An Analysis of Education and Demographic Information on Immigrant Naturalization in America

Chaira Harder, Isabelle Elder, Sirohi Kumar

Abstract

The purpose of our analysis is to investigate the variables that determine if individuals choose to pursue US citizenship status after immigration. Understanding the variables that influence naturalization can inform future policy formation, ensure legal and human rights considerations, and uphold equitable citizenship access. We used data from the IPUMS Higher Education database, which collects data from the Science and Engineers Statistical Data System, to analyze the determinants that increase likelihood of an individual being a naturalized citizen, including race/ethnicity, gender, employment, education level, and number of children. After filtering our data and using variable selection techniques to find an appropriate model, we found that all named predictors except for the number of children have significant impacts on the odds of an individual being a naturalized citizen. The most significant changes in someone's odds of being a naturalized citizen are increases that come from being, interestingly, unemployed, and being male. This research adds to a growing understanding of how different demographic and socioeconomic factors influence the citizenship process, which is especially relevant in this political climate.

INTRODUCTION

The US is home to more international migrants than any other country, with 46.2 million immigrants living in America in 2022.1 Naturalization is the process that grants US citizenship to immigrants that have resided in the US for a minimum of five years and meet certain eligibility requirements.(2) Many immigrants view naturalization as an important milestone that indicates their integration into the United States, and it comes with rights and benefits, including voting and access to certain public-sector jobs among others.(2) Sociodemographics influence the naturalization process, with census data indicating that men, those with higher education levels, and being employed are factors that increase likelihood.(3,4)

Despite naturalization rates hitting a record high since 2008, information about the impact of specific factors or characteristics that lead to individuals choosing naturalization is lacking.(4) This report aims to analyze how different levels of education, employment, gender, race/ethnicity, and number of children affected naturalization. We hypothesize that all our predictor variables will have an impact, with higher educational levels, being employed, and having fewer children, being male, and being an underrepresented minority all increasing the likelihood of naturalization. Further insight about the variables that influence naturalization can inform future immigration policy, and help the US work towards equitable citizenship access.

METHODS

Data

This study uses data from IPUMS-Higher Education. This data set compiles information from the Science and Engineers Statistical Data System, the gold standard for longitudinally studying the education and employment of the US science and engineering workforce. Data from three different surveys sponsored by the National Science Foundation (NSF), including the National Surveys of College Graduates (NSCG), Recent College Graduates, and Doctorate Recipients, has been integrated from 1993 to present day. Of note, the NSRCG was discontinued in 2010, as survey items were folded into the other 2 surveys. However, the data set comprehensively brings together microdata to maintain continuity across variables. These surveys are administered biennially using a national widespread collection approach, where non-institutionalized individuals under the age of 76 who hold a bachelor's degree or higher are asked but not required to respond. A single observation is defined as an eligible individual graduate in the US surveyed by one of the three NSF surveys, who is a naturalized US citizen.

Variables

The outcome variable in our analysis is a variable indicating whether or not the respondent is a US citizen. We wrangled this categorical variable to be binary, where a 1 indicates if

the citizen is naturalized and a 0 represents anything else (citizen by birth, non-citizen, or born abroad to American parents). We filtered the data to remove observations which had an NA value for citizenship status, which brought our data from 115,152 observations (total people) to 103,695 people (citizens). Gender is a binary categorical variable for male or female. Other variables of interest include race/ethnicity, a categorical variable that includes specific categories for Asian and White, with other race categories with fewer observations grouped into "Under-represented minorities" or "other". Our educational variable is a categorical response to their highest degree received, as in a Bachelors, Masters, Doctorate, or Professional degree. Employment is grouped as employed, not employed, or not in the labor force.

Model Selection

We used forward, backward, and stepwise sequential variables selection, with AIC as the goodness-of-fit statistic to determine the best model to use. All three methods found the best model was CTZUS ~ RACETH + GENDER + DGRDG + LFSTAT. All three directions eliminated 'number of children' as a possible predictor. To run this variable selection method, we had to filter the data to remove any observations with NA values for any of these predictors, because the method requires all data be the same length. This left us with 42,566 observations to use in our analysis.

Our model effectively meets the assumptions required for our regression analysis. As all our variables are categorical, linearity is already met. Our data comes from a database of multiple sources, and comprises a small portion of the total US population. Additionally, the surveys used to collect data were administered nationally, in order to collect data from a range of individuals, which means we can assume randomness. Given that our data samples a small portion of the total population, approximately 115,000 out of 336 million Americans(5), and surveys characteristics, we can assume that one person's responses do not affect another and our data collected is independent. Based on these assumptions, we moved forward with this model in our analysis.

