PREDICTORS OF LOW RATES OF HISPANIC WORKERS’ PARTICIPATION IN STEM OCCUPATIONS IN THE US

Mario A. Camargo Espana

Thesis Project

Submitted in partial fulfillment of the

Requirements for the degree of

Master of Science in Data Science

May 2023

Candice Bradley, First Reader

William Mickelson, Second Reader

**Abstract**

The number of Hispanics in engineering departments at companies and universities differs from the proportion of Hispanics in the US population. Hispanics represented only 8% of the STEM (Science, Technology, Engineering and Mathematics) workforce in 2020, despite being 17% of the overall workforce. The ethnic disparity in STEM occupations contributes to economic inequalities in the US, since STEM occupations offer salaries that are more than double the national average wage for non-STEM occupations. In fact, the national average wage for STEM occupations in 2019 was $86,980, while the national average wage for non-STEM occupations was $38,160.

A logistic regression model was used to better understand the effect of demographic variables on STEM workers’ ethnic representation. First, data was gathered from household surveys (2009-2019) from IPUMS USA, a database supported by the University of Minnesota. Then the logistic model was applied and validated, and the resulting coefficients of ten different variables were used to compare their importance for Hispanic and non-Hispanic workers in STEM occupations. The results showed that (1) years of education and gender are the most relevant factors for the model, (2) having the US as their place of birth and being a US citizen increases the chances of working in STEM for Hispanic but decreases the chances of working in STEM for non-Hispanics, (3) the STEM participation gap between Hispanic males and females is considerably lower than the gap between non-Hispanic males and females, and (4) Hispanic participation rates did not increase significantly during the studied period. In 2009, a change in the variable *Hispan* from 0 (non-Hispanic) to 1 (Hispanic), had a decrease of around 24% in odds of working on STEM. That number decreased to 23% in 2019.

Government agencies, NGOs and private organizations can leverage these findings to design programs oriented to attract Hispanic talent into STEM fields. Such initiatives should consider the diversity of the Hispanic population in the US to impact the most vulnerable Hispanic groups.

**Dedication**

The completion of this document, and my master’s degree studies were possible thanks to the love and support of my family, especially my wife and brother. I dedicate this thesis to the two of them, they are the best example of the amazing contributions that Hispanics can bring to STEM professions in the US.

Thank you for living this adventure with me.

**Acknowledgments**

I would like to thank my thesis advisor, Dr. Candice Bradley, for her support and guidance. I would like to thank my second reader, William Mickelson, for his comments. I am also grateful to Northwestern University School of Professional Studies.

**Contents**

**Chapter 1: Introduction………………………………………………………………....1**

Statement of the Problem………………………………………………………….2

Justification……………………………………………………………………......2

**Chapter 2: Review of the Literature……………………………………………………3**

**Chapter 3: Methods……………………………………………………………………...7**

**Chapter 4: Results……………………………………………………………………....20**

**Chapter 5: Conclusions………………………………………………………………...28**

**References……………………………………………………………………………….30**

**Appendices……………………………………………………………………………....33**

**Chapter 1: Introduction**

As a Hispanic who came to the US in 2016 to work in a STEM (Science, Technology, Engineering, and Mathematics) occupation, I have noticed with concern that the number of Hispanics in different engineering departments has always been small compared to the proportion of Hispanics from the total US population. I have found the same disparity in the group of students in the Master of Science in Data Science program that I am now completing. In my experience, most Hispanic workers have low-paying jobs (for example, servers, cashiers, and construction workers). My personal experience is reflected in the national statistics. The ethnic composition of STEM workers in the US differs considerably from the ethnic composition of the general US employed population. Hispanics were only 8% of the STEM workforce in 2020, despite being 17% of the overall workforce (Student Research Foundation 2020).

The national average wage for STEM occupations in 2019 was $86,980, more than double the national average wage for non-STEM occupations ($38,160; U.S. Bureau of Labor Statistics 2020). Therefore, the low participation of Hispanics in STEM negatively impacts the buying power of that population. That income disparity will most likely continue to grow. Since we live in a knowledge-based economy with high demand for STEM workers, it is expected that STEM jobs will increase by 8% by 2029 over the 6.5% of all US occupations (nearly 10 million) in 2019. That is more than double the 3.4% expected net growth for non-STEM occupations for the same period (US Bureau of Labor Statistics 2020).

I used a logistic regression model to better understand the effect of ten variables on the chances of having a STEM career in the US. I ran the model twice, once for the Hispanic population in the US and once for the non-Hispanic population in the US. The data for the study come from the IPUMS USA database, which includes personal and household records from US population surveys from 2009 to 2019 (Ruggles et al. 2021). Then I assessed different options for the model using AIC and ANOVA tables (Harrell 2015). Finally, I used the coefficients to understand the differences between the variables for Hispanic and non-Hispanic workers in STEM occupations. The results help explain differences between the two populations and offer insights into the most relevant variables.

Statement of the Problem

Hispanic workers in the US are underrepresented in STEM occupations, and the specific reasons are not completely clear. However, factors such as gender, age, degree area, and socioeconomic status (presented in chapter 2) have been suggested to impact participation of Hispanics in STEM. Predictive models can be used to find patterns and analyze how distinct factors intersect to impact the probability of occurrence for an event (Harrell 2015).

The coefficient values and odd ratios of the variables studied in this thesis can expand our understanding of inequality in STEM jobs in the US. This type of analysis is especially relevant for Hispanics in the US because they are a heterogeneous population composed of people from different countries and having diverse cultures. Hispanics are often multiracial, have lived in the country for different lengths of time, belong to different generations, and live in all regions in the US [(Aspray 2016)](https://paperpile.com/c/chKrl6/eHyP).

Justification

Expanding our understanding of the effect of various sociodemographic factors on the Hispanic participation in STEM jobs can help to create solutions and initiatives that address this issue and contribute to a more equal society. Personally, I am interested in eventually creating a nonprofit organization that helps Hispanics integrate into the STEM workforce. It is therefore important for me to understand who requires the most help, and on which problems to focus.

**Chapter 2: Review of the Literature**

Compared to whites and Asians, Hispanics have lower employment levels in STEM occupations. Hispanics with science and engineering degrees also have lower levels of employment in STEM. Additionally, Hispanic STEM workers have lower median incomes than white and Asian STEM workers [(Landivar 2013)](https://paperpile.com/c/chKrl6/jYEo). The underlying causes of those disparities have not been completely identified. However, previous research has associated certain factors with different levels of participation in the STEM workforce.

In the first place, **Gender** remains one of the most important factors contributing to inequality in the STEM workforce. Women make up almost half of the current workforce but hold only 25% of STEM jobs. Although Hispanic women represented about 7% of the total workforce in 2020, their participation in STEM occupations was less than 2% (Student Research Foundation 2020). Data from the National Science Foundation showed that Hispanic males are less likely than Hispanic females to earn a STEM bachelor's degree. In 2016, Hispanic men represented 44% of total Hispanic sciences and engineering graduates (National Science Foundation 2017). And the numbers are similar for higher STEM degrees: Hispanic males received 46% of STEM doctoral degrees (National Science Foundation 2017). However, all women (Hispanic and non-Hispanic) still face several challenges to obtaining a job or a PhD in STEM. Some of the challenges are: (1) girls do not study for STEM degrees, (2) girls have negative experiences with STEM areas at school, (3) there are not many role models in STEM for girls, (4) STEM courses do not have content focused on girls, and (5) those courses are biased in favor of males [(Blickenstaff 2005)](https://paperpile.com/c/chKrl6/FTcE).

The data show some positive trends in favor of Hispanic females, such as increasing enrollment in higher education (6.8%) compared to Hispanic males (4%; National Science Foundation 2017). In fact, Hispanic women earn a majority of all undergraduate and advanced degrees within the Hispanic population. They graduate at a higher rate than Hispanic men at all levels of education, from high school through post-doctoral programs, they also have an important participation in health professions, from MDs to laboratory technicians. However, Hispanic women are poorly represented in STEM fields, especially in engineering and computer science, and in the reported intention to major in STEM fields. On a positive note, around 37% of Hispanic females declared a STEM discipline as their career choice in 2012, an increase compared to 32% in 2003 (National Science Foundation 2016).

