Labs: Trees, Hierarchical Clustering, Outlier detection, Cook's Distance

Project Dataset One-on-One with the instructor

Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 3 - Lab 6, 21st October 2022

Work on the remaining scripts (Group 3)

- https://rpi.box.com/s/2xx9ul1fmc6bf5ff8h4jreae69emikmf

- **What is expected**: You are asked to Explore, Inspect the code/scripts that are on Group 3 in the Rstudio environment and get familiar with those scripts.

Scripts – work through these

Reminder to finish these code examples See in folder group2/ Lab1

Go over the following scrips,

Lab1_bronx1.R.

Lab1_bronx2.R

Lab1_ctree2.R

Lab1_kknn1.R

Lab1_kknn2.R

Lab1_kknn3.R

Lab1_kmeans1.R

Lab1_nyt.R

Search before you ask! You might need to search your code errors online when you are debugging your code!.

script fragments in R available on:

https://rpi.box.com/s/2xx9ul1fmc6bf5ff8h4jreae69emikmf

NOTE: you are allowed to work in small groups and discuss during this lab.

Scripts – work through these

Next…

See in folder group2/ Lab3

Go over the following scrips,

Lab3_ctree1.R

Lab3_ctree2.R

Lab3_ctree3.R

…..

And the remaining code snippets in group2/Lab 2 and Lab3

Search before you ask! You might need to search your code errors online when you are debugging your code!

script fragments in R available on:
https://rpi.box.com/s/2xx9ul1fmc6bf5ff8h4jreae69emikmf

NOTE: you are allowed to work in small groups and discuss during this lab.

Scripts – work through these

Next…

See in folder group2 and group3/ Labs

Go over the following scrips,

Lab3_ctree1.R

Lab3_ctree2.R

Lab3_ctree3.R

…..

And the remaining code snippets in group2/Lab 2 and Lab3

Search before you ask! You might need to search your code errors online when you are debugging your code!

script fragments in R available on:
https://rpi.box.com/s/2xx9ul1fmc6bf5ff8h4jreae69emikmf

NOTE: you are allowed to work in small groups and discuss during this lab.

Outlier Detection

- Outliers are atypical (extreme) values.
- Although they are not errors (sometimes!)
- We need to pay careful attention to outliers.
- Sometimes, it is necessary to identify them and remove from the data and this is when we are looking for typical cases.
- Our objective is to see the performance at the extremes.
- Outliers provide very valuable and useful information.

Using Interquartile Range (IQR) to detect Outliers

- IQR is the measure of the variability, that is more robust than the standard deviation for skewed data.
- IQR is the difference between the Q3 – Q1

  (Q1=25th percentile and Q3=75th)
- It is important to check the Interquartile Range (IQR) for your data.
- Check how many data points are outside the IQR. Easiest way to check IQR is by generating boxplots, but this visual inspection is questionable and very subjective from person to person
- So we will use a metric generated using IQR.

Boxplot

- Boxplot is a useful graphical tool for mapping the data behavior in the middle as well as at the end of the distributions.
- Boxplots uses the median (Q2) and the lower quartile and upper quartile (alternative names are)

lower quartile= Q1=25th percentile

upper quartile = Q3= 75th percentile

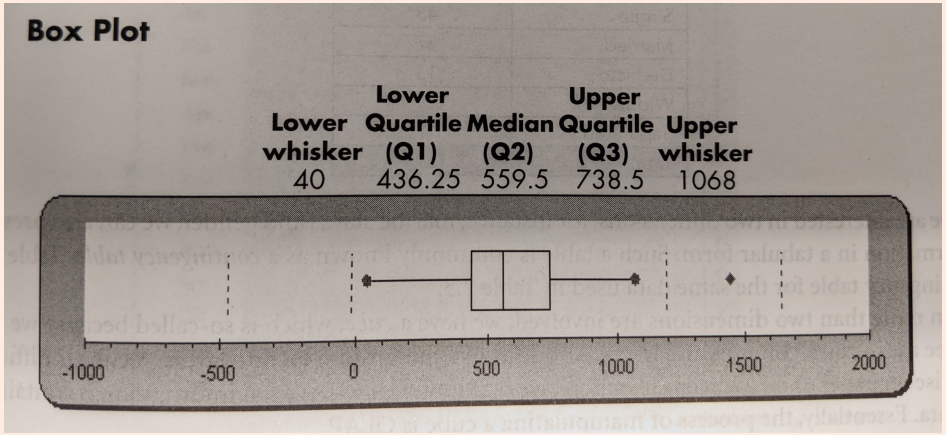


Image Credit/Resource: Data Mining Methods and Techniques Textbook

Outliers: outside the fence

- Boxplot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median value.
- Following quantities are also called "***fences".***
- ***"Fences"*** are essential for identifying the extreme values in the tails of the distributions.

