Redo exam 1

Thakur Prasad Ghimire

10/18/2021

rm(list = ls(all=TRUE))  
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

load("C:/Homework EcoB2000/Redo Exam 1/HHPulse1.Rdata")  
load("C:/Homework EcoB2000/Redo Exam 1/HHPulse2.Rdata")

NewDF<-data.frame(HHPulse1,HHPulse2)

#cross tabulation before Na value cleaned

xtabs(~EEDUC+RECVDVACC+REGION,data=NewDF)

## , , REGION = Northeast  
##   
## RECVDVACC  
## EEDUC NA yes got vaxx no did not get vaxx  
## less than hs 0 31 16  
## some hs 1 107 19  
## HS diploma 27 1028 190  
## some coll 20 1576 222  
## assoc deg 15 928 120  
## bach deg 35 2901 162  
## adv deg 20 2961 99  
##   
## , , REGION = South  
##   
## RECVDVACC  
## EEDUC NA yes got vaxx no did not get vaxx  
## less than hs 0 103 48  
## some hs 5 211 107  
## HS diploma 32 2052 652  
## some coll 62 3868 858  
## assoc deg 30 1903 387  
## bach deg 63 5780 546  
## adv deg 76 5582 315  
##   
## , , REGION = Midwest  
##   
## RECVDVACC  
## EEDUC NA yes got vaxx no did not get vaxx  
## less than hs 4 41 20  
## some hs 5 114 56  
## HS diploma 29 1326 370  
## some coll 35 2306 511  
## assoc deg 29 1352 269  
## bach deg 65 3598 343  
## adv deg 41 2977 160  
##   
## , , REGION = West  
##   
## RECVDVACC  
## EEDUC NA yes got vaxx no did not get vaxx  
## less than hs 2 115 31  
## some hs 4 220 87  
## HS diploma 25 1691 435  
## some coll 61 4272 805  
## assoc deg 36 2083 356  
## bach deg 75 5993 514  
## adv deg 54 5207 239

#To show EEDUC,RECVDVACC and Region data intraction.

str(select(NewDF,EEDUC,RECVDVACC,REGION))

## 'data.frame': 69114 obs. of 3 variables:  
## $ EEDUC : Factor w/ 7 levels "less than hs",..: 6 3 6 7 4 6 7 4 5 7 ...  
## $ RECVDVACC: Factor w/ 3 levels "NA","yes got vaxx",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ REGION : Factor w/ 4 levels "Northeast","South",..: 2 2 2 2 2 2 4 2 2 2 ...

#Count how much is RECVDAVCC has NA values , and each values,because YES has more frequency than NO, I decide NA to YES.

table(NewDF$RECVDVACC)

##   
## NA yes got vaxx no did not get vaxx   
## 851 60326 7937

# Change the NA values to YES.

NewDF$RECVDVACC[NewDF$RECVDVACC=="NA"]<-"yes got VAxx"

## Warning in `[<-.factor`(`\*tmp\*`, NewDF$RECVDVACC == "NA", value =  
## structure(c(2L, : invalid factor level, NA generated

NewDF$RECVDVACC<-factor(NewDF$RECVDVACC)  
NewDF$RECVDVACC<-as.numeric(NewDF$RECVDVACC)

#New cross tabulation ,after NA is cleaned.

xtabs(~EEDUC+RECVDVACC+REGION,data=NewDF)

## , , REGION = Northeast  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 31 16  
## some hs 107 19  
## HS diploma 1028 190  
## some coll 1576 222  
## assoc deg 928 120  
## bach deg 2901 162  
## adv deg 2961 99  
##   
## , , REGION = South  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 103 48  
## some hs 211 107  
## HS diploma 2052 652  
## some coll 3868 858  
## assoc deg 1903 387  
## bach deg 5780 546  
## adv deg 5582 315  
##   
## , , REGION = Midwest  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 41 20  
## some hs 114 56  
## HS diploma 1326 370  
## some coll 2306 511  
## assoc deg 1352 269  
## bach deg 3598 343  
## adv deg 2977 160  
##   
## , , REGION = West  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 115 31  
## some hs 220 87  
## HS diploma 1691 435  
## some coll 4272 805  
## assoc deg 2083 356  
## bach deg 5993 514  
## adv deg 5207 239

#from question 1 i got is RECVADVACC=EEDUC+REGION+REGION\*EEDUC

#H0: There is no effect of Education and Region to RECVDVACC #H1: There is an effect of Education and Region to RECVADVACC simultaan test using ANOVA

anov1<-aov(formula=RECVDVACC~EEDUC\*REGION,data=NewDF)  
summary(anov1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## EEDUC 6 271 45.20 458.923 < 2e-16 \*\*\*  
## REGION 3 16 5.23 53.059 < 2e-16 \*\*\*  
## EEDUC:REGION 18 7 0.38 3.811 7.86e-08 \*\*\*  
## Residuals 68235 6721 0.10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 851 observations deleted due to missingness

