Logistic Regression: Social Determinants of Health Database

Factors Determining if a County is Medically Underserved

By: Olivia Schultheis

Contents

[Page Numbers of Figures 2](#_Toc183787871)

[Page Numbers of Tables 2](#_Toc183787872)

[Introduction to Database: 4](#_Toc183787873)

[Overview of Response Variable: HRSA\_MUA\_COUNTY 4](#_Toc183787874)

[Calculation of Response: HRSA\_MUA\_COUNTY (see Appendix for Additional Tables) 6](#_Toc183787875)

[Geography: Viewing Medically Underserved Areas (MUAs) on a Map 8](#_Toc183787876)

[Checking Assumptions for Logistic Regression 10](#_Toc183787877)

[Initial Logistic Regression Model 15](#_Toc183787878)

[Considering Impact of Other Races: Are Races Other than Asian Significant in Determining whether a County is Medically Underserved? 25](#_Toc183787879)

[Checking Assumptions for Logistic Regression Model with Additional Races 26](#_Toc183787880)

[Conclusion 27](#_Toc183787881)

[Appendix A: Weighted Value Tables for Four Contributors to IMU Score 29](#_Toc183787882)

[Appendix B: Omitted Counties for Logistic Regression Model 34](#_Toc183787883)

# Page Numbers of Figures

[Figure 1: Map of U.S. by Medical Underservice based on MUA Designation 9](#_Toc183787884)

[Figure 2: Map of Alaska with Counties Colored by Medically Underserved or Not 9](#_Toc183787885)

[Figure 3: Map of Hawaii with Counties Colored by Medically Underserved or Not 10](#_Toc183787886)

[Figure 4: VIF for Initial Logistic Regression 11](#_Toc183787887)

[Figure 5: Scatterplots of the Log Odds Against the Independent Variables in Logistic Regression with Outliers 11](#_Toc183787888)

[Figure 6: Scatterplots of the Log Odds Against the Independent Variables in Logistic Regression without Outliers 13](#_Toc183787889)

[Figure 7: VIF values for logistic Regression without ACS\_PCT\_HOUSEHOLDER\_BLACK and without ACS\_PCT\_HOUSEHOLDER\_AIAN (American Indian/Alaskan Native) 14](#_Toc183787890)

[Figure 8: Scatterplot of Log-Odds Against Independent Variables when the only numeric variables are percentage with a graduate degree and percentage of householders who are only Asian 14](#_Toc183787891)

[Figure 9: Boxplot of the Distribution for the percent of 25+ year old county population with a graduate degree 16](#_Toc183787892)

[Figure 10: Boxplot of the Distribution of Percentage of Counties' Householders whose only race is Asian 17](#_Toc183787893)

[Figure 12: Scatterplot of Logit against Quantitative Predictors for only Counties in Puerto Rico 22](#_Toc183787894)

[Figure 13: Scatterplot of Deviance Residuals for Logistic Regression of Only Counties in Puerto Rico 23](#_Toc183787895)

[Figure 14: Cook's Distance Plot of Data Points in a Logistic Regression - only counties in Puerto Rico 24](#_Toc183787896)

[Figure 15: Histogram of Predictions from Logistic Regression for only Counties in Puerto Rico 25](#_Toc183787897)

[Figure 16: VIF Values for Logistic Regression with Various Races 26](#_Toc183787898)

[Figure 17: Scatterplot of Logit against Predictors for Regression analyzing the percentage of the population of each race, or the percentage of the population that is more than one race 27](#_Toc183787899)

[Figure 18: APPENDIX B: Counties that are Omitted from Logistic Regression due to lack of data 34](#_Toc183787900)

# Page Numbers of Tables

[Table 1: Number of Medically Underserved and Not Medically Underserved Counties in 2019 5](#_Toc183787901)

[Table 2: Contributors to IMU (Index of Medical Underservice) Score 6](#_Toc183787902)

[Table 3: Poverty Guidelines per Number of people in Household for states in the contiguous U.S. 7](#_Toc183787903)

[Table 4: R-Squared Values of Linear Regression between the Logit and the Independent Variables in Logistic Regression 12](#_Toc183787904)

[Table 5: Lower and Upper Boundaries for Outliers in the Variables Related to Race. Note that these are percentages 12](#_Toc183787905)

[Table 6: Frequency of Regions for data points that are outliers (based on IQR rule) for the variable of the percentage of householders whose only race is black 13](#_Toc183787906)

[Table 7: New R-Squared values for Linear Regression between the Logit and the Independent Variables without the variables or percentage of householders who are either only black or only American Indian/Alaskan Native 14](#_Toc183787907)

[Table 8: Summary Measures for Percentage of 25-year old and above population with a graduate degree 15](#_Toc183787908)

[Table 9: Summary Measures for the Percentage of Householders with only Race Asian within the Counties 16](#_Toc183787909)

[Table 10: R Output for Logistic Regression Model 18](#_Toc183787910)

[Table 11: Distribution of Number of Counties in U.S. Territories and Not in Territories 18](#_Toc183787911)

[Table 12: Coefficients and Odds Ratios for the Variable "Region" 18](#_Toc183787912)

[Table 13: Coefficients and Odds Ratios for the Rural-Urban Continuum Code, given by the USDA in 2013, from Area Health Resource Files 19](#_Toc183787913)

[Table 14: Coefficients and Odds Ratios for the percentage of householders whose only race is Asian and the percentage of the 25+ year old population with a graduate degree 19](#_Toc183787914)

[Table 15: APPENDIX A: Standardized Values for Percentage of County Below Poverty in IMU Score 30](#_Toc183787915)

[Table 16: APPENDIX A: Standardized Values for Percent of Population Aged 65 and Older in IMU Score 31](#_Toc183787916)

[Table 17: APPENDIX A: Standardized Values for County's Infant Mortality Rate in IMU Score 32](#_Toc183787917)

