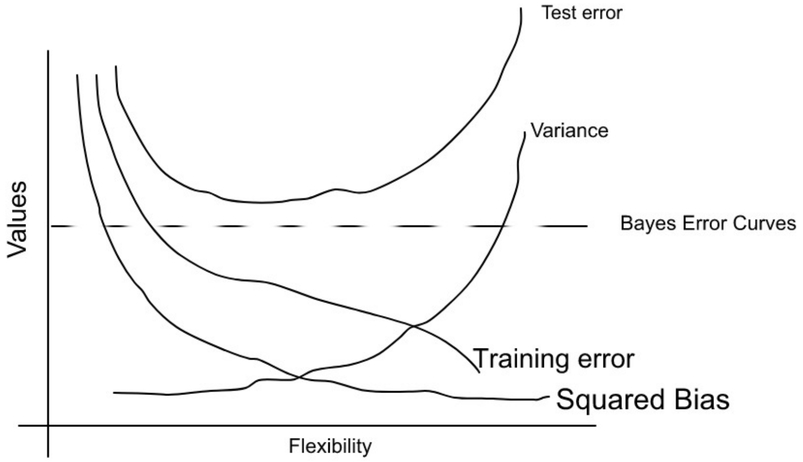
Sam Lin
Stats 202
Hw. 1

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 = $x_{0,}$ but X can only be $x_5$, 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 = $x_0$) = $\frac{1}{3}\sum I(y_i = Red) = \frac{1}{3}(1+0+1) = \frac{2}{3}$

P(Y = Green | X = $x_0$) = $\frac{1}{3}\sum I(y_i = Green) = \frac{1}{3}(0+1+0) = \frac{1}{3}$

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) $\frac{1}{|C_K|}\sum_{i,i'\in C_K}\sum_j (x_{ij} - x_{i'j})^2 = \frac{1}{|C_K|}\sum_{i,i'\in C_K}\sum_j {x_{ij}}^2 - 2x_{ij}x_{i'j} + {x_{i'j}}^2$

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

$\sum_{i,i'\in C_K}\sum_j 2{x_{ij}}^2 - 2x_{ij}x_{kj}$

Given $|C_K|x_{kj} = \sum_{i\in C_K} x_{ij}$, then the simplified form is $\sum_{i\in C_K}\sum_j 2{x_{ij}}^2 - |C_K|\sum_j 2{x_{kj}}^2$

Now the right side of the equation

$$2\sum_{i\in C_K}\sum_{j=1}^{p}(x_{ij} - x_{kj})^2 = 2\sum_{i\in C_K}\sum_{j=1}^{p}(x_{ij})^2 - 4\sum_{i\in C_K}\sum_{j=1}^{p}x_{ij}x_{kj} + 2\sum_{i\in C_K}\sum_{j=1}^{p}{x_{kj}}^2$$

Given $|C_K|x_{kj} = \sum_{i\in C_K} x_{ij}$, then the simplified form is

$$2\sum_{i\in C_K}\sum_{j=1}^{p}(x_{ij})^2 + 2\sum_{i\in C_K}\sum_{j=1}^{p}{x_{kj}}^2 - 4|C_K|\sum_{j=1}^{p}{x_{kj}}^2$$

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

$$2\sum_{i\in C_K}\sum_{j=1}^{p}(x_{ij})^2 - 2|C_K|\sum_{j=1}^{p}{x_{kj}}^2$$

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

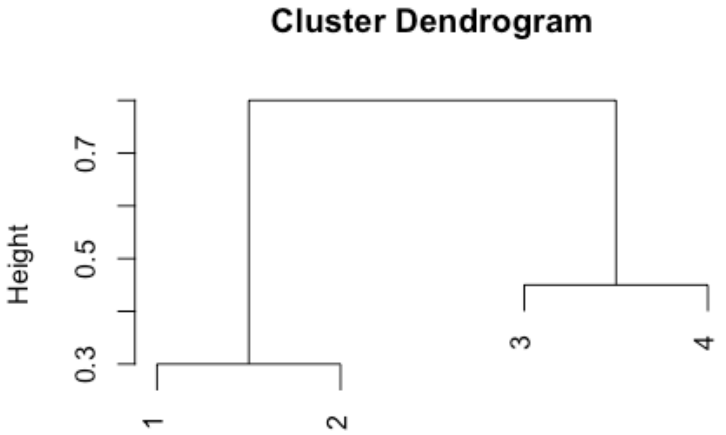
| | (1,2) | 3 | 4 |
|---|---|---|---|
| (1,2) | | 0.5 | 0.8 |
| 3 | 0.5 | | 0.45 |
| 4 | 0.8 | 0.45 | |

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.

