HUDM5126 Linear Models and Regression Analysis Homework 9

Yifei Dong

10/29/2020

0. Data Preparation

```
setwd("~/Documents/Teachers College/Linear Models and Regression/Week 9/hw9/hw9_R")
getwd()
```

```
## [1] "/Users/yifei/Documents/Teachers College/Linear Models and Regression/Week 9/hw9/hw9_R"
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1. KNNL 10.5

Refer to **Brand Preference** Problem 6.5b.

```
data1 <- read.table(paste("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/",
                          "textdatasets/KutnerData/Chapter%20%206%20Data%20Sets/CH06PR05.txt",
                          sep = ""))
data1 <- data1 %>%
  select("Y" = V1, "X1" = V2, "X2" = V3)
data1
```

```
##      Y X1 X2
## 1   64  4  2
## 2   73  4  4
## 3   61  4  2
## 4   76  4  4
## 5   72  6  2
## 6   80  6  4
## 7   71  6  2
## 8   83  6  4
## 9   83  8  2
## 10  89  8  4
## 11  86  8  2
## 12  93  8  4
```

```
## 13  88 10  2
## 14  95 10  4
## 15  94 10  2
## 16 100 10  4
```

a. Prepare an added-variable plot for each of the predictor variables.

```
library(car)
```

```
## Loading required package: carData
```

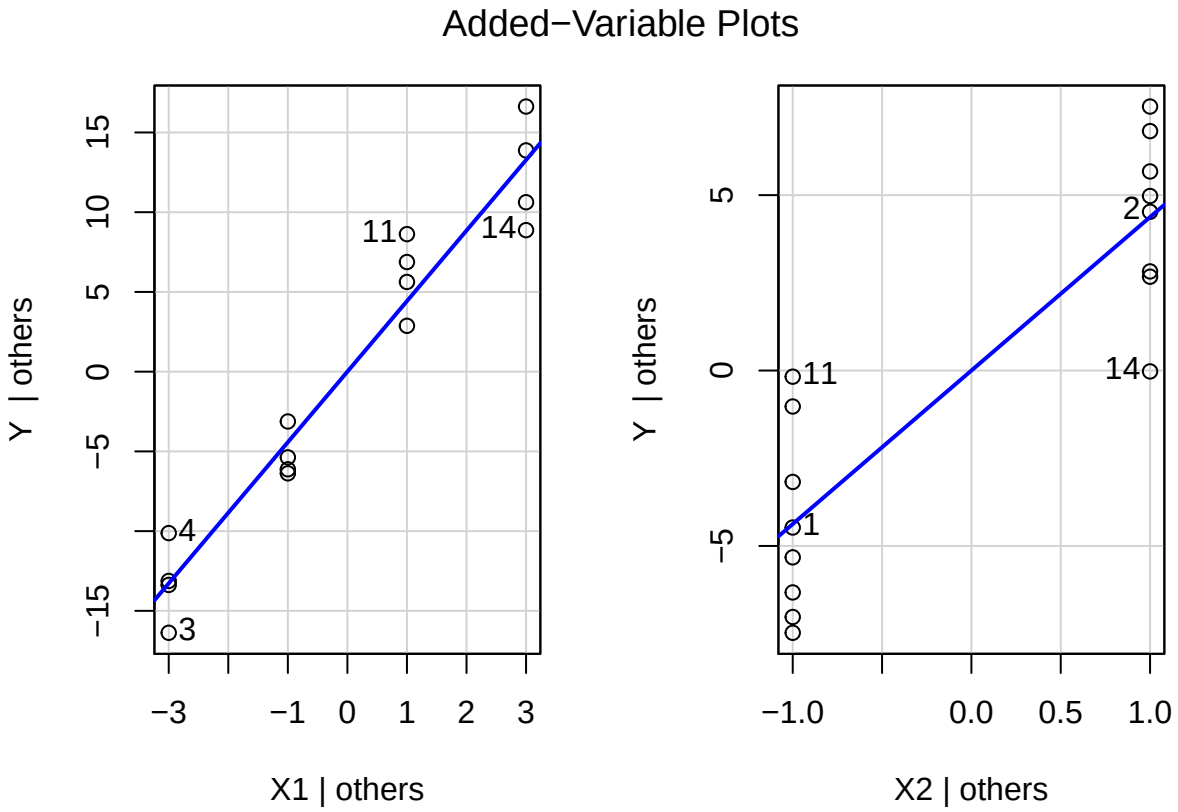
```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
reg <- lm(Y~X1+X2, data = data1)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## X1            4.4250     0.3011  14.695 1.78e-09 ***
## X2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

```
avPlots(reg)
```



b. Do your plots in part (a) suggest that the regression relationships in the fitted regression function in Problem 6.5b are inappropriate for any of the predictor variables. Explain.

