcomp(X)

plot(pr.out$x[,1:2], col=cbind(rep(1, 20), rep(2, 20), rep(3, 20)))
```



(4c)

Perform K-means clustering of the observations with K = 3.

```
km.out = kmeans(X, 3, nstart = 20)
table(km.out$cluster, cbind(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1 20  0  0
##   2  0 20  0
##   3  0  0 20
```

The clusters perform really well, and we get 3 separate clusters.

(4d) K = 2

```
km.out = kmeans(X, 2, nstart = 20)
table(km.out$cluster, cbind(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1 20  0  0
##   2  0 20 20
```

There are two clusters, one which combines two of the original clusters.

(4e) K = 4

```
km.out = kmeans(X, 4, nstart = 20)
table(km.out$cluster, cbind(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1  0  7  0
##   2  0  0 20
##   3 20  0  0
##   4  0 13  0
```

One of the original clusters splits into two new clusters while the other two remain intact.

(4f) K = 3, PCA vectors,

```
km.out = kmeans(pr.out$x[,1:2], 3, nstart = 20)
table(km.out$cluster, cbind(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1  0  0 20
##   2 20  0  0
##   3  0 20  0
```

The clusters also perform really well, and we get 3 separate clusters as seen in the original data.

(4g) K = 3, scaled

```
km.out = kmeans(scale(X), 3, nstart = 20)
table(km.out$cluster, cbind(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1  3  9  5
##   2 12  4  9
##   3  5  7  6
```

Scaling worsens the clustering because scaling the observations down to the same stdev shrinks the distance between the observations.

5) 3.7.3

5a)

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

Given fixed values for IQ and GPA, the effect of the coefficients on gender is $(\hat{\beta}_3 + \hat{\beta}_5 X_1)X_3$. If X_1 (GPA) is high enough, then this will be a negative effect for females, where $\hat{\beta}_3 = 35$ and $\hat{\beta}_5 = -10$, e.g if GPA $= 4.0$ then the net effect for females is negative.

5b)

```
50 + (20 * 4.0) + (0.07* 110) + (35*1) + (0.01 * 4.0 * 110) + (-10 * 1 * 4.0)
```

```
## [1] 137.1
```

5c)

False, the significance of a regression term has nothing to do with the size of the coefficient itself, but instead the p-value of the coefficient.

6) 3.7.10

6a)

```
library(ISLR)
fit = lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

6b)

As the price increases by 1, the sales of car seats decreases by 54, and this is a significant relationship.

From the regression p-values, we can see that the UrbanYes variable does not have a significant effect on sales.

If the store is in the US, the sales of car seats increases by 1201, and this is a significant relationship.

For a store that is not in the US and not urban, and if the price is 0, then they sell 13043 car seats.

6c)

$Sales = 13.043469 + -0.054459 * Price + -0.021916 * UrbanYes + 1.200573 * USYes$

6d)

We can reject the null hypothesis for Price and USYes because the p-values are below 0.05.