  Lower Inner Fence: Q1 – 1.5 x (IQR)

  Lower Inner Fence: Q3 + 1.5 x (IQR)

Rule of Thumb

- Values are outside the Lower Inner Fence and Uppers Inner Fence can be considered as extreme values.
- Note: existing many outlier points indicates that the dataset may not be suitable for reliable modeling.
- There are other more prominent methods to detect outliers than using IQR lower and inner fencing technique.
- Cook's Distance is also used to identify the outlier points in data.

Cook’s Distance

- Cook’s distance is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis.
- Cook's distance can be used in several ways:
  [1] to indicate influential data points that are particularly worth checking for validity;
  [2] or to indicate regions of the design space where it would be good to be able to obtain more data points.

Cook's Distance

- Each element in the Cook's distance *D* is the normalized change in the fitted response values due to the deletion of an observation. The Cook's distance of observation *i* is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\,MSE},$$

Where:

- $\hat{y}_j$ is the j^{th} fitted response value.
- $\hat{y}_j(i)$ is the j^{th} fitted response value, where the fit does not include observation *i*.
- *MSE* is the mean squared error.
- *p* is the number of coefficients in the regression model.

Regression Model using lm()

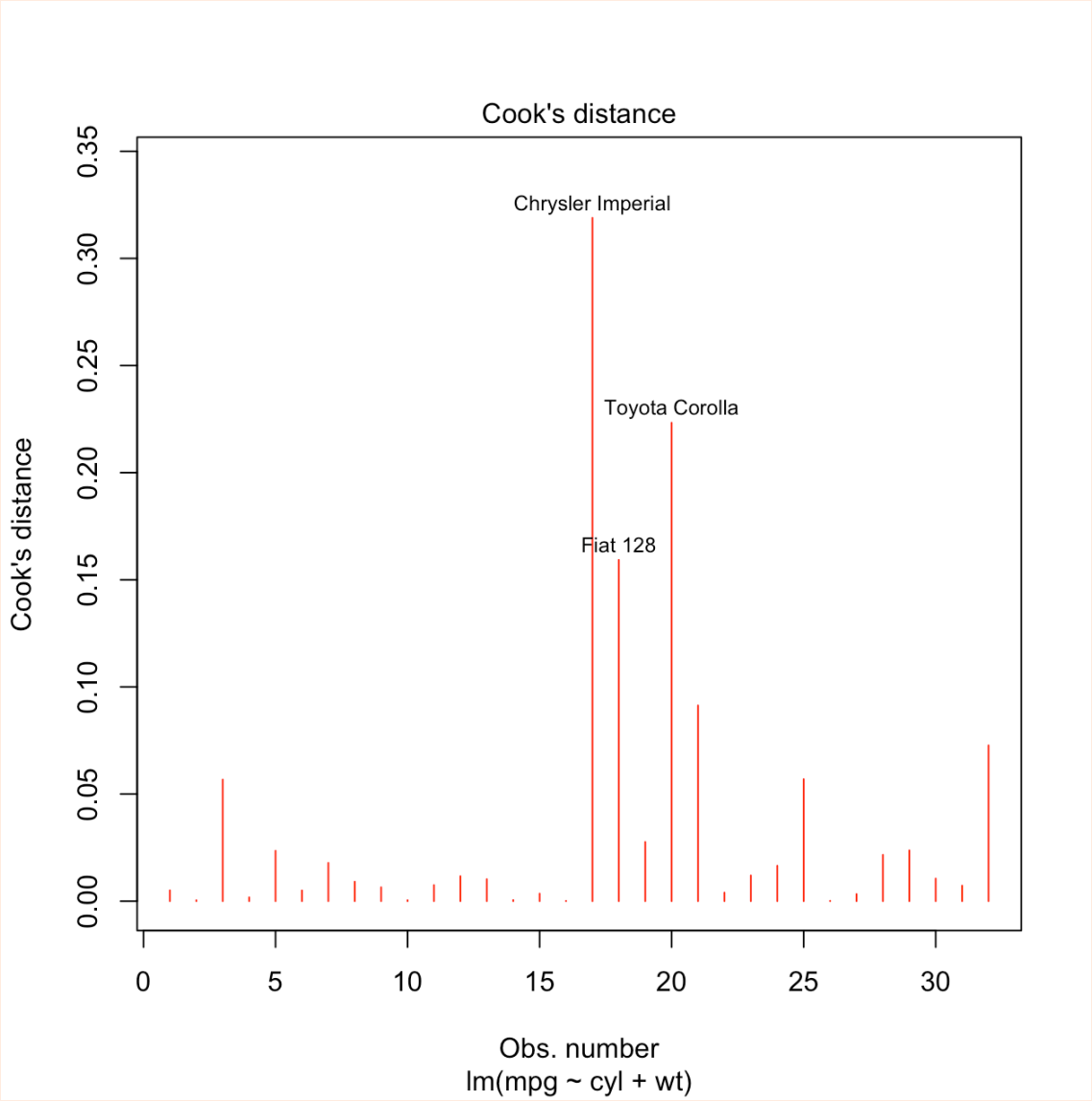
We can generate the regression model using lm() function,
Our response variable is mile per gallon (mpg)
and predictor variables are number of cylinders (cyl) and weight (wt)

```
# Cook's Distance' example using mtcars
mtcars
head(mtcars)
str(mtcars)
model1 <- lm(mpg ~ cyl + wt , data = mtcars)
model1
help("cooks.distance")
plot(model1, pch = 18, col= 'red', which = c(4))

# we can use the cooks.distance() function to identify the Cook's distance to
# each observation
cooks.distance(model1)
CooksDistance <- cooks.distance(model1)
```