Sam Lin
Stats 202
Hw. 1

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 | |



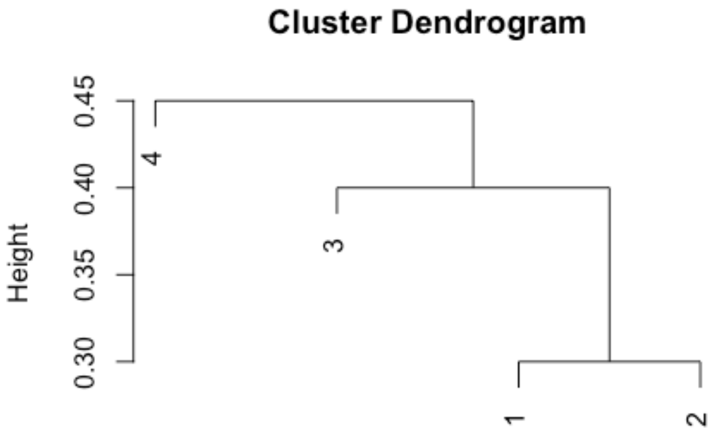
b) For i = 4, most similar pair is observation 1 and 2. Fuse these together for a new dissimilarity matrix. However, this time we will take the minimal inter cluster dissimilarity

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

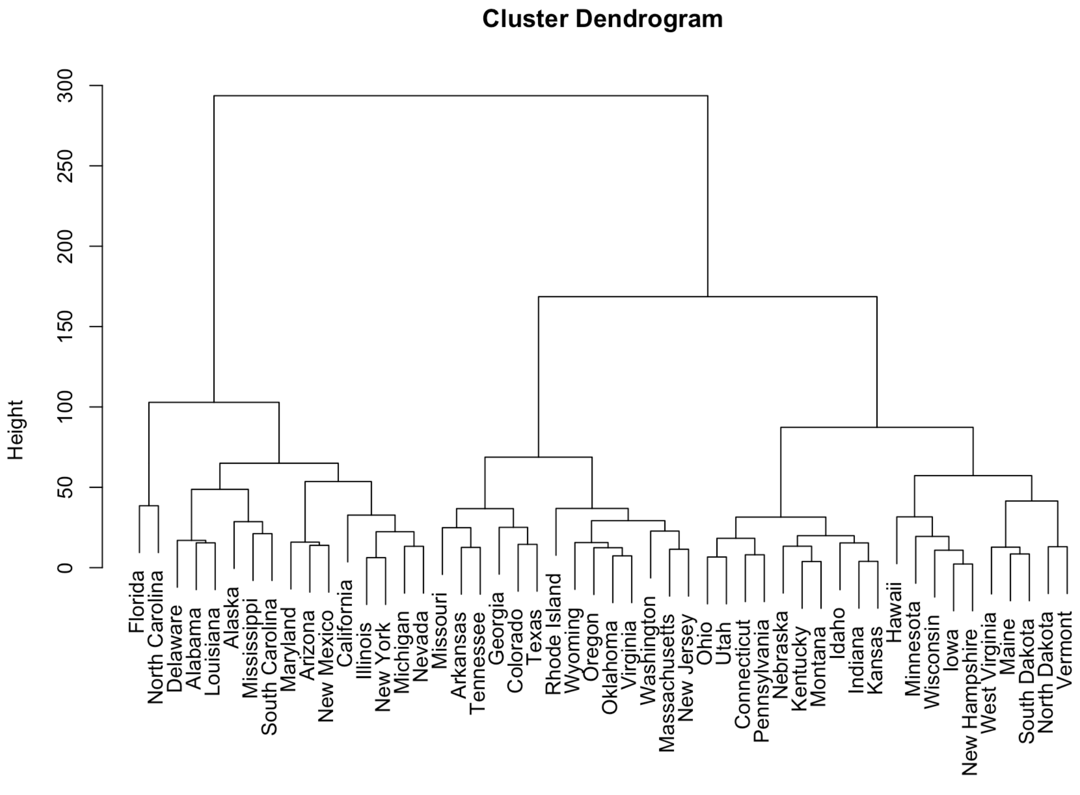
For i = 3, most similar pair is observation (1,2) and 3. The new dissimilarity matrix is

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

Sam Lin
Stats 202
Hw. 1



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)

