Homework Lab#7

Thakur Prasad Ghimire

11/4/2021

Study Group members’ name 1. Thakur Prasad Ghimire 2.Nicholas Alonso 3.Bolina Oxilus 4.Diep luu 5.Joel rosas 6.Hassan Fayyaz

load("C:/Homework EcoB2000/lab#7/Household\_Pulse\_data\_v2 (2).RData")

#Analyze the get vaxx variable

table(Household\_Pulse\_data$RECVDVACC)

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

# Explain what you’re doing with NA.?

First lets count the missing values in the Get vaccine variable

sum(is.na(Household\_Pulse\_data$RECVDVACC))

## [1] 0

#Then lets replace NA strings with NULL according to 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

Household\_Pulse\_data= Household\_Pulse\_data %>%   
 mutate(RECVDVACC = na\_if(RECVDVACC, "NA"))

#count missing values in the RECVDVACC variable

sum(is.na(Household\_Pulse\_data$RECVDVACC))

## [1] 851

#Create the dependent variable vaxx

Household\_Pulse\_data$vaxx <-(Household\_Pulse\_data$RECVDVACC == "yes got vaxx")  
Household\_Pulse\_data$vaxx = as.factor(Household\_Pulse\_data$vaxx)  
table(Household\_Pulse\_data$vaxx)

##   
## FALSE TRUE   
## 7937 60326

#First decide on how you’re defining your subgroup (all adults or 12+? Within certain age? Other?) - select 12-17 subgroup

dim(Household\_Pulse\_data) #data dimension

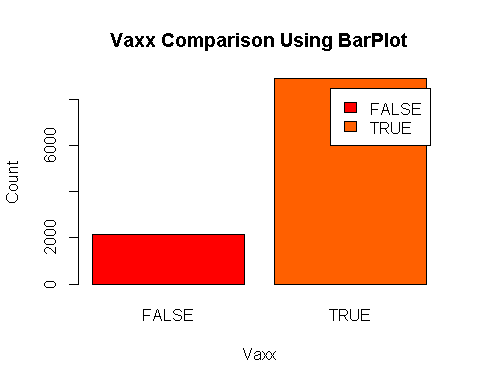
## [1] 69114 45

df\_household = Household\_Pulse\_data[Household\_Pulse\_data$KIDS\_12\_17Y!="NA",]  
dim(df\_household) #data dimension

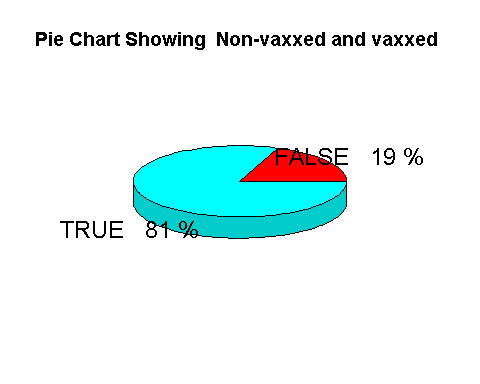
## [1] 11068 45

Then find some basic statistics – what fraction are not vaxxed?

a=table(df\_household$vaxx)  
barplot(a,main="Vaxx Comparison Using BarPlot",  
 ylab="Count",  
 xlab="Vaxx",  
 col=rainbow(16),  
 legend=rownames(a))

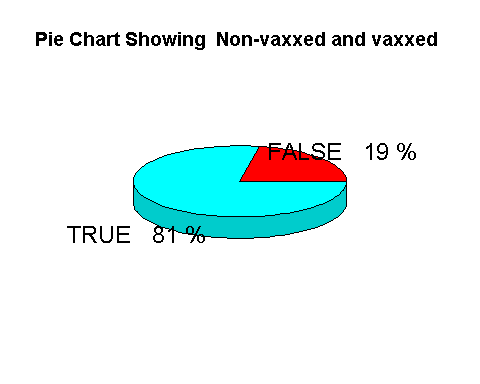


pct=round(a/sum(a)\*100)  
lbs=paste(c("FALSE","TRUE")," ",pct,"%",sep=" ")  
library(plotrix)  
pie3D(a,labels=lbs,  
 main="Pie Chart Showing Non-vaxxed and vaxxed ")



