Homework 8 with lab

Madhu Jagdale

Prepared for ITMD/ITMS/STAT 514, Spring 2021

Packages

```
library(tidyverse)  # includes tibbles, ggplot2, dyplr, and more.
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
library("ggplot2")
library("tigerstats")
library(dplyr)
```

In addition, I'd like to ask `R` to print decimal numbers with 2 digits:

```
options(scipen=2)
```

Part I: Tests for a cybersecurity data set

Let's revisit cybersecurity breach report data downloaded 2015-02-26 from the US Health and Human Services. From the *Office for Civil Rights* of the *U.S. Department of Health and Human Services*, I obtained the following information:

"As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals.

"Since October 2009 organizations in the U.S. that store data on human health are required to report any incident that compromises the confidentiality of 500 or more patients / human subjects (45 C.F.R. 164.408). These reports are publicly available. Our data set was downloaded from the Office for Civil Rights of the U.S. Department of Health and Human Services, 2015-02-26."

Load this data set and store it as `cyber.data`, using the following code:

```
cyber.data<-read.csv(url("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBrea
str(cyber.data)
```

```
## 'data.frame':    1151 obs. of  10 variables:
##  $ X                              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name.of.Covered.Entity         : chr  "Brooke Army Medical Center" "Mid America Kidney Stone Ass
##  $ State                          : chr  "TX" "MO" "AK" "DC" ...
##  $ Covered.Entity.Type            : chr  "Healthcare Provider" "Healthcare Provider" "Healthcare Pr
```

```
##  $ Individuals.Affected            : int  1000 1000 501 3800 5257 857 6145 952 5166 5900 ...
##  $ Breach.Submission.Date          : chr  "2009-10-21" "2009-10-28" "2009-10-30" "2009-11-17" ...
##  $ Type.of.Breach                  : chr  "Theft" "Theft" "Theft" "Loss" ...
##  $ Location.of.Breached.Information: chr  "Paper/Films" "Network Server" "Other, Other Portable Elec
##  $ Business.Associate.Present      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Web.Description                 : chr  "A binder containing the protected health information (PHI
```

As you know, this data set contains *all* reports regarding health information data breaches from 2009 to 2015. Let's pretend this is just a *sample* from the population of *all data breaches*, related or not to health information.

Question 1.

Compare the number of individuals affected by data breaches (column `Individuals.Affected`) in two states, Arkansas (`State=="AR"`) and California (`State=="CA"`). This can be done by performing a test of difference in means, for example. Repeat the same test for another pair of states, California ("CA") and Illinois ("IL").

Please note, in order to answer this question completely, you will need to run several lines of code, extract subsets of the data appropriately, run a statistical hypothesis test, and interpret the results. Draw a conclusion. Partial answers to the question will are insufficient.

```
table( cyber.data$State )
```

```
##
##  AK  AL  AR  AZ  CA  CO  CT  DC  DE  FL  GA  HI  IA  ID  IL  IN  KS  KY  LA  MA
##   5  17   7  27 128  20  18   6   1  69  41   1   7   3  57  37   7  26   9  35
##  MD  ME  MI  MN  MO  MS  MT  NC  ND  NE  NH  NJ  NM  NV  NY  OH  OK  OR  PA  PR
##  17   1  25  27  24   6   6  34   3   6   4  17  11   9  72  34   8  16  45  28
##  RI  SC  SD  TN  TX  UT  VA  VT  WA  WI  WV  WY
##   7  13   2  33 100  11  22   1  28  11   5   4
```

```
Individuals.AR <- cyber.data[which(cyber.data$Individuals.Affected & cyber.data$State == 'AR'), ]
count(Individuals.AR)
```

```
##   n
## 1 7
```

```
Individuals.CA <- cyber.data[which(cyber.data$Individuals.Affected & cyber.data$State == 'CA'), ]
count(Individuals.CA)
```

```
##     n
## 1 128
```

```
set.seed(7)
my.sample <- sample(1:nrow(Individuals.AR), 7)
my.sample
```

```
## [1] 2 3 4 6 7 1 5
```

```r
set.seed(7)
my.sample1 <- sample(1:nrow(Individuals.CA), 7)
my.sample1
```

```
## [1]  42  83  31  92 103  66  15
```

```r
m <- mean(my.sample)
m1 <-mean(my.sample1)
mean_diff <- mean(m)-mean(m1)
mean_diff
```

```
## [1] -57.71429
```

```r
t.test(my.sample,my.sample1)
```

```
##
##  Welch Two Sample t-test
##
## data:  my.sample and my.sample1
## t = -4.5917, df = 6.0509, p-value = 0.003645
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -88.40762 -27.02095
## sample estimates:
## mean of x mean of y
##   4.00000  61.71429
```

