Exploring the Association between Poverty and Heart Disease Mortality at the County Level in the United States

Group 61: Elizabeth Yee, Ivory Parker Jr., Michael Cornejo, Dan Condon

Introduction

The objective of our project is to explore the relationship between poverty and heart disease mortality at the county level in the United States. Heart disease is a leading cause of death in the United States and may be strongly influenced by socioeconomic factors such as poverty. Heart disease and stroke costs the United States healthcare system $216 billion per year. Understanding the relationship between poverty and heart disease mortality is important for identifying potential areas for intervention and policy action.

Since heart disease is a leading cause of death in the United States, there is a need to understand the relationship between those causes and potential socioeconomic factors such as poverty, ethnicity, and geographic location. In order to identify potential areas for intervention and policy action, there is a need to understand the relationship between poverty and heart disease mortality at the local county level. By developing a better understanding of the patterns and trends in poverty and heart disease mortality at the county level, we hope to contribute to efforts to reduce the burden of heart disease in the United States. The findings from our analysis could inform efforts to address poverty levels in specific counties and highlight the need for improved strategies for education, job training, financial assistance, economic development, or providing access to healthcare and social services.

If we are able to demonstrate a relationship between socioeconomic factors and cardiovascular disease mortality, it could have significant implications for public policy makers. Based on our analysis, public policy makers will be able to make informed decisions on where to allocate resources and invest in more job training and more job opportunities in areas with higher levels of poverty. By addressing poverty proactively, public policy makers could potentially avoid larger public health issues in the future, reducing healthcare costs and improving overall health outcomes. Our initial hypothesis is that we expect there to be a significant association between poverty rates and heart disease mortality at the county level in the United States.

Literature Review

Coronary heart disease (CHD) remains the leading cause of mortality in low-income US counties. In a paper by O'connor and Wellenius, the authors found that the prevalence rates of diabetes and coronary heart disease were higher among people living in rural areas compared to urban areas. As annual household income decreased, the prevalence rates of the two diseases increased. Those in the lowest income bracket were almost three times more likely to report coronary heart disease compared with those in the highest bracket. Even after controlling for many of the common risk factors such as income, age, gender, ethnicity, BMI, and tobacco use, people in rural environments are still more likely to be diagnosed with CHD than people in urban locations.

For those under the age of 70, developing heart disease is associated with an increased risk of falling into both income poverty and multidimensional poverty. Callander and Schofield found that 31% of those who developed heart disease fell into income poverty and by comparison only 15% of people who did not develop heart disease did. For individuals over 70, those who developed heart disease had a reduced risk of poverty.

Finally a paper by Hamad *et al*. found that individuals with low socioeconomic status (SES) are disproportionately affected by CHD. Low SES was defined by income below 150% of the federal poverty level or an educational level less than a high school diploma. Approximately 31.2 million US adults aged 35 to 64 years had low SES, of whom approximately 16 million (51.3%) were women. Adults with low SES have double the rate of myocardial infarctions than individuals with higher SES.

Methods

Our approach to analyzing the relationship between poverty and heart disease mortality at the county level in the United States involved several steps. First, we obtained data from the American Community Survey (ACS) on poverty rates at the county level as published by the US Census Bureau. We also obtained data from the Rates and Trends in Heart Disease Mortality dataset from the Centers for Disease Control and Prevention (CDC) regarding data on heart disease mortality rates by county, race/ethnicity, and sex.

Before conducting our analysis, we used the DPLYR package in R Studio to clean and organize the two datasets. The ACS data had several rows that were not necessary for our analysis, so those were removed. The removal of these rows did not affect the overall analysis of the data because these rows only served as descriptors of the dataset.

The dataset column headers were also too wordy and descriptive. The dataset initially contained over 500 columns, so columns were filtered down to only the relevant variables. To make referencing the column names easier, titles of the columns were renamed. Finally, the county names and states were initially combined within the same column, so these were separated into two individual columns. This made it easier for the team to filter through different states.

After cleaning the ACS dataset, the team prepared and cleaned the Heart Disease Mortality dataset. This process involved removing observations that contained any missing data or null values. Again, the team removed irrelevant columns and unwanted rows. Dummy variables for different demographics such as genders, regions of the country, and ethnicities were created. Region of the country was determined per Figure B.1. Lastly, the state names were changed from an abbreviation to the full state name.

Both the ACS and CDC datasets had their county and state name combined into one new column. From there, the datasets were merged based on that new county plus state name column. During the merge process, Alaska in particular was difficult to merge, as it used different nomenclature (‘Borough’, ‘Municipality’, ‘City’) than the rest of the states (‘County’). This merging error was resolved by dropping the Alaskan nomenclatures and replacing it with the same verbiage that was utilized by the rest of the states.

After the full datasets were merged together, some exploratory plots were formed. In addition, a new dataset was formed with solely the overall data. The combined dataset was divided into training and test groupings. The model was trained with 70% of the data, keeping all variables that are statistically significant and removing variables that exhibited multicollinearity. During this process, any outliers using Cook’s Distance were noted and adjustments were made to the model as necessary. An initial linear regression model was generated from this overall dataset. Each team member then worked on crafting a more well-tailored linear regression model. To compare the performance of the different models, adjusted R-squared and root mean squared error were used as deciding metrics. Three separate linear regression models were created for the different variables in each of the following 3 factors: gender, ethnicity and region. In addition, separate models were developed for each ethnicity and gender.

The top model was identified via a backwards elimination regression technique and was selected due to its relatively high adjusted R-square and low root mean squared error values. First, a linear regression model was fitted with all broad predictor variables included. All the variables with p-values above the threshold of 0.05 were then removed. Finally, the linear model was refitted using only the remaining variables. After this was done, the model was checked for variables which exhibited multicollinearity and subsequently removed all variables with a VIF greater than 5. This model then attempted to predict the results of the test data, which comprised the final 30% of the overall data. The models' accuracy was then observed using adjusted R-squared and root mean squared error.