[Table 18: APPENDIX A: Standardized Values for Ratio of Full-Time Equivalent Physicians per 1,000 Population in IMU Score 33](#_Toc183787918)

# Introduction to Database:

The data for this analysis comes from the [Social Determinants of Health Database (SDOH)](https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html) released by the Agency for Human Research and Quality (AHRQ). The AHRQ is a federal agency that aims to improve accessibility and quality of health care by developing tools and data for individuals and professionals to make informed health-related decisions. AHRQ works with the U.S. Department of Health and Human Services to ensure its evidence is strong and understandable. One major goal of AHRQ is to generate materials to help healthcare personnel apply research to their practices.[[1]](#footnote-1)

The SDOH database provides data at three geographic levels: county, zip code, and census tract. The data files are easily linkable. A major benefit of this database is that data is compiled from many sources, including multiple other federal agencies, such as the CDC, USDA, and even the EPA. Data are provided for the years 2009-2020 over five domains. The five domains are **social** context(age, race, veteran status, etc.), **economic** context (income, unemployment), **education** (test scores, level of degree attained), **physical infrastructure** (housing, crime, public transportation, presence of stores/fitness centers, etc.), and **healthcare** (insurance, Medicare).[[2]](#footnote-2) In this analysis, I look at only the county-level data, as I believe it is the most interpretable level for data analysis across the entire United States.

Each county-level data file from 2009-2020 has at least 650 variables, and some of the year’s files have over 1,000 variables. It is possible to merge the 11 years of possible data into one dataset with only the common fields. A logistic regression model will be applied to the 2019 county-level data, which is being used because it is the most recent data. Further, the assumption of independence for logistic regression will be examined. Since measurements for some variables are repeated for the counties over the years, running a model with all 11 years of data for the same counties would violate this assumption of independence.

# Overview of Response Variable: HRSA\_MUA\_COUNTY

The goal of this logistic regression is to determine what factors or characteristics of counties in the United States influence whether the counties are deemed medically underserved. This is an important question that could influence where people would prefer to live, as healthcare is a priority for many people. Also, the results could highlight areas of improvement for policymakers and suggest geographical areas that need attention for healthcare improvement. Of course, the results could also contradict expectations of which variables would impact a county being medically underserved, which is also important information.

The response variable for the logistic regression model is titled **“HRSA\_MUA\_COUNTY”**. This variable is present for all years from 2009-2019. It is not included in 2020’s dataset. The data for the response variable is provided by the Health Resources and Services Administration (HRSA). MUA stands for medically underserved areas, which indicates an area has a shortage of access to primary care. In the datasets, **“HRSA\_MUA\_COUNTY”** is a binary variable that can only take on a value of 0 (not underserved) or 1 (underserved).

The distribution of counties being medically underserved or not underserved for 2019 is almost a 50/50 split. As seen in the table below, there are slightly more underserved (1,625) counties than not underserved (1,605) counties in 2019’s dataset.

|  |  |
| --- | --- |
| Number of Counties **Not** Medically Underserved in 2019 | Number of Counties Medically Underserved in 2019 |
| 1605 | 1625 |

Table : Number of Medically Underserved and Not Medically Underserved Counties in 2019

While there are 3,232 counties in the dataset, two counties (American Samoa and Mariana Islands, which are both in territories) do not have the response variable provided. These two counties were omitted from the study.

Behind the scenes, each state’s Primary Care Office (PCO) applies for MUA designation for a county. The applications are viewed by the Health Resources & Service Administration’s (HRSA) Office of Shortage Destination, which decides if the application meets the criteria for a Medically Unsupervised Area. [[3]](#footnote-3)

The following section details how the response variable is calculated, which is important to understand the factors that go into a county’s classification of being medically underserved or not. Note that the following variables directly impact the classification, and variables related to these are not too meaningful in a logistic regression because it is known that they impact the response. This analysis aims to discover if additional, unrelated factors influence the underserved/not underserved status.

# Calculation of Response: HRSA\_MUA\_COUNTY (see Appendix for Additional Tables)

There is a calculation to determine whether the indicator **“HRSA\_MUA\_COUNTY”** is 0 or 1 for each record in the database. An Index of Medical Underservice (IMU) score is calculated for each county using four measures. For each of the four measures, a table converts the raw value to a weighted value. These four standardized tables are featured in Appendix A. Then, the four weighted values are added to get the IMU score. The four contributors to IMU are shown below, along with the maximum contribution each can have to the Index of Medical Underservice: [[4]](#footnote-4)

|  |  |
| --- | --- |
| Measure | Maximum Number of Points contributed to IMU |
| Primary Care Physicians per 1,000 Population Ratio | 28.7 |
| Percent of Population at or below 100% of the Federal Poverty Level (FPL) | 25.1 |
| Percent of Population aged 65+ | 20.2 |
| Infant Mortality Rate | 26 |

Table : Contributors to IMU (Index of Medical Underservice) Score

Areas that are **shortage** **zones** for a lack of access to primary care have **low** **IMU** scores. Areas that are **not underserved** have **higher IMU** scores. Specifically, a county with an **IMU score of 62.0 or lower** **is designated an MUA (Medically Underserved Area).** On the contrary, a county with an **IMU score greater than 62.0 is not considered an MUA.** The IMU score ranges from 0 to 100, with 0 indicating being completely underserved and 100 indicating that being medically underserved is not a concern for that county.[[5]](#footnote-5)

For clarification, being at or below 100% of the poverty level means that a household’s income is less than or equal to the guideline amount (i.e. Federal Poverty Level) determined by the Department of Health and Human Services to indicate poverty. New values for FPL are released each year for the 48 contiguous states, with separate measures for Alaska and Hawaii. There are different poverty guidelines depending on family/household size. Below is a chart showing the values signifying 100% of the poverty level in 2019 for the 48 contiguous states and Washington D.C: [[6]](#footnote-6)

A white sheet with black text

Description automatically generated

Table : Poverty Guidelines per Number of people in Household for states in the contiguous U.S.