[Above hypothesis is for the number of individuals affected by data breaches (column `Individuals.Affected`) in two states, Arkansas (`State=="AR"`) and California (`State=="CA"`).Mean difference is -57.71429 which is true mean of this population lies in this interval of -88.40762 and -27.02095 for 95% of the times. p-value = 0.003645]

```r
Individuals.CA1 <- cyber.data[which(cyber.data$Individuals.Affected & cyber.data$State == 'CA'), ]
count(Individuals.CA1)
```

```
##     n
## 1 128
```

```r
Individuals.IL <- cyber.data[which(cyber.data$Individuals.Affected & cyber.data$State == 'IL'), ]
count(Individuals.IL)
```

```
##    n
## 1 57
```

```r
set.seed(50)
my.sample <- sample(1:nrow(Individuals.CA1), 50)
my.sample
```

```
##  [1] 112  11  52  95 125 114  46 119  67   8  16  18  91  21 116  84  63  56  37
## [20]  98  71  28  93  90  10  57  62  32  13 109  89  34  47  31 104   7  85  82
## [39]  68  87  70  39  72   6  40  53  80   4  26  17
```

```r
set.seed(50)
my.sample1 <- sample(1:nrow(Individuals.IL), 50)
my.sample1
```

```
##  [1] 48 11 52 31 54 50 46  3  8 16 18 27 21 51 20 37 34  7 28 29 26 10 32 13 25
## [26]  2 15 36 56 44 40 45 12 47  4  6 39 49 22  5 43 24 23 30  1 33 35 19 41 17
```

```r
m <- mean(my.sample)
m1 <-mean(my.sample1)
mean_diff <- mean(m)-mean(m1)
mean_diff
```

```
## [1] 31.4
```

```r
t.test(my.sample,my.sample1)
```

```
##
##  Welch Two Sample t-test
##
## data:  my.sample and my.sample1
## t = 5.6731, df = 67.718, p-value = 3.168e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  20.35452 42.44548
## sample estimates:
## mean of x mean of y
##      59.1      27.7
```

[Above hypothesis is for the number of individuals affected by data breaches (column `Individuals.Affected`) in two states, Illinois (`State=="IL"`) and California (`State=="CA"`).Mean difference is 31.4 which is true mean of this population lies in this interval of 20.35452 and 42.44548 for 95% of the times. p-value = 3.168e-07]

Question 2.

Explore the variable `Type.Of.Breach` collected in this data set:

- What proportion of data entries in `cyber.data` have `Type.of.Breach  == "Hacking/IT Incident"` ?

```r
prop.table(table(cyber.data$Type.of.Breach == "Hacking/IT Incident"))
```

```
##
##      FALSE       TRUE
## 0.93310165 0.06689835
```

[Total 6.6% of the data is `Type.of.Breach  == "Hacking/IT Incident"`]