I see a tilted straight-line scatterplot for both added variable plot of X_1 controlling for X_2 and added variable plot of X_2 controlling for X_1. Therefore, both two variables are important and appropriate.

c. Obtain the fitted regression function in Problem 6.5b by separately regressing both Y and X_2 on X_1, and then regressing the residuals in an appropriate fashion.

```
regYX1 <- lm(Y~X1, data = data1)
summary(regYX1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.475 -4.688 -0.100  4.638  7.525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.775      4.395  11.554 1.52e-08 ***
## X1             4.425      0.598   7.399 3.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.349 on 14 degrees of freedom
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7818
```

```
## F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

```
regX2X1 <- lm(X2~X1, data = data1)
summary(regX2X1)
```

```
##
## Call:
## lm(formula = X2 ~ X1, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.000e+00  8.783e-01   3.416  0.00418 **
## X1          -2.483e-17  1.195e-01   0.000  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.069 on 14 degrees of freedom
## Multiple R-squared:  6.163e-32,  Adjusted R-squared:  -0.07143
## F-statistic: 8.628e-31 on 1 and 14 DF,  p-value: 1
```

```
# Obtain residuals
eYX1 <- residuals(regYX1)
eX2X1 <- residuals(regX2X1)
# Fitting regression of eYX1 v.s eX2X1
reg2 <- lm(eYX1~eX2X1)
summary(reg2)
```

```
##
## Call:
## lm(formula = eYX1 ~ eX2X1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0000     0.6488   0.000        1
## eX2X1         4.3750     0.6488   6.743 9.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.595 on 14 degrees of freedom
## Multiple R-squared:  0.7646, Adjusted R-squared:  0.7478
## F-statistic: 45.47 on 1 and 14 DF,  p-value: 9.427e-06
```

The slope of the regression through the origin of $e_i(Y|X_1)$ on $e_i(X_2|X_1)$ is the prartial regression coefficient for X_2. It is the same as the β_2 in the fitted regression function in Problem 6.5b, which is 4.375.

2. KNNL 10.9

Refer to **Brand preference** Problem 6.5.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.10$. State the decision rule and conclusion.

```
# Studentized deleted residuals
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
studres(reg)
```

```
##           1           2           3           4           5           6
## -0.04085498  0.06128781 -1.36059879  1.38602483 -0.36694571 -0.66490618
##           7           8           9          10          11          12
## -0.76716157  0.50461264  0.46506694 -0.60436295  1.82302030  0.97784298
##          13          14          15          16
## -1.13966417 -2.10272640  1.48973208  0.24572878
```

Test for outliers (Bonferroni adjustment): Outlier if:

$$|t_i| \geqslant t\big(1 - (\frac{\alpha}{2n}), n - p - 1\big)$$

```
# Cut off point
(n = nrow(data1)) # 16
```

```
## [1] 16
```

```
# df = n-p-1 = 16-3-1 = 12
qt(0.9969, 12)
```

```
## [1] 3.31212
```

```
# [1] 3.31212
```

```
which(abs(studres(reg)) >= 3.31212)
```

```
## named integer(0)
```

Therefore, we conclude there are no outliers based on the Bonferroni outlier test.

b. Obtain the diagonal elements of the hat matrix, and provide an explanation for the pattern in these elements.

