Company: National Institute of General Medical Sciences

Project Title: Using ACS data to understand the demographic and socioeconomic makeup of biomedical research workers in the United States

Background:

The National Institute of General Medical Science (NIGMS) supports basic research that increases understanding of biological processes and lays the foundation for advances in disease diagnosis, treatment, and prevention. Understanding the effectiveness of the Institute's programs is necessary to promote scientific discovery, advancement, and achievement; enhance the stewardship of public resources; and excel as a federal science agency in the responsible management of scientific and organizational results. The Division of Data Integration, Modeling, and Analytics (DIMA) – NIGMS' primary division for all data-driven discussions, decisions, and actions – is tasked with conducting multiple types of analyses and evaluations to inform programmatic and business process improvements and enhancements for the Institute. Although analyses and evaluations primarily rely on internal data, DIMA staff will occasionally use external data to help frame and understand various research problems.

Websites:

https://www.nigms.nih.gov/

https://www.nigms.nih.gov/about-nigms/what-we-do/data-integration-modeling-and-analytics

Project Overview and Problem Definition:

NIGMS administers numerous programs designed to foster the training and development of a strong and diverse biomedical research workforce. These programs generally aim to progress individuals through various career stages with the ultimate goal of contributing to a vibrant, robust biomedical research enterprise. An initial component for understanding the effectiveness of these programs is to understand the preexisting makeup of the biomedical research workforce. This is information that NIGMS currently does not have at a granular level. For example, NIGMS does not have estimates of the distribution of wage income of biomedical researchers stratified across demographic and other socioeconomic attributes, industry (i.e., private sector vs. government vs. self-employed), or geography. Having a statistically robust picture of the demographic and socioeconomic makeup of the biomedical research workforce in the United States is highly desirable for the purposes of ongoing planning, evaluation, and programmatic evolution.

Project Goals:

Estimate descriptive demographic and socioeconomic statistics of the U.S. biomedical research workforce, stratified by industry and state. Although measures of central tendency are necessary, insights into distributions are desirable. Generally, we are interested in better understanding the makeup of workers conducting biomedical research in the U.S.

Potential Roadblocks and Barriers to Success:

- Students lacking familiarity with the use of survey data, especially weighted survey responses, for descriptive analysis may have additional barriers to success in this project.

Preferred Methodology:

- The Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) is the prime dataset for this project. Although weighting needs to be considered during analysis, the process itself is relatively easy and straightforward. Most well-known statistical software have sufficient survey data analysis functions for these data.

Data Requirements and Availability:

- The team will need to acquire the data from the Census Bureau's website:https://www2.census.gov/programs-surveys/acs/data/pums/

Analytics requirements:

- A descriptive analysis will suffice. The complexity will be in proper stratification and then presentation of the groups. A general understanding of survey data, sampling weights, and stratification will be necessary.
- Time permitting: use an additional set of time points (i.e., earlier years' data files) to do a regression model and identify potential trends or statistically significant changes in the descriptive statistics.

Preferred Tooling:

- R Statistical Software is preferred, but Python is acceptable as long as documentation is robust enough to detail the process to a person with data science experience but lacks fluency in Python.
- Familiarity with BI platforms such as Tableau or PowerBI is a plus, especially for any outputs that involve interactive visualizations.

Project Schedule:

- Weeks 1-4: Data product orientation
    - Download and merge the nationwide person-level 5-year data files.
    - Familiarize yourself with the ACS instrument and the PUMS data.
        - Including sample design (e.g., timing, geographic entities (PUMAs), etc.)
        - Review the questionnaire to better understand the data.
    - Downloadable files: https://www2.census.gov/programs-surveys/acs/data/pums/
        - Data files: https://www2.census.gov/programs-surveys/acs/data/pums/2022/1-Year/csv_pus.zip

    - Select occupation and industry codes to identify biomedical research workers.
        - DIMA can work with students to ensure that correct codes are being used.
        - Occupation titles of interest: Biological Scientists, all others; Microbiologists; Biochemists and Biophysicists; Medical Scientists, except Epidemiologists; Life Scientists, all others; Chemists; Biological Technicians; Chemical Technicians.
- Weeks 5-9: Begin descriptive analysis
    - Design table shells to help orient the analysis.
    - Using weights properly to stratify, and estimate various statistics of demographic and socioeconomic attributes for biomedical research workers by industry and state.
    - Establish a procedure to verify estimates.
        - https://data.census.gov/ is a good resource and will allow the students to produce various tables that will allow them to verify different statistics throughout your process. The website will not be able to produce more granular, stratified statistics, but can serve as a way to easily verify more general statistics.
    - Students should use this time to explore the statistical capabilities of the survey data.
        - E.g., are sample sizes large enough in each state to estimate all desirable demographic and socioeconomic statistics? If not, are there meaningful ways to categorize and group states in order to pool the data?
    - Time permitting: regression analysis
- Weeks 10-12: Prepare and finalize deliverables.
    - Develop table shells for the final product.
        - At this point, there should be a solid understanding of what statistics the data are capable of estimating, in terms of varying sample sizes of stratified groups.
    - Estimate statistics and populate the table shells.
        - Verify throughout.
    - Presentation of results to NIGMS leadership and other audiences (e.g., GWU)

Confidentiality Concerns:

NIGMS would like to review content prior to wider distribution to ensure accurate and appropriate messaging. Otherwise, data are publicly accessible so there are no major concerns about NDAs or other confidentiality agreements.

Budget

- Data are publicly available/open source and free.
- All software planned for use in analysis is also open source and free, with the exception of dashboarding tools which are optional.

Contact:

Andrew Miklos, Ph.D., https://www.nigms.nih.gov/about/Pages/miklos.aspx, andrew.miklos@nih.gov