Exploratory Data Analysis

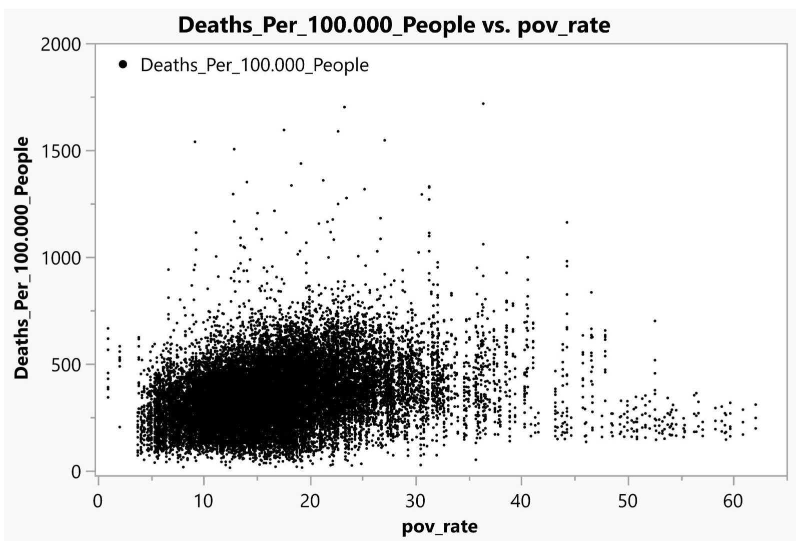

Figure 1 - Deaths per 100,000 people vs Poverty Rate

Figure 1 above displays a scatter plot of Deaths per 100,000 People vs poverty rate, with each point representing a county of the United States. The x-axis represents the poverty rate in percentage, and the y-axis represents the number of deaths per 100,000 people due to heart disease. There appears to be a fairly positive linear relationship between the two variables, up to a poverty rate of approximately 30%. However, beyond this point, the distribution of the data becomes more skewed. There are still observations of low deaths per 100,000 people at very high poverty rates.

This suggests that while poverty rates may be an important factor in heart disease mortality, other factors may be at play in determining the death rate for individuals living in areas with higher poverty rates. Further investigation outside of this project may be necessary to identify additional factors and how to understand their interaction with poverty rates to impact the risk of heart disease.

To understand why there is a skew in the data, a deeper dive was required. From some cursory investigation, the team was able to identify that the vast majority of these skewed data points at a high poverty rate and low mortality rate were from counties in Puerto Rico. The U.S. Census Bureau reports that in 2021 the poverty rate in Puerto Rico was 40.5%. The population is also not that large, just over 3 million citizens. This low population density coupled with the territory's high average poverty rate would explain the skewed data points.

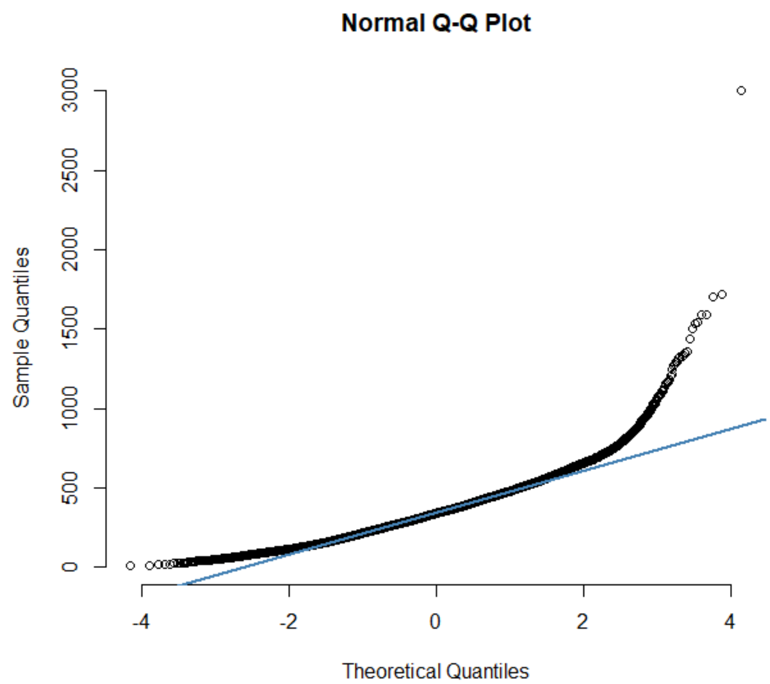

Figure 2 - Q-Q Plot

Figure 2 shows that the data is skewed towards the lower end of the distribution curve. There seems to be a concentration of having higher mortality rates, which is causing our dataset to not adhere to a normal distribution.