From thee result of anov1 , i can conclude that there is an effect of Education and Region to RECVDVACC.

using by partial t- test

reg<-lm(formula=RECVDVACC~EEDUC+REGION,data=NewDF)  
summary(reg)

##   
## Call:  
## lm(formula = RECVDVACC ~ EEDUC + REGION, data = NewDF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.30467 -0.16101 -0.08871 -0.04779 0.98424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.251546 0.015880 78.812 < 2e-16 \*\*\*  
## EEDUCsome hs 0.008687 0.018719 0.464 0.643   
## EEDUCHS diploma -0.070710 0.016006 -4.418 9.99e-06 \*\*\*  
## EEDUCsome coll -0.117375 0.015819 -7.420 1.18e-13 \*\*\*  
## EEDUCassoc deg -0.130100 0.016024 -8.119 4.77e-16 \*\*\*  
## EEDUCbach deg -0.203756 0.015760 -12.928 < 2e-16 \*\*\*  
## EEDUCadv deg -0.235785 0.015781 -14.941 < 2e-16 \*\*\*  
## REGIONSouth 0.044432 0.003732 11.905 < 2e-16 \*\*\*  
## REGIONMidwest 0.040921 0.004109 9.959 < 2e-16 \*\*\*  
## REGIONWest 0.026838 0.003745 7.166 7.83e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3139 on 68253 degrees of freedom  
## (851 observations deleted due to missingness)  
## Multiple R-squared: 0.0409, Adjusted R-squared: 0.04077   
## F-statistic: 323.4 on 9 and 68253 DF, p-value: < 2.2e-16

#confident interval with 99% level two sides

confint(reg,level=0.99)

## 0.5 % 99.5 %  
## (Intercept) 1.21064055 1.29245232  
## EEDUCsome hs -0.03953167 0.05690551  
## EEDUCHS diploma -0.11193906 -0.02948099  
## EEDUCsome coll -0.15812207 -0.07662770  
## EEDUCassoc deg -0.17137500 -0.08882498  
## EEDUCbach deg -0.24435268 -0.16315909  
## EEDUCadv deg -0.27643538 -0.19513436  
## REGIONSouth 0.03481783 0.05404569  
## REGIONMidwest 0.03033736 0.05150523  
## REGIONWest 0.01718994 0.03648540

#2

xtabs(~EEDUC+RECVDVACC+GENID\_DESCRIBE,NewDF)

## , , GENID\_DESCRIBE = NA  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 11 2  
## some hs 13 2  
## HS diploma 70 15  
## some coll 96 22  
## assoc deg 47 7  
## bach deg 139 20  
## adv deg 129 19  
##   
## , , GENID\_DESCRIBE = male  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 120 34  
## some hs 275 103  
## HS diploma 2200 628  
## some coll 4665 880  
## assoc deg 2107 330  
## bach deg 7552 561  
## adv deg 6943 292  
##   
## , , GENID\_DESCRIBE = female  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 142 62  
## some hs 344 158  
## HS diploma 3732 974  
## some coll 7078 1469  
## assoc deg 4027 778  
## bach deg 10401 956  
## adv deg 9494 458  
##   
## , , GENID\_DESCRIBE = transgender  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 4 6  
## some hs 6 1  
## HS diploma 22 6  
## some coll 48 5  
## assoc deg 12 3  
## bach deg 43 2  
## adv deg 32 10  
##   
## , , GENID\_DESCRIBE = other  
##   
## RECVDVACC  
## EEDUC 1 2  
## less than hs 13 11  
## some hs 14 5  
## HS diploma 73 24  
## some coll 135 20  
## assoc deg 73 14  
## bach deg 137 26  
## adv deg 129 34

# I change NA gender to other Gender

NewDF$GENID\_DESCRIBE[NewDF$GENID\_DESCRIBE=="NA"]<-"other"  
NewDF$GENID\_DESCRIBE<-factor(NewDF$GENID\_DESCRIBE)

H0: There is no effect odf education and GENDER\_DESCRIBE to RECVDVACC H1: There is an effect of Education and GENDER\_DESCRIBE to RECVDVACC

#simultaan test using ANOVA

anov2<-aov(formula=RECVDVACC~EEDUC+GENID\_DESCRIBE,data=NewDF)  
summary(anov2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## EEDUC 6 271 45.20 457.82 < 2e-16 \*\*\*  
## GENID\_DESCRIBE 3 4 1.48 14.95 9.96e-10 \*\*\*  
## Residuals 68253 6739 0.10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 851 observations deleted due to missingness

#From the result ANOVA2 i can conclude that there is an effect on Education and GENID\_DESCRIBE to participant of RECVDVACC

#partial test using t-test

reg2<-lm(formula=RECVDVACC~EEDUC+GENID\_DESCRIBE,data=NewDF)  
summary(reg2)