Alaska and Hawaii generally have higher Federal Poverty Levels than shown above for the 48 contiguous states, indicating a higher cost of living in these two states.

To begin calculating IMU — for each county — the number of people with an income at or below the Federal Poverty Level is determined and divided by the county’s total civilian (non-institutionalized) population before multiplying by 100. This results in the population percentage at or below 100% of the FPL. This percentage is found in Table 14 of Appendix A and matched to its weighted value.

Next, the percentage of the civilian (non-institutionalized) population aged 65 or older is determined. This percentage is found in Table 15 of Appendix A below to get its weighted value.

Third, the infant mortality rate for each county is calculated using the state’s Department of Health data. This is computed by dividing the number of infant deaths by the number of live births and multiplying by 1000. This yields an estimate of the number of deaths per 1,000 live births in a county. The corresponding weighted value can be found in Table 16 of Appendix A.

Finally, the ratio of primary care providers per 1,000 population is calculated by dividing all primary care physicians in a county by the civilian population and multiplying by 1,000. The physicians counted in this study must be full-time equivalent (FTE) primary care physicians. The calculated value of primary care physicians per 1,000 population is found in Table 17 of Appendix A to determine the matching weighted value.

All tables in Appendix A are taken from Appendix II of a document titled “Guidelines for MUA/MUP Designation” from the Community Clinic Association of Los Angeles County’s website.[[7]](#footnote-7)

After the four weighted values are determined from the tables in Appendix A, they are added. Again, a value of 62.0 or less qualifies a county as a medically underserved area.

# Geography: Viewing Medically Underserved Areas (MUAs) on a Map

Below is a map of the United States, colored by medically underserved status. Areas that are green on the map are not medically underserved, and areas that are red on the map are underserved. The Northeast region, as well as much of the West Coast, is mostly *not* underserved. The Southern region of the United States shows the highest overall degree of medical underservice. The central United States shows a mix of counties being both medically underserved and not underserved. The data defines the region as being “Northeast”, “Midwest”, “South”, or “West” within the U.S.

**Map of US: Counties Colored by Medical Underservice or Not**

A map of the united states

Description automatically generated

Figure : Map of U.S. by Medical Underservice based on MUA Designation

While not intuitive, counties in Alaska and Hawaii are assigned a region of “West” in the dataset.

Alaska shows a mix of medically underserved and not medically underserved counties. Many of the counties in southern Alaska are medically underserved, while the northern and central parts of the state do not appear medically underserved, based on MUA score.

**Alaska:**

:A map of the united states

Description automatically generated

Figure : Map of Alaska with Counties Colored by Medically Underserved or Not

2 of the 5 Hawaiian counties in the dataset are medically underserved, while 3 are not. Honolulu County, Kalawao County (label not shown), and Maui County are **not** medically underserved, while Kauai County and Hawaii County are medically underserved.

**Hawaii:**

A map of hawaii with red and green outline

Description automatically generated

Figure : Map of Hawaii with Counties Colored by Medically Underserved or Not

# Checking Assumptions for Logistic Regression

Assumptions for a logistic regression model with variables hypothesized to impact the probability that a county is medically underserved are examined. These variables are the rural-urban continuum code from 2013 (AHRF\_USDA\_RUCC\_2013, which is the most recent rural-urban designation in the dataset), the percentage of the 25+ year-old county population with a graduate degree (ACS\_PCT\_GRADUATE\_DGR), the percentage of householders whose only race is black (ACS\_PCT\_HOUSEHOLDER\_BLACK), whose only race is Asian (ACS\_PCT\_HOUSEHOLDER\_ASIAN), and whose only race is American Indian/Alaskan Native (ACS\_PCT\_HOUSEHOLDER\_AIAN).

The following assumptions will be examined:

1. **Response variable is binary**
2. **Observations are independent: one county’s values should not impact another county’s values.**
3. **VIF Values for Numeric Variables to examine multicollinearity**

A black text on a white background

Description automatically generated

Figure : VIF for Initial Logistic Regression

All the VIF values are between 1 and 2. Since a VIF of 5 is often a threshold to signal initial concern of multicollinearity and a value of 10 signals high multicollinearity, these values are all low enough to not be concerned about multicollinearity.

1. **Linear Relationship between Explanatory Variables and Logit of Response**

**This assumption is violated due to the percent of householders that only identify as American Indian/Alaskan Native (AIAN in the dataset) and only identify as black/African American.**

**ORIGINAL CHECK OF ASSUMPTIONS (OUTLIERS ARE PRESENT)**

A graph of a graph showing a number of blue lines

Description automatically generated with medium confidence

Figure : Scatterplots of the Log Odds Against the Independent Variables in Logistic Regression with Outliers

values for a **linear** regression conducted with log odds as the dependent variable and the quantitative, independent variables of interest for the logistic regression are shown below. The log odds come from a logistic regression model run with region, rural-urban continuum code, percentage of county population with a graduate degree, and percentage of householders in a county identifying as only Black, Asian, or American Indian/Alaskan Native (AIAN):

|  |  |  |
| --- | --- | --- |
| Which variable against log odds? |  | Correlation |
| ACS\_PCT\_HOUSEHOLDERS\_BLACK | 0.1026 | 0.3203 |
| ACS\_PCT\_HOUSEHOLDERS\_ASIAN | 0.3581 | -0.5984 |
| ACS\_PCT\_HOUSEHOLDERS\_AIAN | 0.0292 | 0.1709 |
| ACS\_PCT\_GRADUATE\_DGR | 0.4889 | -0.6992 |

Table : R-Squared Values of Linear Regression between the Logit and the Independent Variables in Logistic Regression

While a high is not expected for this type of data, Figure 5 shows that outliers (for Black and American Indian/Alaskan Native race) are influencing the linearity assumption between the logit and the independent variables of the percentage of householders who only identify as Black or who only identify as American Indian/Alaskan Native.

