SDS 291 Data Appendix

Isabel Gomez, Ester Zhao, Karina Lieb

March 20, 2019

Data wrangling

```r
#First load ipums
library(ipumsr)

# Note that you can pass in the loaded DDI into the library(read_ipums_micro()`
nhis_ddi <- read_ipums_ddi("nhis_00002.xml")

nhis_data <- read_ipums_micro(nhis_ddi, verbose = FALSE)

#Finding the variables that have a label
nhis_data %>%
  select_if(is.labelled)

# Convert the labels to factors (and drop the unused levels) and filter for only 2017
nhis_data<-nhis_data%>%
  mutate(FSRAWSCORE = droplevels(as_factor(FSRAWSCORE)),
         REGION = droplevels(as_factor(REGION)),
         RACE = droplevels(as_factor(RACEBR)))

#Filter for only 2017
nhis_data <- nhis_data %>%
  filter(YEAR == 2017)
```

```r
#Filter GOTSTAMPFAM to only include those who answered "Yes" Filter for those only who received SNAP be

#GOTSTAMPFAM Codes: 10 = NO, 21 = Yes in last calendar year, 22 = Yes in last month.
#This will remove all observations where GOTSTAMPFAM does not equal 21 or 22
nhis_data<-nhis_data %>%
  filter(GOTSTAMPFAM==21 | GOTSTAMPFAM==22)
```

```r
#Convert FSRAWSCORE into a categorical variable, called FoodSecurity, with 3 levels: Secure (0-2), Low 

#FSRAWSCORE Codes:

nhis_data<- nhis_data %>%
  mutate(Food_Security = if_else(FSRAWSCORE %in% c(0,1,2), "Secure", "Food Insecure")) %>%
  mutate(FOOD_SECURITY=as.factor(Food_Security))
```

```r
#---------- WRANGLE CONFOUNDING VARIABLES -------------

#REGION: a categorical variable with string labels as values.

#AGE: A numerical variable indicating age of the individual surveyed.

#EDUC: Create a categorical variable education, with three levels. Did not graduate high school (0), hi
nhis_data<-nhis_data %>%
  mutate(education=if_else(EDUC %in% c(500, 601, 602, 603), 3, #bachelor's or higher
                           if_else(EDUC %in% c(402, 403), 2, #vocational degree
                                   if_else(EDUC %in% c(301, 302, 401), 1, 0)))) %>%
    mutate(EDUCATION=as.factor(education))#high school or GED. Includes college drop outs.
                                  #else, set to 0 (did not complete high school)


#RACE: Has been set to labels. Filter out those who chose not to answer race.
nhis_data<-nhis_data %>%
  filter(RACE != "Unknown-refused" | RACE != "Unknown-not ascertained" | RACE != "Unknown-don't know")


  #INCOME: Make into a categorical variable. Low-Income includes any family making $0 - $49,999, Middle
nhis_data<-nhis_data %>%
  mutate(Fam_Income=if_else(INCFAM07ON %in% c(10,11,12), "Low Income",
                            if_else(INCFAM07ON %in% c(21,22,23), "Middle Income",
                                    if_else(INCFAM07ON %in% c(96,99), "N/A", "High Income")))) %>%
  mutate(FAM_INCOME=as.factor(Fam_Income))

#For Income, do we want remove the undefined and unknown, re-labelled as N/A.

  #HEALTH: Make into a binary inducator variable Health_Issues, which has a value of (0) if the individ
 nhis_data <- nhis_data %>%
   filter(HEALTH %in% c(1,2,3,4,5))

nhis_data <- nhis_data %>%
   mutate(Health_Issues=if_else(HEALTH %in% c(1,2,3), 0, 1)) %>% #0 = health, 1= poor
   mutate(HEALTH_ISSUES = as.factor(Health_Issues))



  #NCHILD: already contains total number of children. Cap at 9 total. The max amount of children in the


  #FAMOLDNO: number of persons in the household aged 65 or older. Data collection cap at 5 total. The m

  #NUMPREC: number of individuals in the household total. "A 3 digit numeric value". Should we categori

 #GOTSTAMPFAM: Mutate into a factor RECEIVED_STAMPS with two levels, "YES" or "NO"
  nhis_data <- nhis_data %>%
    mutate(Received_Stamps = if_else(GOTSTAMPFAM %in% c(21,22,20), "Yes", "No")) %>%
    mutate(RECEIVED_STAMPS=as.factor(Received_Stamps))
```