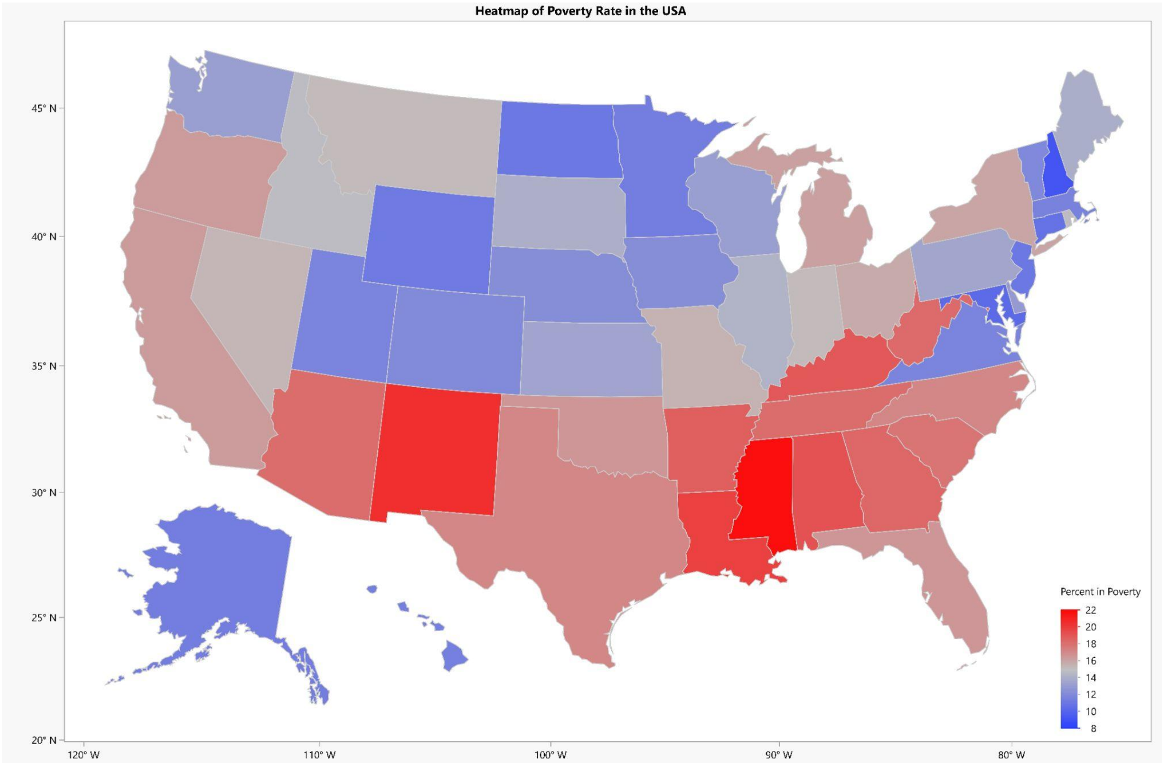

Figure 3 - Heat Map of Poverty Rates in the US

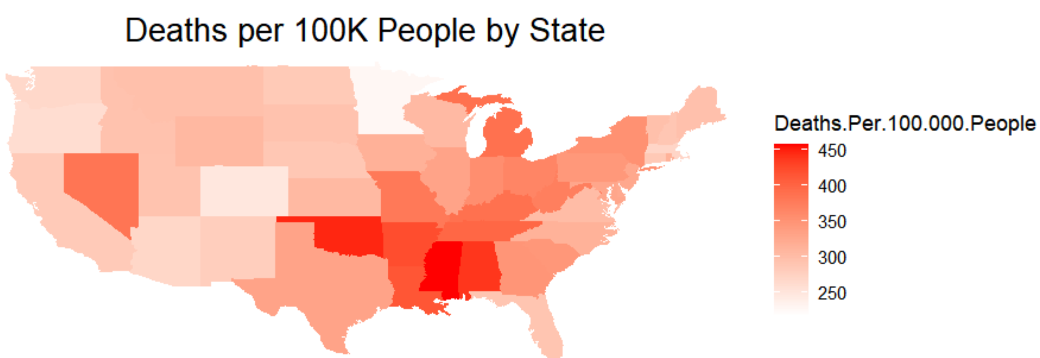

Figure 4 - Heat Map of Heart Disease Mortality in the US

Figures 3 and 4 represent the poverty rate in each state and the heart disease mortality rate in each state, respectively. Some states in the West-South Central region and East-South Central region exhibit a positive relationship between poverty rates and heart disease mortality. These regions simultaneously have high levels of poverty and high levels of mortality. These regions tend to be more rural and have lower wages. Additionally, the populations in more rural regions may not have access to the same quality healthcare as more urban regions, resulting in higher mortality rates. Conversely, many states in the West Mountain region exhibit lower levels of poverty and also illustrate lower levels of heart disease mortality.

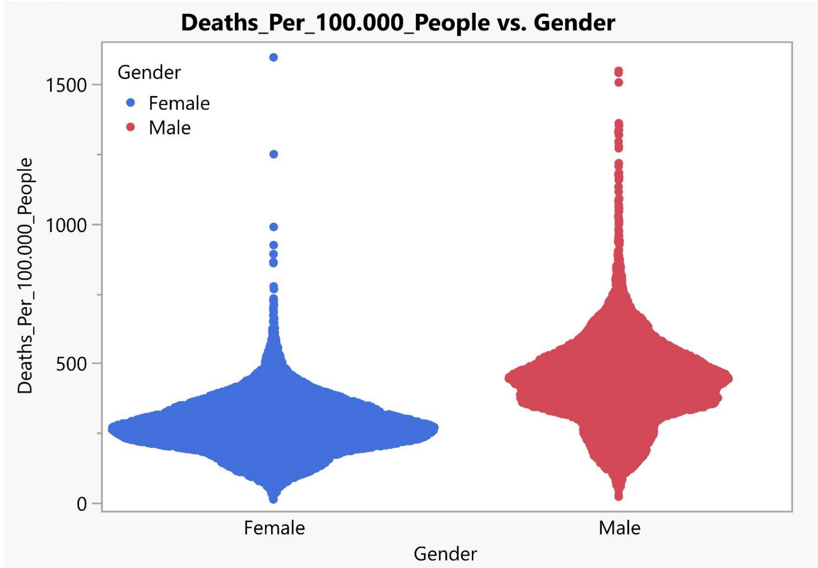

Figure 5 - Deaths per 100,000 people by Gender

Figure 5 illustrates the effects of gender on death rates. Males have an average death rate of 427.8 per 100,000 while females have an average death rate of 277.2 per 100,000. Males have a higher death rate overall when compared to females. This effect is confirmed by the linear regression on gender (Appendix A.1). Being female is associated with a lower death rate of 150.5 less deaths per 100,000 people.