##   
## Call:  
## lm(formula = RECVDVACC ~ EEDUC + GENID\_DESCRIBE, data = NewDF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33397 -0.15868 -0.08264 -0.05007 0.96077   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.273055 0.015709 81.041 < 2e-16 \*\*\*  
## EEDUCsome hs 0.010998 0.018744 0.587 0.557   
## EEDUCHS diploma -0.068264 0.016035 -4.257 2.07e-05 \*\*\*  
## EEDUCsome coll -0.114376 0.015851 -7.216 5.43e-13 \*\*\*  
## EEDUCassoc deg -0.128105 0.016057 -7.978 1.51e-15 \*\*\*  
## EEDUCbach deg -0.201259 0.015794 -12.743 < 2e-16 \*\*\*  
## EEDUCadv deg -0.233825 0.015814 -14.786 < 2e-16 \*\*\*  
## GENID\_DESCRIBEfemale 0.010841 0.002486 4.361 1.30e-05 \*\*\*  
## GENID\_DESCRIBEtransgender 0.035421 0.022315 1.587 0.112   
## GENID\_DESCRIBEother 0.049916 0.008935 5.587 2.32e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3142 on 68253 degrees of freedom  
## (851 observations deleted due to missingness)  
## Multiple R-squared: 0.0393, Adjusted R-squared: 0.03917   
## F-statistic: 310.2 on 9 and 68253 DF, p-value: < 2.2e-16

#confident interval with 99% level two-sides

confint(reg2,level=0.99)

## 0.5 % 99.5 %  
## (Intercept) 1.232590875 1.31351925  
## EEDUCsome hs -0.037285787 0.05928131  
## EEDUCHS diploma -0.109568243 -0.02695896  
## EEDUCsome coll -0.155206990 -0.07354478  
## EEDUCassoc deg -0.169467076 -0.08674339  
## EEDUCbach deg -0.241942792 -0.16057604  
## EEDUCadv deg -0.274559690 -0.19309007  
## GENID\_DESCRIBEfemale 0.004436984 0.01724528  
## GENID\_DESCRIBEtransgender -0.022060320 0.09290234  
## GENID\_DESCRIBEother 0.026901654 0.07293064

#Q3

library(class)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

load("C:/Homework EcoB2000/Redo Exam 1/Household\_Pulse\_data.RData")

# I choose the RACE and States because it is interesting how people on each state response to Vaccine.

dataKN<-data.frame(select(Household\_Pulse\_data,RRACE,RECVDVACC,EST\_ST))  
summary(dataKN)

## RRACE RECVDVACC EST\_ST   
## White:56938 NA : 851 California : 5359   
## Black: 5412 yes got vaxx :60326 Texas : 3766   
## Asian: 3561 no did not get vaxx: 7937 Florida : 2728   
## Other: 3203 Washington : 2634   
## Massachusetts: 1965   
## Oregon : 1934   
## (Other) :50728

r<-sample(x=nrow(dataKN),size = 0.01\*nrow(dataKN),replace = FALSE)  
smpl<-dataKN[r,]  
  
smpl$RECVDVACC[smpl$RECVDVACC=="NA"]<-"yes got vaxx"  
smpl$RECVDVACC<-factor(smpl$RECVDVACC)  
smpl$RRACE<-as.integer(smpl$RRACE)  
smpl$EST\_ST<-as.integer(smpl$EST\_ST)  
  
str(smpl)

## 'data.frame': 691 obs. of 3 variables:  
## $ RRACE : int 1 3 1 1 1 1 1 1 3 1 ...  
## $ RECVDVACC: Factor w/ 2 levels "yes got vaxx",..: 2 1 1 1 1 1 1 1 1 1 ...  
## $ EST\_ST : int 26 11 20 2 9 21 3 15 11 24 ...

raNUm<-sample(x=nrow(smpl),size = 0.70\*nrow(smpl),replace = FALSE)  
train\_kn<-smpl[raNUm,]  
test\_kn<-smpl[-raNUm,]  
train\_lab<-train\_kn$RECVDVACC  
knn1<-knn(train = train\_kn[,-2],test = test\_kn[,-2],cl =train\_lab,k=7)  
  
confusionMatrix(table(test\_kn$RECVDVACC,knn1))

## Confusion Matrix and Statistics  
##   
## knn1  
## yes got vaxx no did not get vaxx  
## yes got vaxx 182 0  
## no did not get vaxx 26 0  
##   
## Accuracy : 0.875   
## 95% CI : (0.8222, 0.9167)  
## No Information Rate : 1   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : 9.443e-07   
##   
## Sensitivity : 0.875   
## Specificity : NA   
## Pos Pred Value : NA   
## Neg Pred Value : NA   
## Prevalence : 1.000   
## Detection Rate : 0.875   
## Detection Prevalence : 0.875   
## Balanced Accuracy : NA   
##   
## 'Positive' Class : yes got vaxx   
##

Confusion matrix show accurency 92% which mean knn work very well to classify the case.

Thakur Prasad Ghimire

City college of New York

Student ID is:23987365

detach()