Chapter 2 Lab

Tianwei Liu

Preparation

```
## If any of the packages do not exist in your environment (meaning you have not previously downloaded
## them) you should first install them using (for example):
## install.packages("haven")

## Load packages used in this session of R
require(haven)
require(knitr)
require(car)

opts_chunk$set(echo = TRUE)
options(digits = 3)

#insert path to your working directory below
opts_knit$set(root.dir ="~/Desktop/GU/Stats/Lab1")
# You cannot use setwd() with knitr, so use this command to set root directory where data is saved
dta <- 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?

```
dta$Arlington <- (dta$precinct == "AR49" | dta$precinct == "AR22" | dta$precinct == "AR2" |
  dta$precinct == "AR18" | dta$precinct == "41" | dta$precinct == "4" | dta$precinct == "16" |
  dta$precinct == "17" | (dta$precinct == "2" & dta$state == 4 & !is.na(dta$state)) |
  dta$precinct == "31" | dta$precinct == "48")

dta$PrinceWilliam <- (
  dta$precinct == "PW 101" | dta$precinct == "PW 104" | dta$precinct == "PW101" |
  dta$precinct == "PW 401" | dta$precinct == "PW104" | dta$precinct == "PW402" |
  dta$precinct == "PW406"|    dta$precinct == "401" | dta$precinct == "402" |
  (dta$precinct == "104" & dta$state == "4" & !is.na(dta$state))
)
## How many observations are from Arlington
table(dta$Arlington)
```

```
##
## FALSE  TRUE
##  1884   475
```

```
## How many observations are from Prince William
table(dta$PrinceWilliam)
```

```
##
## FALSE  TRUE
##  2171   188
```

Therefore, there are 475 observations in Arlington and 188 observations in Prince William County.

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

```
dta$DC <- (dta$state == 1)
table(dta$DC)
```

```
##
## FALSE  TRUE
##  1580   768
```

```
dta$MD <- (dta$state == 2)
table(dta$MD)
```

```
##
## FALSE  TRUE
##  1979   369
```

```
dta$OH <- (dta$state == 3)
table(dta$OH)
```

```
##
## FALSE  TRUE
##  1801   547
```

```
dta$VA <- (dta$state == 4)
table(dta$VA)
```

```
##
## FALSE  TRUE
##  1684   664
```

Overall, there are 768 observations in DC; 369 observations in MD; 547 observations in OH; and 664 observations in VA.

3) Convert the year_born variable into age. Be sure to check for and correct for data errors. What is the average age of all observations in the data set? The minimum and maximum?

```
dta$age <- 2016 - dta$year_born
summary(dta$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      17      30      41      43      55     152     482
```

```
dta$age[dta$year_born <= 1920] <- NA
dta$age[dta$year_born > 2016] <- NA
summary(dta$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      17      30      41      43      55      95     484
```

After correction, the average age is 41. Minimum is 17 and maximum is 95.

4) 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.

```
dta$male <- (dta$gender == 1)
table(dta$male)
```

```
##
## FALSE  TRUE
##  1067   886
```

```
table(dta$gender)
```

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

Other than male, there are 5 people selected "other".

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

```
summary(dta$therm_trump)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0     0.0    17.8    25.0   100.0     292
```

```
summary(dta$therm_clinton)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    20.0    70.0    57.1    90.0   200.0     231
```

```
dta$therm_clinton[dta$therm_clinton > 100] <- NA
summary(dta$therm_clinton)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    20.0    70.0    57.1    90.0   100.0     232
```

6) 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?

```
table(dta$education)
```

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

```
dta$education[dta$education == 4] <- NA
dta$education[dta$education == 5] <- 4
dta$education[dta$education == 6] <- 5
dta$education[dta$education == 7] <- 6
summary(dta$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       1       4       5       5       6       6     415
```

Additional Comments: It may be convenient to save the data, as we'll use this in the lab for next chapter. Here, we save the data as .Rdata format.

```
# save the data as .Rdata format
save.image(file = "data_chapter3_lab.Rdata")
# To load this .Rdata file late,  use the load function: load("data_chapter3_lab.Rdata")
```