Select only the variables we'll need for analysis

```
nhis_data <- nhis_data %>%
  select(NHISPID, AGE, FOOD_SECURITY, NCHILD, NUMPREC, FAMOLDNO, HEALTH_ISSUES, RACE, EDUCATION, FAM_IN
```

Structure and Names

```r
str(nhis_data, give.attr = FALSE)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    10912 obs. of  12 variables:
##  $ NHISPID        : atomic  0020170000080101 0020170000080102 0020170000470101 0020170000470102 ...
##  $ AGE            :Class 'labelled'  atomic [1:10912] 27 10 36 36 8 7 84 33 39 11 ...
##  $ FOOD_SECURITY  : Factor w/ 2 levels "Food Insecure",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ NCHILD         :Class 'labelled'  atomic [1:10912] 1 0 2 2 0 0 0 3 3 0 ...
##  $ NUMPREC        : atomic  2 2 5 5 5 5 5 5 5 5 ...
##  $ FAMOLDNO       : atomic  0 0 1 1 1 1 1 0 0 0 ...
##  $ HEALTH_ISSUES  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ RACE           : Factor w/ 8 levels "White","Black/African American",..: 2 2 1 1 1 1 1 6 6 6 ...
##  $ EDUCATION      : Factor w/ 4 levels "0","1","2","3": 2 1 2 4 1 1 2 3 3 1 ...
##  $ FAM_INCOME     : Factor w/ 4 levels "High Income",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ REGION         : Factor w/ 5 levels "Northeast","North Central/Midwest",..: 3 3 4 4 4 4 4 1 1 1 .
##  $ RECEIVED_STAMPS: Factor w/ 1 level "Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

There are 11 variables in the data we are using, and 10926 observations. The variables are:

1. `AGE` states the surveyed individual's age (number)
2. `FOOD_SECURITY` states the household's food security as a binary indicator variable with two levels, "Secure", and "Insecure".
3. `NCHILD` states how many children have been born to the household in the last year (categorical).
4. `NUMPREC` states how many individuals total live in the household (categorical).
5. `FAMOLDNO` states how many individuals over the age of 65 are within the household. (number)
6. `HEALTH_ISSUES` is a binary indicator variable that states whether an individual is healthy (HEALTH_ISSUES=0), or has health issues (HEALTH_ISSUES=1).
7. `RACE` states the race of the individual surveyed as a categorical variable with 8 levels.
8. `EDUCATION` states the highest level of education the surveyed individual received, as a categorical variable with 4 levels.
9. `FAM_INCOME` states the income level of the household as a categorical variable with 4 levels.
10. `REGION` states the region of the US in which the household is located, as a categorical variable with 5 levels.
11. `RECEIVED_STAMPS` states whether or not the household used Food Stamps within the last year, and we have filtered out those households which did not use Food Stamps in the last year.

Variable analysis

```r
favstats(~AGE, data=nhis_data)
```

```
##  min Q1 median Q3 max     mean       sd     n missing
##    0 11     28 50  85 31.50761 22.67857 10912       0
```

The minimum age is 0, which is logical because it includes small children. The maximum is 85 - we may have expected the maximum to be a bit higher, but this is also reasonable. The mean age is 31, which also seems reasonable when surveying households.

```r
tally(~FOOD_SECURITY, data = nhis_data)
```

```
## FOOD_SECURITY
## Food Insecure        Secure
##          3234          7678
```

We see here that from families that participated in SNAP, about 13.51% reported having Very Low Food Security, 29.64% reported having Low Food Security, and 70.36% reported feeling Food Secure. This seems fairly in line with our expectations, as the majority of families that particpated in SNAP are reporting that the system worked well for them, while a percentage reported a lower success rate.

