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. The table which completely attempts to replicate Cara Brown's is the following one.

Orientation	Marital Status	Female / 45 to 54	Female / 55 to 64	Male / 45 to 54	Male / 55 to 64
Heterosexual	Married	64174.51	60259.78	86543.60	81978.36
	Separated or Widowed	54393.82	46941.19	55600.00	62420.02
	Single	44698.13	55798.07	48805.56	35599.97
Homosexual		62139.39	63564.50	36897.44	41275.86

You will note that this table shows similar results to Cara Brown's as (1) in general male heterosexuals make more than anyone else, homosexual men earn less than heterosexual women, and straight women make less than homosexual women (in almost all cases, at least). However, there are multiple complications which arise because of different institutional contexts in Canada between 1998 and 2016. First of all, marriage for same-sex couples was legalized in the early 2000s, which means that it is not correct to only consider singles for potentially being homosexual. Second, we do not have last names in our dataset, which means that it is very difficult to leave out those who are not family. Further, I am restricted to households where only 2 people live (after filtering for the age groups I'm interested in) since I cannot know to "pair" people if they are not

I created a table that better shows what possibly could represent a sample that shows presumed homosexuals. First, I do not leave out anyone based on their marital status, as it communicates very little regarding presumed sexual orientation. I sadly cannot control for potential fraternal relationships, since I do not have the last name or a variable which tells me who is related to whom, so I have to leave that bias inside my table. Finally, I consider a couple more age groups to have a better informed perspective about all this. I also consider the common law marital status, which now does exist and most likely holds a big number of presumed homosexuals, as marriage might still be prejudiced against same sex couples.

Orientation	Marital Status	Female / 35 to 44	Female / 45 to 54	Female / 55 to 64	Male / 35 to 44	Male / 45 to 54	Male / 55 to 64
Het.	Married	64600.42	64383.06	60326.58	86029.47	86590.01	82348.51
	Separated or Widowed	50551.29	51484.72	45484.38	53926.83	70554.81	73057.94
	Single	46124.30	46537.06	58251.85	54263.74	53783.13	42548.35
Hom.	Married	54800.00	64138.89	68091.58	58391.91	70080.00	60583.33

Orientation	Marital Status	Female / 35 to 44	Female / 45 to 54	Female / 55 to 64	Male / 35 to 44	Male / 45 to 54	Male / 55 to 64
	Separated or Widowed	60120.00	55750.00	45000.00	49750.00	55260.87	47300.00
	Single	52333.33	59171.80	60933.73	57729.94	48304.19	39409.09

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,
  message = F
)

# Load my libraries

library(tidyverse) # For data wrangling
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 now 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. I also count the **distinct** number of both economic and census families within that household.

```
households<-
  df %>%
  group_by(HH_ID) %>%
  summarise(people = n(),
            sexes = n_distinct(sex),
            singles = sum(single),
            families_econ = n_distinct(EF_ID),
            families_census= n_distinct(CF_ID)) %>%
  filter(people == 2) %>%
  mutate(
    hmsxl = case_when(
      singles == 2 & people > sexes ~ 'Homosexual',
      TRUE ~ 'Heterosexual'
  ))
```

Now, I will have a (presumed) `hmsxl` indicator which captures, in a way, the same sample that Cara Brown did. The most important problem is that I cannot know if the people who are contained in that sample are related by blood or not, at least not in an exact way. This is because I cannot be sure that the people who I have labeled as homosexual are not related in some way. This is because both definitions of family in this census, do not consider "extended" family which doesn't live in the same household. However, it is still plausible for two cousins to live together and also both them could be single. However, without last names or a better family indicator, I cannot really do anything with the economic families.

I do a robustness check below, where I group each economic and census family next to each other and present the results. I determine that there are no economic or census families scattered across different households. It is unlikely to believe that there are no families in Canada which are scattered across different households.

```
econ_families <-
  hierarchical_data %>%
  group_by(EF_ID) %>%
  summarise(n_households = n_distinct(HH_ID)) %>%
  filter(n_households != 1)

econ_families
```

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

The same for census families:

```
census_families <-
  hierarchical_data %>%
  group_by(CF_ID) %>%
  summarise(n_households = n_distinct(HH_ID)) %>%
  filter(n_households != 1)

census_families
```

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

Joining my household-level data to the personal-level data

I know join my household data to the person data by using an inner join, as follows. I do an inner join so that I only get the households which I considered back in my household level to be "eligible" for potentially knowing if there are homosexual or heterosexual couples living in there.

```
df <-
  df %>%
  inner_join(households, by = 'HH_ID')

df %>% select(PP_ID, hmsxl, income, sex)
```

```
## # A tibble: 13,856 x 4
##     PP_ID hmsxl        income sex
##     <dbl> <chr>         <dbl> <dbl+lbl>
##  1 231101 Heterosexual 130000 2 [Male]
##  2 231102 Heterosexual  52000 1 [Female]
##  3 501101 Heterosexual  56000 1 [Female]
##  4 501102 Heterosexual  93000 2 [Male]
##  5 551101 Heterosexual  65000 2 [Male]
##  6 551102 Heterosexual  76000 1 [Female]
##  7 641101 Heterosexual  38000 1 [Female]
##  8 641102 Heterosexual 120000 2 [Male]
##  9 761101 Heterosexual 110000 1 [Female]
## 10 761102 Heterosexual  17000 1 [Female]
## # ... with 13,846 more rows
```

Thus, this table has a binary indicator for presumed homosexuality, as well as their income and other variables of interest.