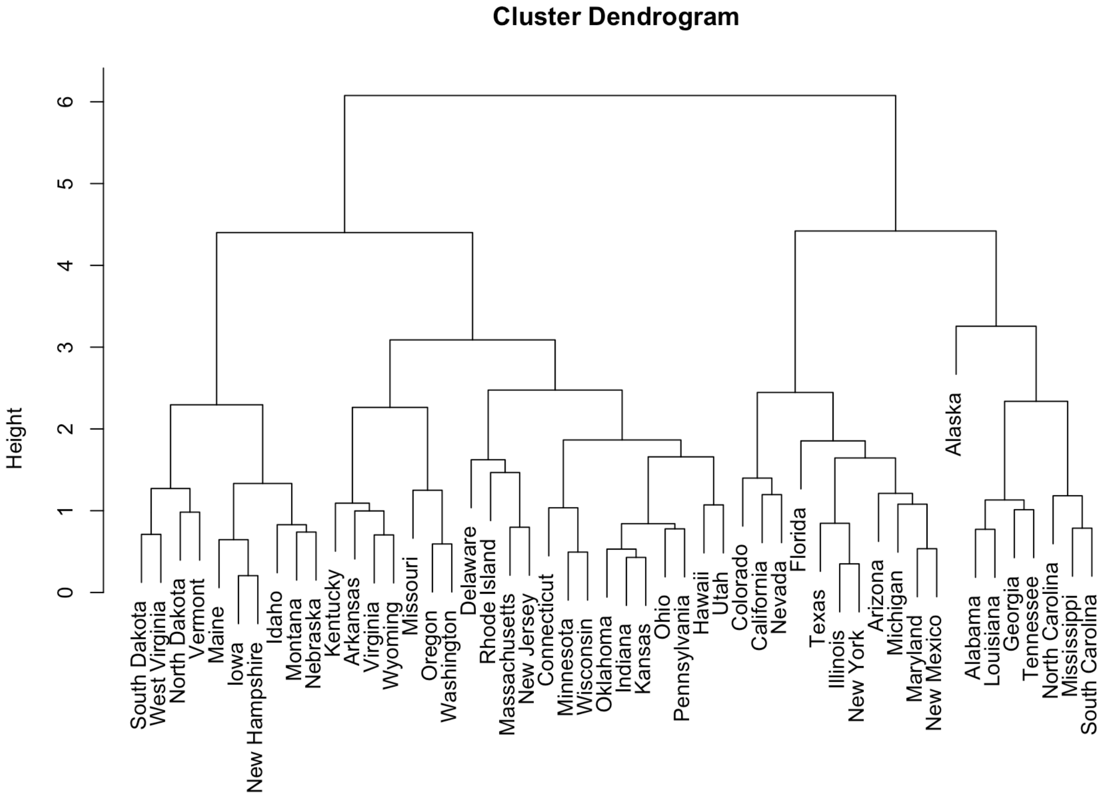


Sam Lin
Stats 202
Hw. 1

b)

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

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)

Sam Lin
Stats 202
Hw. 1

a) There is not enough information. RSS is defined as $e_1^2 + \ldots + e_n^2$ where $e_i = y_i -$ (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.

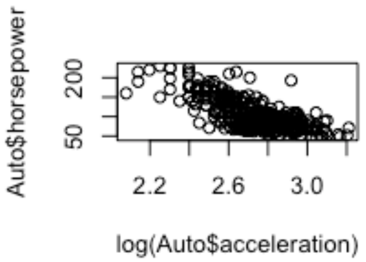
d) Since our test data relies on training data, there is not enough information.

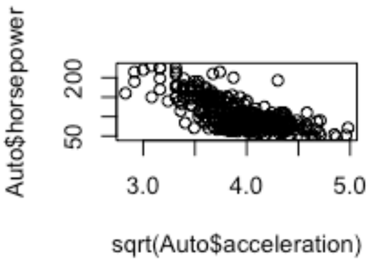
9. (Ex 9e, 9f Pg 122)