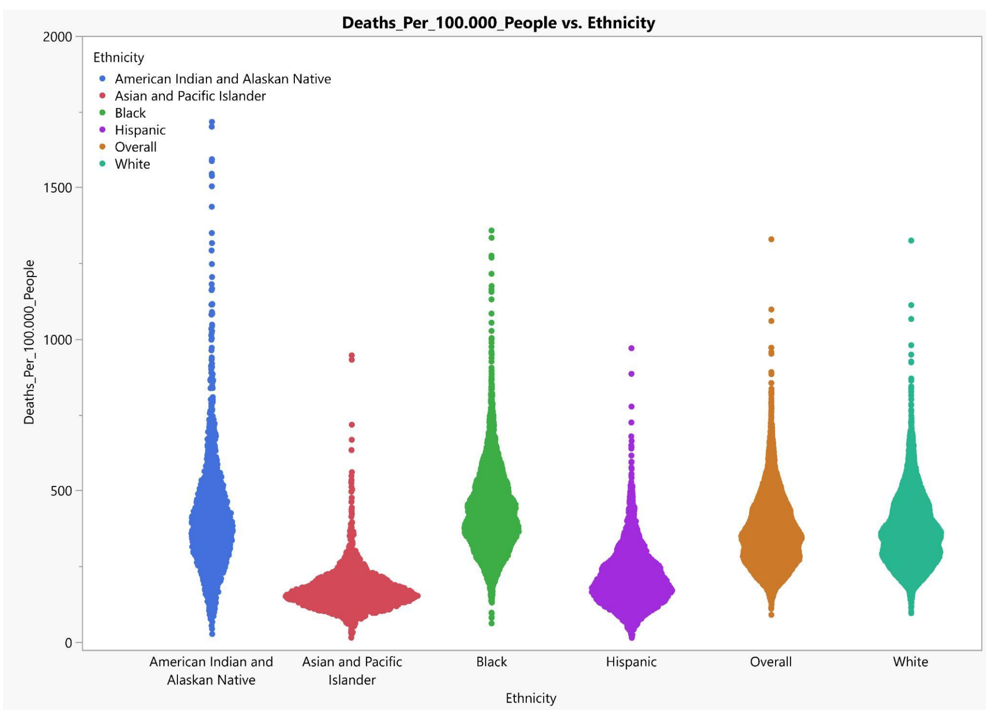

Figure 6 - Deaths per 100,000 people by Ethnicity

Figure 6 shows the mortality (based on deaths per 100,000) of selected ethnic groups. The following average death rates (per 100,00) were observed: American Indian and Alaskan

Native at 454.0, Asian and Pacific Islander (AAPI) at 174.8, Black at 439.7, Hispanic at 218.2, Overall at 354.5, and White at 370.0. Compared to white Americans, Black and Native Americans were more likely to have a higher death rate at 69.6 and 84.1 deaths per 100,000 respectively. Hispanic and AAPI groups had a death rate of 151.0 and 195.2 deaths per 100,000 lower than white Americans.

The American Indian and Alaskan Native groups were found to have a much larger spread of mortality. This could be due to there being a much smaller population of this group within each county, so this data could be skewed. Additionally, the team observed some outliers within this group. Cowley County in Kansas had a mortality rate of over 3,000 per hundred thousand. The AAPI group has mortality rates concentrating at much lower rates than other ethnic groups. This suggests that the AAPI group are dying from cardiovascular disease at much lower rates than the general population. The black population mortality rates, on average, appear to be slightly higher than the mortality rates of the other ethnic groups.

Modeling Results

Model	Adj. R-Squared	R-Squared Rating	Gender? (Y/N)	Ethnicity? (Y/N)	Region? (Y/N)	Poverty Rate? (Y/N)
Gender Only	0.2256	Weak	Y	N	N	N
Ethnicity Only	0.3119	Weak	N	Y	N	N
Region Only	0.1065	Weak	N	N	Y	N
Poverty Only	0.0739	Weak	N	N	N	Y
Overall Model	0.6345	Moderate	Y	Y	Y	Y
Females	0.5762	Moderate	Y	Y	Y	Y
Males	0.5411	Moderate	Y	Y	Y	Y
White	0.6748	Moderate	Y	Y	Y	Y
Black	0.478	Moderate	Y	Y	Y	Y
Hispanic	0.3585	Moderate	Y	Y	Y	Y
Asian	0.3286	Moderate	Y	Y	Y	Y
Native American	0.3959	Moderate	Y	Y	Y	Y

Table 1 - Summary of Multiple Linear Regression Models

Multiple models were produced based on the exploratory data analysis. Table 1 shows the performance of the different models. Models generated from lone demographic descriptors were not good models based on their adjusted R^2 values. The overall linear regression model showed the highest adjusted R^2 value, with the exception of a model that was developed for the White demographic only.

```
Call:
lm(formula = Deaths_Per_100.000_People ~ pov_rate + Male + Black +
    Hispanic + Asian_and_Pacific_Islander + Native_American +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-442.98  -50.67   -4.45   44.95 1113.02

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 174.5987     4.4122  39.572  < 2e-16 ***
pov_rate                      4.2514     0.1672  25.424  < 2e-16 ***
Male                        152.0228     1.9760  76.933  < 2e-16 ***
Black                        57.3025     2.4834  23.074  < 2e-16 ***
Hispanic                   -147.5586     3.0396 -48.545  < 2e-16 ***
Asian_and_Pacific_Islander -178.4688     3.8746 -46.061  < 2e-16 ***
Native_American              85.1938     4.6483  18.328  < 2e-16 ***
Middle_Atlantic              50.2634     4.9765  10.100  < 2e-16 ***
East_North_Central           54.6031     4.1518  13.152  < 2e-16 ***
West_North_Central           36.5137     4.2434   8.605  < 2e-16 ***
South_Atlantic               21.0269     3.9084   5.380 7.63e-08 ***
East_South_Central          104.7618     4.5459  23.045  < 2e-16 ***
West_South_Central           97.5222     3.9458  24.715  < 2e-16 ***
Mountain                      7.3660     4.5264   1.627    0.104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.66 on 9387 degrees of freedom
Multiple R-squared:  0.6351,    Adjusted R-squared:  0.6345
```

Figure 7 - Regression Model

Figure 7 shows the top linear regression model that was extracted from the analysis of the dataset. The base case for this model was a White female from New England. This model showed that the most important predictors are poverty rate, male, Black, Hispanic, Asian and Pacific Islander, Native American, Middle Atlantic, East and West North Central, South Atlantic, East and West South Central. The R-squared value was 0.6351 and the adjusted R-squared value is 0.6345. This indicates that the model explains 63.51% of the variance in our dependent variable, the heart disease mortality rate, which is a moderate to strong level of predictive power. The results of training and testing our data gave a mean square error of 9379.48 and a root mean square error of 96.85. This indicates that the difference between our predicted and actual values averages 96.85 deaths per 100,000.

Discussion

Our analysis indicated that poverty rates alone were not as significantly predictive as we initially hypothesized when compared to other factors, such as gender, ethnicity, and region, in determining death rates for most population groups. Although there is evidence of an overall positive correlation between high poverty rates and high death rates, other factors weighed heavier in our model. This could be due to the fact that there are other factors that contribute to heart disease mortality specifically to each group which were not accounted for in this study. A possible area for future research could be to expand on the factors that are measured per population group in the United States.