```
tally(~NCHILD, data=nhis_data)
```

```
## NCHILD
##    0    1    2    3    4    5    6    7
## 7283 1524 1129  588  260   99   19   10
```

```
favstats(~NCHILD, data=nhis_data)
```

```
##  min Q1 median Q3 max      mean      sd     n missing
##    0  0      0  1   7 0.6657808 1.14498 10912       0
```

Because this variable can be interpreted as both numeric and categorical, we will do both the favstats() and tally() to analyze the distribution. We would predict that the majority of families would have only had one or two children added to the household in the last year. We can see from the favstats result that the mean value for children born in the last year is 1.67, which supports our expectation. The tally also supports that, showing that for those families that it applied to, 66.7% had one child in the last year.

```
tally(~NUMPREC, data=nhis_data)
```

```
## NUMPREC
##    1    2    3    4    5    6    7    8    9   10   11   12
## 1070 1736 1958 2330 1760 1028  518  284  125   80   11   12
```

```
favstats(~NUMPREC, data=nhis_data)
```

```
##  min Q1 median Q3 max     mean       sd     n missing
##    1  2      4  5  12 3.921554 1.935038 10912       0
```

Because this variable can be interpreted as both numeric and categorical, we will do both the favstats() and tally() to analyze the distribution. We would predict that most households had a total of about 3-5 people living in them. In the favstats result, we see that the mean is 3.92, which is in line with out expectation. We can see from the tally results that households with 3-5 people in them make up 55.4% of the surveyed population.

```
tally(~FAMOLDNO, data=nhis_data)
```

```
## FAMOLDNO
##    0    1    2    3
## 8890 1578  424   20
```

This distribution of number of elderly in the household is expected. We aren't expecting a strong correlation between the number of elderly people in a household and food stamp participation, so we would expect this distribution to reflect the overall population of elderly people in the US. About 80% of houses do not have an elderly member, which makes sense because most Americans are not over the age of 65. There are no unusual values.

```
tally(~HEALTH_ISSUES, data = nhis_data)
```

```
## HEALTH_ISSUES
##    0    1
## 8682 2230
```

We predict that being a recipient of food stamps may be correlated with health issues, because food insecurity and lower income are correlated with lower health. So, we predicted more people in this group will have issues than in the general population, but the majority of people will not have health issues. Our expectation is reflected in the data: about 80% of those on food stamps are healthy.

```
tally(~RACE, data=nhis_data)
```

```
## RACE
##                                    White
##                                     6340
##                   Black/African American
##                                     2748
## American Indian (Includes Eskimo, Aleut)
##                                      319
##                                  Chinese
##                                       46
##                                 Filipino
##                                      101
##                             Asian Indian
##                                       57
##                               Other Race
##                                     1214
##                            Multiple Race
##                                       87
```

These data show that most respondents are white (~58%), which is expected as most Americans are white.We would also expect people of color to be over-represented in the food stamp recipient group, because POC tend to have lower income than white people. This holds true, as about 75% of americans are white overall, and they represent only about 58% of those that receive food stamps. The 8 levels provided for this variable we find a little strange. There is no category for hispanic/latinx people, as well as overall asian people (only options provided are chinese, filipino, or asian indian, which are nationalities rather than race categories). If these were the options given to respondents, this could have caused some confusion in their answers.

```
tally(~EDUCATION, data = nhis_data)
```

```
## EDUCATION
##    0    1    2    3 
## 6121 3663  623  505 
```

The number of people who did not complete or only completed high school is 89.66% of the surveyed population. This is as expected as it is probably difficult for individuals with only a high school degree or less to get high paying job, meaning they may need the financial support of food stamps to survive.

```
tally(~FAM_INCOME, data =nhis_data)
```

```
## FAM_INCOME
##   High Income    Low Income Middle Income           N/A 
##           290          8532          1360           730 
```

The amount of people who identified as low income was 8532, or 78% of the surveyed individual. This is reasonable as the Food Stamps programs is meant to aid individuals who are low income obtain nutritous meas. However, it is unusual that 290, about 2% individuals identified as high income as this program is meant to aid those with financial hardships.