An analysis can be performed using data without outliers by ignoring data points for the percentage of householders that are only black or only American Indian/Alaskan Native which are 1.5 times the interquartile range beyond the first and third quartiles. The values in Table 5 represent **percentages.**

|  |  |  |
| --- | --- | --- |
| Variable | Lower Bound (Q1 – 1.5 \* IQR) | Upper Bound (Q3 + 1.5 \* IQR) |
| ACS\_PCT\_HOUSEHOLDER\_BLACK | -13.8 | 23.9 |
| ACS\_PCT\_HOUSEHOLDER\_AIAN | -0.88 | 1.76 |

Table : Lower and Upper Boundaries for Outliers in the Variables Related to Race. Note that these are percentages

Only the upper bound will be considered since the lower bounds give negative percentages. Outliers for the percentage of householders that are only black are values greater than 23.9% and for the percentage of householders that are only American Indian/Alaskan Native are values greater than 1.76%. Overall, this shows that most (at least 75%) of the data for the percentage of householders that are only American Indian or Alaskan Native are very low, as any value of around 2% or greater is considered an outlier.

**NO OUTLIERS (BASED ON RANGE BETWEEN Q1-1.5\*IQR and Q3 + 1.5\*IQR)**

A graph of a graph showing a number of numbers

Description automatically generated with medium confidence

Figure : Scatterplots of the Log Odds Against the Independent Variables in Logistic Regression without Outliers

A problem is that outliers mostly define any linear relationship that could be salvaged for each of these two variables. The percentage of householders who are only American Indian/Alaskan Native and only black appear to have a completely random relationship with the logit when outliers for these two variables are removed. While we could develop a model that ignores the cluster of points on the left of the original graphs in Figure 5 and only looks at the “outliers” on the right side of the two graphs, this would result in ignoring most of the data.

Below is a frequency table of **REGION** for outliers of the percentage of householders who are only black variable:

|  |  |  |  |
| --- | --- | --- | --- |
| Midwest | Northeast | South | Territory |
| 12 | 4 | 394 | 9 |

Table : Frequency of Regions for data points that are outliers (based on IQR rule) for the variable of the percentage of householders whose only race is black

About 94% of the outliers for the percentage of householders whose only race is black exist in the Southern region, so it would not make sense to run the model for only these points when investigating the impact of a county’s region.

In Figure 6, the logit plots of the graduate degree and percentage of householders who are Asian are reasonably linear. Overall, these plots could be assumed to not violate the assumption of linearity between the logit and the predictor.

Assumptions for a model using region, urban-rural continuum code (as measured by **AHRF\_USDA\_RUCC\_2013**), the percent of householders who are only Asian, and the percent of the population 25 and older with a graduate degree are examined.

1. **The response variable is still binary.**
2. **Observations are still independent**
3. **VIF Values**

A number and numbers on a white background

Description automatically generated

Figure : VIF values for logistic Regression without ACS\_PCT\_HOUSEHOLDER\_BLACK and without ACS\_PCT\_HOUSEHOLDER\_AIAN (American Indian/Alaskan Native)

1. **Plot of Log odds against variables**

A graph of a graph showing a blue line

Description automatically generated with medium confidence

Figure : Scatterplot of Log-Odds Against Independent Variables when the only numeric variables are percentage with a graduate degree and percentage of householders who are only Asian

values for a **linear** regression conducted with log odds as the dependent variable and the quantitative, independent variables of interest for the logistic regression are shown below. The log odds come from a logistic regression model run with region, rural-urban continuum code, percentage of county population with a graduate degree, and percentage of householders in a county identifying as only Asian:

|  |  |  |
| --- | --- | --- |
| Which variable against log odds? |  | Correlation |
| ACS\_PCT\_GRADUATE\_DGR | 0.5147 | -0.71741 |
| ACS\_PCT\_HOUSEHOLDER\_ASIAN | 0.3484 | -0.59025 |

Table : New R-Squared values for Linear Regression between the Logit and the Independent Variables without the variables or percentage of householders who are either only black or only American Indian/Alaskan Native

The relationship between the log-odds and the percentage of the 25+ aged population with a graduate degree became slightly stronger than before when the percentage of householders who were only Asian and the percentage of householders who were only American Indian/Alaskan Native were included in the model as well. The relationship between the log-odds and the percent of householders who are only Asian is slightly weaker —but still very similar to—what was found in the previous model. This provides evidence that this is a valid model to run.

# Initial Logistic Regression Model

First, a logistic regression model was developed on the 2019 county-level data. One variable included in this model was **Region** (which can be “South”, “Midwest”, “Northeast”, or “West” for counties in the 50 states, or “Territory” for U.S. territories). Since the original data left the “**Region”** column blank for territories, the values for **“Territory”** were used for “**Region”** in United States territories. The **baseline category for “Region” is “Midwest.”** Also included is a variable called **“AHRF\_USDA\_RUCC\_2013”,** which is provided by the U.S. Department of Agriculture. This variable is categorical and is a rural-urban continuum code from the year 2013, which can take on a value of 1-9, denoting if a county is in a metro, urban, or rural area within certain population values. For rural and urban areas, the variable also records whether a county is adjacent to a metro area. The specific population ranges that the variable records, as well as adjacency to a metro area, are shown in the R model output. For this original regression, the impact of education and race are also considered. Further, a model could potentially be considered unreliable if both the percentage of householders whose only race was white and the percentage whose only race was black were included because these variables were highly negatively correlated. A discussion of the percentage of householders whose only race is white is not included separately because when it is used in a regression with the other variables of interest (as ACS\_PCT\_HOUSEHOLDER\_BLACK was), it does not show linearity with the logit.