Building the table

I now build my table. I just need to apply some value labels so that it looks nicer.

```
# Applying value lables and/or cleaning the data a little bit.
```

```
df <-
  df %>%
  mutate(
  agegrp = case_when(
    agegrp %in% c(9,10) ~ '45 to 54',
    agegrp == 11 ~ '55 to 64'
  ),
  sex = case_when(
    sex == 1 ~ 'Female',
    sex == 2 ~ 'Male'
  ),
  MaritalStatus = case_when(
    MarStH == 2 ~ 'Married',
    MarStH == 1 ~ 'Single',
    MarStH == 4 ~ 'Separated or Widowed'
  ))  %>%
  rename(Orientation = 'hmsxl')

# Building a table

# In Markdown output

datasummary(income * Orientation * MaritalStatus ~ Mean*sex*agegrp,
            data = df,
            output = 'markdown')
```

	Orientation	MaritalStatus	Female / 45 to 54	Female / 55 to 64	Male / 45 to 54	Male / 55 to 64
income	Heterosexual	Married	64174.51	60259.78	86543.60	81978.36
		Separated or Widowed	54393.82	46941.19	55600.00	62420.02
		Single	44698.13	55798.07	48805.56	35599.97
	Homosexual	Married				
		Separated or Widowed				
		Single	62139.39	63564.50	36897.44	41275.86

```
# In excel output

  datasummary(income * Orientation * MaritalStatus ~ Mean*sex*agegrp,
            data = df,
            output = 'table1.xlsx')
```

My other table:

I now do my other table of means, by considering the other conditions that I talked about in the commentary (first section) of this document.

```
# Cleaning the data with the new conditions

df1<-
  hierarchical_data %>%
  filter(agegrp %in% c(7,8,9,10,11),
         !(empin %in% c(88888888, 99999999)),
         sex != 8,
         !(MarStH %in% c(8)),
         fptwk == 1,
         wkswrk == 6,
         ) %>%
  rename(income = 'empin') %>%
  mutate(single = ifelse(MarStH == 1,1,0))

# Grouping at the household level

households1<-
  df1 %>%
  group_by(HH_ID) %>%
  summarise(people = n(),
            sexes = n_distinct(sex),
            singles = sum(single),
            families_econ = n_distinct(EF_ID),
            families_census= n_distinct(CF_ID)) %>%
  filter(people == 2) %>%
  mutate(
    hmsxl = case_when(
      people > sexes ~ 'Homosexual',
      TRUE ~ 'Heterosexual'
  ))

# Back to individual level

df1 <-
  df1 %>%
  inner_join(households1, by = 'HH_ID')

# Labels for my table

df1 <-
  df1 %>%
  mutate(
  agegrp = case_when(
    agegrp %in% c(9,10) ~ '45 to 54',
    agegrp %in% c(7,8) ~ '35 to 44',
    agegrp == 11 ~ '55 to 64'
  ),
  sex = case_when(
    sex == 1 ~ 'Female',
    sex == 2 ~ 'Male'
  ),
  MaritalStatus = case_when(
    MarStH == 2 ~ 'Married',
    MarStH == 1 ~ 'Single',
    MarStH == 4 ~ 'Separated or Widowed'
  ))  %>%
  rename(Orientation = 'hmsxl')
```

```r
# Building a table

# In Markdown output

datasummary(income * Orientation * MaritalStatus ~ Mean*sex*agegrp,
            data = df1,
            output = 'markdown')
```

	Orientation	MaritalStatus	Female / 35 to 44	Female / 45 to 54	Female / 55 to 64	Male / 35 to 44	Male / 45 to 54	Male / 55 to 64
income	Heterosexual	Married	64600.42	64383.06	60326.58	86029.47	86590.01	82348.51
		Separated or Widowed	50551.29	51484.72	45484.38	53926.83	70554.81	73057.94
		Single	46124.30	46537.06	58251.85	54263.74	53783.13	42548.35
	Homosexual	Married	54800.00	64138.89	68091.58	58391.91	70080.00	60583.33
		Separated or Widowed	60120.00	55750.00	45000.00	49750.00	55260.87	47300.00
		Single	52333.33	59171.80	60933.73	57729.94	48304.19	39409.09

```r
datasummary(income * Orientation * MaritalStatus ~ Mean*sex*agegrp,
            data = df1,
            output = 'table2.xlsx')
```