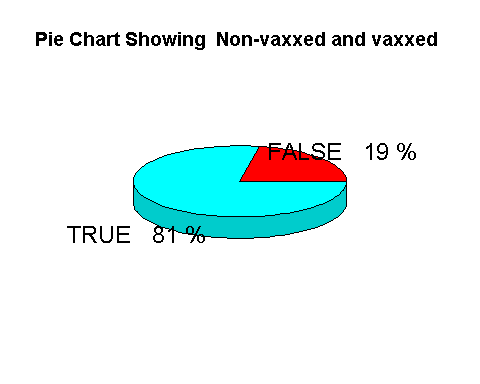
(Later go back to look at simple stats for subgroups to see if there are sharp differences.) other subgroups are: KIDS\_12\_17Y - select 12-17Y subgroup

df\_KidsLT5Y = Household\_Pulse\_data[Household\_Pulse\_data$KIDS\_LT5Y!="NA",]  
#plot pie chart   
b=table(df\_KidsLT5Y$vaxx)  
pct2=round(b/sum(b)\*100)  
lbs=paste(c("FALSE","TRUE")," ",pct,"%",sep=" ")  
library(plotrix)  
pie3D(b,labels=lbs,  
 main="Pie Chart Showing Non-vaxxed and vaxxed ")



other subgroups are: KIDS\_5\_11Y - select 5-11Y subgroup

df\_Kids5\_11Y = Household\_Pulse\_data[Household\_Pulse\_data$KIDS\_5\_11Y!="NA",]  
#plot pie chart   
c=table(df\_Kids5\_11Y$vaxx)  
pct3=round(c/sum(c)\*100)  
lbs=paste(c("FALSE","TRUE")," ",pct,"%",sep=" ")  
library(plotrix)  
pie3D(b,labels=lbs,  
 main="Pie Chart Showing Non-vaxxed and vaxxed ")



**NO, there were no sharp differences**

#### Run several different types of models to explain vaccination rates with some explanatory variables, vaxx ~ TBIRTH\_YEAR + EEDUC + MS + RRACE + RHISPANIC + GENID\_DESCRIBE + REGION.

* Slice the dataframe to include only the features named above

df\_household = subset(df\_household,select = c(vaxx,EEDUC,MS,RRACE,  
 RHISPANIC,GENID\_DESCRIBE,REGION))  
head(df\_household)

## vaxx EEDUC MS RRACE RHISPANIC GENID\_DESCRIBE REGION  
## 5 TRUE some coll never Black Not Hispanic female South  
## 10 TRUE adv deg married White Not Hispanic female South  
## 18 TRUE assoc deg married White Not Hispanic male West  
## 24 TRUE HS diploma never Other Not Hispanic female West  
## 33 TRUE bach deg married Black Not Hispanic male South  
## 48 TRUE assoc deg married Asian Not Hispanic female West

We want to set up the data in a way that is common to all of the models. Some of the estimation procedures are not as tolerant about factors so we need to set those as dummies. Some are also intolerant of NA values. I’ll show the code for the basic set of explanatory variables, which you can modify as you see fit.

str(df\_household)

## 'data.frame': 11068 obs. of 7 variables:  
## $ vaxx : Factor w/ 2 levels "FALSE","TRUE": 2 2 2 2 2 2 2 2 2 1 ...  
## $ EEDUC : Factor w/ 7 levels "less than hs",..: 4 7 5 3 6 5 6 7 6 4 ...  
## $ MS : Factor w/ 6 levels "NA","married",..: 6 2 2 6 2 2 2 2 2 4 ...  
## $ RRACE : Factor w/ 4 levels "White","Black",..: 2 1 1 4 2 3 3 4 4 1 ...  
## $ RHISPANIC : Factor w/ 2 levels "Not Hispanic",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ GENID\_DESCRIBE: Factor w/ 5 levels "NA","male","female",..: 3 3 2 3 2 3 2 3 3 2 ...  
## $ REGION : Factor w/ 4 levels "Northeast","South",..: 2 2 4 4 2 4 1 2 4 4 ...

library(dummies)

## dummies-1.5.6 provided by Decision Patterns

vars=c("EEDUC","MS" ,"RRACE","RHISPANIC","GENID\_DESCRIBE","REGION")  
df\_household[,vars] <- lapply(df\_household[,vars] , factor)  
df\_new<-dummy.data.frame(df\_household,names = vars,sep=".")

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored

dim(df\_new)

## [1] 11068 29

split the data into train and test data 80% of the sample size

set.seed(654321)  
  
smp\_size <- floor(0.80 \* nrow(df\_new))  
train\_ind <- sample(seq\_len(nrow(df\_new)), size = smp\_size)  
train.set <- df\_new[train\_ind, ]  
test.set <- df\_new[-train\_ind, ]  
  
dim(train.set)

## [1] 8854 29

str(train.set)