Note that 10 of the 3,230 counties with the response variable present are missing values for the percentage of householders who are only Asian and the percentage with graduate degrees. All 10 of the counties missing these two variables are in territories: 5 counties in American Samoa, 1 county in Guam, 1 county in the Northern Mariana Islands, and three in US Virgin Islands. The specific counties missing from the logistic regression are provided Appendix B, Figure 18. **As a result of these 10 observations being removed, all territory data comes only from Puerto Rico.**

Below is a summary of the quantitative variables in the model with boxplots to provide an exploratory overview of the variables.

**Summary: Percentage of Population Aged 25+ with a Graduate Degree**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum | Number NAs |
| 0 | **4.9** | **6.4** | **7.7** | **9.2** | **42.7** | **10** |

Table : Summary Measures for Percentage of 25-year old and above population with a graduate degree

75% of the counties have populations where the percentage of 25+ year-olds with a graduate degree is 9.2% or less. The highest percentage of a county’s 25-year-old and above population with a graduate degree out of all the counties is about 42.7%. A boxplot of these results is given below.

**Boxplot:**

A graph with a line

Description automatically generated

Figure : Boxplot of the Distribution for the percent of 25+ year old county population with a graduate degree

**Summary: Percentage of Householders whose only race is Asian**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum | Number of NAs |
| 0 | **0.1** | **0.4** | **1.1** | **1.0** | **45.2** | **10** |

Table : Summary Measures for the Percentage of Householders with only Race Asian within the Counties

**Boxplot:**

**A graph with a line

Description automatically generated**

Figure : Boxplot of the Distribution of Percentage of Counties' Householders whose only race is Asian

Most counties have a very small percentage of householders whose only race is Asian, with the majority of counties having a rate of close to only about 0.5% - 1%. The upper 25% of the data for the percentage of householders whose only race is Asian can range from about 1% to about 45%.

**REGRESSION RESULTS:**

Logistic Regression results are given in the following table. This model includes region, rural-urban continuum code, percentage of population aged 25+ with a graduate degree, and percentage of householders who are only Asian.

**Logistic Regression Model Output from R:**

A close up of text

Description automatically generated

Table : R Output for Logistic Regression Model

All variables used in this model are significant at the 0.05 level for at least one of their categories. Two of the categories within the USDA rural-urban continuum code are not significant — but since six of the eight categories are significant — the variable is included. The standard error of territory is especially high, partially because there are significantly fewer counties in territories than in non-territories. For 2019, the frequency distribution for territories and non-territories is:

|  |  |
| --- | --- |
| Number of counties that are not in U.S. territories | Number of counties in U.S. territories |
| 3142 | 88 |

Table : Distribution of Number of Counties in U.S. Territories and Not in Territories

For the variables of interest, Puerto Rico is the only **territory** without missing values, and there are **78 counties in Puerto Rico observed in the data. As such, the territory interpretation only applies to Puerto Rican counties and not counties in all U.S. territories.** The coefficients of the logistic regression model are most interpretable in terms of odds ratios. Below, the tables summarize the model’s coefficients and odds ratios:

**REGION (Baseline is Midwest Region)**

|  |  |  |
| --- | --- | --- |
| **Region** | **Coefficient** | **Odds-Ratio** |
| Northeast | -0.980 | e-0.98044 = 0.375 |
| South | 1.437 | e1.43707 =4.208 |
| Territory (Puerto Rico) | 3.496 | e3.49640 =33 |
| West | 0.563 | e0.56263 =1.756 |

Table : Coefficients and Odds Ratios for the Variable "Region"

**Rural-Urban Continuum Code as of 2013 (AHRF\_USDA\_RUCC\_2013, Baseline is Counties in Metropolitan Areas with Population of 1 million or more)**

|  |  |  |
| --- | --- | --- |
| **Rural-Urban Code (2013)** | **Coefficient** | **Odds Ratio** |
| Metro Areas, Population 250,000 – 1,000,000 | 0.197 | e 0.197 = 1.218 |
| Metro Areas, Population < than 250,000 | 0.364 | e 0.364 = 1.438 |
| Urban Areas, Population 20,000+, adjacent to a metro area | 0.308 | e0.308 =1.360 |
| Urban Areas, Population 20,000+, **not adjacent to a metro area** | 0.594 | e0.594 = 1.812 |
| Urban Areas, Population 2,500-19,999, adjacent to a metro area | 0.707 | e0.707 = 2.029 |
| Urban Areas, Population 2,500-19,999, **Not adjacent** to a metro area | 0.718 | e0.718 = 2.050 |
| Completely rural or less than 2,500 urban population, adjacent to a metro area | 1.850 | e1.850 = 6.357 |
| Completely rural or less than 2,500 urban population, **Not adjacent** to a metro area | 1.884 | e 1.884 = 6.582 |

Table : Coefficients and Odds Ratios for the Rural-Urban Continuum Code, given by the USDA in 2013, from Area Health Resource Files

**Continuous Variables**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient** | **Odds Ratio** |
| ACS\_PCT\_HOUSEHOLDER\_ASIAN | -0.167 | e -0.167 = 0.846 |
| ACS\_PCT\_GRADUATE\_DGR | -0.117 | e -0.117 = 0.890 |

Table : Coefficients and Odds Ratios for the percentage of householders whose only race is Asian and the percentage of the 25+ year old population with a graduate degree

From the odds-ratio related to REGION, counties in Puerto Rico are much more likely to be medically underserved compared to Midwest regions in the US, which is the baseline category. In fact, **a county in Puerto Rico is about 33 times as likely to be medically underserved than a county in the Midwestern U.S**. A northeastern U.S. County is less likely to be medically underserved than a Midwest county. By reciprocating the odds ratio for the Northeast coefficient, a Midwest County is about 2.67 times as likely to be medically underserved than a county in the Northeast. Based on this model, counties in the South and the West are more likely than counties in the Midwest to be medically underserved. Southern counties are about 4.21 times as likely to be medically underserved than Midwest counties, and Western counties are about 1.76 times as likely to be medically underserved than Midwest counties.