- What are all the different values of `Type.Of.Breach` reported in the data set? How many are hacking/IT incidents?

```
table( cyber.data$Type.of.Breach )
```

```
##
##                                        Hacking/IT Incident
##                                                         77
##                                   Hacking/IT Incident, Other
##                                                          2
## Hacking/IT Incident, Other, Unauthorized Access/Disclosure
##                                                          1
##                                   Hacking/IT Incident, Theft
##                                                          1
## Hacking/IT Incident, Theft, Unauthorized Access/Disclosure
##                                                          3
##        Hacking/IT Incident, Unauthorized Access/Disclosure
##                                                         10
##                                              Improper Disposal
##                                                         42
##                                       Improper Disposal, Loss
##                                                          3
##                                Improper Disposal, Loss, Theft
##                                                          3
##   Improper Disposal, Theft, Unauthorized Access/Disclosure
##                                                          1
##          Improper Disposal, Unauthorized Access/Disclosure
##                                                          2
##                                                       Loss
##                                                         79
##                                                 Loss, Other
##                                                          2
##                                           Loss, Other, Theft
##                                                          1
##                                                 Loss, Theft
##                                                         15
##                         Loss, Unauthorized Access/Disclosure
##                                                          5
##              Loss, Unauthorized Access/Disclosure, Unknown
##                                                          1
##                                               Loss, Unknown
##                                                          2
##                                                      Other
##                                                         89
##                                                Other, Theft
##                                                          5
##               Other, Theft, Unauthorized Access/Disclosure
##                                                          2
##                       Other, Unauthorized Access/Disclosure
##                                                          7
##                                             Other, Unknown
##                                                          2
##                                                      Theft
##                                                        577
##                       Theft, Unauthorized Access/Disclosure
##                                                         24
```

```
##             Theft, Unauthorized Access/Disclosure, Unknown
##                                                          1
##                              Unauthorized Access/Disclosure
##                                                        183
##                             Unauthorized Access/Disclosure
##                                                          1
##                                                    Unknown
##                                                         10
```

[Following are the different values of `Type.Of.Breach` *reported in the data set.]*

```
df.unique <- unique(cyber.data$Type.of.Breach)
df.unique
```

```
##  [1] "Theft"
##  [2] "Loss"
##  [3] "Other"
##  [4] "Unauthorized Access/Disclosure"
##  [5] "Hacking/IT Incident"
##  [6] "Unauthorized Access/Disclosure "
##  [7] "Loss, Theft"
##  [8] "Improper Disposal"
##  [9] "Improper Disposal, Loss"
## [10] "Other, Theft"
## [11] "Loss, Other"
## [12] "Hacking/IT Incident, Unauthorized Access/Disclosure"
## [13] "Improper Disposal, Loss, Theft"
## [14] "Hacking/IT Incident, Theft, Unauthorized Access/Disclosure"
## [15] "Unknown"
## [16] "Theft, Unauthorized Access/Disclosure"
## [17] "Other, Unauthorized Access/Disclosure"
## [18] "Hacking/IT Incident, Other"
## [19] "Other, Unknown"
## [20] "Loss, Unknown"
## [21] "Loss, Unauthorized Access/Disclosure, Unknown"
## [22] "Hacking/IT Incident, Other, Unauthorized Access/Disclosure"
## [23] "Hacking/IT Incident, Theft"
## [24] "Loss, Other, Theft"
## [25] "Other, Theft, Unauthorized Access/Disclosure"
## [26] "Improper Disposal, Theft, Unauthorized Access/Disclosure"
## [27] "Improper Disposal, Unauthorized Access/Disclosure"
## [28] "Loss, Unauthorized Access/Disclosure"
## [29] "Theft, Unauthorized Access/Disclosure, Unknown"
```

```
Hacking <- cyber.data[ which(cyber.data$Type.of.Breach == "Hacking/IT Incident") ,]
count(Hacking)
```

```
##    n
## 1 77
```

[`Type.of.Breach  == "Hacking/IT Incident"` *are 77.]*

Your answer here: what do you see??

- What type of breach is reported in the 748th row of `cyber.data`? How about 349th row? Was row 349 counted in the proportion of Hacking/IT incident breaches you computed above? Why or why not?

```
cyber.data[748,]
```

```
##       X Name.of.Covered.Entity State Covered.Entity.Type Individuals.Affected
## 748 748          UT Physicians    TX Healthcare Provider                  596
##     Breach.Submission.Date Type.of.Breach Location.of.Breached.Information
## 748             2013-08-28    Loss, Theft                            Laptop
##     Business.Associate.Present Web.Description
## 748                      FALSE             \\N
```