```
# We obtain X matrix first
attach(data1)
X = cbind(rep(1, nrow(data1)), X1, X2)
X
```

```
##         X1 X2
##  [1,] 1  4  2
##  [2,] 1  4  4
```

```
##  [3,] 1  4  2
##  [4,] 1  4  4
##  [5,] 1  6  2
##  [6,] 1  6  4
##  [7,] 1  6  2
##  [8,] 1  6  4
##  [9,] 1  8  2
## [10,] 1  8  4
## [11,] 1  8  2
## [12,] 1  8  4
## [13,] 1 10  2
## [14,] 1 10  4
## [15,] 1 10  2
## [16,] 1 10  4
```

```
# Hat matrix
H = X %*% solve(t(X) %*% X) %*% t(X)
# diagonal elements of Hat matrix
diag(H)
```

```
##  [1] 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
## [11] 0.1375 0.1375 0.2375 0.2375 0.2375 0.2375
```

h_{ii} is a measure of the distance between the X values for the ith case and the means of the X values for all n cases. Thus, a large value h_{ii} indicates that the ith case is distant from the center of all X observations. The diagonal element h_{ii} in this context is called the leverage (in terms of the X values) of the ith case. In this case, we have two groups of equally influential obervations, which are 0.2375 and 0.1375.

c. Are any of the observations outlying with regard to their X values according to the rule of thumb stated in the chapter?

```
# Cut off point
2*(3/n)
```

```
## [1] 0.375
```

```
# [1] 0.375
```

```
which(diag(H) > 2*(3/n))
```

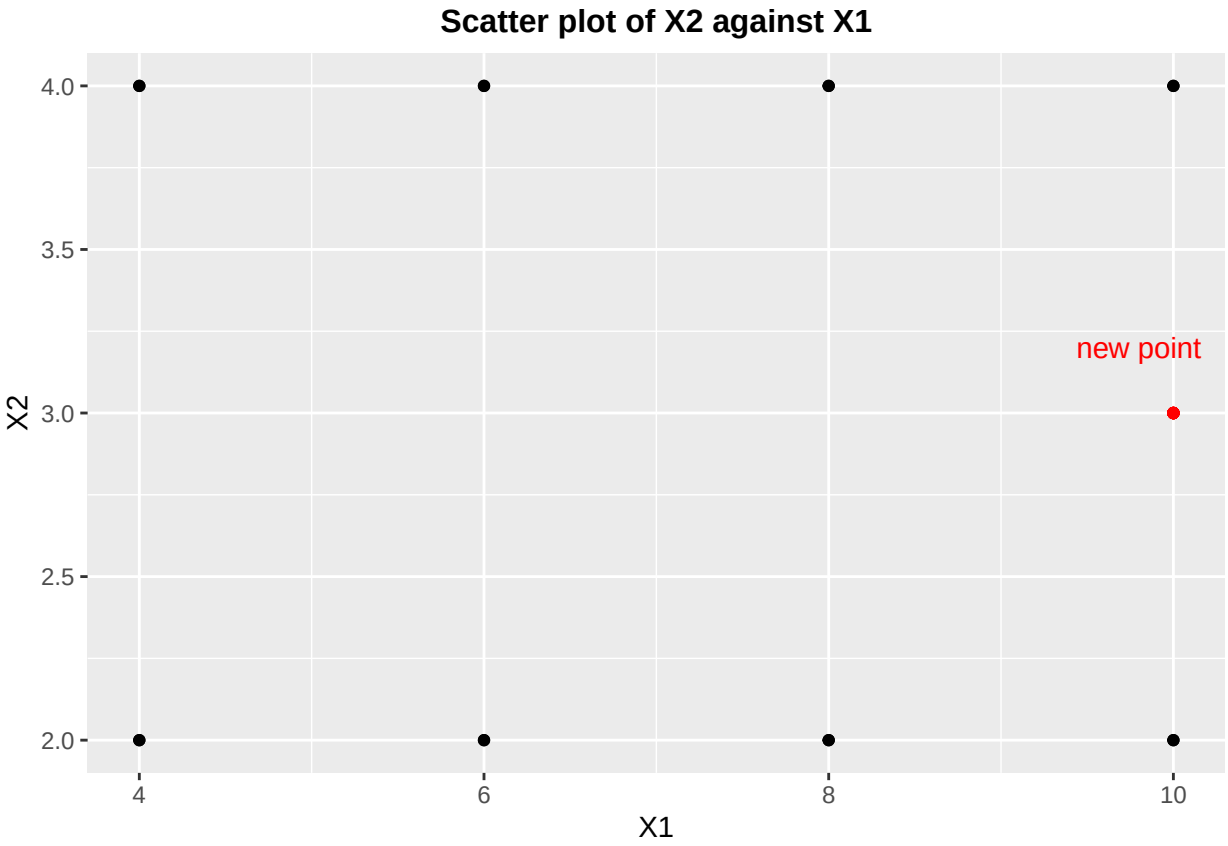
```
## integer(0)
```

The rule of thumb suggests that points with a hat diagonal element greater than $\frac{2p}{n}$ should be considered high leverage points. Here, $\frac{2p}{n} = 0.375$ and no observations outlying with regard to their X values.

d. Management wishes to estimate the mean degree of brand liking for moisture content $X_1 = 10$ and sweetness $X_2 = 3$. Construct a scatter plot of X_2 against X_1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, use (10.29) to determine whether an extrapolation is involved. Do your conclusions from the two methods agree?