Next, the coefficients of **AHRF\_USDA\_RUCC\_2013** increase as the rural-urban continuum code increases, meaning that the counties that are less populated and more rural have a higher probability of being medically underserved. The code is an arbitrary number from 1-9, with 1 indicating metro areas with a population of more than one million, and 9 indicating completely rural counties or urban areas with populations that are less than 2,500 and are not adjacent to a metro area. Codes for metro areas are distinguished only based on population, whereas nonmetro areas are distinguished by the level of urbanization and adjacency to a metro area.[[8]](#footnote-8) A metro refers to an area in which at least one core city and can include other cities within the range. The population density in a metro is higher than in a non-metro area; metros also have more economic and social integration than non-metros.[[9]](#footnote-9) Since the coefficients increase within the metro levels, the urban levels, and rural levels as the population decreases, the model suggests that rural areas are more likely to be medically underserved than metro areas with more than 1 million in population, which matches expectations due to population density.

Within just the metro categories, the likelihood of a county being medically underserved was about 21.8% higher for metro areas with a population between 250,000 to 1 million compared to metro areas with populations over 1 million. The likelihood of a county being medically underserved was about 43.8% higher for metro areas with a population of less than 250,000 compared to metro areas with a population of more than a million people. For urban counties with populations greater than 20,000 (adjacent to a metro area), the likelihood of a county being medically underserved increases by a factor of about 1.36, or about 36%, compared to metro counties with more than 1,000,000 people. For urban counties with a population greater than 20,000 that are *not* adjacent to a metro area, the likelihood of a county being medically underserved increases by a factor of about 1.81, or about 81%, as compared to metro areas with a population greater than one million. A similar pattern exists for urban areas with populations between 2,500 and 19,999. Counties in these areas that are not adjacent to a metro area are more likely (about 2.05 times, or 105% more likely) to be medically underserved ( in reference to metro areas with over 1 million people) than urban counties with populations between 2,500 and 19,999 that are adjacent to a metro area (which are about 2.03, or 103% more likely to be medically underserved than metro areas with population over a million). Additionally, for rural areas with a population less than 2,500, the rural areas adjacent to metro areas are about 6.36 times as likely to be medically underserved as metro areas above population 1 million, and rural areas less than 2,500 population that are not adjacent to metro areas are about 6.58 times as likely to be medically underserved as metro areas above population one million.

The overall pattern is that urban areas are more likely to be medically underserved than metro areas, and rural areas are the most likely areas to be medically underserved. Also, it is less likely for urban and rural areas that are adjacent to metro areas to be medically underserved than urban and rural areas that are **not** adjacent to metro areas.

The percentage of a county’s population that has a graduate degree and is aged 25 and above does appear to be associated with a decreased likelihood that a county is medically underserved. The coefficient of -0.11663 means that — for each percentage point increase in a county’s population aged 25 or above that has a master’s or professional/doctorate degree —the odds of a county being medically underserved decrease by a factor of about 0.8899; in other words, the odds of a county being medically underserved are expected to be about 11% lower for each increase of 1% of the population with a graduate degree that is 25 or older, holding constant region, rural-urban continuum code, and percentage of householders whose only race is Asian.

The logistic regression model predicts that counties with higher percentages of householders whose only race is Asian are expected to have a lower likelihood of being medically underserved. The odds-ratio for **ACS\_PCT\_HOUSEHOLDER\_ASIAN** suggests that, as the percentage of householders in a county who are Asian-alone increases by one percent, the odds of a county being medically underserved decrease by a factor of about 0.8461. The odds of a county being medically underserved are about 15.4% lower for each one percent increase in householders who report Asian being their only race, holding constant region, rural-urban continuum code, percentage of the 25 and above year-old population that has a graduate degree.

This initial model gives a broad overview of significant variables in determining which counties are medically underserved. However, the pseudo-can be calculated by finding:

1 -

An optimal model would have a large drop from null to residual deviance. In this case, the pseudo-is:

= 0.221.

While the variables included in this model were significant, the low pseudo- indicates that there are other variables that impact the likelihood of a county being medically underserved that are not currently in the model. However, for this type of analysis, a high is not anticipated. Rather, the goal is to consider which variables are significant following all assumptions of a logistic regression model.

A possible alternative analysis would be to develop separate logistic regression models for only counties within the 50 U.S. states and for only counties in Puerto Rico to see if variable significance is different between the two models. This question was motivated by the finding that counties in Puerto Rico are much more likely to be medically underserved than counties within the 50 states based on this initial regression model.

When performing this analysis, there were no changes in which variables were significant nor how significant they were for logistic regression of only counties within the continental US, Alaska, and Hawaii. Further, no model assumptions were violated.

However, running a logistic regression with the **same variables as above for only counties in Puerto Rico** results in **no significant variables** (among rural-urban continuum code, percentage of householders with only race Asian, and percentage of population aged 25 and above with a graduate degree).