```
tally(~REGION, data =nhis_data)
```

```
## REGION
##             Northeast North Central/Midwest                 South
##                  1625                  2162                  4736
##                  West               Unknown
##                        2389                          0
```

The number of families receiving food stamps varies by region which is what we expected as there is an income disparity that exist across the United States.

```
tally(~RECEIVED_STAMPS, data =nhis_data)
```

```
## RECEIVED_STAMPS
##   Yes
## 10912
```

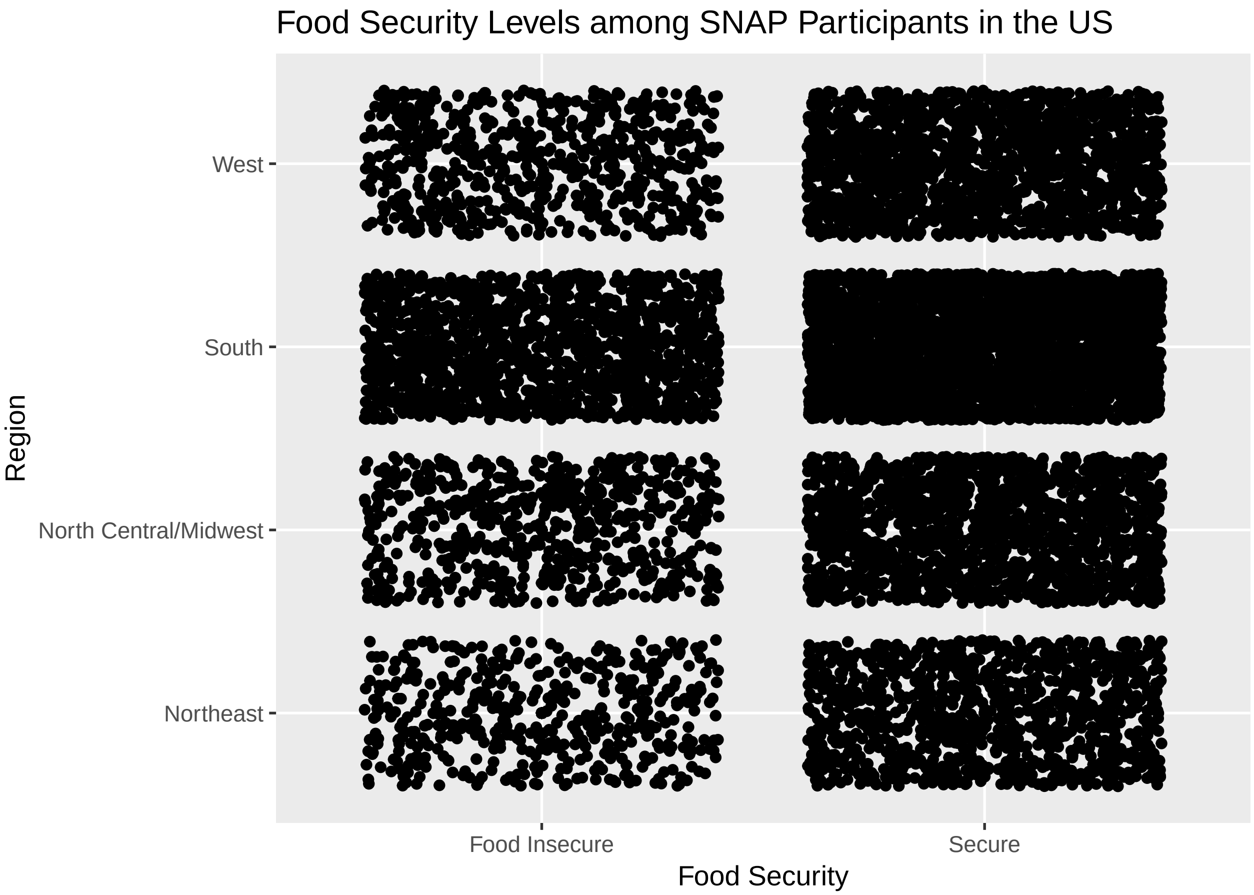
All families in this survey population received food stamps which is what we expected as we filtered out for individuals who only received food stamps.

DRAFT RESULTS

Food insecurity is a large problem in many different communities across the USA. The government provides federally funded food stamps for those who require it using a program called the Supplemental Nutrition Assistance Program (SNAP). This project will focus on the effectiveness of SNAP, in particular whether or not participating families experience varying levels of food insecurity depending on their region of residence and size of household. Our primary hypothesis is that families participating in SNAP will have equal levels of food security across the different regions of residence, adjusting for other confounding variables. Our secondary hypothesis is that for families participating in SNAP, level of food security will be lower for those with larger households, adjusting for other confounding variables including region of residence.

Visualization of primary hypothesis (food security by region)