```
# scatter plot of X2 against X1
library(ggplot2)
g1 <- ggplot(data1, aes(x = X1, y = X2))+
  geom_point()+
  geom_point(aes(x = 10, y = 3), color = "red")+
  annotate("text", label = "new point", x = 9.8, y = 3.2, color = "red")+
```

```
  ggtitle("Scatter plot of X2 against X1")+
  theme(plot.title = element_text(size = 12, face = "bold", color = "black", hjust = 0.5))
g1
```



Based on the scatterplot, the prediction $(X_1 = 10, X_2 = 3)$ is not outlying beyond the range of the data.

Recall, to spot hidden extrapolations, we can utilize the direct leverage calculations in (10.18) for the new set of X values for which inferences are to be made:

$$h_{\text{new.new}} = \mathbf{X}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\text{new}} \tag{1}$$

where $\mathbf{X}_{\text{new}}$ is the vector containing the X values for which an inference about a mean response or a new observation is to be made, and the $\mathbf{X}$ matrix is the one based on the data set used for fitting the regression model. If $h_{\text{new.new}}$ is well within the range of leverage values h_{ii} for the cases in the data set, no extrapolation is involved. On the other hand, if $h_{\text{new.new}}$ is much larger than the leverage values for the cases in the data set, and extrapolation is indicated.

```
# Obtain X vector
Xnew <- rbind(1, 10, 3)
# h new new
hnewnew <- t(Xnew) %*% solve(t(X) %*% X) %*% Xnew
hnewnew
```

```
##       [,1]
## [1,] 0.175
```

Since $h_{\text{new.new}} = 0.175$ is well within the range of leverage values h_{ii} for the cases in the data set, I conclude no extrapolation is involved.

e. The largest absolute studentized deleted residual is for case 14. Obtain the **DFFITS**, **DFBETAS**, and Cook's distance values for this case to assess the influence of this case. What do you conclude?

```
# First, find the largest absolute studentized deleted residual
which.max(abs(studres(reg))) # confirm case 14
```

```
## 14
## 14
```

```
# Built-in function for influential observations
im = influence.measures(reg)
im
```

```
## Influence measures of
##   lm(formula = Y ~ X1 + X2, data = data1) :
##
##      dfb.1_  dfb.X1  dfb.X2   dffit cov.r   cook.d   hat inf
## 1  -0.02155  0.0157  0.0117 -0.0228 1.667 0.000188 0.238
## 2   0.00868 -0.0235  0.0175  0.0342 1.666 0.000422 0.237
## 3  -0.71785  0.5226  0.3895 -0.7593 1.084 0.180392 0.237
## 4   0.19619 -0.5324  0.3968  0.7735 1.068 0.186258 0.238
## 5  -0.11987  0.0442  0.0988 -0.1465 1.426 0.007666 0.138
## 6   0.02413  0.0800 -0.1790 -0.2655 1.322 0.024547 0.138
## 7  -0.25062  0.0924  0.2065 -0.3063 1.277 0.032297 0.138
## 8  -0.01832 -0.0607  0.1358  0.2015 1.384 0.014354 0.138
## 9   0.07315  0.0560 -0.1252  0.1857 1.397 0.012231 0.138
## 10  0.12431 -0.0728 -0.1627 -0.2413 1.347 0.020406 0.138
## 11  0.28674  0.2195 -0.4907  0.7279 0.708 0.149828 0.138
## 12 -0.20113  0.1177  0.2632  0.3904 1.171 0.050983 0.138
## 13  0.01467 -0.4378  0.3263 -0.6360 1.225 0.131821 0.237
## 14  0.83881 -0.8077 -0.6020 -1.1735 0.651 0.363412 0.237
## 15 -0.01917  0.5722 -0.4265  0.8314 1.002 0.210661 0.237
## 16 -0.09802  0.0944  0.0704  0.1371 1.643 0.006758 0.237
```