The fact that the white population is the majority in the US may partially explain why the model for this group had a higher R^2 value than other ethnic groups. The larger sample size of white citizens may have allowed for a more precise estimate of their relationship between heart disease mortality and our other independent variables. By the same notion, it is possible that the relatively low population for Native Americans would influence the model's ability to predict this ethnic group's heart disease mortality rate.

Additionally, the fact that Hispanic and Asian populations had the lowest heart disease mortality rate could have led to mixed results from the model. Since heart disease mortality rate is relatively low for these groups compared to others, it may be more difficult to accurately predict mortality rates using the same model as for other ethnic groups with higher rates.

Alternative models were created during the result-seeking phase of this project. A specific model which focused on only one of either gender, ethnicity, or region was created. These models ignored all other factors including poverty. The goal of this was to try to establish some one to one differences between each value within each grouping.

Another group of alternative models focused on generating a model for each demographic group (gender & ethnicity). These models removed all competing demographic values (i.e. the male model would remove the female values) and compared the isolated demographic value against all other variables covered in the broad linear regression model (Figure 8). All alternative linear regression models can be found in Appendix A with a description of what the model is showcasing.

Lastly, there were some models that were worked on, but never came to be used. For example, a Random Forest Model with the critical variables was created. While a random forest can give accurate predictions, it is very difficult to interpret the results. A "black box" model would be difficult to pitch to insurance companies or public health officials if the designers of the model themselves couldn't interpret the model. In addition, a model with a high R-squared value was generated, but due to the multicollinearity of the variables in the model, some variables had to be removed. After the variables had been removed, the model no longer was as reliable and our final linear regression model was used.

Conclusion

The best broad model for this project incorporated the general poverty rate of a resident's county, the region they were in (excluding this factor if a resident lived in the Pacific Region), and the resident's ethnicity and gender. The model had a moderate to strong level of predictive power, explaining 63.51% of the variance in heart disease mortality rates. This report also found some interesting data which suggested that while poverty is a critical factor in predicting heart disease mortality for most ethnicities, Asian and Hispanic residents do not exhibit a correlation between their specific demographic's poverty rate and heart disease mortality.

A benefit of this analysis is that the government/healthcare industry can use our model as a predictive tool. One could look at the results and focus resources on areas that have similar attributes to "at-risk" areas before the mortality rate begins to climb. The government could view this information as an argument to increase social safety net measures in an attempt to address higher mortality rates. This report's analysis could have implications for other stakeholders such as health insurance companies. They may choose to use the analysis to predict if the population of an area will have an increased likelihood for heart disease, and thus decide if they will raise insurance costs accordingly. Insurers can charge higher premiums to groups with higher levels of predicted heart failure based on the model. Lastly, hospitals/clinics could look at this data and see if they serve in or near an area that is clustered with high mortality rates. These healthcare organizations could take the initiative to perform outreach on the more at-risk counties in an attempt to lessen the most severe mortality rates. Overall, the potential benefits of our analysis could impact the healthcare sector and beyond.

References

Callander, E. J., & Schofield, D. J. (2016). The risk of falling into poverty after developing heart disease: a survival analysis. BMC Public Health, 16(1). [*https://doi.org/10.1186/s12889-016-3240-5*](https://doi.org/10.1186/s12889-016-3240-5)

Hamad, R., Penko, J., Kazi, D. S., Coxson, P., Guzman, D., Wei, P. C., Mason, A., Wang, E. A., Goldman, L., Fiscella, K., & Bibbins-Domingo, K. (2020). Association of Low Socioeconomic Status With Premature Coronary Heart Disease in US Adults. JAMA Cardiology, 5(8), 899. [*https://doi.org/10.1001/jamacardio.2020.1458*](https://doi.org/10.1001/jamacardio.2020.1458)

O'Connor, A., & Wellenius, G. (2012). Rural–urban disparities in the prevalence of diabetes and coronary heart disease. Public Health, 126(10), 813–820. [*https://doi.org/10.1016/j.puhe.2012.05.029*](https://doi.org/10.1016/j.puhe.2012.05.029)

Health and Economic Costs of Chronic Diseases. (2023). [*https://www.cdc.gov/chronicdisease/about/costs/index.htm#:~:text=Heart%20Disease%20and%20Stroke&text=These%20diseases%20take%20an%20economic,lost%20productivity%20on%20the%20job.&text=See%20the%20health%20and%20economic%20benefits%20of%20high%20blood%20pressure%20interventions*](https://www.cdc.gov/chronicdisease/about/costs/index.htm#:~:text=Heart%20Disease%20and%20Stroke&text=These%20diseases%20take%20an%20economic,lost%20productivity%20on%20the%20job.&text=See%20the%20health%20and%20economic%20benefits%20of%20high%20blood%20pressure%20interventions)

Appendix

A. Linear regression models

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Female, data = sub_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-404.76  -82.16   -0.86   79.05 1315.55 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  427.758      1.678  254.88   <2e-16 ***
Female      -150.512      2.406  -62.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.4 on 13428 degrees of freedom
Multiple R-squared:  0.2257,	Adjusted R-squared:  0.2256 
```

Figure A.1 - Linear Regression Based on Gender. “Male” was used as the base case.

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Black + Hispanic + Asian_and_Pacific_Islander +
    Native_American, data = sub_data)

Residuals:
    Min      1Q  Median      3Q     Max
-388.92  -89.74  -17.26   72.24 1262.08

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 369.955      1.674  221.06   <2e-16 ***
Black                        69.608      2.788   24.97   <2e-16 ***
Hispanic                   -151.001      3.387  -44.58   <2e-16 ***
Asian_and_Pacific_Islander -195.173      4.342  -44.95   <2e-16 ***
Native_American              84.068      5.136   16.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131.4 on 13425 degrees of freedom
Multiple R-squared:  0.3121,    Adjusted R-squared:  0.3119
```

Figure A.2 - Linear Regression Based on Ethnicity. “White” was used as the base case.

