Assignment 1b: Replicating Brown (1998)

Daniel Sánchez

Due January 19th, 2023

General comments

Cara Brown's 1998 paper *Sexual Orientation and Labor Economics* published at *Feminist Economics* looks at the wage differential between heterosexual and homosexual Canadians. In this replication, I attempt to capture the sample that she did by looking at 2016 Hierarchical Census Data.

Below, I include my code which creates the tables that I have included above.

Preliminaries

Here I load my libraries and the dataset.

```
# Set my default code chunk options
knitr::opts_chunk$set(
  echo = T,
  warning = F,
  error = F
)

# Load my libraries

library(tidyverse) # For data wrangling
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(haven) # Reading stata files
library(modelsummary) # For tables

# Load my data

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

Homosexual Indicator, Cara Brown approach

I know replicate the homosexual indicator with an attempt at Cara Brown's approach. You will see that I was not able to fully capture the same sample due to a problem with family, which I talk about above.

Cleaning

In this section, I focus on cleaning my dataset for further use. In the following code chunk, I basically take all of my variables of interest then drop out the data which I don't need or that has any kind of missing value. I do this based on the catalogue.

```
df<-
  hierarchical_data %>%
  filter(agegrp %in% c(9,10,11),
         !(empin %in% c(88888888, 99999999)),
         sex != 8,
         !(MarStH %in% c(3,8)),
         fptwk != 1,
         wkswrk == 6,
         ) %>%
  rename(income = 'empin') %>%
  mutate(single = ifelse(MarStH == 1,1,0))
```

I drop the missing values rather than applying an `NA` to them since I only need to create a table of means where I explicitly know what is the data that I need. If I were to run a regression, I'd most likely wouldn't do that.

My code above only considers age groups 9,10 and 11 which are the ones that Cara Brown used for her paper. Further, I consider only people that have worked full time, and also people who have worked for the full year. I drop those who live under common law, as the original paper did not consider that type of people (likely because some institutional context difference). From examining the variable of marital status, `MarStH`, it is evident I won't be able to clearly replicate the table Brown did, since there is no longer a separation between Widowed and separated.

Aggregating at the household level

I use the household identifier, `HH_ID`, to group the dataset at the household level, creating different variables which summarize the individual-level dataset. I then filter for houses that only have two people living in them (after filtering out people who are not in my age group, that is). I must do this simplification in order to be able to accurately identify homosexuals, because when there are houses of more than one person, I cannot be sure if there are homosexual pairs or heterosexual, since we don't know who is related to whom.

I then consider the marital status of the people who are potentially homosexual. As Cara Brown did, I only consider (legally) single people. This is an incorrect approach, since by 2016 marriage between same-sex couples was allowed. I will fix that later.

```
households<-
  df %>%
  group_by(HH_ID) %>%
  summarise(people = n(),
            sexes = n_distinct(sex),
            singles = sum(single)) %>%
  filter(people == 2) %>%
  mutate(
    hmsxl = case_when(
      singles == 2 & people > sexes ~ 'Gay',
```

```
    TRUE ~ 'Not Gay'
))
```