**Age** is another factor explaining Hispanic underrepresentation in STEM. Among high school students in STEM classrooms, Hispanics are about as likely as ORGs (other representative groups) to aspire to STEM careers (47% vs. 50%). However, Hispanic adults are underrepresented in STEM given that they currently comprise 17% of the workforce overall but merely 8% of the STEM workforce (Student Research Foundation 2020; Mellander 2021).

Note that age is also relevant for US students who graduate with STEM majors: younger students are more likely to graduate than older students most likely due to financial challenges or academic skills (Mau 2016). Younger workers have higher levels of employment in STEM occupations compared to older workers who are science and engineering graduates. About 27% of workers from ages 25 to 44 with a science or engineering degree are employed in a STEM occupation. The percentage slightly decreases to about 26% for science and engineers graduates aged 45 to 54, and it goes down to 19% for science and engineers graduates aged 55 to 64 [(Landivar 2013)](https://paperpile.com/c/chKrl6/jYEo). Therefore, the Hispanic community has an opportunity to increase the number of STEM degrees and jobs, give that Hispanics are the youngest racial/ethnic group in the US (Patten 2016).

**Education** is another important factor explaining the low Hispanic representation in STEM, as they received only 5% of the total doctorates in science and engineering with 21,000 degrees. That is a small number when considering that Hispanics represent 17% of the US population. For college education in engineering, both male and female Hispanic students persist at a similar rate to white students. However, unlike business and related fields, a lower percentage of Hispanics are recruited into engineering majors and careers. Therefore, the problem for Hispanic students in STEM disciplines is one of recruitment rather than retention. This situation persists despite engineering being one of the most intended majors among Hispanic high school students [(Aspray 2016)](https://paperpile.com/c/chKrl6/eHyP).

Data from a 2009 high school longitudinal study were used to analyze the role of factors such as utility, interest, and attainment in high school to drive success in future STEM careers [(Gottlieb 2018)](https://paperpile.com/c/chKrl6/0TP4). The results suggest that all those factors are correlated to the STEM career plans of white students, but have a low relationship to STEM career plans of black and Hispanic students. Additionally, higher levels of science attainment significantly increased the odds for planning for a STEM career only for white female students. Finally, higher levels of math achievement increased the odds of planning for a STEM career only for white female students, and white, Hispanic, and black male students [(Gottlieb 2018)](https://paperpile.com/c/chKrl6/0TP4).

**Socioeconomic status.** In the same 2009 high school longitudinal study (Gottlieb 2018) higher levels of SES (socio-economic status) were associated with higher odds of planning for a STEM career for all ORGs, but not for Hispanic male students. The relationship was particularly powerful for Hispanic female students, black female students, and black male students [(Gottlieb 2018)](https://paperpile.com/c/chKrl6/0TP4). Receiving financial aid was associated with higher graduation rates in STEM degrees for Hispanic students in community college. This situation could occur because on many students have to pay for their own education as well as contribute to family finances. Financial considerations also impact the time that Hispanic students take to complete their community college degrees and start their professional careers [(Aspray 2016)](https://paperpile.com/c/chKrl6/eHyP).

**STEM confidence**, defined as the students’ judgments regarding their competence in math and science, was ranked in a study from 1 to 5 (5 being the higher STEM confidence) for different race/gender groups (Lizler 2014). The results showed that all groups have either lower STEM confidence or confidence not significantly different from white men. The groups with lower STEM confidence include all female groups, except for Hawaiian/Pacific Islander and Native American women, and for Asian and Hawaiian/Pacific Islander men (Lizler 2014).

**English proficiency**. The final factor impacting Hispanic participation in STEM careers is language. A large portion of Hispanic students in the US enter public school learning English as a second language. This could have a negative impact in the long term, because there is a strong relationship between low level of English and receiving low grades in high school. Students with low grades take remedial courses instead of college preparatory courses [(Aspray 2016)](https://paperpile.com/c/chKrl6/eHyP). Some programs have been targeted at Hispanic students to encourage them to prepare for college. A specific program with a STEM orientation is College Board’s Equality 2000, which focuses on math instruction to prepare students for college STEM majors and STEM careers. The program is more successful in low socioeconomic level schools [(Aspray 2016)](https://paperpile.com/c/chKrl6/eHyP)**.**

**Chapter 3: Methods**

*Data*

Data were gathered from the IPUMS USA database (originally, the "Integrated Public Use Microdata Series"; Ruggles et al. 2021). IPUMS USA is part of IPUMS International, the world's largest collection of population microdata available for research. It is composed of data collected on annual household surveys from the National Statistical Offices of dozens of participating countries. For the US, data was collected from more than sixty high-precision samples of the population from the American Community Surveys and from the Puerto Rican Community Surveys. The representative samples used in this study are collected with a focus on quality geographic coverage to produce a good picture of the country’s people and homes (United States Census Bureau 2017). As each of these samples contains 0.4% of the population, this research does not contain the same amount of data as the census data. However, the data represent people in the fifty states, Puerto Rico, and the District of Columbia. It is important to mention that the IPUMS data could include bias because asking about citizenship status significantly increases the percentage of questions skipped, with particularly strong effects among Hispanics (Baum et al. 2019).

While the IPUMS dataset contains household surveys from the years 1790 to the present, this study only uses the samples from the years 2009, 2014 and 2019. Data before 2009 was not considered because this research is focused on the analysis of current factors that impact Hispanic participation in the STEM workforce, and data for the years 2020 and 2021 were excluded because of the impact that the COVID-19 pandemic (Ruggles et al. 2021). The demographic, socioeconomic, and geographic variables chosen for this study are based on the review of the literature from the previous chapter. They are Year, Hispanic origin, English proficiency, Citizenship status, Place of birth, Year of education, Gender, Age, Race, and Type of occupation.

With data for the selected variables and the defined time, this study uses a binary logistic regression model to identify what factors help to predict if an individual is a STEM worker or not. A regression approach was preferred because of the type of data available in the IPUMS dataset. It contains a mix of categorical and continuous input variables and a binary target variable. However, data manipulation was needed to get the data ready for analysis. It is important to note that there is no information about the number of times participants answered the survey year after year, and the database records are not identifiable from year to year, which precludes using a longitudinal study in which participants are followed over the years.

The type of occupation was defined as the target variable for the regression model, defined as STEM vs NON-STEM occupation, which is not explicitly included in the IPUMS dataset. To transform the target variable, different occupations listed in the dataset were classified as STEM or NON-STEM. There are twelve different occupational classification systems consisting of between 530 occupation codes. The complete list of occupation codes can be found in the IPUMS website (Ruggles et al. 2021). The categories of occupations classified as STEM professions are Communication Technologies, Computer and Information Sciences, Engineering, Engineering Technologies, Mathematics and Statistics, Military Technologies, Physical Sciences, Nuclear, Industrial Radiology, Biological Technologies, Electrical and Mechanic Repairs and Technologies, Precision Production and Industrial Arts, and Transportation Sciences and Technologies. Additionally, the data was filtered to include only people between 21 and 62 years old, as that range of ages includes most of the active workers.

The samples of the US population were analyzed to determine the factors that influence whether a person works in STEM or not, and to understand the importance of the independent variables on the target variable, working in a STEM field. To do that, it was necessary to performed data exploration for all important factors found in previous research studies mentioned on the review of literature of this document. After that, the binary logistic regression model was implemented for three years, 2009, 2014 and 2019.

*Data exploration*

Variable transformations were needed before descriptive statistics calculation and data visualization. The format of the variable “race, racesing” was converted from numeric to categorical, as race is a qualitative variable. Hispanics from all origins were grouped, as a result, the new form of the binary variable assigns 1 to Hispanics and 0 to non-Hispanics. One last variable (educ2) was added to the database in order to compare different options for the binary logistic regression model. This new variable groups the years of education into three categories, “0” for people who didn't study after high school, “1” for people who didn't study after college and “2” for people who studied after college.