['Type.of.Breach is reported in the 748th row of cyber.data is Loss, Theft.]

```
cyber.data[349,]
```

```
##       X Name.of.Covered.Entity State Covered.Entity.Type Individuals.Affected
## 349 349  Freda J Bowman  MD PA    TX Healthcare Provider                 1300
##     Breach.Submission.Date                                    Type.of.Breach
## 349             2011-09-20 Hacking/IT Incident, Unauthorized Access/Disclosure
##     Location.of.Breached.Information Business.Associate.Present Web.Description
## 349                   Network Server                      FALSE             \\N
```

```
tb. <- strsplit(cyber.data$Type.of.Breach, ', ')
table(unlist(tb.))
```

```
##
##             Hacking/IT Incident               Improper Disposal
##                              94                              51
##                            Loss                           Other
##                             111                             111
##                           Theft  Unauthorized Access/Disclosure
##                             633                             240
## Unauthorized Access/Disclosure                         Unknown
##                               1                              16
```

['Type.of.Breach is reported in the 349th row of cyber.data is Hacking/IT Incident, Unauthorized Access/Disclosure. It does not counted in the proportion of Hacking/IT incident breaches because it counts only the breach type of Hacking/IT Incident. When we did split the breach type with comma then then it calculated all Hacking/IT Incidents.]

- Perform a hypothesis test on whether there is a difference in proportion of Hacking/IT incidentes between the state of Illinois and the state of California. Write your conclusion interpreting the resutls of the statistical test.

```
Individuals.IL1 <- cyber.data[which(cyber.data$Type.of.Breach == "Hacking/IT Incident" & cyber.data$Sta
count(Individuals.IL1)
```

```
##   n
## 1 8
```

```r
Individuals.IL2 <- cyber.data[which(cyber.data$State == 'IL'), ]
count(Individuals.IL2)
```

```
##    n
## 1 57
```

```r
Individuals.CA2 <- cyber.data[which(cyber.data$Type.of.Breach == "Hacking/IT Incident" & cyber.data$Sta
count(Individuals.CA2)
```

```
##   n
## 1 6
```

```r
Individuals.CA3 <- cyber.data[which(cyber.data$State == 'CA'), ]
count(Individuals.CA3)
```

```
##     n
## 1 128
```

```r
res <- prop.test(x = c(8, 6), n = c(57, 128))
```

```
## Warning in stats::prop.test(x = x, n = n, p = p, alternative = alternative, :
## Chi-squared approximation may be incorrect
```

```r
res
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of c8 out of 576 out of 128
## X-squared = 3.6807, df = 1, p-value = 0.05505
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01652701  0.20347876
## sample estimates:
##    prop 1    prop 2
## 0.1403509 0.0468750
```

[There is a difference in proportion of Hacking/IT incidents between the state of Illinois and the state of California.Proportion of Hacking/IT incidents between the state of Illinois is 14% and Proportion of Hacking/IT incidents between the state of California is 4.6%.]

Part II: Review of basic concepts in statistical learning

You will spend some time thinking of some real-life applications for statistical learning.

Question 3.

Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

[Predict the type of an animal by considering their features like stipes, color, weight. *Predict what will be the life expectancy of a person by taking information of their health. *Predict what kind of an email(spam) is by considering their keywords.]*

Question 4.

Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

[Predict price of houses, predictors: size of yard, number of rooms, number of bathrooms, area,nearby schools, neighborhood. Deciding an income of a person, inference between salary (response) and factors as years of education, age, years of work experience, field of study. Predict car values based on predictors as mileage, make, model, engine size, interior style and cruise control.]

Question 5.

Describe three real-life applications in which cluster analysis might be useful.

[Food market analyses groups the people depending on what type of food most of the people refers depending on their past purchases. Hospital can have number of serious condition wards depends on the accidents happened in the particular area. Depending on the patients history doctors can group the people for better health.]

Question 6.

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

[Advantages: Better for large number od data, represents more complex-systems and non-linear relationship between predictor and response, can generate a wider range of possible shapes to estimate p.]

[Disadvantages: Difficult to interpret, large variance, not useful for small number of data.]

[A more flexible approach is preferable when the dataset has large number of observations,the system is underfitted, or when the data has non-linear characteristcs.]

[A less flexible approach is preferable when the dataset has few observations, or when more interpretability is desirable, or when the data tends to a linear behavior, high variance of error terms]