```
#scatterplot
qplot(x=REGION,y=FOOD_SECURITY,data=nhis_data)+geom_jitter()+coord_flip() + labs( x = "Region", y = "Fo
```



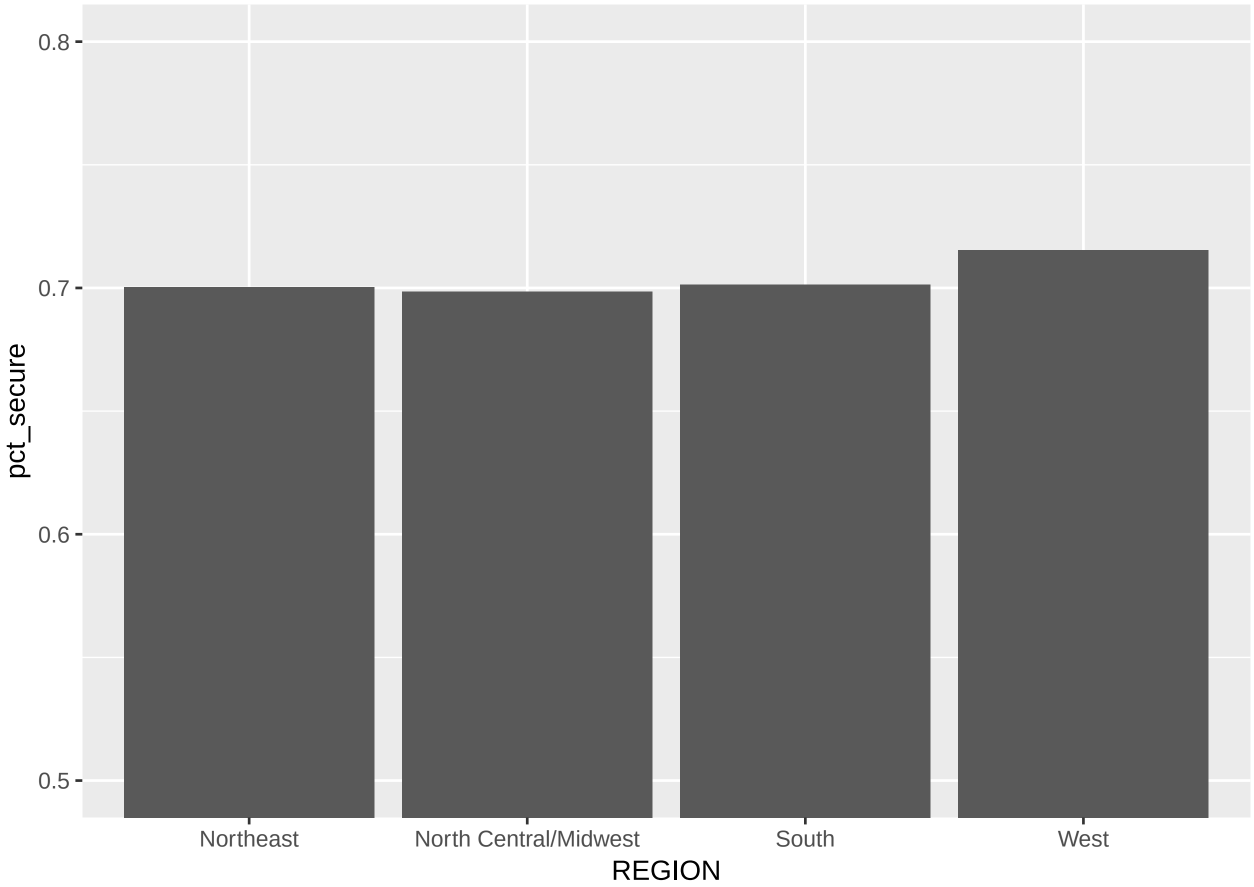
```
bar_data1 <- nhis_data %>%
  filter(FOOD_SECURITY == "Secure") %>%
  group_by(REGION) %>%
  summarise(num_secure = n())