Other relevant facts about the meaning of certain values include the definitions for values in variable race -*racesing-*, code 1 means white, 2 black, 3 American Indian, 4 Chinese, 5 Japanese, 6 other Asian, 7 other races, 8 two major races and 9 three major races. Sex is coded as 1 for male and 2 for female.

Descriptive statistics for all nine variables including STEM were calculated for each year, those statistics include the size of the sample, total values missing information, all available distinct values for each variable, the total sum, and mean.

TABLE 1. Descriptive statistics for discrete variables in 2009.

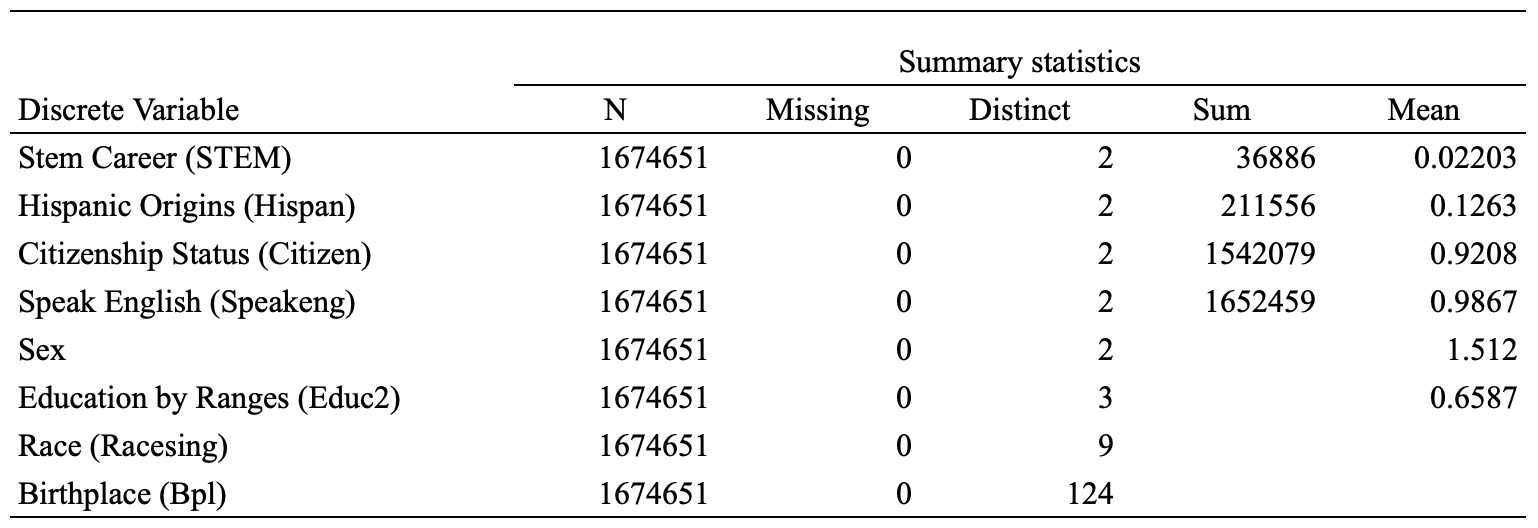
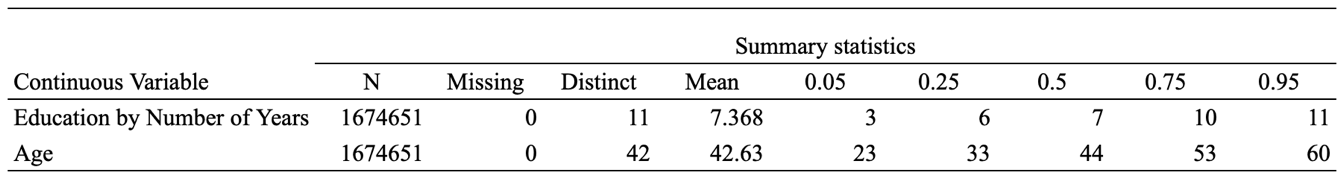


TABLE 2. Descriptive statistics for continuous variables in 2009.



There is no missing data for any of the variables in the samples in 2009, the total samples were 1’674.651, the average age was 42.63 (all other means describe the average of variable codes rather than real values averages) years and distinct values match the definition that IPUMS offers for their dataset (Table 2).

TABLE 3. Descriptive statistics for discrete variables in 2014.

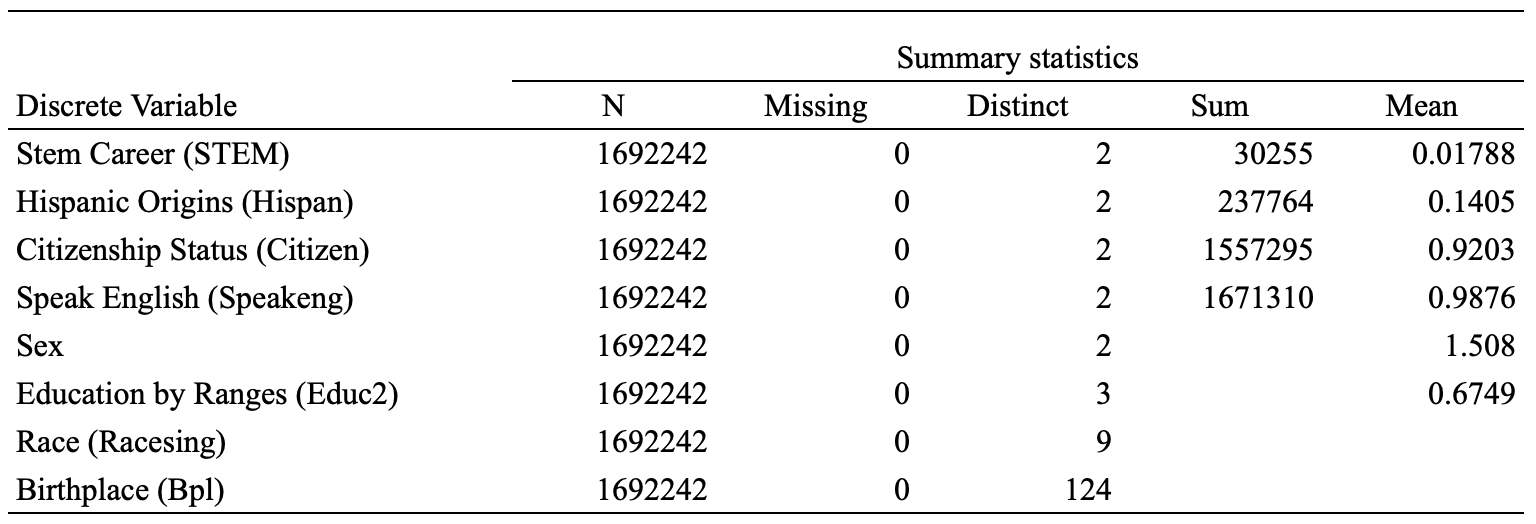
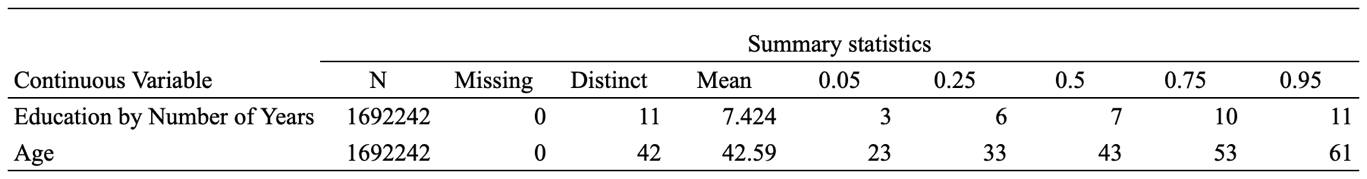


TABLE 4. Descriptive statistics for continuous variables in 2014.



