THE HENRY J. KAISER FAMILY FOUNDATION

Filling the need for trusted information on national health issues...

April 2015 | Issue Brief

Kaiser Family Foundation ACA Eligibility Analysis: Technical Appendix B: Immigration Status Imputation

To impute documentation status for each person in the sample, we draw on the methods underlying the 2013 analysis by the State Health Access Data Assistance Center (SHADAC) and the recommendations made by Van Hook et. al..[1,2] This approach uses the Survey of Income and Program Participation (SIPP) to develop a model that predicts immigration status; it then applies the model to a second data source, controlling to state-level estimates of total undocumented population from Department of Homeland Security. Below we describe how we developed the regression model and applied it to the Current Population Survey. We also describe how the model may be applied to other data sets. The programming code, written using the statistical computing package R v.3.1.1, is available upon request for people interested in replicating this approach for their own analysis.

Data Sources

We used the second wave of the 2008 Survey of Income and Program Participation (SIPP) panel data to build the regression model. The SIPP Wave Two dataset contains questions on migration history at the person level.

The regression model is designed to be applied to other datasets in order to impute legal immigration status. The code mentioned above includes programming to apply the model to either the SIPP Core file or the Current Population Survey (CPS) (for years 2007 on). Because the SIPP Core file and CPS contain different survey questions and variable specifications, we create unique regression models to apply the model to each dataset. For the updated analysis underlying *The Coverage Gap: Uninsured Poor Adults in States that Do Not Expand Medicaid*, we apply the regression model to the 2014 CPS-ASEC.

Due to underreporting of legal immigration status in the SIPP, in imputing immigration status we control to state and national-level estimates of the undocumented population from the Department of Homeland Security, Office of Immigration Statistics. DHS reports estimates for the nation and for states with the highest population of unauthorized immigrants.[3] It also includes estimates by age categories.

Construction of Regression Model

We use the SIPP Wave Two to create a binomial, dependent variable that identifies a respondent as a potential unauthorized immigrant. The dependent variable is constructed based on the following factors:

1) Respondent was not a United States (US) citizen,
2) Respondent did not have permanent resident status upon US entry,
3) Respondent's immigration status did not change to permanent resident since US entry, and
4) Respondent does not have other indicators that imply legal status.[4]

We use the following independent variables to predict unauthorized immigrant status:[5]

1) Place of birth,
2) Year of US entry,
3) Whether respondent moved into current residence within the last twelve months,
4) Job industry classification,
5) State of residence,
6) Family Poverty Level,
7) Ownership or rental of residence,
8) Presence of at least one citizen in household,
9) Number of occupants in the household (< or >= six occupants),
10) Whether all household occupants are related,
11) Number of workers in household,
12) Health insurance coverage status,
13) Ethnicity, and
14) Age.

The regression model was sub-populated to remove respondents who could not be considered unauthorized. People who could not be considered unauthorized include people who 1) were born in the US, 2) are US citizens, or 3) have other indicators that imply legal status.[3]

Imputing Unauthorized Immigrants in Other Datasets

We use the DHS estimates as targets for the total number of unauthorized immigrants that the imputation generates. We stratify the targets by state (among the nine states with the highest population of unauthorized immigrants) and by six age categories, for a total of 60 strata. State categories include: Arizona, California, Florida, Georgia, Illinois, New Jersey, New York, North Carolina, Texas, and all other states. Age categories include 18 and under, 18-24 years, 25-34 years, 35-44 years, 45 to 54 years, and 55 and above. We impute immigration status within each stratum.[6]

To generate the imputed immigration status variable, we first calculated the probability that each person in the dataset was unauthorized based on the SIPP regression model. Next, we isolated the dataset to each individual stratum described above. Within each stratum, we sampled the data using the probability of being unauthorized for each person. After sampling, we summed the person weights until reaching the DHS population estimate for each stratum. The records that fell within the DHS population estimate were considered to be unauthorized immigrants. We repeated the process of sampling using the probability of being unauthorized and subsequently summing the person weights to reach DHS targets five times, creating five different unauthorized variables per record. These five imputed authorization status variables were then incorporated into a standard multiple imputation algorithm, closely matching the imputed variable analysis techniques used by the Centers for Disease Control and Prevention for the National Health Interview Survey.[7]

To easily apply the regression model to other data sets, we created a function that applies this approach to a chosen data set. The function first loads the dataset of choice; then standardizes the data to match the independent variables from the SIPP regression model; and finally applies the multiple imputation to generate a variable for legal immigration status.

---

[1] State Health Access Data Assistance Center. 2013. "State Estimates of the Low-income Uninsured Not Eligible for the ACA Medicaid Expansion." Issue Brief #35. Minneapolis, MN: University of Minnesota. Available at: http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2013/rwjf404825.

[2] Van Hook, J., Bachmeier, J., Coffman, D., and Harel, O. "Can We Spin Straw into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches" *Demography*. Forthcoming.

[3] DHS updates these estimates periodically. We use the estimates applicable to the year for the data sets to which we apply the regression model. The most recent estimates are: N Rytina, B Baker. *Estimates of the Unauthorized Immigrant Population Residing in the United States*. (Department of Homeland Security, Office of Immigration Statistics), March 2013. Available at: http://www.dhs.gov/publication/estimates-unauthorized-immigrant-population-residing-united-states-january-2012.

[4] Indicators that imply legal status include: (i) respondent entered the US prior to 1980, or (ii) respondent is enrolled in any of the following public programs: Medicare, military health insurance, public assistance, supplemental security income, or social security income.

[5] The first three listed independent variables are excluded when using the regression model to analyze the SIPP Core Data because they are not included in Core SIPP files.

[6] For more information, see SHADAC 2013, footnote 6. The table created in the function contains data from 2007-2011. Nevada and Washington were not included in the strata as they are not consistently identified as one of the top ten states with high unauthorized immigrant populations from year to year.

[7] For more detail, see documentation available at: National Health Interview Survey. *2013 Imputed Family Income/Personal Earnings Files*. August 14, 2014. Available at: http://www.cdc.gov/nchs/nhis/2013imputedincome.htm