A look at a plot of the logit against the quantitative variables shows why a logistic model performs poorly on only Puerto Rican counties.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

Figure : Scatterplot of Logit against Quantitative Predictors for only Counties in Puerto Rico

No linear relationship between the logit and predictors can be salvaged for counties in **Puerto Rico.**

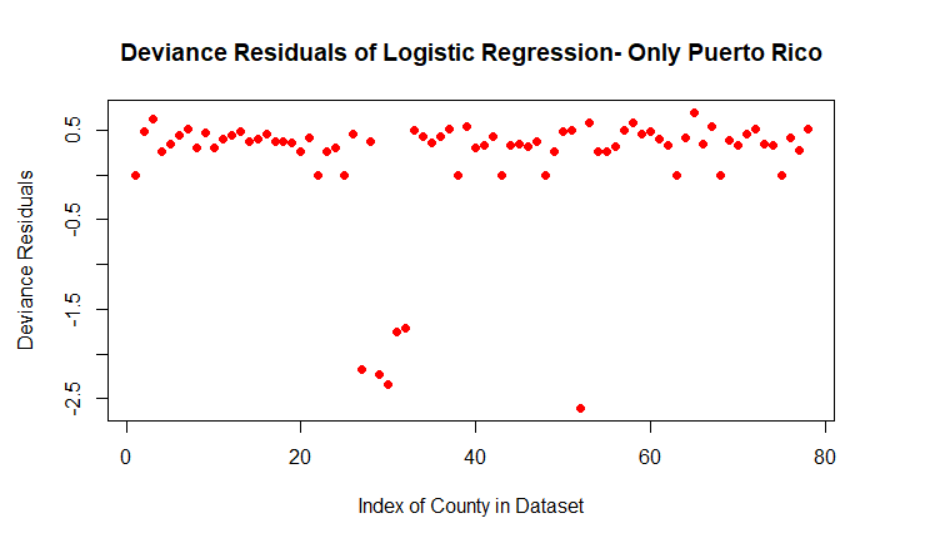


Figure : Scatterplot of Deviance Residuals for Logistic Regression of Only Counties in Puerto Rico

There are six data points with particularly low deviance residuals out of the 78 total points for counties in Puerto Rico. These are the same points that appear unusual in a plot of Cook’s Distance values, which consider points that are influential to a regression, whether the influence comes from being an outlier, high leverage point, or both. A high Cook’s Distance value for a point indicates that, in this case, logistic regression coefficients could be substantially changed if the point is removed. It turns out that the 6 data points with negative deviance residuals and high Cook’s Distance values are the 6 out of 78 Puerto Rican territories that are **not medically underserved.** The raw residuals ( = are also negative for the 6 points where the deviance residuals are negative. Since the 6 territories that are **not medically underserved have HRSA\_MUA\_COUNTY = 0 for the response,** subtracting the actual value of 0 from a predicted probability close to 1 produces a negative raw residual.

A graph with numbers and lines

Description automatically generated

Figure : Cook's Distance Plot of Data Points in a Logistic Regression - only counties in Puerto Rico

The threshold denoting a high Cook’s Distance value and represented by the red line is an arbitrary —but frequently used — cutoff of .[[10]](#footnote-10)

A graph of a number of people

Description automatically generated

Figure : Histogram of Predictions from Logistic Regression for only Counties in Puerto Rico

The bar chart above shows that the model predicts a median probability of a county in Puerto Rico being underserved of about 0.93. Since most counties in Puerto Rico are medically underserved, the model predicts a high enough value for the probability of being medically underserved that — at almost any reasonable threshold to use for classification —the Puerto Rico counties that are **not medically underserved will nevertheless be classified as medically underserved**. With such a vast lack of balance in the response variable for territories, it is difficult to run a model to determine the impact of variables on whether counties in territories are medically underserved.

From this analysis, it appears that counties in Puerto Rico are much more likely to be medically underserved than non-territories and continue analysis looking only at the contiguous United States, as well as Alaska and Hawaii. There is much more data for counties in the 50 US states, and it is still very useful to know what variables influence whether a county in the 50 states is medically underserved.

Since there is additional data from the years 2009-2018, the distributions of medically underserved/not medically underserved territories were examined. The frequencies for these other years were very similar to the 2019 data.

# Considering Impact of Other Races: Are Races Other than Asian Significant in Determining whether a County is Medically Underserved?

From the previous logistic regression, a higher percentage of householders whose only race is Asian in a county seems to reduce the likelihood of a county being medically underserved. Any linear relationship between the percentage of householders who are only black or only American Indian/Alaskan Native and the logit of our original model was mostly defined by outliers. However, there is data for other races in the dataset and a variable that indicates the percentage of the population of multiple races. Another point is that the relationship of a variable with the logit can change, depending on what other variables it is combined with in a logistic regression model.

In this section, the impact of other races will be explored.

First, region and rural-urban continuum code will be retained in this analysis. Then, the percentage of **the** **population** that is only Black, Hispanic, Native Hawaiian and Pacific Islander, Asian, American Indian and Alaskan Native, as well as the percentage of the population that is multiple races will be examined. Note that these variables are different in the dataset than the percentage of householders of certain races. An analysis of the assumptions for the races of Black and American Indian and Alaskan Native will be performed again to examine if the linearity assumption between these variables and the logit improves when data on other races is included in the model.

# Checking Assumptions for Logistic Regression Model with Additional Races

At this point, the main assumptions that must be checked are VIF values and linearity between the predictors and the logit. As discussed earlier, the data are independent and have a binary response.

**VIF Values**

A number and numbers on a white background

Description automatically generated

Figure : VIF Values for Logistic Regression with Various Races

None of these values are high, so multicollinearity does not appear to be an issue.

**Relationship between Predictors and the Logit**

A graph of a graph showing the difference between a number of people

Description automatically generated with medium confidence

Figure : Scatterplot of Logit against Predictors for Regression analyzing the percentage of the population of each race, or the percentage of the population that is more than one race

The plots of the logit against the predictor variables suggest that linearity cannot be assumed between the logit and any other races except for Asian. As a result, the logit model will not be redeveloped with these additional variables.