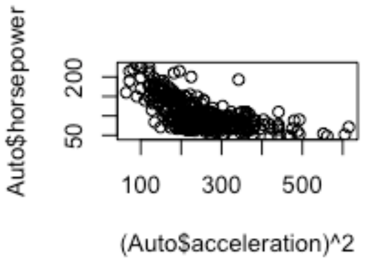
e) Weight and displacement were not as statistically significant as weight and acceleration.

```
> fit <- lm(year ~ weight*acceleration+weight*displacement, data = Auto[,1:8])
> summary(fit)

Call:
lm(formula = year ~ weight * acceleration + weight * displacement,
    data = Auto[, 1:8])

Residuals:
    Min      1Q  Median      3Q     Max
-7.3458 -2.5215  0.0231  2.8886  6.3389

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.550e+01  6.448e+00  13.259   <2e-16 ***
weight              -3.082e-03  2.420e-03  -1.273   0.2036
acceleration        -5.646e-01  3.192e-01  -1.769   0.0777 .
displacement        -9.811e-03  1.039e-02  -0.945   0.3454
weight:acceleration  2.391e-04  1.102e-04   2.169   0.0307 *
weight:displacement -7.449e-07  3.115e-06  -0.239   0.8111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.369 on 386 degrees of freedom
Multiple R-squared:  0.1742,    Adjusted R-squared:  0.1635
F-statistic: 16.29 on 5 and 386 DF,  p-value: 1.386e-14
```

f)