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Middle_Atlantic + East_North_Central +
    West_North_Central + South_Atlantic + Pacific + East_South_Central +
    West_South_Central + Mountain, data = sub_data)

Residuals:
    Min      1Q  Median      3Q     Max
-448.27 -103.00  -14.57   92.12 1371.36

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         256.201      8.760  29.245  < 2e-16 ***
Middle_Atlantic      61.675     10.090   6.112 1.01e-09 ***
East_North_Central   95.560      9.446  10.116  < 2e-16 ***
West_North_Central   88.536      9.480   9.339  < 2e-16 ***
South_Atlantic       83.211      9.227   9.018  < 2e-16 ***
Pacific              31.323      9.858   3.178  0.00149 **
East_South_Central  205.269      9.634  21.307  < 2e-16 ***
West_South_Central  148.846      9.284  16.033  < 2e-16 ***
Mountain             38.576      9.764   3.951 7.82e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 149.7 on 13421 degrees of freedom
Multiple R-squared:  0.107,     Adjusted R-squared:  0.1065
```

Figure A.3 - Linear Regression Based on Region. “New England” was used as the base case.

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Male + White + Black +
    Hispanic + Asian_and_Pacific_Islander + allpop + pov_rate +
    Pop18_64_pov + Pov_rate_18_64 + Pop65 + Pop65_pov + male_pop +
    Pov_rate_male + Pov_rate_white + Pov_rate_white_alone + Middle_Atlantic +
    East_North_Central + West_North_Central + South_Atlantic +
    East_South_Central + West_South_Central, data = sub_data)

Residuals:
    Min      1Q  Median      3Q     Max
-496.67  -50.71   -5.46   44.33 1232.40

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.607e+02  4.886e+00  53.362  < 2e-16 ***
Male                        1.518e+02  1.633e+00  92.967  < 2e-16 ***
White                      -8.467e+01  3.800e+00 -22.282  < 2e-16 ***
Black                      -2.494e+01  3.992e+00  -6.246 4.35e-10 ***
Hispanic                   -2.252e+02  4.174e+00 -53.953  < 2e-16 ***
Asian_and_Pacific_Islander -2.483e+02  4.636e+00 -53.555  < 2e-16 ***
allpop                     -9.582e-04  1.681e-04  -5.700 1.22e-08 ***
pov_rate                    1.276e+01  8.276e-01  15.414  < 2e-16 ***
Pop18_64_pov               -9.446e-05  1.243e-04  -0.760   0.4475
Pov_rate_18_64             -5.283e+00  6.242e-01  -8.464  < 2e-16 ***
Pop65                       2.002e-04  1.036e-04   1.931   0.0535 .
Pop65_pov                   3.182e-03  5.957e-04   5.342 9.35e-08 ***
male_pop                    1.817e-03  3.284e-04   5.533 3.21e-08 ***
Pov_rate_male              -4.554e+00  7.380e-01  -6.171 6.98e-10 ***
Pov_rate_white             -4.474e+00  4.427e-01 -10.106  < 2e-16 ***
Pov_rate_white_alone        6.458e+00  4.457e-01  14.492  < 2e-16 ***
Middle_Atlantic             4.233e+01  3.877e+00  10.919  < 2e-16 ***
East_North_Central          3.946e+01  3.072e+00  12.846  < 2e-16 ***
West_North_Central          2.039e+01  3.146e+00   6.483 9.31e-11 ***
South_Atlantic              9.139e+00  2.821e+00   3.240   0.0012 **
East_South_Central          8.336e+01  3.517e+00  23.701  < 2e-16 ***
West_South_Central          8.962e+01  2.877e+00  31.148  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.48 on 13408 degrees of freedom
Multiple R-squared:  0.6447,    Adjusted R-squared:  0.6441
```

Figure A.4-Highest Adjusted R^2 model which was not used due to multicollinearity issues

```
Call:
lm(formula = Deaths_Per_100.000_People ~ pov_rate + Male + Black +
    Hispanic + Asian_and_Pacific_Islander + Native_American +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-442.98  -50.67   -4.45   44.95 1113.02

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                174.5987     4.4122  39.572  < 2e-16 ***
pov_rate                     4.2514     0.1672  25.424  < 2e-16 ***
Male                       152.0228     1.9760  76.933  < 2e-16 ***
Black                       57.3025     2.4834  23.074  < 2e-16 ***
Hispanic                  -147.5586     3.0396 -48.545  < 2e-16 ***
Asian_and_Pacific_Islander -178.4688    3.8746 -46.061  < 2e-16 ***
Native_American             85.1938     4.6483  18.328  < 2e-16 ***
Middle_Atlantic             50.2634     4.9765  10.100  < 2e-16 ***
East_North_Central          54.6031     4.1518  13.152  < 2e-16 ***
West_North_Central          36.5137     4.2434   8.605  < 2e-16 ***
South_Atlantic              21.0269     3.9084   5.380 7.63e-08 ***
East_South_Central         104.7618     4.5459  23.045  < 2e-16 ***
West_South_Central          97.5222     3.9458  24.715  < 2e-16 ***
Mountain                     7.3660     4.5264   1.627    0.104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.66 on 9387 degrees of freedom
Multiple R-squared:  0.6351,    Adjusted R-squared:  0.6345
```

Figure A.5 - The overall model (removed high P value: Pacific Region)

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_female + Black +
    Hispanic + Asian_and_Pacific_Islander + Native_American +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain + Pacific, data = female_combined)

Residuals:
    Min      1Q  Median      3Q     Max
-287.17  -39.46   -4.97   33.75  847.45

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 186.3594     7.2107  25.845  < 2e-16 ***
Pov_rate_female               3.0620     0.1633  18.747  < 2e-16 ***
Black                        57.4421     2.5498  22.528  < 2e-16 ***
Hispanic                   -106.7951     3.1762 -33.624  < 2e-16 ***
Asian_and_Pacific_Islander -136.9617     3.9778 -34.432  < 2e-16 ***
Native_American              45.7635     5.0790   9.010  < 2e-16 ***
Middle_Atlantic              44.5137     7.8851   5.645 1.75e-08 ***
East_North_Central           53.4584     7.3591   7.264 4.39e-13 ***
West_North_Central           26.9412     7.4289   3.627  0.00029 ***
South_Atlantic               20.7183     7.2512   2.857  0.00429 **
East_South_Central           90.1715     7.6867  11.731  < 2e-16 ***
West_South_Central           89.9082     7.3092  12.301  < 2e-16 ***
Mountain                     13.8268     7.6394   1.810  0.07037 .
Pacific                       8.1284     7.7768   1.045  0.29598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.51 on 4566 degrees of freedom
Multiple R-squared:  0.5774,    Adjusted R-squared:  0.5762
```

Figure A.6-Female LM