# Conclusion

Among data for about 911 variables for over 3,000 territories in the U.S. and its territories, the region of a county within the United States, as well as if a county is in Puerto Rico, had a statistically significant impact on whether the county is medically underserved. Southern counties are more likely to be medically underserved than counties in the Midwest, and Northeastern counties are less likely to be medically underserved than counties in the Midwest. Counties in Puerto Rico are about 33 times as likely to be medically underserved than counties in the Midwestern United States. Only 6 of 78 Puerto Rican Counties were not medically underserved in 2019. Additionally, being medically underserviced is more likely as the population changes from metro areas to urban and then rural. Overall, the likelihood that a county is medically underserved increases as a county becomes more rural. Within the same level of rural/urban status and population, counties adjacent to metropolitan areas are slightly less likely to be medically underserved than counties not adjacent to metropolitan areas, as compared to the baseline of metropolitan counties with populations of 1 million or higher. So— even if counties are in urban or rural areas — being close to a metro area seems to slightly help the degree of medical service available. Finally, higher values for the percentage of householders whose only race is Asian and the percentage of the 25+ year old population in a county with a graduate degree are associated with a decreased likelihood of a county being medically underserved. Comparing the impacts of these factors on medical underservice within counties in U.S. states compared to Puerto Rico was inconclusive since about 92% of Puerto Rican counties were medically underserved in 2019.

# Appendix A: Weighted Value Tables for Four Contributors to IMU Score

**Weighted Values for IMU Calculation**

**Percentage of Population Below Poverty Level [[11]](#footnote-11)**

A table with numbers and lines

Description automatically generated

Table : APPENDIX A: Standardized Values for Percentage of County Below Poverty in IMU Score

**Percentage of Population Aged 65 and Older [[12]](#footnote-12)**

A table with numbers and lines

Description automatically generated

Table : APPENDIX A: Standardized Values for Percent of Population Aged 65 and Older in IMU Score

**Infant Mortality Rate [[13]](#footnote-13)**

A table of numbers with numbers on it

Description automatically generated with medium confidence

Table : APPENDIX A: Standardized Values for County's Infant Mortality Rate in IMU Score

**Ratio of Full-Time Equivalent Primary Care Physicians Per 1,000 People [[14]](#footnote-14)**

A table of numbers with numbers on it

Description automatically generated

Table : APPENDIX A: Standardized Values for Ratio of Full-Time Equivalent Physicians per 1,000 Population in IMU Score

# Appendix B: Omitted Counties for Logistic Regression Model

A screenshot of a computer

Description automatically generated

Figure : APPENDIX B: Counties that are Omitted from Logistic Regression due to lack of data

These counties had missing values for percentage of householders of only race Asian and percentage of 25+ year old population that had graduate degrees. They are the ten omitted counties from the logistic regression model

1. Agency for Healthcare Research and Quality. (2022, February). *About AHRQ*. Retrieved from ahrq.gov: https://www.ahrq.gov/cpi/about/index.html [↑](#footnote-ref-1)
2. Agency for Healthcare Research and Quality (2023, June). *Social Determinants of Health Database.* Retrieved from ahrq.gov: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html [↑](#footnote-ref-2)
3. Community Clinic Association of Los Angeles County. *Guidelines for MUA/MUP Designation. Appendix II-MUA-MUP-General Procedures for Requesting Designation Guidelines* (p.1). Retrieved from Community Clinic Association of Los Angeles County: https://ccalac.org/wordpress/wp-content/uploads/Appendix-II-MUA-MUP-Guidelines.doc [↑](#footnote-ref-3)
4. NORC at the University of Chicago. AHRQ Social Determinants of Health (SDOH) Database: Data Source Documentation. In *ahrq.gov* (p. 329). https://www.ahrq.gov/sites/default/files/wysiwyg/sdoh/SDOH-Data-Sources-Documentation-v1-Final.pdf [↑](#footnote-ref-4)
5. NORC at the University of Chicago. AHRQ Social Determinants of Health (SDOH) Database: Data Source Documentation. In *ahrq.gov* (p. 329). [↑](#footnote-ref-5)
6. Office of the Assistant Secretary for Planning and Evaluation (2019). *2019 Poverty Guidelines. Retrieved from aspe.hhs.gov:* https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines/prior-hhs-poverty-guidelines-federal-register-references/2019-poverty-guidelines [↑](#footnote-ref-6)
7. Community Clinic Association of Los Angeles County. *Guidelines for MUA/MUP Designation. Appendix II-MUA-MUP-General Procedures for Requesting Designation Guidelines.* (pp. 5-8). Retrieved from Community Clinic Association of Los Angeles County: https://ccalac.org/wordpress/wp-content/uploads/Appendix-II-MUA-MUP-Guidelines.doc [↑](#footnote-ref-7)
8. NORC at the University of Chicago. AHRQ Social Determinants of Health (SDOH) Database: Data Source Documentation. In *ahrq.gov* (pp. 89-90). [↑](#footnote-ref-8)
9. Summers, Owen. “What Is the Difference between Metro and Non-Metro?” *NCESC.com*, 23 June 2024, www.ncesc.com/geographic-faq/what-is-the-difference-between-metro-and-non-metro/. [↑](#footnote-ref-9)
10. Altman, N., Krzywinski, M. Analyzing outliers: influential or nuisance?. *Nat Methods* **13**, 281–282 (2016). https://doi.org/10.1038/nmeth.3812 [↑](#footnote-ref-10)
11. Community Clinic Association of Los Angeles County. *Guidelines for MUA/MUP Designation. Appendix II-MUA-MUP-General Procedures for Requesting Designation Guidelines.* (p. 5). [↑](#footnote-ref-11)
12. Community Clinic Association of Los Angeles County. *Guidelines for MUA/MUP Designation. Appendix II-MUA-MUP-General Procedures for Requesting Designation Guidelines.* (p. 6). [↑](#footnote-ref-12)
13. Community Clinic Association of Los Angeles County. *Guidelines for MUA/MUP Designation. Appendix II-MUA-MUP-General Procedures for Requesting Designation Guidelines.* (p. 7). [↑](#footnote-ref-13)
14. Community Clinic Association of Los Angeles County. *Guidelines for MUA/MUP Designation. Appendix II-MUA-MUP-General Procedures for Requesting Designation Guidelines.* (p. 8). [↑](#footnote-ref-14)