There is no missing data for any of the variables in the samples in 2014, the total samples were 1’692.242, the average age was 42.59 (slightly lower than 2009), the Hispanic population increased compared to 2009 as well as the average years of education (Table 3).

TABLE 5. Descriptive statistics for discrete variables in 2019.

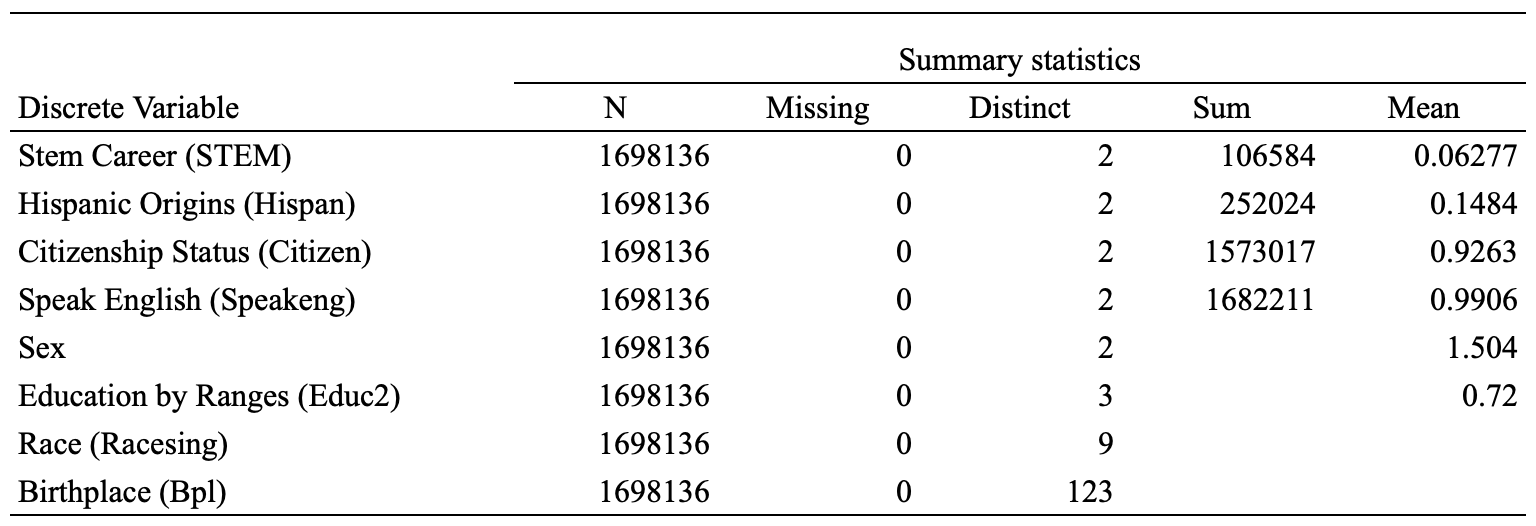
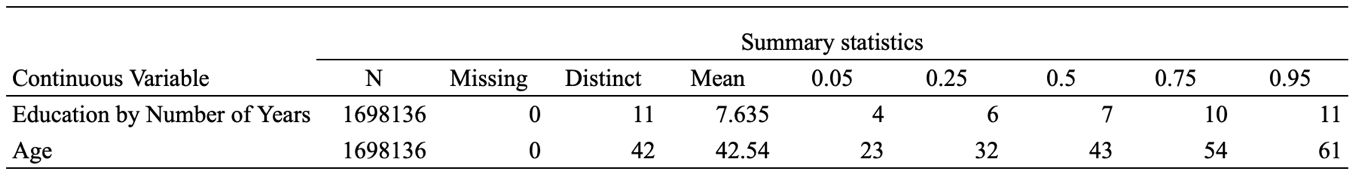


TABLE 6. Descriptive statistics for continuous variables in 2019.



There is no missing data for any of the variables in the samples in 2019, the total samples were 1’698.136, the average age was 42.54 (slightly lower than 2014), the Hispanic population also increased compared to 2009 and 2014 as well as the average years of education. The percentage of workers on STEM careers had a significant increase compared to the previous 2 samples (Table 5).

As part of the data exploration process, data visualizations were created to assess the association between the variables Hispanic (*hispan)*, citizenship (*citizen)*, education (*educ)* and English proficiency (*speakeng)* with the binary options for the target variable STEM. Age was used in the horizontal axis to better understand how the variables perform over the working life of the population in a particular year. The *hispan* variable presents Hispanics = 1 and non-Hispanics = 0. The *citizen* variable shows 1 for citizens and 0 for non-citizens. The variable *speakeng* shows 1 for English speakers and 0 for non-English speakers.

A picture containing diagram, line, plot, text

Description automatically generated

Figure 1. Percentage of population in STEM careers by age for Hispanics and non-Hispanics in 2009

A picture containing text, diagram, line, plot

Description automatically generated

Figure 2. Percentage of population in STEM careers by age for Hispanics and non-Hispanics in 2014

A picture containing text, diagram, plot, line

Description automatically generated

Figure 3. Percentage of population in STEM careers by age for Hispanics and non-Hispanics in 2019

The graphs that visualize Hispanic and non-Hispanic participation in STEM in 2009, 2014 and 2019 shows that both groups of the population have been increasing the participation in STEM, the population increase the participation in STEM in their thirties and start decreasing in their forties.

A picture containing text, diagram, line, plot

Description automatically generated

Figure 4. Percentage of population in STEM careers by age for citizens and non-citizens in 2009

A picture containing text, line, diagram, plot

Description automatically generated

Figure 5. Percentage of population in STEM careers by age for citizens and non-citizens in 2014

A picture containing diagram, text, line, plot

Description automatically generated

Figure 6. Percentage of population in STEM careers by age for citizens and non-citizens in 2019

The graphs that visualize Citizens and non-citizens participation in STEM in 2009, 2014 and 2019 shows that non-citizens have a higher participation at the beginning of their professional life, however, as they get older, the participation is lower than for citizens.

A picture containing text, line, diagram, plot

Description automatically generated

Figure 7. Percentage of population in STEM careers by age for English speakers and non-English speakers in 2009

A picture containing text, diagram, line, plot

Description automatically generated

Figure 8. Percentage of population in STEM careers by age for English speakers and non-English speakers in 2014

A picture containing text, diagram, plot, line

Description automatically generated

Figure 9. Percentage of population in STEM careers by age for English speakers and non-English speakers in 2019

The graphs that visualize English speakers and non-English speakers’ participation in STEM in 2009, 2014 and 2019 shows that non-English speakers have a very low participation rate when compared to English speakers.

A picture containing text, line, plot, diagram

Description automatically generated

Figure 10. Percentage of population in STEM careers by age for population with high school education, college education and graduate education in 2009

A picture containing text, line, plot, diagram

Description automatically generated

Figure 11. Percentage of population in STEM careers by age for population with high school education, college education and graduate education in 2014

A picture containing text, line, plot, diagram

Description automatically generated

Figure 12. Percentage of population in STEM careers by age for population with high school education, college education and graduate education in 2019

The graphs that visualize participation in STEM for different education levels in 2009, 2014 and 2019 shows that higher level of education have higher level of participation for all ages.

*Model formulation and fitting*

The logistic regression model can be expressed as: , where B0 is a constant and B1 is the coefficient attached to predictor variable X1. Four alternative models were tested with different variables and interactions. Model 1 was fitted using the independent variables X: X1 = Hispanic origins, X2 = Citizenship, X3 = Race, X4 = Speak English, X5 = Birthplace (USA), X6 = Sex, X7 = Education (in years), X8 = Age (in years). Model 2 included the variable Education (in ranges) instead of the original variable Education (in years). Model 3 expands the relationship between Citizenship and Hispanic origins (as it was found on the exploratory phase that citizenship had an important effect in the participation of Hispanics in STEM jobs) and Model 4 expands the relationship between Birthplace and Hispanic origins (as it was also found on the exploratory phase that birthplace had an important effect in the participation of Hispanics in STEM jobs).

