Assignment 1b: Replicating *Sexual Orientation and Labor Economics* by Cara Brown (1998)

Daniel Sánchez

1/12/23

This document walks through the replication of Cara Brown's 1998 paper *Sexual Orientation and Labor Economics* published at *Feminist Economics*.

Preliminaries

I need the following R libraries to make this code work:

```
library(tidyverse) # For data wrangling
library(haven) # Reading stata files
library(labelled) # For survey labels
library(modelsummary) # For tables
```

I load the data from the given file below:

```
df<-
  read_dta('data/Census_2016_Hierarchial.dta')
```

Data Cleaning

In this subsection I do some data cleaning and exploratory analysis so that I build a fit dataframe for the replication study.

Below, I clean the income and age group variables to have them as I'd need to in the replication. Note I use *haven*'s `as_factor()`, not base R `as.factor()`. We also filter for the ages we'd like to use, then create another age group which combines 45 to 49 with 50 to 54 to create 45-54.

```r
# First, rename the income variable (empin) to something more usable
# Also, we apply value labels as necessary in the mutate() call

df<-
  df %>%
  rename(income = 'empin') %>%
  mutate(agegrp = case_when(
    agegrp ==  88 ~ as.double(NA),
    TRUE ~ as.double(agegrp)
  ))

# Now, replace the income invalid values (88,888,888, 99,999,999) with actual NAs in R.

df <-
  df %>%
  mutate(income_clean = case_when(
      income %in% c(88888888, 99999999) ~ as.numeric(NA),
      TRUE ~ income
    )
  )

df %>% select(PP_ID, agegrp, income_clean) %>% head()
```

```
# A tibble: 6 x 3
  PP_ID agegrp income_clean
  <dbl>  <dbl>        <dbl>
1 11101     10        64000
2 21101      6        47000
3 21102      7           NA
4 21103      1           NA
5 31101     10        43000
6 41101      8        58000
```

```r
# Do the same data cleaning with gender

df$sex <- as_factor(df$sex)

df <-
  df %>%
  mutate(sex = case_when(
    sex == 'Not available' ~ as.factor(NA),
```

```
    TRUE ~ sex
  ))
```

Let us group by person and count how much records I have for that. Test to see if there is more than one observation per person

```
people <-
  df %>%
  group_by(PP_ID) %>%
  summarise(obs= n())

people %>% filter(obs != 1)
```

```
# A tibble: 0 x 2
# ... with 2 variables: PP_ID <dbl>, obs <int>
```

There is only one observation per person in the dataset, as expected

Constructing the presumed homosexual variable

I now construct a dummy variable `hmsxl` which is 1 if the individual is a presumed homosexual and 0 otherwise. The variable is complicated to build as it depends on many things.

First, I will pay attention to the people who live with non-relatives only. I can see this variable from the `cfstat` variable in the census. These are potential couples to someone else, so I will create a dummy for these people, `non_rel`.

```
# Create the variable

df <-
  df %>%
  mutate(non_rel = ifelse(cfstat == 7, 1, 0))

# Now, do the grouping at the houshold level, filtering out the people whose age we're not

households <-
  df %>%
  filter(agegrp %in% c(9,10,11)) %>%
  group_by(HH_ID) %>%
  summarise(non_rel = sum(non_rel),
```

```
            people = n(),
            sexes = n_distinct(sex)) %>%
  arrange(desc(non_rel))

head(households)
```

```
# A tibble: 6 x 4
   HH_ID non_rel people sexes
   <dbl>   <dbl>  <int> <int>
1  98015       7      7     1
2   5948       4      4     1
3 102220       4      4     1
4 126135       4      4     1
5  17901       3      3     1
6  20045       3      5     1
```