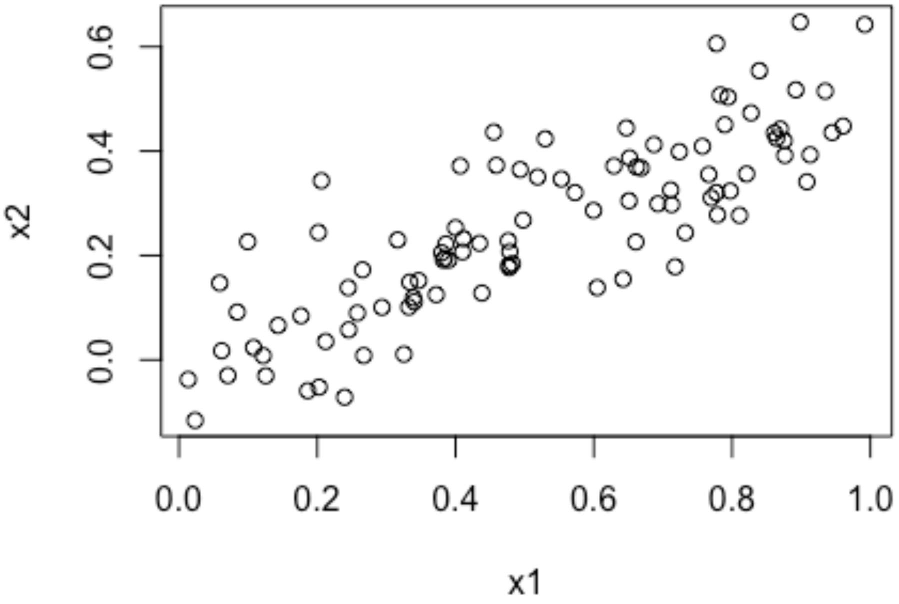




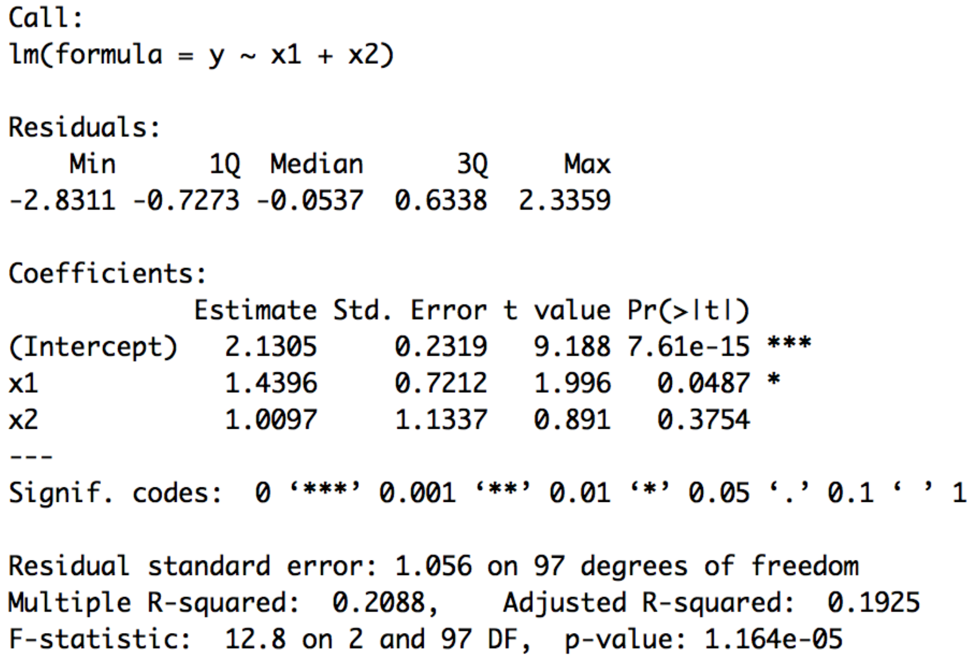
Sam Lin
Stats 202
Hw. 1

10. (Ex 14 Pg 125)
a) Form of linear model: $y = 2 + 2x_1 + 0.3x_2$ + error, where error is a N(0,1) (standard Normal distribution) random variable. Regression coefficients $\beta_i$ for $i = 1,2,3$ 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)

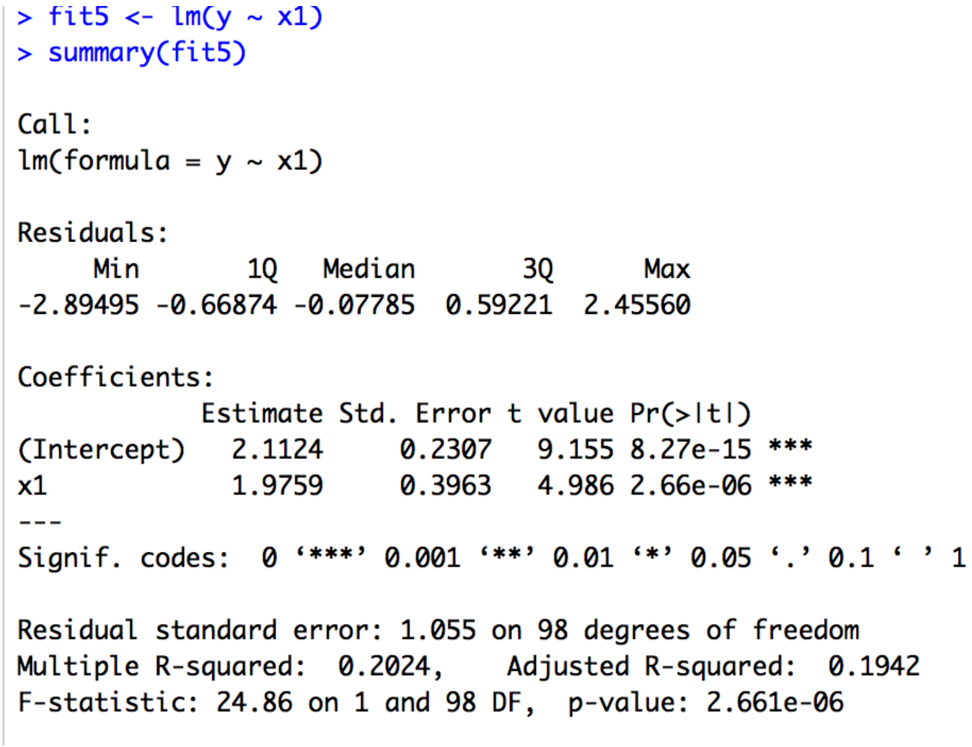


```
Call:
lm(formula = y ~ x1 + x2)

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
```

