KFF ACA Eligibility Analysis, Technical Appendix B: Immigration Status Imputation

To impute documentation status, 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 for each person in the sample; it then applies the model to a second data source, controlling to state-level estimates of total undocumented population as well as the undocumented population in the labor force from the Pew Research Center.[3] Below we describe how we developed the regression model and applied it to the American Community Survey (ACS). We also describe how the model may be applied to other data sets. The programming code, written using the statistical computing package R v.4.0.3, 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 in surveys that do not ask about migration status. The code mentioned above includes programming to apply the model to either the SIPP Core file, ACS, or the Current Population Survey (CPS). Because the SIPP Core file contains different survey questions and variable specifications from the ACS and CPS, we create unique regression models to apply the model to each dataset. For the analysis underlying this brief and other KFF estimates of eligibility for ACA coverage, we apply the regression model to the 2013 ACS and then each subsequent year of the ACS.

Due to underreporting of legal immigration status in the SIPP, in imputing immigration status we control to state and national-level estimates of the total undocumented population and also the undocumented population in the labor force from the Pew Research Center. Pew reports these estimates for all states and the District of Columbia.[4]

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.[5]

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

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

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.

Imputing Unauthorized Immigrants in Other Datasets

We use the Pew estimates as targets for the total number of unauthorized immigrants that the imputation generates. We first apply this strategy to the 2013 ACS, which contains health insurance information prior to the ACA's coverage expansions. We stratify the targets by state and the District of Columbia and by participation in the labor force. We impute immigration status within each of these 102 strata.[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 Pew population estimate for each stratum. The records that fell within the Pew 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 Pew 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]

We used this first pass on the ACS 2013 to inform our sampling targets for the latest available microdata (ACS 2019). Looking at the results of our undocumented imputation on the ACS 2013, we calculated the share of undocumented immigrants lacking health insurance within each of those 102 strata prior to the ACA's coverage expansions and transferred that information into a new dimension of sampling strata for the ACS 2019. We split each of the 102 sampling strata used on the pre-ACA ACS 2013 into uninsured versus insured categories, resulting in 204 sampling strata for subsequent years. We then repeated our imputation on the ACS 2019 with the newly-divided strata, allowing for a small decline in the undocumented uninsured rate based off of the percent drop in the uninsured rate we see in the Kaiser Family Foundation’s Survey of the Low-Income Population and the ACA.[8]

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.

Endnotes

[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. 2015. “Can We Spin Straw into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches” *Demography*. 52(1):329-54.

[3] This data source is a change from previous KFF analyses, which used estimates from the Department of Homeland Security.

[4] Pew updates these estimates periodically. We use the most recent estimates available at the time of our analysis. We draw on Pew directly for all published data and interpolate years missing from their trend. Our analysis uses the year applicable to the year for the data sets to which we apply the regression model. The most recent estimates as of the time of our analysis were: J Passel, D Cohn. *Mexicans decline to less than half the U.S. unauthorized immigrant population for the first time*. (Pew Research Center), June 2019. Available at: https://www.pewresearch.org/fact-tank/2019/06/12/us-unauthorized-immigrant-population-2017/.

[5] 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.

[6] For more information, see SHADAC 2013, footnote 6. The table created for this function contains estimates of the undocumented across 2013-2019.

[7] For more detail, see documentation available at: National Health Interview Survey. *2019 Imputed Income Files.* August 2020. Available at: https://www.cdc.gov/nchs/nhis/2019nhis.htm.

[8] As an example of this calculation, we found that approximately 59% of undocumented uninsured individuals did not have health coverage in 2013. We allow the undocumented rate to drop slightly after 2013. We base the percent drop in the uninsured rate that we see in the Kaiser Family Foundation's Survey of the Low-Income Population and the ACA (which has a direct measure of citizenship) for 2013 to 2014, which is an 11% decline, to estimate an uninsured rate in 2014 for the undocumented (52%). We use the ratio of that drop relative to the drop for citizens (less than half the scale of the drop for citizens) to estimate a 7% drop from 2014 to 2015, getting us to a 49% uninsured rate in 2015 and repeat this until 2019, resulting in the final undocumented uninsured rate of 47% in calendar year 2019. Prior to implementing this new sampling dimension, we found unrealistic drops in the uninsured rate of the undocumented population that we largely attributed to our prediction model's inability to discern this group from legally-present non-citizens, many of whom are eligible for assistance under the ACA's coverage expansions. Although a few states have implemented programs that allow for coverage of the undocumented population, these programs are state-funded and relatively small in scale compared to the nationwide coverage expansions accompanying the ACA.