Analyzing the American Community Survey with *R*

Anthony Damico[1,*] , Rachel Licata[1]

1. Kaiser Family Foundation
*Contact author: adamico@kff.org

Topic Area: Statistics in the Social and Political Sciences

Keywords: American Community Survey, Health Policy, Big Data, SQL, County-Level Data

Research Objective: Designed to replace the information gathered by the decennial Census, the American Community Survey (ACS) is likely to play an increasingly important role in the future of health policy research, especially for regional and state-level studies. Despite powerful statistical analysis, data manipulation, and presentation capabilities, the open-source R Statistical Software Package has not become widely adopted in the fields of health economics, health policy, or healthcare management and evaluation. User guidelines and other technical documents for government-funded and publicly-available data sets rarely provide appropriate syntax examples to R users. The objective of this presentation will be to describe the steps required to import ACS data into R, to prepare that data for analysis using the replicate weight variance and generalized variance formula calculation techniques, and to produce the principal set of statistical estimates sought by health policy researchers.

Study Design: This presentation reviews the step-by-step method explanations and syntax needed to analyze the ACS with the R Statistical Software package. This includes importation instructions, estimate calculations, variance and error term calculations, as well as linkages to other data sets. In order to equip healthcare researchers with the tools needed to analyze this large dataset on their personal computers, each of these steps include a brief discussion of absolute minimum computing requirements, as well as detailed workarounds and shortcuts for the more memory-intensive processes.

Population Studied: The ACS represents all civilian, noninstitutionalized Americans. The examples used in this presentation include state and regional estimates; however, all instructions and syntax are presented with the intention of allowing the researcher to re-define a population of interest with minimal effort.

Principal Findings: Depending on research budget, computing resources, and level of programming skill, conducting analyses of the ACS with R often presents a viable alternative to other statistical analysis packages.

Conclusions: Given the large file size of the ACS, interested health policy researchers may be limited in their ability to analyze this survey with off-the-shelf statistical software packages due to memory overload issues. Although R users often experience similar memory limits and problems, the flexibility of the core R programming language and its integration with both parallel processing engines and Structured Query Language (SQL) allows for the relatively straightforward analysis of large, complex-sample survey data such as the ACS.

Implications for Policy, Practice or Delivery: Providing health policy researchers and statisticians with a strategy to work with the American Community Survey using freely available software will increase their ability to examine a variety of health and demographic indicators at the sub-national level. By outlining the technical steps to analyze this data, researchers could study topics such as regional characteristics of Health Professional Shortage Areas, state demographic factors associated with Medicare Advantage plan premiums, or county-level demographics of uninsured populations. An understanding of the methods needed to work with the ACS will open up the field of geographic region-based analyses to health policy researchers.