## 'data.frame': 8854 obs. of 29 variables:  
## $ vaxx : Factor w/ 2 levels "FALSE","TRUE": 2 2 2 2 2 2 2 2 2 2 ...  
## $ EEDUC.less than hs : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ EEDUC.some hs : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ EEDUC.HS diploma : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ EEDUC.some coll : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ EEDUC.assoc deg : int 1 0 0 1 0 0 0 0 0 1 ...  
## $ EEDUC.bach deg : int 0 1 0 0 0 0 0 1 0 0 ...  
## $ EEDUC.adv deg : int 0 0 0 0 1 1 1 0 1 0 ...  
## $ MS.NA : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MS.married : int 1 1 0 0 1 1 1 1 1 0 ...  
## $ MS.widowed : int 0 0 0 0 0 0 0 0 0 1 ...  
## $ MS.divorced : int 0 0 0 1 0 0 0 0 0 0 ...  
## $ MS.separated : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MS.never : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ RRACE.White : int 1 1 0 1 1 1 1 1 0 1 ...  
## $ RRACE.Black : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ RRACE.Asian : int 0 0 1 0 0 0 0 0 1 0 ...  
## $ RRACE.Other : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ RHISPANIC.Not Hispanic : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ RHISPANIC.Hispanic : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ GENID\_DESCRIBE.NA : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GENID\_DESCRIBE.male : int 0 0 0 0 0 1 0 1 1 1 ...  
## $ GENID\_DESCRIBE.female : int 1 1 1 1 1 0 1 0 0 0 ...  
## $ GENID\_DESCRIBE.transgender: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GENID\_DESCRIBE.other : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ REGION.Northeast : int 0 1 0 0 1 0 0 0 1 0 ...  
## $ REGION.South : int 0 0 0 1 0 1 0 0 0 1 ...  
## $ REGION.Midwest : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ REGION.West : int 1 0 1 0 0 0 0 1 0 0 ...  
## - attr(\*, "dummies")=List of 6  
## ..$ EEDUC : int [1:7] 2 3 4 5 6 7 8  
## ..$ MS : int [1:6] 9 10 11 12 13 14  
## ..$ RRACE : int [1:4] 15 16 17 18  
## ..$ RHISPANIC : int [1:2] 19 20  
## ..$ GENID\_DESCRIBE: int [1:5] 21 22 23 24 25  
## ..$ REGION : int [1:4] 26 27 28 29

## Logistic regression model

model\_logit1 <- glm(vaxx~., family = binomial, data = train.set)  
summary(model\_logit1)

##   
## Call:  
## glm(formula = vaxx ~ ., family = binomial, data = train.set)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8479 0.3325 0.5146 0.7053 1.6013   
##   
## Coefficients: (6 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.39464 0.26494 5.264 1.41e-07 \*\*\*  
## `EEDUC.less than hs` -1.77623 0.23198 -7.657 1.90e-14 \*\*\*  
## `EEDUC.some hs` -1.80998 0.17917 -10.102 < 2e-16 \*\*\*  
## `EEDUC.HS diploma` -1.69197 0.10498 -16.117 < 2e-16 \*\*\*  
## `EEDUC.some coll` -1.36251 0.09782 -13.929 < 2e-16 \*\*\*  
## `EEDUC.assoc deg` -1.07837 0.11120 -9.698 < 2e-16 \*\*\*  
## `EEDUC.bach deg` -0.56299 0.10056 -5.599 2.16e-08 \*\*\*  
## `EEDUC.adv deg` NA NA NA NA   
## MS.NA 0.02947 0.34870 0.085 0.932655   
## MS.married 0.51397 0.08126 6.325 2.53e-10 \*\*\*  
## MS.widowed 0.63581 0.19229 3.307 0.000944 \*\*\*  
## MS.divorced 0.33172 0.10144 3.270 0.001075 \*\*   
## MS.separated 0.08488 0.16370 0.519 0.604081   
## MS.never NA NA NA NA   
## RRACE.White 0.31725 0.10368 3.060 0.002215 \*\*   
## RRACE.Black 0.43487 0.13150 3.307 0.000943 \*\*\*  
## RRACE.Asian 1.83718 0.23405 7.850 4.17e-15 \*\*\*  
## RRACE.Other NA NA NA NA   
## `RHISPANIC.Not Hispanic` -0.42627 0.08682 -4.910 9.13e-07 \*\*\*  
## RHISPANIC.Hispanic NA NA NA NA   
## GENID\_DESCRIBE.NA -0.15438 0.39212 -0.394 0.693795   
## GENID\_DESCRIBE.male 0.71833 0.22841 3.145 0.001661 \*\*   
## GENID\_DESCRIBE.female 0.66533 0.22527 2.953 0.003142 \*\*   
## GENID\_DESCRIBE.transgender 0.38626 0.48873 0.790 0.429329   
## GENID\_DESCRIBE.other NA NA NA NA   
## REGION.Northeast 0.34980 0.10139 3.450 0.000560 \*\*\*  
## REGION.South -0.23360 0.07061 -3.308 0.000939 \*\*\*  
## REGION.Midwest -0.11799 0.08050 -1.466 0.142723   
## REGION.West NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8647.6 on 8801 degrees of freedom  
## Residual deviance: 7894.9 on 8779 degrees of freedom  
## (52 observations deleted due to missingness)  
## AIC: 7940.9  
##   
## Number of Fisher Scoring iterations: 5