Based on the purpose of this investigation, the full model that includes all these variables, was replicated at different points in time to evaluate changes on the population.

**Chapter 4: Results**

*Data exploration results*

The data exploration work done in the methods section sown that non-Hispanics, non-citizens under 40, English speakers, people who were born outside of the US, males, highly educated people. and people between 34 and 45 years had higher participation in STEM jobs in 2009, 2014, and 2019. However, 3 main changes were identified in 2019 when compared to 2009 and 2014. (1) Participation in STEM jobs significantly increased in 2019 for all the ages and variables evaluated, (2) non-citizens increased their participation in STEM jobs when compared to citizens. In 2019, non-citizens workers on their thirties doubled STEM participation compared to citizens, and in 2019 the 45-year-old non-citizens still had higher participation in STEM careers as citizens.

A final observation that can be said from the data exploration work is that Hispanics have a lower participation rate in STEM careers in 2009, 2014 and 2019 even when accounted for other variables. (1) Hispanic citizens had a lower STEM participation rate than non-Hispanic citizens, (2) Hispanics who speak English had a lower participation rate than non-Hispanics who speak English, (3) Hispanic males had a lower participation in STEM careers than non-Hispanic males, and (4) highly educated Hispanics also had a lower participation rate in STEM careers when compared to non-Hispanics in 2009, 2014 and 2019 (Table 7).

A picture containing text, screenshot, number, font

Description automatically generated

*Model results*

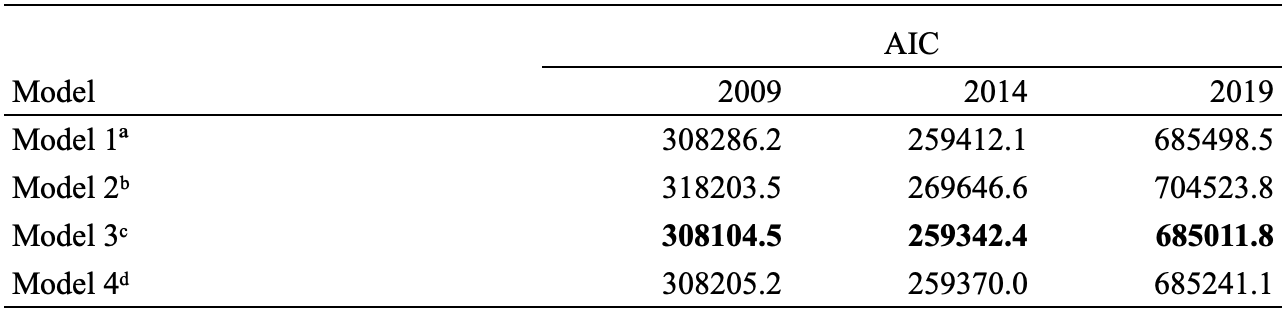
To determine the variables to include in the regression model, and to identify if there are any variables that should be deleted from the model, fast backward step-down was performed with total residual AIC as the stopping rule, using the function *fastbw(f4)* in R, the result validates that all variables are relevant for the model. The conclusion of the analysis for the fitting of the binary logistic regression model for the data in 2009 shows a good performance.

Citizenship continued having different effects for Hispanics and non-Hispanics in 2014.Being Hispanic also decreased STEM participation regardless of the level of English or years of education (figure 13). The resulting AIC for the models were f1=524144.6, f2=634739.9, f3=524143.8 and f4=524141.3. Therefore, model f4 also presents the best performance for 2014.

The resulting AIC for the models can be found below, Model 3 achieved the best performance, as it has the lowest AIC for all years, 2009 (308104.5), 2014 (259342.4) and 2019 (685011.8) (Table 8).

The variables with positive coefficients are *speakeng*, *educ*, and *Asian races*. This suggests that as those independent variables increase, the dependent variable STEM also tends to increase, in other words, people in the sample with Asian origins that speak English well and are highly educated, have higher chances to work on a STEM job. Variables with negative coefficients are being born in the US, being Hispanic, being a citizen, being black and being American Indian. This suggests that the dependent variable STEM tends to decrease when those variables increase. The P values for all independent variables are greater than the usual significance level of 0.05 (statistically significant), meaning that changes in those variables are associated with changes in the dependent variable *STEM* at the population level (the complete sample of the US population). Therefore, all variables can be added to the logistic regression model.

TABLE 8. AIC for the evaluated models for 2009, 2014 and 2019



ªModel 1 includes the variables Hispanic, Citizenship, Race, Birthplace, Sex, Education (in years) and Age.

ᵇModel 2 includes the variables Hispanic, Citizenship, Race, Birthplace, Speak English, Sex, Education (in ranges) and Age.

ᶜModel 3 includes the variables Race, Birthplace, Speak English, Sex, Education (in years), Age and Hispanic and Citizenship combined.

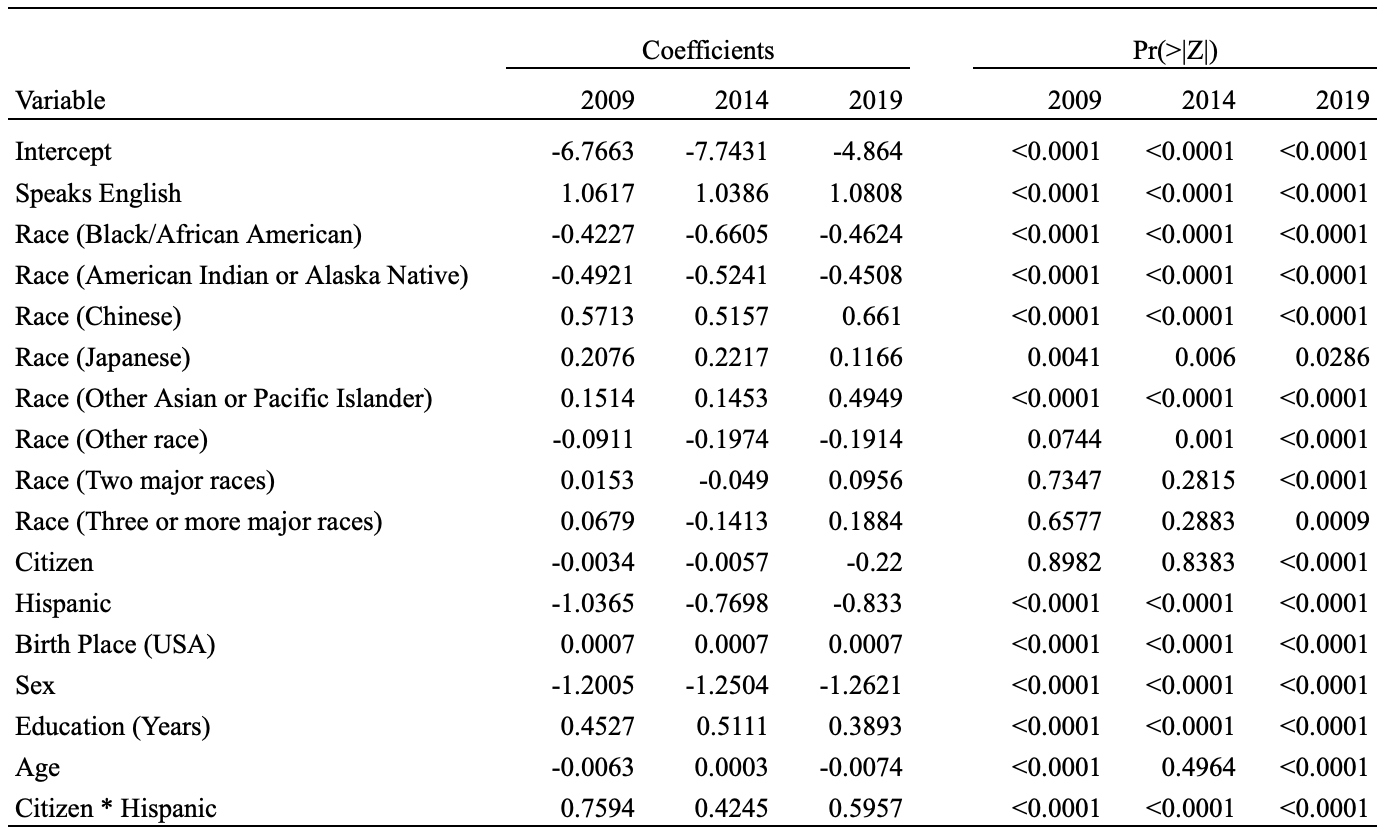
ᵈModel 4 includes the variables Race, Citizenship, Speak English, Sex, Education (in years), Age and Hispanic and Birthplace combined.