Sam Lin
Stats 202
Hw. 1

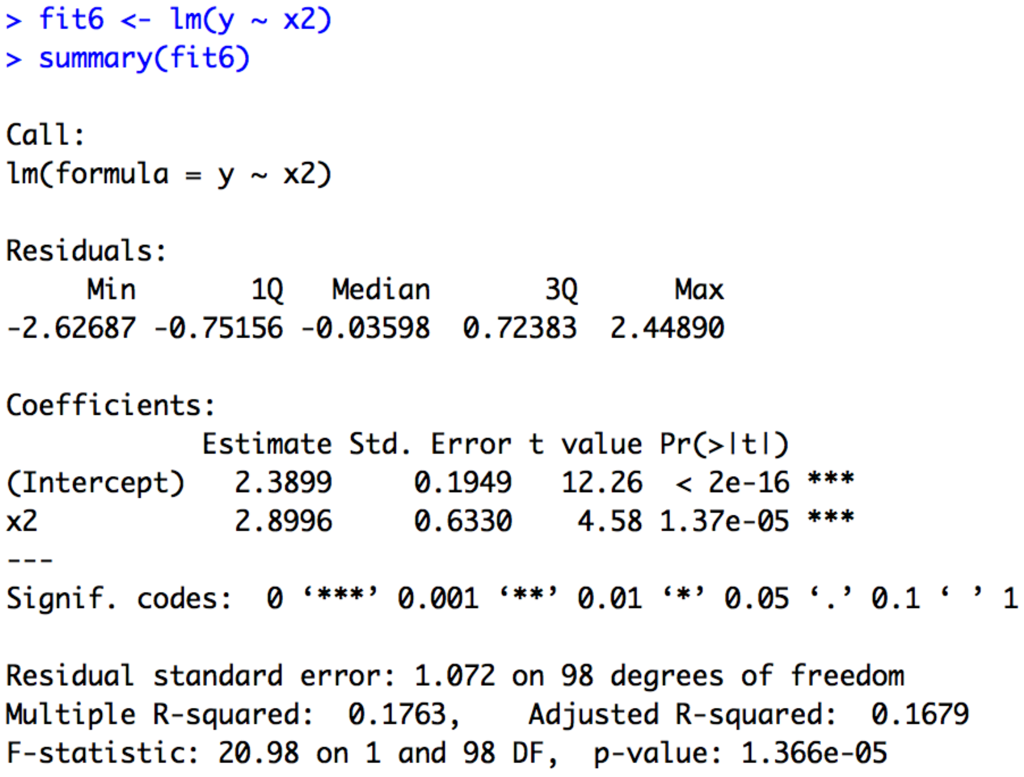
$B_0$, $B_1$, $B_2$ is 2.1305, 1.4396, 1.0097 respectively. Only $B_0$ is close to $\hat{B}_0$.We can reject $H_0$ for $B_1$, because its p-value is less than 0.05, but not for $B_2$ since 0.3754 is not less than 0.5.

d) x1 here is different from the previous question. Since the p-value is very low, we may reject $H_0$.



```
> fit5 <- lm(y ~ x1)
> summary(fit5)

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
```

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

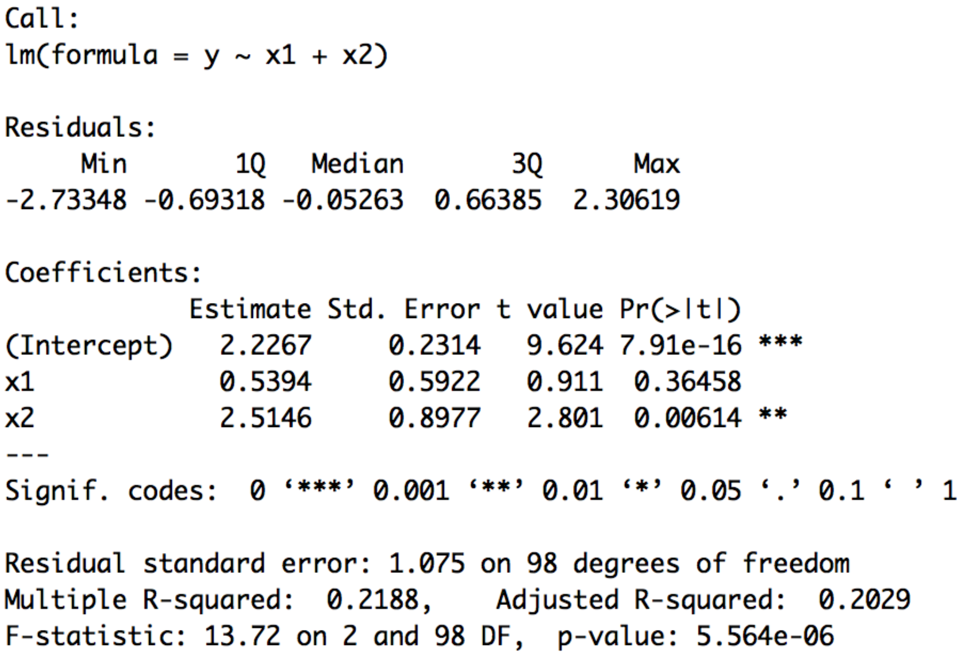


```
> fit6 <- lm(y ~ x2)
> summary(fit6)

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
```