bar_data2 <- nhis_data %>%
  filter(FOOD_SECURITY == "Food Insecure") %>%
  group_by(REGION) %>%
  summarise(num_insecure = n())

bar_data <- full_join(bar_data1, bar_data2, by = "REGION")

bar_data <- bar_data %>%
  mutate(pct_secure = num_secure/(num_secure + num_insecure)) %>%
  mutate(pct_insecure = num_insecure/(num_secure + num_insecure))

#bad barchart
ggplot(data=bar_data, aes(x=REGION, y=pct_secure)) +
  geom_bar(stat="identity") +
  coord_cartesian(ylim=c(0.5, 0.8))
```



```
#good barchart needs stacked percentages. format:
#ggplot(datm,aes(x = REGION, y = percentage???,fill = FOOD_SECURITY)) +
    #geom_bar(position = "fill",stat = "identity")

#Create FOOD_SECURITY related variable that groups by region, and creates a percentage of the whole reg
```

Fitted model of primary hypothesis (food security by region)

Hypothesis Testing:

Our primary hypothesis is that of families participating in SNAP, level of food security will vary for at least one region, adjusting for other confounding variables.

$H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$

Assumptions for logistic regression

```
nhis_data2 <- nhis_data %>%
  mutate(FOOD_SECURE_NUM=if_else(FOOD_SECURITY=="Food Insecure", 0, 1))
#linearity doesnt really apply because the x is categorical (no specific order to regions). we would no

#independence is violated. choose a household head. individual covariants apply to whole household. use

#visualization: bar plot that karina made but with zoomed in y axis
```

Since the response variable for this study is binary, the logistic equation is automatically linear. Furthermore, the data was collected all across the United States through a survey, therefore this passes the random assumption. Finally, since the individuals who were surveyed did not know each other, the independence assumption was also passed as they would not have influenced each other.

Fitting a logistic model