Interpretation

Cook’s Distance visualization for the regression model: lm(mpg ~ cyl +wt)



cooks.distance() function

We can use the cooks.distance() function to identify the Cook's distance to each observation in our dataset

```
> cooks.distance(model1)
          Mazda RX4         Mazda RX4 Wag            Datsun 710        Hornet 4 Drive  Hornet Sportabout             Valiant 
       0.0050772590        0.0004442585        0.0567764620        0.0018029260        0.0235271472        0.0050205614 
         Duster 360            Merc 240D             Merc 230             Merc 280            Merc 280C           Merc 450SE 
       0.0178733213        0.0091033181        0.0065061176        0.0004643600        0.0075293380        0.0116847953 
         Merc 450SL          Merc 450SLC  Cadillac Fleetwood Lincoln Continental   Chrysler Imperial            Fiat 128 
       0.0102875723        0.0005228914        0.0035498738        0.0001501537        0.3189363624        0.1592990291 
         Honda Civic       Toyota Corolla        Toyota Corona   Dodge Challenger          AMC Javelin           Camaro Z28 
       0.0276449872        0.2233281268        0.0913548207        0.0040263378        0.0120218543        0.0165559199 
     Pontiac Firebird            Fiat X1-9       Porsche 914-2        Lotus Europa      Ford Pantera L        Ferrari Dino 
       0.0569730451        0.0001790454        0.0033281614        0.0216355209        0.0237336584        0.0105550987 
       Maserati Bora          Volvo 142E 
       0.0072685192        0.0727399065 
```

```r
# we can use the cooks.distance() function to identify the Cook's distance to
# each observation
cooks.distance(model1)
CooksDistance <- cooks.distance(model1)

# Now we will round off the values to 5 desimal points so that it is easy to read
# we can use the round() function to round off values in R.
 round(CooksDistance, 5)
```

```r
# Now we will round off the values to 5 desimal points so that it is easy to read
# we can use the round() function to round off values in R.
 round(CooksDistance, 5)
```

```
> CooksDistance
          Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive Hornet Sportabout             Valiant          Duster 360
       0.0050772590        0.0004442585        0.0567764620        0.0018029260        0.0235271472        0.0050205614        0.0178733213
          Merc 240D           Merc 230           Merc 280           Merc 280C          Merc 450SE          Merc 450SL         Merc 450SLC
       0.0091033181        0.0065061176        0.0004643600        0.0075293380        0.0116847953        0.0102875723        0.0005228914
 Cadillac Fleetwood Lincoln Continental   Chrysler Imperial            Fiat 128          Honda Civic       Toyota Corolla       Toyota Corona
       0.0035498738        0.0001501537        0.3189363624        0.1592990291        0.0276449872        0.2233281268        0.0913548207
    Dodge Challenger         AMC Javelin          Camaro Z28      Pontiac Firebird           Fiat X1-9      Porsche 914-2       Lotus Europa
       0.0040263378        0.0120218543        0.0165559199        0.0569730451        0.0001790454        0.0033281614        0.0216355209
     Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E
       0.0237336584        0.0105550987        0.0072685192        0.0727399065
```

