Chapter 2 Lab

Allen Church

Load necessary packages

library(haven)

Set working directory and load lab data

lab <- read\_dta("Ch2\_lab\_survey\_data.dta")

1. Use the following to create dummy variables for Arlington and Prince William Counties. How many observations are from each county?

#Create dummy variable for Arlington county, selecting corresponding precinct codes with OR operators  
lab$Arlington <- (  
 lab$precinct == "AR49" | lab$precinct == "AR22" | lab$precinct == "AR2" |  
 lab$precinct == "AR18" | lab$precinct == "41" | lab$precinct == "16" |  
 lab$precinct == "4" | lab$precinct == "17" | (lab$precinct == "2" &  
   
lab$state == 4& !is.na(lab$state))|  
 lab$precinct == "31" | lab$precinct == "48")  
  
#Count observations from each county, in this case TRUE corresponds to Arlington  
table(lab$Arlington)

##   
## FALSE TRUE   
## 1884 475

#Create dummy variable for Prince William county, selecting corresponding precinct codes with OR operators  
lab$PrinceWilliam1 <- (  
 lab$precinct == "PW 101" | lab$precinct == "PW 104" | lab$precinct == "PW 401" | lab$precinct == "PW101" |  
 lab$precinct == "PW104" | lab$precinct == "PW402" | lab$precinct == "PW406" | lab$precinct == "401" | lab$precinct == "402" |   
(lab$precinct == "104" & lab$state == 4) )  
  
#Count observations from each county, in this case TRUE corresponds to Prince William  
table(lab$PrinceWilliam1)

##   
## FALSE TRUE   
## 2171 188

1. Create dummy variables for each state/DC. How many observations are in DC, Maryland, Ohio and Virginia?

#Create dummy variables for each state  
lab$DC <- (lab$state == 1)  
lab$Maryland <- (lab$state == 2)  
lab$Ohio <- (lab$state == 3)  
lab$Virginia <- (lab$state == 4)  
  
#Tabulate observations for each state  
table(lab$state)

##   
## 1 2 3 4   
## 768 369 547 664

1. Convert the year\_born variable into age. Be sure to check for and correct for lab errors. What is the average age of all observations in the lab set? The minimum and maximum?

#Since the survey was taken in 2016, subtract 2016 - year born to obtain age  
#Create new age column  
lab$age <- 2016 - lab$year\_born  
  
#The first summary of the age column shows there is a max age of 152 and 482 NA rows  
summary(lab$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 17.00 30.00 41.00 43.17 55.00 152.00 482

#Subset the lab dataframe and exclude values where age is above 100  
newdata <- subset(lab, age < 100)  
  
#The summary of the newdata age column shows a new maximum of 95, which is possible  
summary(newdata$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 17.00 30.00 41.00 43.07 55.00 95.00

1. What is the distribution of the gender variable? Create a male dummy variable and indicate the distribution of this variable. Compare distribution of your male variable to the distribution of the gender variable.

#Create male and female variables  
lab$male <- (lab$gender == 1)  
lab$female <- (lab$gender == 2)

Distribution of male variable

table(lab$male)

##   
## FALSE TRUE   
## 1067 886

Distribution of gender variable. The below table shows that 5 respondents did not identify with the binary gender definition

table(lab$gender)

##   
## 1 2 3   
## 886 1062 5

1. Provide descriptive stats for Trump and Clinton feeling thermometer. Is there anything you need to adjust?

#Summarize Clinton feeling thermometer, see there is a max of 200  
summary(lab$therm\_clinton)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00 20.00 70.00 57.12 90.00 200.00 231

#Turn values over 100 into NA and summarize again  
lab$therm\_clinton[lab$therm\_clinton > 100] <- NA  
summary(lab$therm\_clinton)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00 20.00 70.00 57.06 90.00 100.00 232

#Summarize Trump feeling thermometer  
summary(lab$therm\_trump)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00 0.00 0.00 17.76 25.00 100.00 292

1. What is the distribution of the education variable? Is there any adjustment you would need to make if you will use this as a continuous variable in a regression model?

#Below shows 7 values for education question  
lab1 <- read\_dta("Ch2\_lab\_survey\_data.dta")  
table(lab1$education)

##   
## 1 2 3 4 5 6 7   
## 17 125 245 11 134 677 746

#Below we adjust education to exclude the Other response in answer 4, and re-assign the other responses accordingly  
#Additionally, the Other response only had 11 values  
lab1$education[lab1$education == 4] <- NA  
lab1$education[lab1$education == 5] <- 4  
lab1$education[lab1$education == 6] <- 5  
lab1$education[lab1$education == 7] <- 6  
  
#Generate a new summary table  
table(lab1$education)

##   
## 1 2 3 4 5 6   
## 17 125 245 134 677 746

#A quick histogram to show the distribution of education  
#Below (as well as the table above) shows a low number of responses from 1 - Some high school, which could indicate a bias in sampling, or simply reflect the education of the sample.  
hist(lab1$education)