*Model coefficients*

According to the general results for the coefficients (Hispanics and non-Hispanics) in all years evaluated (2009, 2014 and 2019), the highest values for the predictors were obtained by the variable related to the total number of years of education and the binary variable sex. The results for education were expected, and it helps to validate the model outcomes. Apart from education, other factors that were important for predicting Hispanic participation in the STEM workforce were birthplace and race. These results were not surprising, but they are interesting because they are not related to STEM skills but to other social factors that cannot be changed or developed.

Hispanic origins and being a US citizen have negative coefficients and odds ratios under 1. Hispanics present lower levels of participation in STEM even for the same levels of education, English fluency, birthplace, and citizenship status (Table 9).

TABLE 9. Coefficients for the Logistic Regression Model applied to data for 2009, 2014 and 2019



*Odds ratio*

Odds ratio can be used to interpret the coefficients of the logistic regression model results, the odds are a transformation from probability, however, probability ranges from 0 to 1 while odds range from 0 to positive infinity (Harrell 2015). Odds ratios were calculated for the results of the logistic regression model with the lowest AIC (Model 3), the values used for each of the variables were: (1) Speak English = True, (2) U.S. Citizenship = True, (3) Having the U.S. as birthplace = True, (4) Having Hispanic origins = True, (5) Sex = Male, and (6) Race = White.

In 2009, 2014 and 2019 the most important positive predictors based on were years of education, speaking English, being a U.S. citizen, and having the U.S. as birthplace. As it is shown below, the four variables mentioned, present odds ratios higher than 1, the ranking of importance in terms of odds ratio doesn´t change for the years 2009, 2014 and 2019 (Figure 13).

Chart, bar chart, histogram

Description automatically generated

Figure 13. Top variables by odds ratio positive effect for the years 2009, 2014 and 2019. The Education variable presents a positive effect when the number of years increase. Speaking English, being an American citizen and having USA as birthplace also present a positive effect on the target variable (having a job in STEM).

Odds ratio for the coefficient of the binary variable Hispanic origins in 2009 suggests that according to the model, when the variable changes from 0 (non-Hispanic) to 1 (Hispanic), there is a 24% decrease in probability of working on STEM in 2009, 29% in 2014 and 21% in 2019. Odds ratios are calculated as *exp(model coefficient) - 1*. Other coefficients that decreased the odds of working in STEM are: (1) Changes from male to female decreased odds to work in STEM in 2009 by 70%, 71% in 2014 and 73% in 2019. (2) Increasing the value of variable age from 33 to 53 decreases the odds of working on STEM jobs in 12% in 2009, 7% in 2014 and 15% in 2019. The summary of those odds can be found in figure 14.

Chart, bar chart

Description automatically generated

Figure 14. Top variables by odds ratio negative effect for the years 2009, 2014 and 2019. Being a woman presents the most negative effect over the analyzed variables. Being Hispanic is the third most negative effect (after race, that is visualized in Figure 3). Finally, being older also has a negative effect.

Odds ratio for the coefficient of the variable race present different outcomes, on one hand, certain races present odds ratios higher than 1, these means that those races presented higher participation in STEM jobs in 2009, 2014 and 2019 when compared to White individuals. Those races are Chinese, Japanese, and other Asians or Pacific islanders, the odds ratios can be found in figure 15.

Chart, bar chart

Description automatically generated

Figure 15. Top values for the variable "Race" by odds ratio positive effect for the years 2009, 2014 and 2019. Chinese presents the highest effect over having an STEM job compared to the White race, followed by Japanese and Other Asian or Pacific Islander (This last race presented an impressive increase by 2019).

On the other hand, some races present odds ratios lower than 1 when compared to White individuals, that means that there is a decrease in odds of working on STEM jobs for Black/African Americas of 34% in 2009, 48% in 2014 and 37% compared to Whites. For American Indians or Alaska Natives there was a decrease in odds of 39% in 2009, 41% in 2014 and 36% in 2019. Finally, for the “other races” category, there was a decrease in odds of having an STEM career of 9% in 2009, 18% in 2014 and 17% in 2019 when compared to White individuals. The summary of the mentioned odds ratios can be found in figure 16.

Chart, bar chart

Description automatically generated

Figure 16. Top values for the "Race" variable by odds ratio negative effect for the years 2009, 2014 and 2019. The races Black/African American and American Indian or Alaska Native presented the highest negative effect on having a job in STEM compared to the White race.

**Chapter 5: Conclusions**

There are four main conclusions that can be drawn from the exploratory analysis of the datasets for the years 2009, 2014 and 2019. The most important conclusion is that Hispanics have lower STEM participation rates for all years included in the model. Second, STEM participation rates are higher for Hispanics and non-Hispanics who speak English fluently and have more years of education. However, non-Hispanics have higher STEM participation in each of those groups when compared to Hispanics. Third, birthplace and citizenship are also relevant factors to predict STEM careers for both Hispanic and non-Hispanic populations. One key difference between the two groups is that for Hispanics, having the US as their place of birth is a positive predictive factor for a job in STEM, and having as birthplace another country is a negative predictive factor for a job in STEM. The opposite is true for non-Hispanics: being born in a country outside of the US is a positive predictive factor to have a job in STEM. Being a US citizen increases the chances of working in STEM for Hispanic individuals and decreases the chances of working in STEM for non-Hispanics, with completely different outcomes. Finally, although males have higher participation rates regarding of their ethnicity, the gap between Hispanic males and females is considerably lower than the gap between non-Hispanic males and females.

There are three main conclusions based on the results of the logistic regression model applied to the data from the IPUMS database: (1) Total years of education and sex are the most important factors in the ranking of importance of the model for all three years included, (2) Hispanic origins are not in the top factors of importance for the model, the variable *Hispan* ranks 6th out of 8th individual factors included in the ranking, and (3) Finally, Hispanic participation rates did not increase significantly during the studied period. In 2009, a change in the variable *Hispan* from 0 (non-Hispanic) to 1 (Hispanic), had a decrease of around 24% in odds of working on STEM. That number decreased to 23% in 2019. This shows that there are ongoing barriers for Hispanics pursuing STEM jobs in the US.

**References**

Aspray, William. 2016. *Women and Underrepresented Minorities in Computing: A Historical and Social Study*. Cham: Springer International Publishing.

Baum, Matthew, Bryce Dietrich, Rebecca Goldstein, and Maya Sen. 2019. “Estimating the Effect of Asking About Citizenship on the US Census: Results from a Randomized Controlled Trial.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3381224.

Blickenstaff, Jacob C. 2005. “Women and Science Careers: Leaky Pipeline or Gender Filter?.” *Gender and Education*, no. 17(4): 369-386.

Jessica J. 2018. “STEM Career Aspirations in Black, Hispanic, and White Ninth-Grade Students.” *Journal of Research in Science Teaching* 55, no. 10 (12): 1365-1392.

Hanson, Sandra L. 2013. “STEM Degrees and Occupations Among Latinos: An Examination of Race/Ethnicity and Gender Variation.” *Race, Gender & Class* 20 (1/2): 214-231.