### Confusion matrix

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

pred\_vals <- predict(model\_logit1, test.set[,-1], type = "response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

pred\_model\_logit1 <- as.factor(pred\_vals > 0.5)  
table(pred = pred\_model\_logit1, true = test.set$vaxx)

## true  
## pred FALSE TRUE  
## FALSE 3 9  
## TRUE 415 1773

confusionMatrix(pred\_model\_logit1,test.set$vaxx,positive = "TRUE")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction FALSE TRUE  
## FALSE 3 9  
## TRUE 415 1773  
##   
## Accuracy : 0.8073   
## 95% CI : (0.7902, 0.8236)  
## No Information Rate : 0.81   
## P-Value [Acc > NIR] : 0.6399   
##   
## Kappa : 0.0034   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.994949   
## Specificity : 0.007177   
## Pos Pred Value : 0.810329   
## Neg Pred Value : 0.250000   
## Prevalence : 0.810000   
## Detection Rate : 0.805909   
## Detection Prevalence : 0.994545   
## Balanced Accuracy : 0.501063   
##   
## 'Positive' Class : TRUE   
##

## Linear model; OLS model (convert vaxx into numeric)

train.set$vaxx = ifelse(train.set$vaxx =="TRUE",1,0)  
test.set$vaxx = ifelse(test.set =="TRUE",1,0)  
model\_lpm1 <- lm(vaxx~., data = train.set)  
summary(model\_lpm1)

##   
## Call:  
## lm(formula = vaxx ~ ., data = train.set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.05924 0.03483 0.12826 0.23498 0.61223   
##   
## Coefficients: (6 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.733047 0.041832 17.524 < 2e-16 \*\*\*  
## `EEDUC.less than hs` -0.249538 0.038429 -6.493 8.85e-11 \*\*\*  
## `EEDUC.some hs` -0.266496 0.029533 -9.024 < 2e-16 \*\*\*  
## `EEDUC.HS diploma` -0.241840 0.014465 -16.718 < 2e-16 \*\*\*  
## `EEDUC.some coll` -0.173165 0.012387 -13.980 < 2e-16 \*\*\*  
## `EEDUC.assoc deg` -0.121953 0.014422 -8.456 < 2e-16 \*\*\*  
## `EEDUC.bach deg` -0.049710 0.011287 -4.404 1.07e-05 \*\*\*  
## `EEDUC.adv deg` NA NA NA NA   
## MS.NA 0.002644 0.060184 0.044 0.964959   
## MS.married 0.091196 0.013343 6.835 8.76e-12 \*\*\*  
## MS.widowed 0.109403 0.028166 3.884 0.000103 \*\*\*  
## MS.divorced 0.064588 0.016290 3.965 7.40e-05 \*\*\*  
## MS.separated 0.016700 0.027373 0.610 0.541812   
## MS.never NA NA NA NA   
## RRACE.White 0.053799 0.016601 3.241 0.001197 \*\*   
## RRACE.Black 0.070662 0.020709 3.412 0.000648 \*\*\*  
## RRACE.Asian 0.173842 0.022950 7.575 3.96e-14 \*\*\*  
## RRACE.Other NA NA NA NA   
## `RHISPANIC.Not Hispanic` -0.067461 0.012351 -5.462 4.84e-08 \*\*\*  
## RHISPANIC.Hispanic NA NA NA NA   
## GENID\_DESCRIBE.NA -0.018203 0.065128 -0.279 0.779871   
## GENID\_DESCRIBE.male 0.117197 0.037220 3.149 0.001645 \*\*   
## GENID\_DESCRIBE.female 0.110865 0.036869 3.007 0.002646 \*\*   
## GENID\_DESCRIBE.transgender 0.055124 0.079377 0.694 0.487414   
## GENID\_DESCRIBE.other NA NA NA NA   
## REGION.Northeast 0.042092 0.013104 3.212 0.001322 \*\*   
## REGION.South -0.035977 0.010203 -3.526 0.000424 \*\*\*  
## REGION.Midwest -0.018200 0.011614 -1.567 0.117145   
## REGION.West NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3791 on 8779 degrees of freedom  
## (52 observations deleted due to missingness)  
## Multiple R-squared: 0.08162, Adjusted R-squared: 0.07931   
## F-statistic: 35.46 on 22 and 8779 DF, p-value: < 2.2e-16