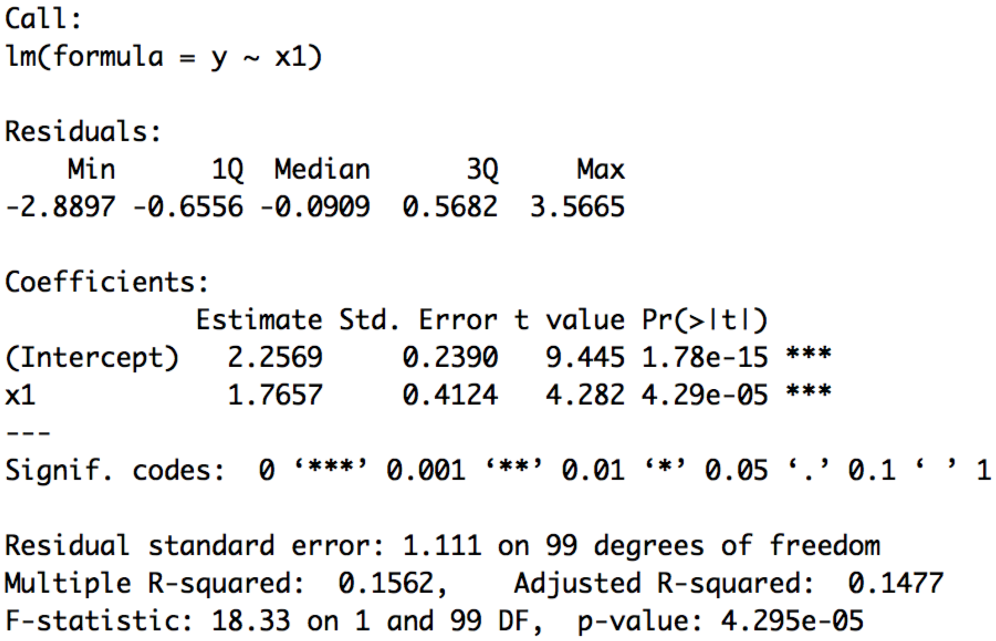
Sam Lin
Stats 202
Hw. 1

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)



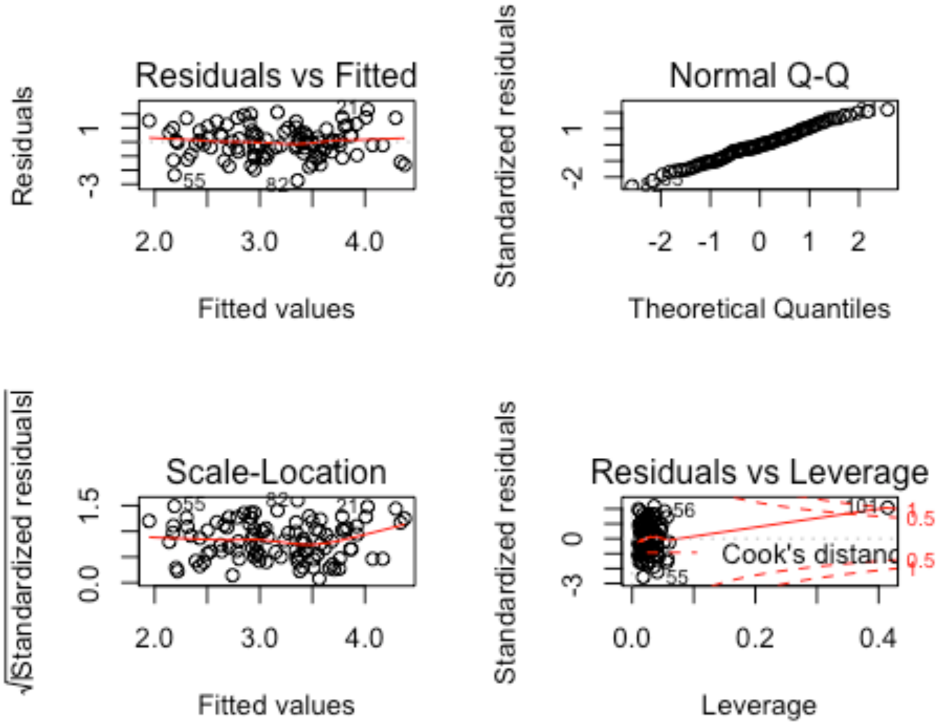
```
Call:
lm(formula = y ~ x1 + x2)

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
```