Harrell, Frank E. 2015. “Case Study in Binary Logistic Regression, Model Selection and Approximation: Predicting Cause of Death.” *Regression Modeling Strategies. Springer Series in Statistics*. Student Research Foundation n.d. “Untitled.” Student Research Foundation. Accessed March 9, 2022. https://www.studentresearchfoundation.org/wp-content/uploads/2020/04/Hispanics\_STEM\_Report\_Final-1.pdf.

Landivar, Liana C. 2013. *Disparities in STEM Employment by Sex, Race, and Hispanic Origin*: *American Community Survey Reports.* Washington, DC: US Census Bureau.

Litzler, Elizabeth, Cate C. Samuelson, and Julie A. Lorah. 2014. “Breaking it Down: Engineering Student STEM Confidence at the Intersection of Race/Ethnicity and Gender.” *Res High Educ*, no. 55 (03), 810–832.

Mau, Wei-Cheng J. 2016. “Characteristics of US Students That Pursued a STEM Major and Factors That Predicted Their Persistence in Degree Completion.” *Universal Journal of Educational Research* 4 (6): 1495-500.

Mellander, Gustavo A. 2021. “Hispanics and STEM Opportunities.” *The Hispanic Outlook in Higher Education*, (07), 6-7. https://www.hispanicoutlook.com/articles/hispanics-and-stem-opportunities.

National Science Foundation. 2016. “Science and Engineering Indicators 2016.” Science and Engineering Indicators 2016. https://www.nsf.gov/statistics/2016/nsb20161/uploads/1/12/chapter-2.pdf.

National Science Foundation (NSF). 2017. “Women, minorities, and persons with disabilities in science and engineering: 2017.” Nacional Science Foundation. https://www.nsf.gov /statistics/2017/nsf17310/data.cfm.

Patten, Eileen. 2016. “The Nation's Latino Population Is Defined by Its Youth.” Pew Research Center. https://www.pewresearch.org/hispanic/2016/04/20/the-nations-latino-population-is-defined-by-its-youth/.

Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. IPUMS USA: Version 11.0 [dataset]. https://doi.org/10.18128/D010.V11.0.

Starkweather, Jon, and Amanda K. Moske. n.d. “Multinomial Logistic Regression.” Multinomial Logistic Regression. Accessed March 9, 2022. https://it.unt.edu/sites/default/files/mlr\_jds\_aug2011.pdf.

U.S. Bureau of Labor Statistics. 2021. “Employment in STEM occupations: US Bureau of Labor Statistics.” Bureau of Labor Statistics. https://www.bls.gov/emp/tables/stem-employment.htm.

Appendix A

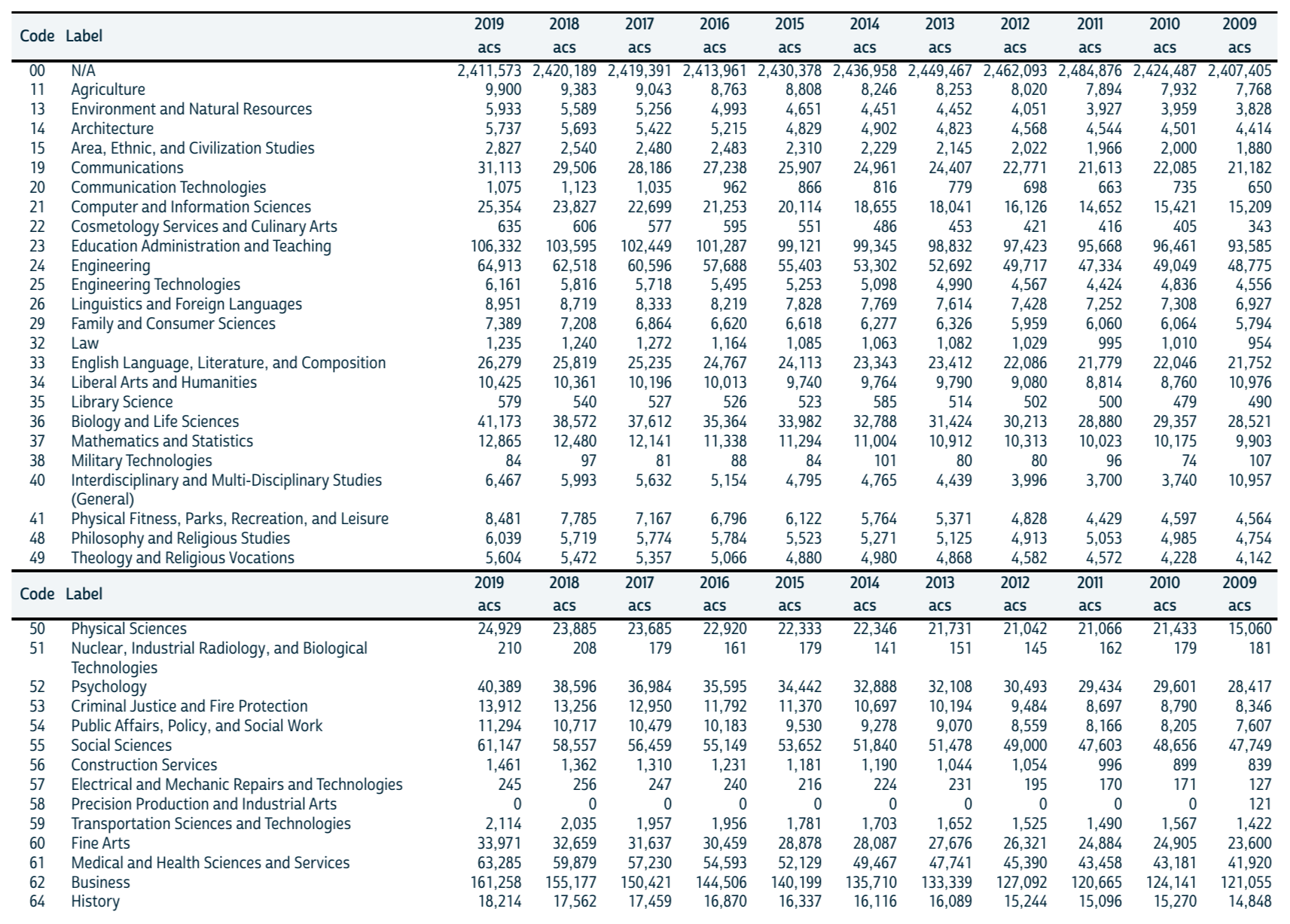
Data dictionary

Table 10. Variables included in the analysis as described by IPUMS USA

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Year | Categorical | Reports the four-digit year when the household was enumerated or included in the census. |
| State | Categorical | Identifies the state in which the housing unit was located, using the coding scheme developed by the Inter-University Consortium for Political and Social Research (ICPSR). The ICPSR scheme orders states first by geographic division and then alphabetically within each division. |
| Sex | Categorical | Reports whether the person was male or female. |
| Age | Discrete | Reports the person's age in years as of the last birthday. |
| Hispanic origin | Categorical | Identifies persons of Hispanic/Spanish/Latino origin and classifies them according to their country of origin when possible. Origin is defined by the Census Bureau as ancestry, lineage, heritage, nationality group, or country of birth. The categories for this variable are: Non-Hispanic, Mexican, Puerto Rican, Cuban, Other and Not reported. |
| Citizenship status | Categorical | Reports the citizenship status of respondents, distinguishing between naturalized citizens and non-citizens. |
| English proficiency | Categorical | Indicates whether the respondent was able to speak English. |
| Years of education | Discrete | Indicates respondents' educational attainment, as measured by the highest year of school or degree completed. Note that completion differs from the highest year of school attendance; for example, respondents who attended 10th grade but did not finish were classified as having completed 9th grade. |
| STEM degree | Categorical | Indicates if the person holds a bachelor’s degree in a STEM field or not. |
| Birthplace | Categorical | Indicates the US state, the outlying US area or territory, or the foreign country where the person was born. |
| STEM occupation | Categorical | Indicates if the person works in a STEM field or not. |