```
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_male + Black +
    Hispanic + Asian_and_Pacific_Islander + Native_American +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain + Pacific, data = male_combined)

Residuals:
    Min      1Q  Median      3Q     Max
-484.75  -59.65   -3.77   51.44 1061.57

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 315.2813    11.3922  27.675  < 2e-16 ***
Pov_rate_male                 5.0434     0.2818  17.900  < 2e-16 ***
Black                        56.1463     4.0606  13.827  < 2e-16 ***
Hispanic                   -184.7804     4.8967 -37.736  < 2e-16 ***
Asian_and_Pacific_Islander -219.6663     6.3428 -34.632  < 2e-16 ***
Native_American             113.4502     7.2978  15.546  < 2e-16 ***
Middle_Atlantic              66.1721    12.5115   5.289 1.29e-07 ***
East_North_Central           65.3598    11.7033   5.585 2.47e-08 ***
West_North_Central           52.8064    11.7806   4.482 7.55e-06 ***
South_Atlantic               31.9072    11.5803   2.755  0.00589 **
East_South_Central          132.2776    12.2066  10.837  < 2e-16 ***
West_South_Central          115.8065    11.6529   9.938  < 2e-16 ***
Mountain                      9.0082    12.2314   0.736  0.46148
Pacific                       3.7710    12.3502   0.305  0.76012
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112 on 4807 degrees of freedom
Multiple R-squared:  0.5423,    Adjusted R-squared:  0.5411
```

Figure A.7-Male LM

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_white + Male +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain + Pacific, data = white_combined)

Residuals:
    Min      1Q  Median      3Q     Max
-323.10  -40.13   -4.56   34.27  790.18

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         167.327      7.558  22.139  < 2e-16 ***
Pov_rate_white        5.061      0.218  23.219  < 2e-16 ***
Male                168.338      2.116  79.571  < 2e-16 ***
Middle_Atlantic      55.633      8.579   6.485 9.90e-11 ***
East_North_Central   50.086      7.615   6.577 5.36e-11 ***
West_North_Central   21.143      7.457   2.835   0.0046 **
South_Atlantic       39.870      7.528   5.296 1.24e-07 ***
East_South_Central  120.537      7.815  15.423  < 2e-16 ***
West_South_Central   97.549      7.647  12.756  < 2e-16 ***
Mountain            -19.564      7.917  -2.471   0.0135 *
Pacific             -16.860      8.476  -1.989   0.0467 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.55 on 4315 degrees of freedom
Multiple R-squared:  0.6755,    Adjusted R-squared:  0.6748
```

Figure A.8-White LM

```
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_black + Male +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain + Pacific, data = black_combined)

Residuals:
    Min      1Q  Median      3Q     Max
-457.73  -57.62   -3.36   50.30  829.33

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        183.7605    15.9701  11.507  < 2e-16 ***
Pov_rate_black       1.2932     0.1472   8.784  < 2e-16 ***
Male               168.2761     4.2813  39.305  < 2e-16 ***
Middle_Atlantic     96.3734    17.6719   5.453 5.44e-08 ***
East_North_Central 112.8791    16.6710   6.771 1.60e-11 ***
West_North_Central 109.9851    18.1225   6.069 1.49e-09 ***
South_Atlantic     101.0985    16.1052   6.277 4.07e-10 ***
East_South_Central 186.1551    16.5329  11.260  < 2e-16 ***
West_South_Central 197.3497    16.3974  12.035  < 2e-16 ***
Mountain            54.6798    19.6193   2.787  0.00536 **
Pacific             43.3539    18.6492   2.325  0.02017 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105.5 on 2427 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.478
```

Figure A.9-Black LM

```
Call:
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_hispanic_or_latino + 
    Male + Middle_Atlantic + East_North_Central + West_North_Central + 
    South_Atlantic + East_South_Central + West_South_Central + 
    Mountain + Pacific, data = hispanic_combined)

Residuals:
    Min      1Q  Median      3Q     Max 
-271.89  -50.34   -9.48   42.81  691.57 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 115.0969    15.2163   7.564 7.14e-14 ***
Pov_rate_hispanic_or_latino   0.1079     0.2336   0.462 0.644115    
Male                         87.2973     4.4534  19.603  < 2e-16 ***
Middle_Atlantic              59.1359    15.6151   3.787 0.000159 ***
East_North_Central           50.3238    15.2901   3.291 0.001023 ** 
West_North_Central           12.2577    16.4736   0.744 0.456953    
South_Atlantic              -11.6861    15.0250  -0.778 0.436836    
East_South_Central          -13.0620    19.2798  -0.677 0.498205    
West_South_Central          116.6543    14.5668   8.008 2.47e-15 ***
Mountain                     73.0670    15.0238   4.863 1.29e-06 ***
Pacific                      42.2192    15.3780   2.745 0.006122 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.16 on 1366 degrees of freedom
Multiple R-squared:  0.3631,	Adjusted R-squared:  0.3585 
```

Figure A.10 Hispanic LM

```
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_asian + Male +
    Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain + Pacific, data = asian_combined)

Residuals:
    Min      1Q  Median      3Q     Max 
-144.25  -33.76   -6.80   23.34  495.20 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        111.1765    11.2308   9.899  < 2e-16 ***
Pov_rate_asian       0.2898     0.1837   1.578 0.115037    
Male                75.5669     4.3363  17.426  < 2e-16 ***
Middle_Atlantic     24.7799    12.2117   2.029 0.042797 *  
East_North_Central  13.3600    12.2388   1.092 0.275359    
West_North_Central  -2.5548    15.2614  -0.167 0.867098    
South_Atlantic      -0.5717    11.8269  -0.048 0.961462    
East_South_Central -19.2928    15.8423  -1.218 0.223687    
West_South_Central  34.8048    12.2627   2.838 0.004661 ** 
Mountain            24.0955    12.9833   1.856 0.063869 .  
Pacific             43.3228    11.9786   3.617 0.000319 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.1 on 738 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.3376,	Adjusted R-squared:  0.3286
```