```
# Obtain DFFITS, DFBETAS and Cook's distance values for this case
# DFFITS
im$infmat[14, 4]
```

```
## [1] -1.173531
```

```
# We have a small sized data set, the cut off point for DFFITS can be 1
abs(im$infmat[14, 4]) > 1
```

```
## [1] TRUE
```

This value is somewhat larger than our guideline of 1, we might consider the observation to be influential. However, the value is close enough to 1 that the case may not be influential enough to require remedial action.

```
# DFBETAS
# b0
im$infmat[14, 1]
```

```
## [1] 0.8388068
```

```
# b1
im$infmat[14, 2]
```

```
## [1] -0.8076796
```

```r
# b2
im$infmat[14, 3]
```

```
## [1] -0.6020088
```

```r
# Cut-off point for DFBETAS is
(co.dfb = 2/sqrt(n))
```

```
## [1] 0.5
```

```r
# check if case 14 influences b0
abs(im$infmat[14, 1]) > co.dfb
```

```
## [1] TRUE
```

```r
# check if case 14 influences b1
abs(im$infmat[14, 2]) > co.dfb
```

```
## [1] TRUE
```

```r
# check if case 14 influences b2
abs(im$infmat[14, 3]) > co.dfb
```

```
## [1] TRUE
```

We conclude that 14th observation does influence both β_1 and β_2 if we use cut off point $\frac{2}{n}$. However, the textbook recommend considering a case influential if the absoulte value of **DFBETAS** exceeds 1 for small to medium data sets and $\frac{2}{n}$ for large data sets (p.405). For this small data set, it looks that the **DFBETAS** values do not exceed 1 so that case 14 may not be so influential.

```r
# Cook's distance
im$infmat[14, 6]
```

```
## [1] 0.3634123
```

```r
# cut off point for Cook's distance
qf(0.5, 3, 13)
```

```
## [1] 0.8315874
```

14th observation's Cook's distance is 0.3634, which is smaller than the cut off point 0.8316. This case may not be very influential.

f. Calculate the average absolute percent difference in the fitted values with and without case 14. What does this measure indicate about the influence of case 14?

```r
# Fitted values with case 14
fit1 <- fitted(reg)
fit1
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 64.10 72.85 64.10 72.85 72.95 81.70 72.95 81.70 81.80 90.55 81.80 90.55 90.65
##    14    15    16
## 99.40 90.65 99.40
```

```r
# Fitted values without case 14
reg2 <- lm(Y~X1+X2, data = data1[-14, ])
fit2 <- predict(reg2, newdata = data1)
fit2
```

```
##         1         2         3         4         5         6         7         8
##  63.45082  72.92213  63.45082  72.92213  72.73361  82.20492  72.73361  82.20492
##         9        10        11        12        13        14        15        16
##  82.01639  91.48770  82.01639  91.48770  91.29918 100.77049  91.29918 100.77049
```