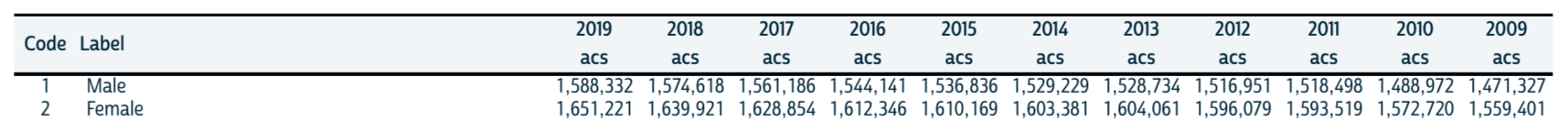
Appendix B

Description of the complete dataset for the target variable STEM taken from IPUMS



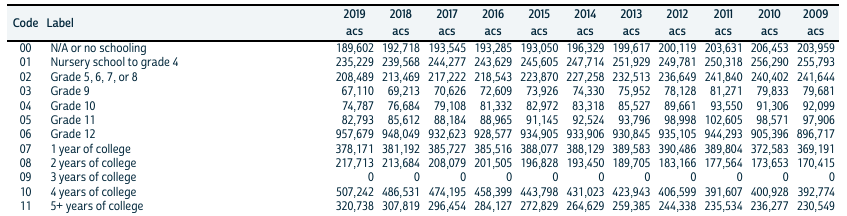
Appendix C

Description of the complete dataset for the variable sex



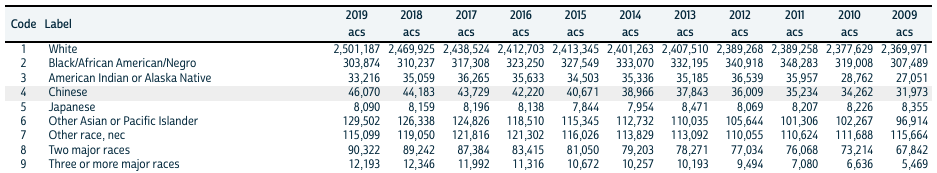
Appendix D

Description of the complete dataset for the variable education



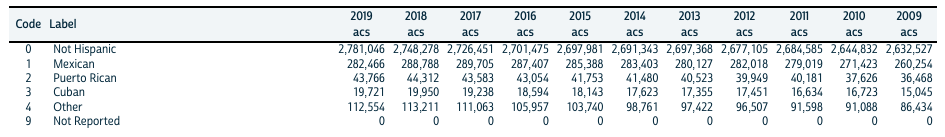
Appendix E

Description of the complete dataset for the variable race



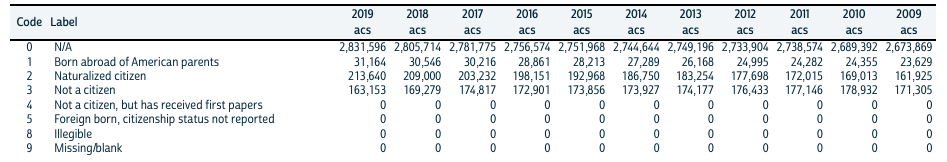
Appendix F

Description of the complete dataset for the variable Hispanic origins



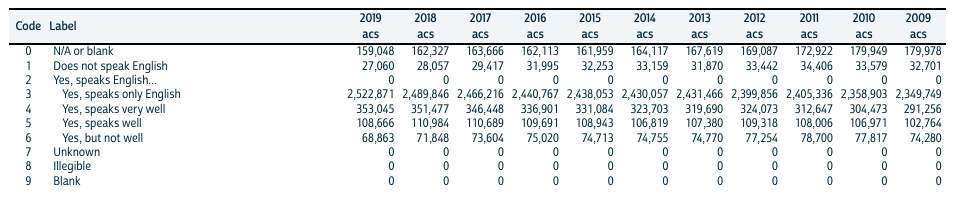
Appendix G

Description of the complete dataset for the variable citizen



Appendix H

Description of the complete dataset for the variable English proficiency



Appendix I

R code used to extract, clean and organize the data that feeds the model.

Data from IPUMS was downloaded on .csv format, the code below stores the data in a R data frame and writes distinct .csv files for each year. The R function read\_csv() can save .csv files into R data frames.

library(tidyverse)

data\_df = read\_csv('~/data\_source/usa\_all\_data.csv')

year\_list = unique(data\_df$YEAR)

for(n in 1:length(year\_list)){

temp = filter(data\_df,YEAR==year\_list[n])

write\_csv(temp,sprintf('~/data\_source/usa\_year\_%d.csv',as.integer(year\_list[n])))

}

The population group between the ages of 21 and 62 years is filtered from the total population for each year. The column OCC that contains multiple profession codes is transformed using the mutate() function in R, this function transforms the column by storing all profession codes related to STEM careers into 1, and all other professional codes into 0. The same function, mutate, is used to group all Hispanic Origins in a single value (1), and all other ethnicities into 0.

year\_df = read\_csv('~/data\_source/usa\_year\_2009.csv')

usa\_2009 = filter(year\_df,AGE>20)

usa\_2009 = filter(usa\_2009,AGE<63)

write\_csv(usa\_2009,'~/data\_source/usa\_2009.csv')

usa\_pop\_2009 = read\_csv('~/data\_source/usa\_2009.csv') %>%

  mutate(OCC = ifelse(OCC %in% c(110,10001,11102,1040,1010,10201,1060,11003,1200,1220,12101,12302,12403,1320,13301,13402,1350,1360,1400,1410,1420,1430,1440,1450,1460,15001,15202,15101,15302,1600,1610,1640,16501,1700,1710,1720,1740,1760),1,0))

write\_csv(usa\_pop\_2009,'~/data\_source/usa\_pop\_2009.csv')

library(rms)

usa\_df = read\_csv('~/data\_source/usa\_pop\_2019.csv') %>%

  mutate(hispan = ifelse(HISPAN == 0, 0, 1)) %>%

  mutate(educ2 = ifelse(EDUC <= 6, 0, ifelse(EDUC < 11, 1, 2))) %>%

  mutate(racesing = as.factor(RACE)) %>%

  mutate(speakeng = ifelse(SPEAKENG == 1, 0, 1)) %>%

  mutate(citizen = ifelse(CITIZEN == 3,0,1))

A R data frame was created with all selected variables, a summary of the main statistical measures of the data set are calculated using the function summary, to start the exploratory data analysis. The R function summary() automatically calculates the following statistics for the data set: the minimum value, the 1st quartile (25th percentile), the median, the 3rd quartile (75th percentile), and the maximum value.

t3 = usa\_df[,c('OCC','hispan','citizen','racesing','speakeng','BPL','SEX','EDUC','educ2','AGE')]

describe(t3)

s = summary(OCC ~ hispan + citizen + racesing + speakeng + BPL + SEX + EDUC + AGE, data=t3)

par(mar=c(4,5,1,3),oma=c(4,5,1,3))

plot(s, main= ' ', subtitles= FALSE)

The Logistic Regression Model was applied to different combinations of the variables to determine the combination of variables that best adjust to the model. The lrm() function in R fits a binary dependent variable and independent variables into a Logistic Regression model.

dd = datadist(t3)

options(datadist='dd')

f = lrm(OCC ~ hispan + citizen + racesing + BPL + SEX + EDUC + AGE, data = t3)

f2 = lrm(OCC ~ hispan + citizen + racesing + speakeng + BPL + SEX + educ2 + AGE, data = t3)

f3 = lrm(OCC ~ speakeng + racesing + (citizen + hispan)^2 + BPL + SEX + EDUC + AGE, data = t3)

f4 = lrm(OCC ~ speakeng + racesing + (BPL + hispan)^2 + citizen + SEX + EDUC + AGE, data = t3)

Finally, the Akaike Information Criterion, or AIC for short, is the method used for scoring and selecting the best model for this study. This can be accomplished using the R function AIC().

print(AIC(f))

print(AIC(f2))

print(AIC(f3))

print(AIC(f4))