```
#m1<-glm(FOOD_SECURITY ~ REGION  + AGE + NCHILD + NUMPREC + FAMOLDNO + HEALTH_ISSUES + RACE + EDUCATION
m1<-glm(FOOD_SECURITY ~ REGION, data = nhis_data, family=binomial)
summary(m1)
```

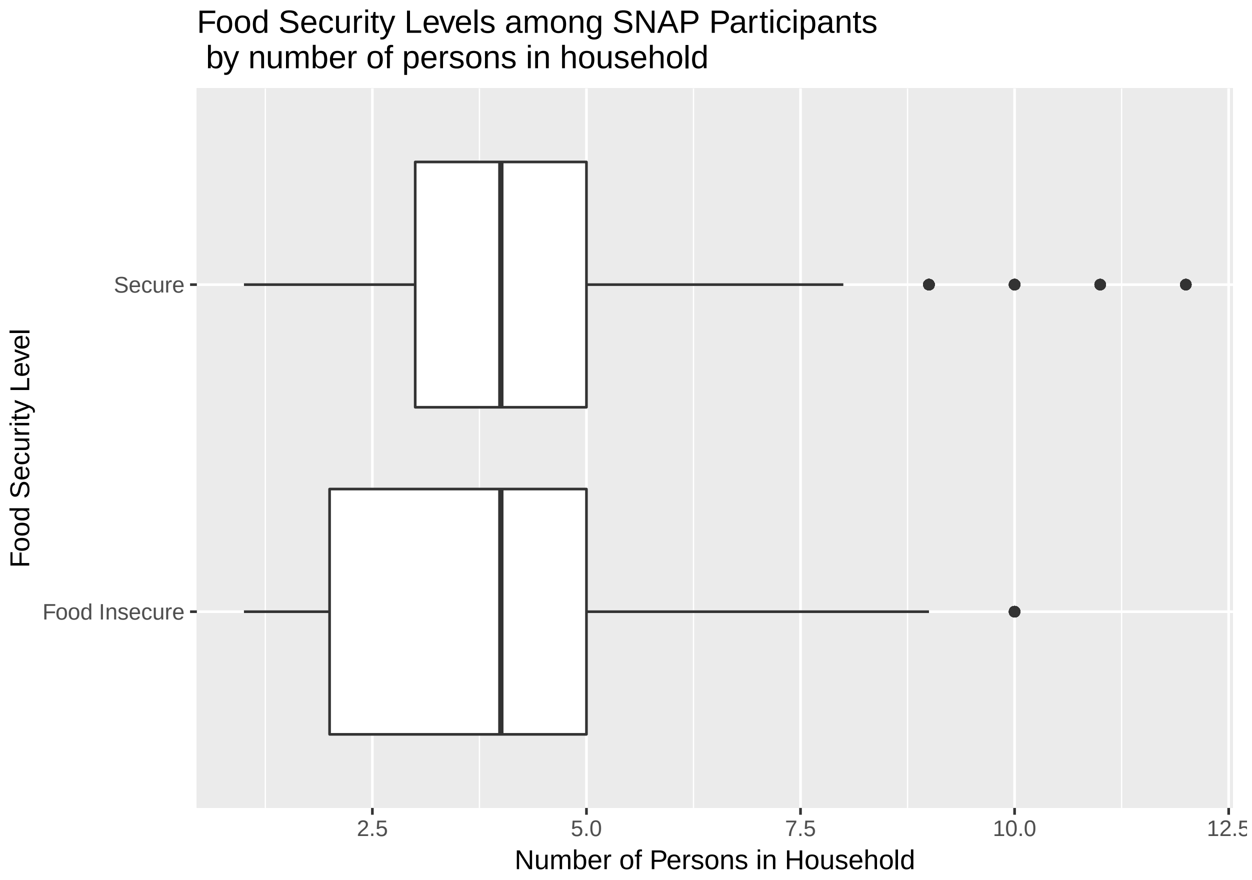
```
##
## Call:
## glm(formula = FOOD_SECURITY ~ REGION, family = binomial, data = nhis_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5853  -1.5484   0.8425   0.8425   0.8473
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.848763   0.054149  15.675   <2e-16 ***
## REGIONNorth Central/Midwest -0.008943   0.071611  -0.125    0.901
## REGIONSouth                 0.004373   0.062769   0.070    0.944
## REGIONWest                  0.072807   0.070625   1.031    0.303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13264  on 10911  degrees of freedom
## Residual deviance: 13262  on 10908  degrees of freedom
## AIC: 13270
##
## Number of Fisher Scoring iterations: 4
```

Model: $FoodSecurity = 0.849 - 0.0089NorthCentral/Midwest + 0.00437South + 0.073West$

According to this model, it is 1% less likely that you will be food secure in the North Central/Midwest region than in the Northeast region.

Visualization of secondary hypothesis (food security by household size)