```
# average absolute percent difference
sum(abs((fit2-fit1)/fit1)*100)/n
```

```
## [1] 0.677679
```

This mean difference is 0.68%. On the basis of this direct evidence about the effect of case 14 on the inferences to be made, it was satisfied that case 14 does not exercise undue influence so that no remedial action is required for handling this case.

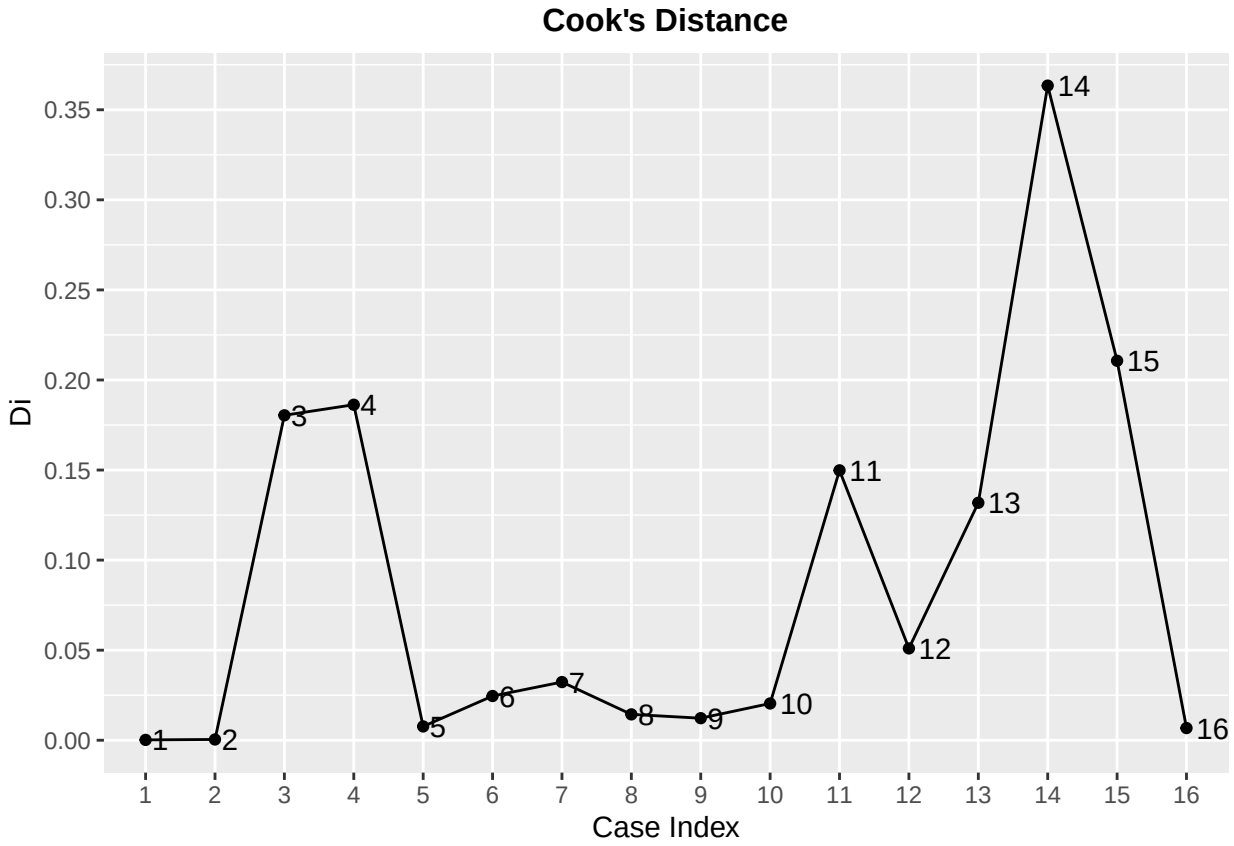
g. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

```
# Extract Cook's distance for each case
data2 <- as.data.frame(im$infmat[, 6])
# change colum names
data2 <- data2 %>%
  dplyr::select("Di" = `im$infmat[, 6]`)
data2
```

```
##              Di
## 1  0.0001877130
## 2  0.0004223542
## 3  0.1803921815
## 4  0.1862582123
## 5  0.0076655286
## 6  0.0245466787
## 7  0.0322971439
## 8  0.0143542862
## 9  0.0122308711
## 10 0.0204060192
## 11 0.1498281704
## 12 0.0509831969
## 13 0.1318214458
## 14 0.3634123447
## 15 0.2106609008
## 16 0.0067576676
```

```
# Create index
index = seq(1, 16, by = 1)
data2 <- cbind(data2, index)
```

```
# Prepare an index plot
g2 <- ggplot(data2, aes(x = as.factor(index), y = Di, label = row.names(data2)))+
  geom_point()+
  geom_text(hjust = -0.2, nudge_x = 0.05)+
  geom_line(group = 1)+
  scale_x_discrete("Case Index")+
  scale_y_continuous("Di", breaks = seq(0, 0.4, by = 0.05))+
  ggtitle("Cook's Distance")+
  theme(plot.title = element_text(size = 12, face = "bold", color = "black", hjust = 0.5))
g2
```



No other case influential based on this measure. Case 14 has the highest Cook's distance value, and then case 15 and case 4.