Figure A.11-Asian LM

```
lm(formula = Deaths_Per_100.000_People ~ Pov_rate_indian_or_alaskan +
    Male + Middle_Atlantic + East_North_Central + West_North_Central +
    South_Atlantic + East_South_Central + West_South_Central +
    Mountain + Pacific, data = native_american_combined)

Residuals:
    Min      1Q  Median      3Q     Max 
-578.93 -108.31  -16.35   70.88 1047.77 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 178.805    182.196   0.981  0.32689    
Pov_rate_indian_or_alaskan    1.065      0.588   1.812  0.07064 .  
Male                        215.139     16.694  12.887  < 2e-16 ***
Middle_Atlantic               7.088    190.845   0.037  0.97039    
East_North_Central          178.057    185.457   0.960  0.33749    
West_North_Central          231.663    183.343   1.264  0.20700    
South_Atlantic              105.983    186.387   0.569  0.56988    
East_South_Central          526.330    188.670   2.790  0.00548 ** 
West_South_Central          139.506    183.005   0.762  0.44625    
Mountain                     70.083    183.151   0.383  0.70215    
Pacific                      51.552    182.707   0.282  0.77795    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 181.8 on 482 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4082,	Adjusted R-squared:  0.3959 
```

Native American LM

```
Call:
lm(formula = Deaths_Per_100.000_People ~ pov_rate, data = overall_combined)

Residuals:
    Min      1Q  Median      3Q     Max 
-306.41  -52.00   -7.11   46.06  695.64 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 305.7844     3.6577   83.60   <2e-16 ***
pov_rate      3.0375     0.1912   15.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.81 on 3150 degrees of freedom
Multiple R-squared:  0.07419,	Adjusted R-squared:  0.0739 
```

Figure A.12-Poverty only LM

B. Regional Grouping of States

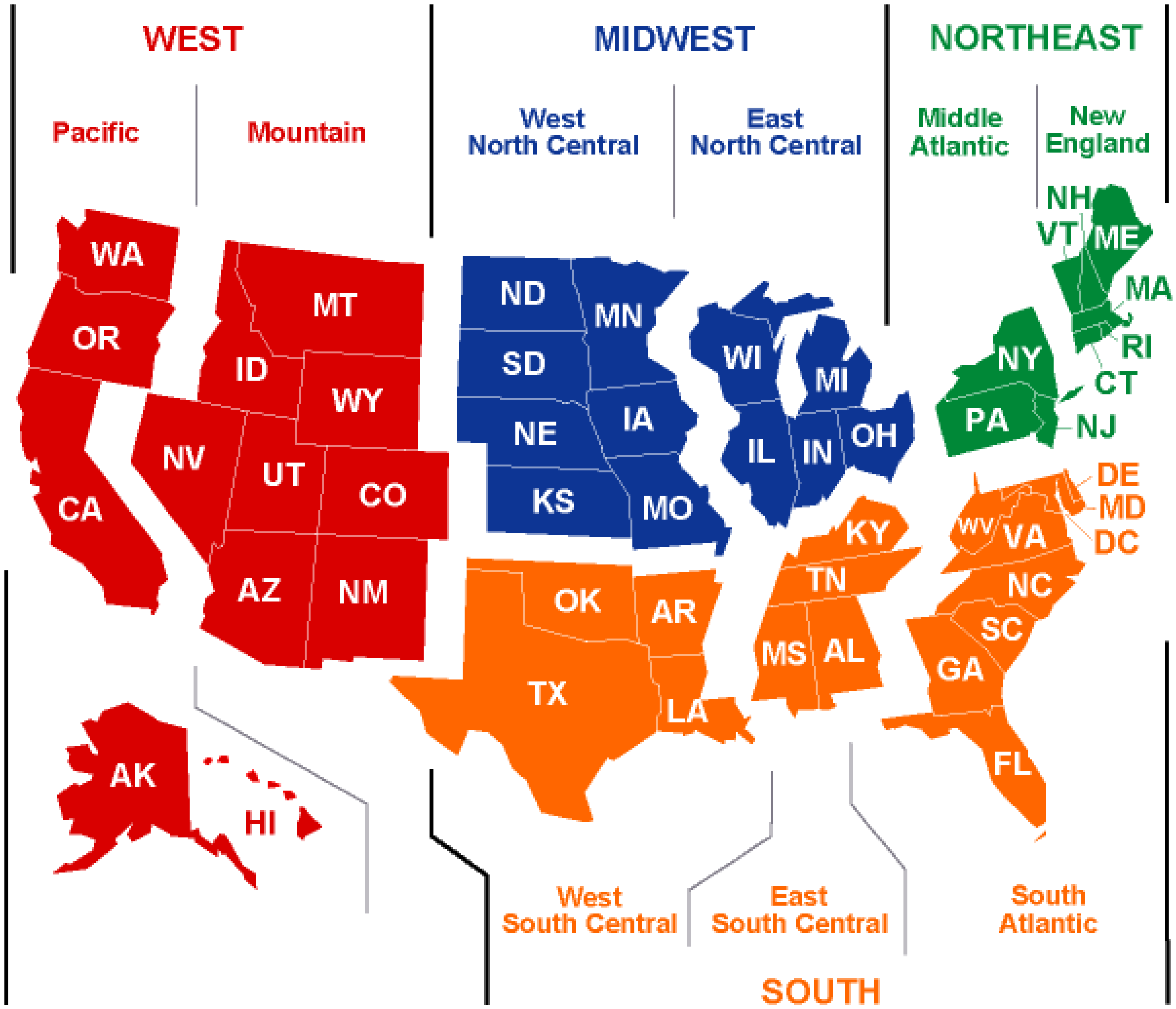

Figure B.1-Grouping of states into regions

Source: U.S. Energy Information Administration and CDC