RESULTS

$$\log\left(\frac{p}{1-p}\right) = 0.76524 - 3.71494 \cdot \text{RACETH2} - 2.38423 \cdot \text{RACETH3} + 0.33245 \cdot \text{GENDER2} + 0.24602 \cdot \text{DGRDG2} + 0.3$$

If we interpret these coefficients on the odds scale, we find that:

- `Intercept`: The odds of being a naturalized citizen are $e^{0.76524} = 2.14951$ when you are an Asian female with a bachelor's degree and you are employed.
- `RACETH2`: the odds of being a naturalized citizen get $e^{-3.7194} = 0.02425$ times lower when you are white.
- `RACETH3`: the odds of being a naturalized citizen get $e^{-2.38423} = 0.09218$ times lower when you are an underrepresented minority.

- `GENDER2`: the odds of being a naturalized citizen increase by $e^{0.33245} = 1.39438$ times when you are a man.
- `DGRDG2`: the odds of being a naturalized citizen increase by $e^{0.24602} = 1.27893$ when you have a Master's degree.
- `DGRDG3`: the odds of being a naturalized citizen increase by $e^{0.30555} = 1.35737$ when you have a Doctorate degree.
- `DGRDG4`: The odds of being a naturalized citizen decrease by $e^{0.10952} = 1.11574$ when your education level is "Professional". However, this is one of the least significant effects.
- `LFSTAT2`: - The odds of being a naturalized citizen increase by $e^{0.48005} = 1.61616$ when you are unemployed.
- `LFSTAT3`: The odds of being a naturalized citizen decrease by $e^{0.08273} = 1.08625$ when you are not in the labor force. This effect is also not significant.

Having generated this model, we can see that several of these coefficients are powerful predictors. After interpreting these coefficients on the odds scale by exponentiating our log odds results, we found the two predictors that change the odds the most. If the individual is unemployed, which increases the odds of being a naturalized citizen by 1.62 compared to someone who is employed, was a highly significant predictor. Then, the odds of being a naturalized citizen increase by 1.4 times for men compared to women. Taken together, this indicates that the most significant changes in someone's odds of being a naturalized citizen comes from whether someone is unemployed, followed by whether they're male.

We then look at all the significant predictors. Specifically, the race of the individual, whether they are male, whether they are unemployed, and if they have a doctorate or a professional degree significantly change the odds of them being a naturalized citizen. Race/ethnicity was a highly significant predictor, and explains that the odds of someone being a naturalized citizen are lower for white individuals or those from underrepresented minorities compared to individuals who are Asian. Our degree level variable was also significant, with a Masters degree and then a Doctorate degree increasing the odds that someone is a naturalized citizen at respectively higher levels. Of note, the odds of being a naturalized citizen then decrease by 1.12 when someone's education level is "Professional". However, this is one of the least significant effects, at a p-value of 0.186 above the standard significance cutoff, and should be interpreted with caution. Also of note, while being unemployed increases the odds that someone is a naturalized citizen, not being in the labor force decreases the odds by 1.08 compared to an employed individual. This effect was not significant for those not in the labor force, with a p-value of 0.183, and should be interpreted with caution.

Based on this data and this model, we can conclude that an Asian male, who has a Doctorate degree but is unemployed, has the highest odds of being a naturalized citizen. A female in the underrepresented minority group who has a Professional degree and is not in the labor force has the lowest odds of being a naturalized citizen, but this conclusion should be interpreted with caution given that having a Professional and not being in the labor force effects are not significant.

DISCUSSION

Initially, we predicted that all our predictor variables would have an impact. However, our variable selection method showed that the number of children did not have a significant effect in terms of predicting the odds of US citizenship. We also predicted that higher education levels would have a significant effect. Our results confirmed this, as the odds of being a naturalized citizen increased with a Masters and increased more with a Doctorate degree compared to a Bachelor's degree. These results seem likely given the prioritization of higher education, and how this could contribute to the citizenship process. Our results also confirmed that being male increased the odds as well. We predicted that being an underrepresented minority would increase the likelihood of naturalization. However, our results showed that this actually significantly lowers the odds of being a naturalized citizen compared to being Asian, and being white also significantly lowers the odds.