```
qplot(x=FOOD_SECURITY,y=NUMPREC,data=nhis_data, geom="boxplot") +
  coord_flip() +
  labs( x = "Food Security Level", y = "Number of Persons in Household", title = "Food Security Levels a
```



Fitted model of secondary hypothesis (food security by household size)

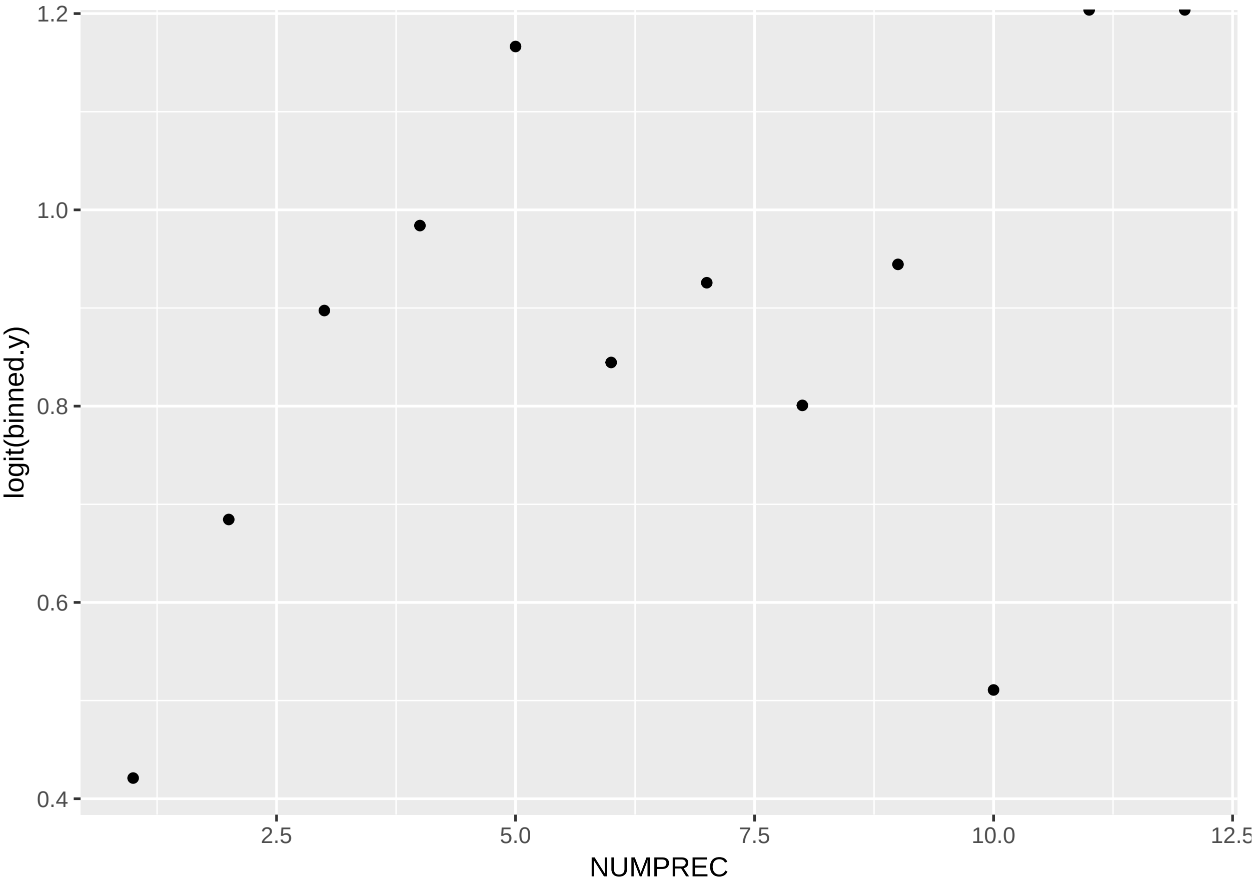
```
m2<-glm(FOOD_SECURITY ~ NUMPREC, data = nhis_data, family=binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = FOOD_SECURITY ~ NUMPREC, family = binomial, data = nhis_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7466  -1.5025   0.8117   0.8594   0.9090
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.60256    0.04722  12.762  < 2e-16 ***
## NUMPREC      0.06773    0.01108   6.113 9.78e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13264  on 10911  degrees of freedom
## Residual deviance: 13226  on 10910  degrees of freedom
## AIC: 13230
##
## Number of Fisher Scoring iterations: 4
```

Assumptions for logistic regression (secondary hypothesis)

```
nhis_data2 <- nhis_data2 %>%
  group_by(NUMPREC) %>%
  summarize(binned.y = mean(FOOD_SECURE_NUM))

ggplot(nhis_data2, aes(x = NUMPREC, y = logit(binned.y))) +
  geom_point()
```


logit(binned.y)
NUMPREC