Interpretation

# we can sort the values in ascending order
sort(round(CooksDistance, 5))

```
>  sort(round(CooksDistance, 5))
Lincoln Continental          Fiat X1-9      Mazda RX4 Wag           Merc 280       Merc 450SLC     Hornet 4 Drive      Porsche 914-2
            0.00015            0.00018            0.00044            0.00046            0.00052            0.00180            0.00333
 Cadillac Fleetwood    Dodge Challenger            Valiant        Mazda RX4           Merc 230      Maserati Bora          Merc 280C
            0.00355            0.00403            0.00502            0.00508            0.00651            0.00727            0.00753
          Merc 240D         Merc 450SL       Ferrari Dino        Merc 450SE        AMC Javelin        Camaro Z28         Duster 360
            0.00910            0.01029            0.01056            0.01168            0.01202            0.01656            0.01787
       Lotus Europa   Hornet Sportabout     Ford Pantera L        Honda Civic          Datsun 710   Pontiac Firebird         Volvo 142E
            0.02164            0.02353            0.02373            0.02764            0.05678            0.05697            0.07274
     Toyota Corona           Fiat 128     Toyota Corolla   Chrysler Imperial
            0.09135            0.15930            0.22333            0.31894
```

Outlier Detection: Cook's Distance

Lab / In-Class work:

```
# Outlier Detection using "Cooks Distance"
# Multivariate Regression using Cook's Distance
# Cook's Distance is an estimate of the influence of a data point.
# Cook's Distance is a summary of how much a regression model changes when the ith observation is removed from the data.
library(ISLR)
# Let's look at the baseball hitters dataset in ISLR package.
head(Hitters)
dim(Hitters)
is.na(Hitters) # check for the missing values.
# Now we will remove the NA (missing values) using the na.ommit() function
HittersData <- na.omit(Hitters)
dim(HittersData) # checking the dimensions after removing the NAs.
glimpse(HittersData)
head(HittersData)
# Now we will implement a multivariate regression model using all the features in the dataset to
# predict the salary of the Baseball player
SalaryPredictModel1 <- lm(Salary ~., data = HittersData)
summary(SalaryPredictModel1)
# Multiple R-squared:  0.5461,  Adjusted R-squared:  0.5106
```

```r
# Cook's Distance.
cooksD <- cooks.distance(SalaryPredictModel1)
influential <- cooksD[(cooksD > (3 * mean(cooksD, na.rm = TRUE)))]
influential
# We see that 18 players have a Cook's Distance greater than 3x the mean.
# Let's exclude these 18 players and rerun the model to see if we have a better fit in our regression model.
names_of_influential <- names(influential) # checking the names of the influential/outlier players
names_of_influential
outliers <- HittersData[names_of_influential,]
Hitters_Without_Outliers <- HittersData %>% anti_join(outliers)

# Model 2: without the outliers
SalaryPredictModel2 <- lm(Salary ~. , data = Hitters_Without_Outliers)
summary(SalaryPredictModel2)
# Multiple R-squared:  0.6721,  Adjusted R-squared:  0.6445
# We have improved from an Adjusted R-Squared of 0.5106 to 0.6445 with the removal of only 18 observations
```

Trees for the Titanic

Reminder:

data(Titanic)

rpart, ctree, hclust, for:
Survived ~ .

Read the titanic dataset documentation in Rdocumentation:
https://www.rdocumentation.org/packages/titanic/versions/0.1.0

Feedback: Assignment 5 Presentation

- Make sure to meet with the instructor during today's class One-on-One time to get the Assignment 5 (Project Proposal Presentation) Feedback.
- You **MUST** get the Assignment 5 feedback from the instructor before you continue to work on the project.
- It is important to get the feedback and discuss your project with the instructor.

Assignment 4

Reminder: Assignment 4 is available on LMS.