Using Ratio Estimator Approach To Estimate ACS Data*

David Qi

October 3, 2024

This paper uses Ratio Estimator Approach on ACS 2022 data. Resulting in a estimation with large error. We then discusses possible reasons for error.

1 Introduction

In this paper, we make use R (R Core Team (2023)) and the tidyverse package (Wickham et al. (2019)) with supporting package dplyr(Wickham et al. (2023)) to analysis data form the 2022 ACS(American Community Surveys). The data set is from IPUMS (Ruggles et al. (2021)). The table in our study are generated by knitr (Xie (2014)). We have used some information form the book Telling Stories With Data(Alexander (2023)) on IPUMS and Ratio Estimator Approach.

The remainder of this paper is structured as follows: We talk about how the data is gained under Data section. And describe the method of data estimation we use in the Method section. Then, we provide a table that compares our estimation with actual data in the Results section. We then discusses reason of error in estimation Discussion section.

2 Data

2.1 Overview of dataset

The data can be obtained form (Ruggles et al. (2021)) by the following steps:

1. Go to the IPUMS USA Home page(https://usa.ipums.org/usa/)
2. Click on "REGISTER", then follow the instructions and create an account

*Code and data are available at: https://github.com/UTDQi/2022ACS_Analysis

3. Log in to the created account, and go back to the Home page.
4. Click on "SELECT DATA"
5. Under the "HOUSEHOLD" drop down menu, select "GEOGRAPHIC"
6. Click on the $\bigoplus$ symbol beside the "STATEICP" Variable to add this variable to cart.
7. Under the "PERSON"drop down menu, select "EDUCATION"
8. Click on the $\bigoplus$ symbol beside the "EDUC" Variable to add this variable to cart.
9. Click on "SELECT SAMPLES" and select only the 2022 ACS sample, then click on "SUBMIT SAMPLE SELECTIONS"
10. Click on "CREATE DATA EXTRACT", on the next page, change the data format according to your needs.
11. Click on "SUBMIT EXTRACT"
12. When the status of the data is "Completed", click on the download button beside it.
13. The data is now downloaded, you may click on "Basic" under the "CODEBOOK" section beside this data to see how to understand this data.

3 Method

3.1 Ratio Estimator Approach

Ratio Estimator Approach is a method to estimate unknown data from what we know. We use this method to recreate the number of respondents for each state form the ratio between number of respondents and number of respondents that had a doctoral degree as their highest educational attainment in California.

The number of respondents that had a doctoral degree as their highest educational attainment in California is:

$$D_{71} = 6336$$

The number of respondents in California is:

$$R_{71} = 391171$$

Which give us the Ratio:

$$Ratio_{71} = \frac{391171}{6336}$$

Then for any state S with number of respondents that had a doctoral degree as their highest educational attainment D_S, we estimate the number of respondents in that state as:

$$\hat{R_S} = \lfloor D_S \times \frac{391171}{6336} \rfloor$$

Where $\lfloor\ \rfloor$ is the floor function that rounds a number to the largest integer smaller than it.

4 Results

We used the Ratio Estimator Approach to as defined in the Method section. And compared it with the actual number of respondents in that state. (Table 1) is the full table of estimated and actual number of participants in each state:

Table 1: Comparison of estimated and actual data

State	Estimated respondents	Actural respondents	Error
Connecticut	37042	37369	-327
Maine	10186	14523	-4337
Massachusetts	124340	73077	51263
New Hampshire	15064	14077	987
Rhode Island	10927	10401	526
Vermont	8087	6860	1227
Delaware	9384	9641	-257
New Jersey	88779	93166	-4387
New York	174656	203891	-29235
Pennsylvania	100015	132605	-32590
Illinois	89952	128046	-38094
Indiana	38277	69843	-31566
Michigan	61182	101512	-40330
Ohio	74888	120666	-45778
Wisconsin	31671	61967	-30296
Iowa	15928	33586	-17658
Kansas	19817	29940	-10123
Minnesota	35314	58984	-23670
Missouri	38339	64551	-26212
Nebraska	9445	19989	-10544
North Dakota	3704	8107	-4403
South Dakota	4383	9296	-4913
Virginia	94520	88761	5759
Alabama	28399	51580	-23181
Arkansas	15496	31288	-15792
Florida	168606	217799	-49193
Georgia	89581	109349	-19768
Louisiana	27782	45040	-17258
Mississippi	16237	29796	-13559
North Carolina	87729	109230	-21501
South Carolina	39944	54651	-14707
Texas	198548	292919	-94371
Kentucky	27658	46605	-18947

Table 1: Comparison of estimated and actual data

State	Estimated respondents	Actural respondents	Error
Maryland	99274	62442	36832
Oklahoma	17348	39445	-22097
Tennessee	51921	72374	-20453
West Virginia	9816	18135	-8319
Arizona	55317	74153	-18836
Colorado	63651	59841	3810
Idaho	10804	19884	-9080
Montana	6976	11116	-4140
Nevada	17410	30749	-13339
New Mexico	21608	20243	1365
Utah	26423	35537	-9114
Wyoming	4445	5962	-1517
California	391171	391171	0
Oregon	39944	43708	-3764
Washington	73776	80818	-7042
Alaska	3148	6972	-3824
Hawaii	13211	14995	-1784
District of Columbia	19200	6718	12482

The mean and of error($e_S = \hat{R_S} - R_S$) is:

$$\bar{e} = -12785.39$$

And standard deviation is:

$$\sigma(e) = 21219.03$$

Also if we define mean percentage error ($\bar{P_e}$)by:

$$\bar{P_e} = mean(R_s/e_s)$$

We have

$$\bar{P_e} = -0.1955785$$

5 Discussion

We see that the error is large for most states, and the average percentage error is at 19%. This is because educational status is different across states, different states have different ratio of people with doctor's degree. So our initial hypothesis to use ratio estimation was incorrect.

Form the negative mean error we see that California is a state with higher ratio of people who have a doctors degree. So using ratio estimation via the data form California will underestimate the number of respondents of most states. Some states, however, have even higher ratio of people with doctor degree, our estimation overestimated number of respondents for those.

6 Conclusion

The result suggest that Ratio estimator using the data of people with doctor degree from a single state does not estimate the data well. Future studies can use more sophisticated ways of estimation, such as regression or neural networks.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. "IPUMS USA: Version 11.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/d010.v11.0.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.