Estimating Respondents with Doctoral Degrees Using the 2022 ACS Data

Group 38

2024-10-03

Table of contents

1. Introduction

This document aims to analyze the 2022 American Community Survey (ACS) dataset from IPUMS. We will estimate the total number of respondents in each state who have a doctoral degree using Laplace's ratio estimator. Finally, we'll compare these estimates to the actual number of respondents.

2. Instructions for Obtaining the Data

To access the 2022 ACS data:

1. Visit the [IPUMS website]
2. Register for an account if you haven't already.
3. Once logged in, go to the IPUMS USA section and select the 2022 ACS dataset.
4. Add the following variables to your extract:

- `STATEICP`: State identifier based on IPUMS coding.
- `EDUCD`: Educational attainment detail.

5. Download the `.csv.gz` file and decompress it using:

```
gunzip usa_00002.csv.gz
```

3. Overview of the Ratio Estimator Approach

The ratio estimator is a technique commonly used to estimate the size of a population based on known sample characteristics. Here, we use the following formula to estimate the total respondents in a state:

$$\text{Estimated Total Respondents} = \frac{\text{Respondents with Doctoral Degrees}}{\text{Total Respondents in the State}} \times \text{Known Total in California}$$

where:

- The numerator is the number of respondents with a doctoral degree in each state.
- The denominator is the total number of respondents in each state.
- The known total for California is provided as 391,171.

4. Data Analysis

```
# Load necessary libraries
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(readr)
library(here)
```

```
here() starts at /Users/tina/STA304/2022_ACS_Ratio
```

```r
system("gunzip usa_00002.csv.gz")

# Read the dataset
acs_data <- read_csv(here("usa_00002.csv"))
```

```
Rows: 3373378 Columns: 3

-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (3): STATEICP, EDUC, EDUCD

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Filter for respondents with doctoral degrees
doctoral_data <- acs_data %>%
  filter(EDUCD == 116) %>%
  group_by(STATEICP) %>%
  summarize(doctoral_count = n())

# Count total respondents in each state
state_totals <- acs_data %>%
  group_by(STATEICP) %>%
  summarize(total_count = n())

# Join the two tables
state_data <- doctoral_data %>%
  left_join(state_totals, by = "STATEICP")

# Calculate the ratio between doctoral respondents and total respondents
state_data <- state_data %>%
  mutate(ratio = doctoral_count / total_count)

# Estimate the total number of respondents using the ratio and known total for California
```

```r
california_total <- 391171
state_data <- state_data %>%
  mutate(estimated_total = ratio * california_total)

# Display the state data with estimates
print(state_data)
```

```
# A tibble: 51 x 5
   STATEICP doctoral_count total_count  ratio estimated_total
      <dbl>          <int>       <int>  <dbl>           <dbl>
 1        1            600       37369 0.0161           6281.
 2        2            165       14523 0.0114           4444.
 3        3           2014       73077 0.0276          10781.
 4        4            244       14077 0.0173           6780.
 5        5            177       10401 0.0170           6657.
 6        6            131        6860 0.0191           7470.
 7       11            152        9641 0.0158           6167.
 8       12           1438       93166 0.0154           6038.
 9       13           2829      203891 0.0139           5428.
10       14           1620      132605 0.0122           4779.
# i 41 more rows
```

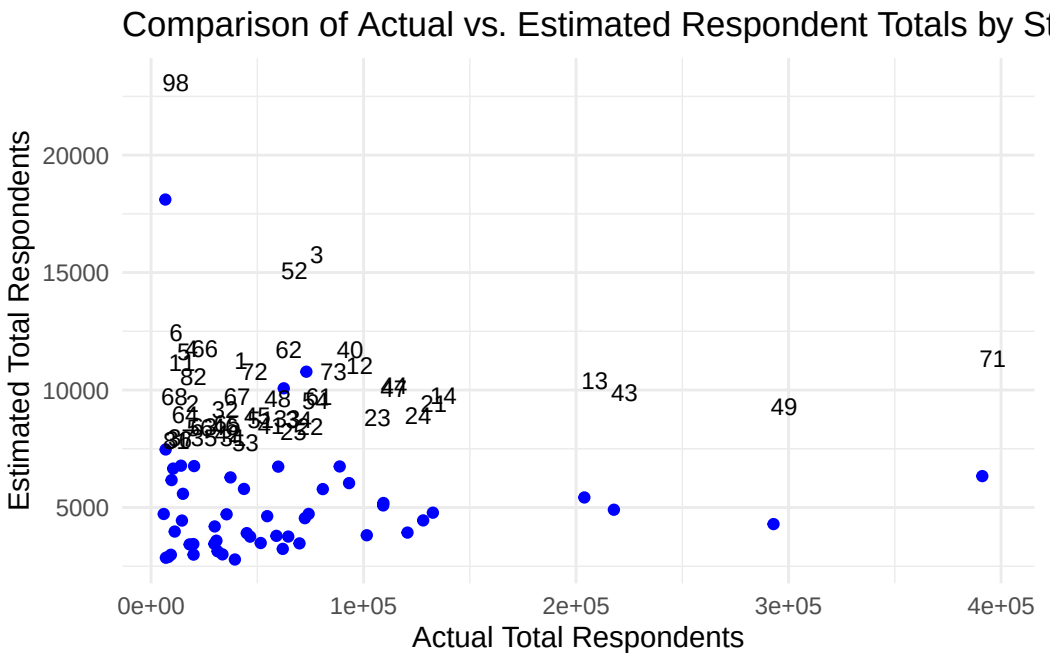
5. Comparison to Actual Respondent Totals

```r
# Calculate the difference between actual and estimated totals
state_data <- state_data %>%
  mutate(difference = total_count - estimated_total)

# Display a summary of the differences
summary(state_data$difference)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-11391   14092   41132   61048   72230  384835
```

```r
# Visualize the comparison between actual and estimated counts
library(ggplot2)

ggplot(state_data, aes(x = total_count, y = estimated_total, label = STATEICP)) +
  geom_point(color = 'blue') +
```

```
  geom_text(nudge_x = 5000, nudge_y = 5000, size = 3) +
  labs(title = "Comparison of Actual vs. Estimated Respondent Totals by State",
       x = "Actual Total Respondents",
       y = "Estimated Total Respondents") +
  theme_minimal()
```



6. Observations and Explanation

The discrepancies between the estimated and actual total respondents in each state can be attributed to various factors:

- **Sample Representation**: The proportion of respondents with doctoral degrees may not be consistent across states, leading to differences in estimates.
- **Variation in Ratios**: States may have varying educational demographics that aren't captured accurately when using a single ratio.
- **Size of the California Sample**: The estimation depends on the accuracy of the known respondent total in California, which may not generalize well across other states.

7. Conclusion

The use of Laplace's ratio estimator provides a quick method to estimate population sizes based on sample characteristics. While useful, the results highlight that estimates can differ significantly from actual values, indicating the importance of considering state-specific demographics and context.

Appendix

The dataset and code used in this analysis are available on GitHub.