6e)

```
fit = lm(Sales ~ Price + US, data = Carseats)
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

7) 3.7.14

7a)

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

Linear model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

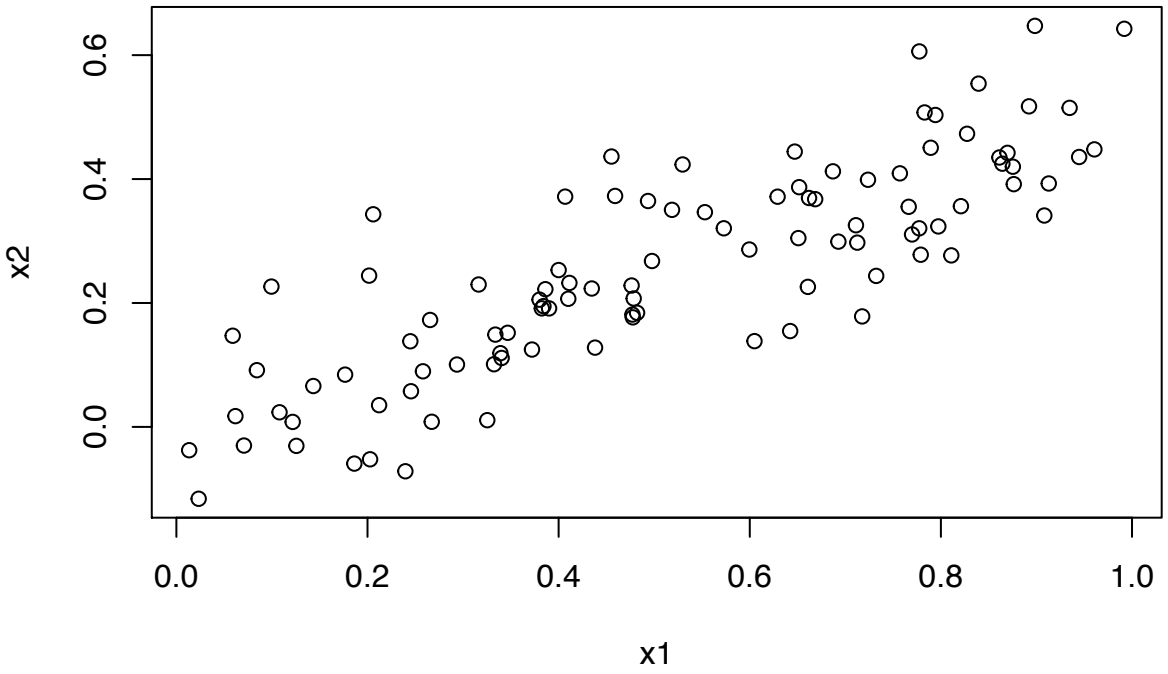
Regression coefficients: $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3$

7b)

```
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
plot(x1, x2)
```



7c)

```
fit = lm(y ~ x1 + x2)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
```

```
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

$\hat{\beta}_0 = 2.1305, \hat{\beta}_1 = 1.4396, \hat{\beta}_2 = 1.0097$

β_0 has the closest estimate to the true model and is also reflected in the low p-value. β_1 has the next closest estimate, and we can reject the null because the p-value is less than 0.05. For β_2 we cannot reject the null because the p-value is greater than 0.05, and we can see that the value is off from the true original value.

d)

```
fit = lm(y ~ x1)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

We can reject the null hypothesis $\beta_1 = 0$ because the p-value for that coefficient is less than 0.05.

e)

```
fit = lm(y ~ x2)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

We can reject the null hypothesis $\beta_2 = 0$ because the p-value for that coefficient is less than 0.05.

f)

The results do not contradict each other because x_1 and x_2 are collinear, and x_2 is generated directly from x_1. Therefore they are both good predictors for y, are unreliable when combined in the linear model. Here we get x_1 to be more significant in the combined model, probably because it has the larger coefficient in the model compared to x_2.

g)

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

Model from pt (c):

```
fit = lm(y ~ x1 + x2)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```
which.max(hatvalues(fit))
```

```
## 101
## 101
```

In this model, x2 has become the significant variable and x1 is no longer significant. We see that the new point did not decrease the R-squared value, so it is not an outlier. We can see that the new point (point #101) has high leverage.

Model from pt (d):

```
fit = lm(y ~ x1)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```
which.max(hatvalues(fit))
```

```
## 27
## 27
```

We can see that the new point decreases the R-squared of the model, so it is an outlier. We can see that the new point (point #101) does not have high leverage for this model.

Model from pt (e):

```
fit = lm(y ~ x2)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

```
which.max(hatvalues(fit))
```

```
## 101
```

```
## 101
```

We can see that the new point does not the R-squared of the model, so it is not an outlier. We can see that the new point (point #101) has high leverage.