Replication Package for "Statistical Bias in Racial and Ethnic Disparity Estimates Using BIFSG"

Elena Derby
Connor Dowd
Jacob Mortenson

April 8, 2025

1 Overview

This replication package provides code for the data analysis reported in "Statistical Bias in Racial and Ethnic Disparity Estimates Using BIFSG". The majority of data used in this paper comes from administrative tax data and proprietary data from the US Department of Housing and Urban Development. These data were accessed and handled solely by the authors, who are all employees of the Joint Committee on Taxation in the US House of Representatives. All results and replication codes have been reviewed to ensure that no confidential taxpayer information is disclosed. This document describes the data construction process, and describes the data analysis code in the replication package.

2 Data Construction and Availability

This paper uses administrative tax data from 2017. Under section 6103 of the Internal Revenue Code, this data cannot be shared publicly. The paper also uses proprietary data from the LIHTC Tenant Form, provided to the Joint Committee on Taxation by the US Department of Housing and Urban Development under an agreement stating that the data cannot be shared publicly.

The geographic input data we use to construct BIFSG probabilities come from the 2019 American Community Survey (ACS) five-year sample based on approximately 780,000 households, which consists of racial and ethnic distributions (percent composition) by zip code from 2015 through 2019 (Census, 2019). The Census conducts the ACS throughout the year, and after each calendar year responses are grouped across time to generate and publish estimates for one-year and five-year periods. We use first name data published by Rosenman et al. (2022) as second input in the process of constructing BIFSG probabilities. These data come from the voter files of six states − Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina − and consist of probability distributions of belonging to each of five racial and ethnic groups associated (independently) with first names, middle names,

and surnames. The race and ethnicity categories are limited to Non-Hispanic (NH) White, NH Black, Hispanic, NH Asian, and other. Finally, We use the 2010 Census Surname Table as a third input source in constructing BIFSG probabilities (Census, 2010). The data from the Surname Table consist of probability distributions of belonging to six distinct racial and ethnic groups associated with each of the 2,254 surnames used by 100 or more people in the United States. Individuals with names listed in the file represent approximately 90 percent of individuals surveyed in the 2010 Census, with the remaining ten percent of records (30 million) consisting of individuals with one of 6.1 million surnames that belong to fewer than 100 individuals and are subsequently classified as "All Other Names."

3 Data Analysis

There are five files provided in this replication package, which are used for the following purposes:

1. **Generate_bifsg_probs.R**: This is an R file used to generate BIFSG probabilities for every individual in our LIHTC Tenant sample using their first name, last name, and zip code as inputs. The input file is a data set containing the names, zip codes, and taxpayer identification numbers of LIHTC residents who we have matched to administrative tax records. The output file is the same group of individuals in the sample data (including spouses listed on returns) indexed by taxpayer identification number, and their calculated BIFSG probabilities of belonging to one of six ethnic/racial groups.

2. **Gen_regdata_Stage1.R**: This is an R file used to merge taxpayer information with BIFSG probabilities in order to create the data set used for our analyses. The input files are the BIFSG data file produced using "Generate_bifsg_probs.R" and administrative tax data for individuals matched to the LIHTC Tenant data. This code is also used to create **Figure 1**: First Stage BIFSG Validation by Race/Ethnicity and **Table 1**: BIFSG Mean Predicted Probabilities across Race/Ethnicity.

3. **EITC_validation.R**: This is an R file used to analyze how BIFSG affects estimates of differences in EITC claiming between groups by race and ethnicity. It is used to produce **Figure 2:** EITC Claiming (Percent of Population and Average Dollar Amount), **Figure 3:** Differences in Unconditional EITC Claimed between Non-White and White Taxpayers, and **Figure 4:** Differences in Unconditional EITC Claimed between Non-White Taxpayers of Different Race/Ethnicity.

4. **Online_Appendix_B.R**: This is an R file used to analyze how BIFSG affects estimates of differences in audit rates between groups by race and ethnicity. It is used to produce the figures in **Online Appendix B**.

5. **Online_Appendix_C.R**: This is an R file used to test the bounding methodology introduced in Elzayn et al. (2024). It is used to produce the figures in **Online Appendix C**.

References

Elzayn, Hadi, Evelyn Smith, Thomas Hertz, Cameron Guage, Arun Ramesh, Robin Fisher, Daniel E. Ho, and Jacob Goldin. 2024. "Measuring and mitigating racial disparities in tax audits." *The Quarterly Journal of Economics*.

Rosenman, Evan T. R., Santiago Olivella, and Kosuke Imai. 2022. "Race and ethnicity data for first, middle, and last names." Harvard Dataverse V9.

United States Census Bureau. 2010. "Frequently Occurring Surnames from the 2010 Census." `https://www.census.gov/topics/population/genealogy/data/2010_surnames.html`.

United States Census Bureau. 2019. "2015-2019 American Community Survey 3-year Public Use Microdata Samples." `https://data.census.gov/`.