While these results were not exactly in line with our predictions, given that naturalized citizens are often immigrants and there is bias within the naturalization process, this is not entirely shocking. Finally, we predicted being employed would increase the odds of being a naturalized citizen, but we found that unemployment increased the odds compared to being employed, and not being in the labor force decreased the odds compared to being employed. These results were the most surprising to us, as we expected employment to increase odds. We hypothesize that this could have some interaction with education, as the citizenship process could be looking to those with higher education to be seeking jobs and meet the criteria to fill them. Additionally, we believe that our predictors with non-significant effects could be due to less observations containing individuals who have a Professional degree or are not in the labor force, but this could also be due to how the survey defined these variables. This could also be explained by the naturalization process prioritizing individuals who are working or plan to work, or who hold higher education degrees compared to technical degrees.

LIMITATIONS

When choosing our model, the variable selection process necessitated removing observations with missing data, which decreased our number of datapoints due to how many observations were missing the total children variable. Due to the at-will responses to the surveys used to collect data used in this process, the sample may not necessarily be representative of the true educational and demographic characteristics of immigrants in the US, as certain factors may predispose people to answering the surveys. However, we think this analysis using this data likely still provides insight into potential impacts of relevant variables. Additionally, with the way race/ethnicity was grouped, the generalizations within these variables may mean details about these impacts are not fully captured within this data.

All the variables in our dataset are categorical, with some redundancy, so some information may be vague. Our total children variable can only have 2 of 4 values used in the entire dataset because of redundancies in the question design, as it is not possible to have Option

02: 1-3 children and also Option 03: 2 or more children. Some of the categories, such as for DRDGR4 being "Professional", are unclear in their meaning. We assume Professional means someone who is not necessarily holding a higher education degree, but has a technical degree for a specific profession, but this was unclear in the survey design and answers. Finally, our analysis does not prove causality between these variables, but rather, identifies potential correlations.

FUTURE DIRECTION

Future research should consider further analysis into the employment aspect of the citizenship process, as this result was the most surprising. Additionally, expanding the race/ethnicity variable to include more groups could provide more information on how this variable truly impacts naturalization. Continued analysis of the factors that influence the naturalization process will add to the body of research, and can provide insight into the equitability of naturalization for different individuals.

References

1. "World Population Prospects - Population Division." United Nations, United Nations, 2022, population.un.org/wpp/.
2. "U.S. Naturalization Policy." CRS Reports, Congressional Research Service, 15 Apr. 2024, crsreports.congress.gov/product/pdf/R/R43366.
3. Mossaad, Nadwa, et al. Determinants of Refugee Naturalization in the United States, Proceedings of the National Academy of Sciences of the United States of 4. America, 27 Aug. 2018, www.pnas.org/doi/abs/10.1073/pnas.1802711115.
4. Batalova, Jeanne. "Frequently Requested Statistics on Immigrants and Immigration in the United States." Migrationpolicy.Org, 4 Apr. 2024, www.migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states-2024#characteristics.
5. U.S. and World Population Clock. United States Census Bureau. (2024, May 9). https://www.census.gov/popclock/

IPUMS Citation

. Minnesota Population Center. IPUMS Higher Ed: Version 1.0 [ImmigrationData]. Minneapolis, MN: University of Minnesota, 2016. https://doi.org/10.18128/D100.V1.0

Data Appendix

Assumptions

Linearity Linearity is given because the predictors of our logistic model are all categorical.

Independence Given that our data samples a small portion of the population (about 115,000 out of 331 million Americans citizens), and that the data is surveying given characteristics (one person's response will not affect another's), we can assume that the data collected is independent.

Randomness This data comes from a database from various sources, but makes up a small portion of the total population. Further, the surveys collected data from individuals across the population (various colleges, etc).

Checking Multicollinearity In Our Final Model

	GVIF	Df	GVIF^(1/(2*Df))
RACETH	1.025394	2	1.006289
GENDER	1.045592	1	1.022542
DGRDG	1.021472	3	1.003547
LFSTAT	1.029840	2	1.007378

Summary Statistics for Our Final Model

Table: GLM Model Summary

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7652447	0.0410283	18.651650	0.0000000
RACETH2	-3.7149428	0.0384179	-96.698342	0.0000000
RACETH3	-2.3842291	0.0398215	-59.872926	0.0000000
GENDER2	0.3324470	0.0312551	10.636584	0.0000000
DGRDG2	0.2460189	0.0370379	6.642347	0.0000000
DGRDG3	0.3055525	0.0376099	8.124259	0.0000000
DGRDG4	-0.1095246	0.0827711	-1.323223	0.1857614
LFSTAT2	0.4800546	0.0871502	5.508356	0.0000000
LFSTAT3	-0.0827315	0.0621524	-1.331108	0.1831537