IPUMS*

Aviral Bhardwaj

Instructions to download IPUMS data

1. Go to the IPUMS website
2. Click on "Select Samples" to choose the samples we want to download
3. Unselect "Default sample from each year"
4. Choose "2022 ACS"
5. Click on "Submit Sample Selections"
6. Navigate to the section called "Select Harmonized Variables"
7. Hover over "Household" and click on "Geographic"
8. Select "STATEICP" by clicking on the "+" icon.
9. Hover over "Person" and click on "Demographics"
10. Select "SEX" by clicking on the "+" icon.
11. Hover over "Person" and click on "Education"
12. Select "EDUC" by clicking on the "+" icon.
13. Click on "View Cart" and then "Create Data Extract"
14. Choose the format "CSV" under "Data Format"
15. Click on "Submit Extract Request"
16. Create an account or log in to download the data

Instructions to load IPUMS data

```
# Unzip gz file (uncomment the following code only if the data is in a gz file)

# gzfile <- "usa_00002.csv.gz"
# gunzip(gzfile, remove = FALSE)
```

*Code and data are available at:

Data Preparation

```
ipums_data <- read_csv("usa_00002.csv")

# Convert data to tibble
ipums_data <- as_tibble(ipums_data)
```

Respondents with Doctoral Degree by State

```
# Filter data for respondents with doctoral degree
doctoral_data <- ipums_data |>
  filter(EDUCD == 116) |>
  group_by(STATEICP) |>
  summarise(total_doctoral_count = n())
```

Ratio of Respondents with Doctoral Degree by State

```
# Total number of respondents by state
total_data <- ipums_data |>
  group_by(STATEICP) |>
  summarise(total_respondants = n())

# Merge the two datasets
state_data <- left_join(doctoral_data, total_data, by = "STATEICP")

# California total doctoral degree holders
california_data <- state_data |>
  filter(STATEICP == 71)

california_total_respondents <- 391171

# Ratio of respondents with doctoral degree for california
california_ratio <- california_data$total_doctoral_count / california_total_respondents

# Apply the ratio to get the estimated total respondents
#state_data <- state_data |>
```

```
 # mutate(estimated_total_respondents = total_respondants * california_ratio)

state_data <- state_data |>
  mutate(estimated_total_respondents = total_doctoral_count / california_ratio)
```

Comparison of Estimates and Actual Values

State	Total Doctoral Respondents	Total Respondents	Estimated Total Respondents	Difference
3	2014	73077	124340.024	51263.0243
52	1608	62442	99274.458	36832.4583
98	311	6718	19200.470	12482.4705
40	1531	88761	94520.644	5759.6441
62	1031	59841	63651.720	3810.7205
66	350	20243	21608.247	1365.2465
6	131	6860	8087.658	1227.6580
4	244	14077	15064.035	987.0347
5	177	10401	10927.599	526.5990
71	6336	391171	391171.000	0.0000

Discussion

The differences between the estimated and actual number of respondents in each state can arise from the assumption that the ratio of doctoral degree holders to total respondents in California is representative of all states. However, this assumption doesn't account for state-specific variations in educated population or socioeconomic factors.

California has unique characteristics, such as a large population, diverse industries, and numerous research institutions, which may lead to a higher concentration of doctoral degree holders compared to other states. States with smaller populations, fewer universities, or different economic structures might have lower proportions of doctoral degree holders.

Additionally, the distribution of educational attainment across the U.S. is not uniform. States with rural populations, different job markets, or less access to higher education may have lower rates of doctoral degrees, skewing the estimates.

The ratio estimator is a simplified approach, useful for generating rough estimates, but it overlooks local factors that significantly affect educational profiles across states, leading to deviations between the estimated and actual respondent numbers.