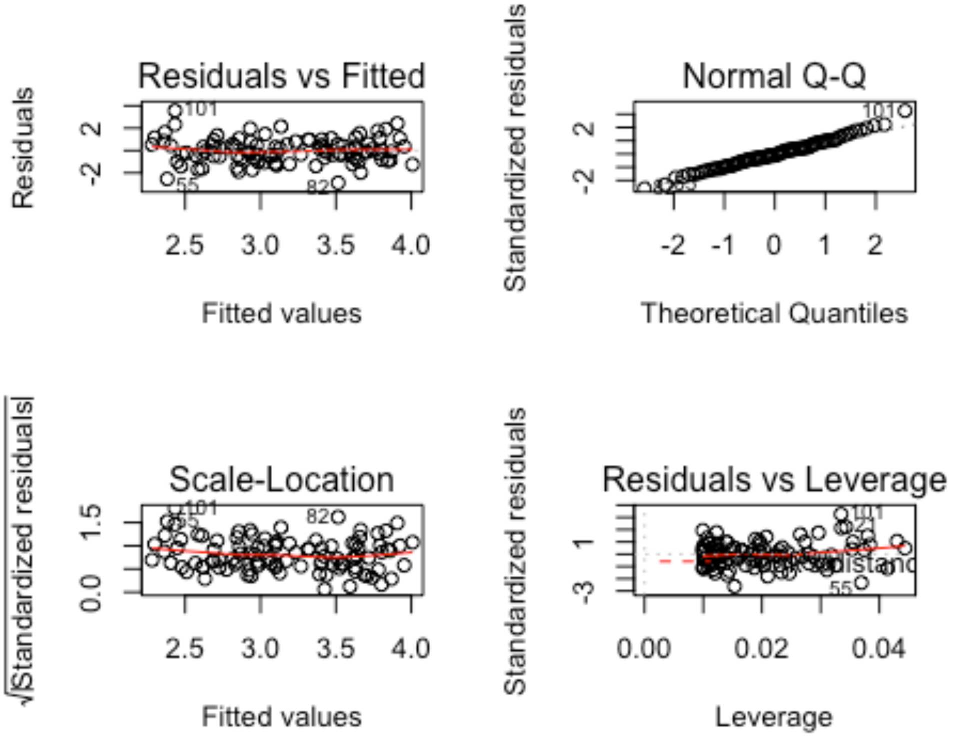
```
Call:
lm(formula = y ~ x1)

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
```

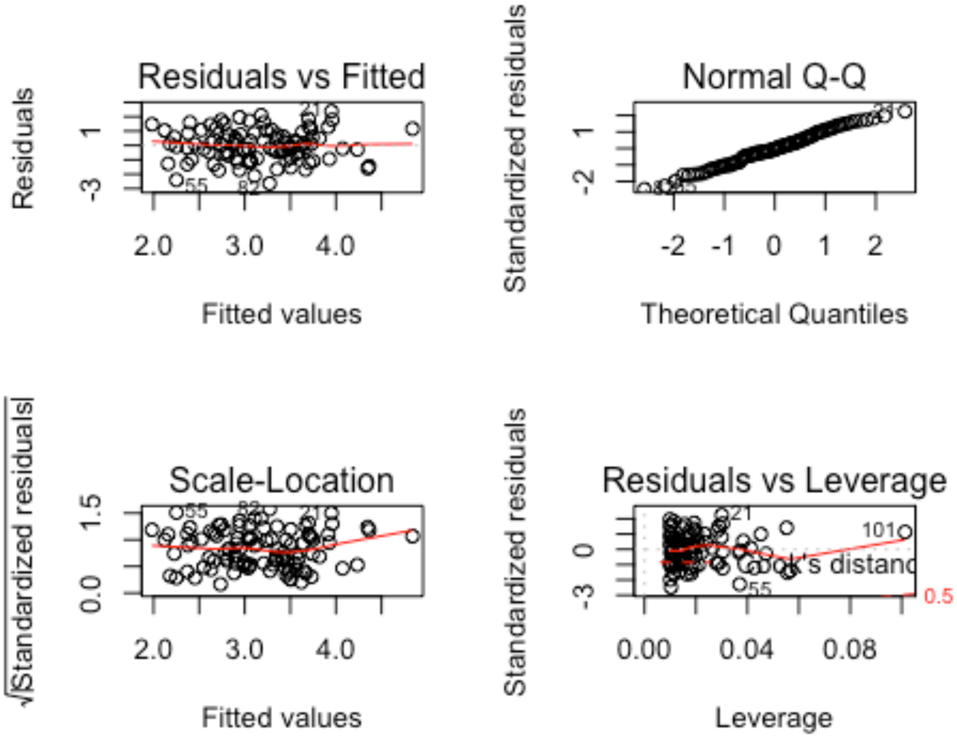
Sam Lin
Stats 202
Hw. 1

```
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
```



Sam Lin
Stats 202
Hw. 1





Sam Lin
Stats 202
Hw. 1

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.