FinalProjectReport

Chang Lu

2024-11-7

Abstract

This report delves into the analysis of the U.S. Chronic Disease Indicators (CDI) – 2023 Release, a comprehensive dataset on chronic disease metrics across the United States. The primary focus is on alcohol-related issues, particularly chronic liver disease mortality, exploring geographic, demographic, and temporal patterns through extensive exploratory data analysis (EDA), modeling, and visualization.

Introduction

For my final project, I have chosen the dataset titled "U.S. Chronic Disease Indicators (CDI) – 2023 Release." While the dataset's extensive nature presents challenges, it also offers rich opportunities for insightful analysis, which is promising for my development in data analysing skills.

The Data Source

I got access to the data("U.S. Chronic Disease Indicators (CDI) – 2023 Release.")(Link Text) in the website of data.gov(Link text). As the page described, this dataset is intended for public access and use.

EDA

data cleaning

Several columns in the dataset contain only NA values, so I excluded those columns from the analysis.

Upon examining the dataset, I noticed that the "DataValueAlt" column contains a value for each valid row. Therefore, I removed rows with NA in this column to proceed with the analysis.

Alcohol issue

The original dataset encompasses a variety of chronic disease topics. For this analysis, I focused on issues related to alcohol, creating a subset of the data by filtering rows where the "Topic" is "Alcohol."

The "Question" column outlines various alcohol-related issues. For my analysis, I focused specifically on "Chronic liver disease mortality" as the area of interest.

```
cldm_data <- alcohol_data %>%  filter(Question == "Chronic liver disease mortality") # cldm means chron

summary(cldm_data)
```

Understand Different Datatypes

The dataset includes three types of data: "Number," "Age-adjusted Rate," and "Crude Rate." I began by focusing on the "Number" data to draw initial conclusions and build models to predict mortality counts.

Number data analysis

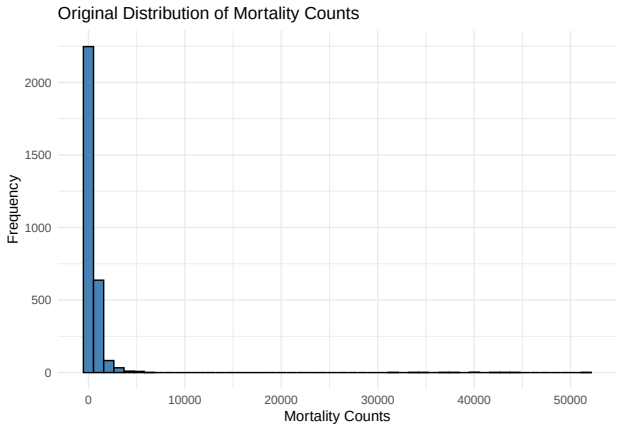


Figure 1: Origin Distribution of Mortality Counts

As is shown in Figure 1, the distribution is highly skewed. To address this, I applied a logarithmic transformation to the "DataValueAlt" column and plotted the transformed data to observe the new distribution.

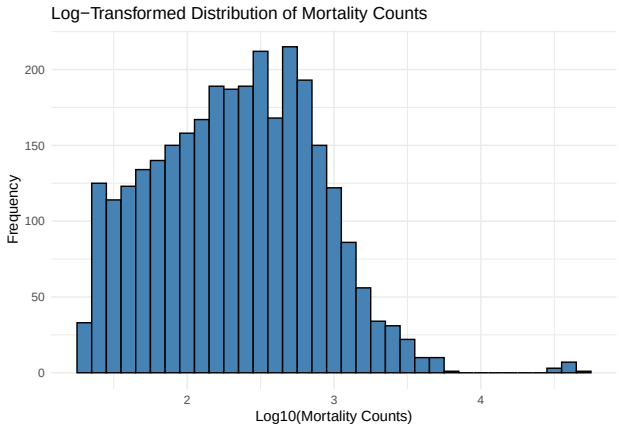


Figure 2: Log-Transformed Distribution of Mortality Counts

As is shown in figure 3, the log-transformed mortality counts exhibit a nearly normal distribution. This transformation makes it easier to analyze and model the data, particularly if working with highly skewed data in its raw form.

location Analysis

Firstly, I used "LocationDesc" and the transformed "LogDataValueAlt" to analyze variations in mean mortality counts across different states and regions in the USA.

Secondly, we can defer sevaral results from figure 4:

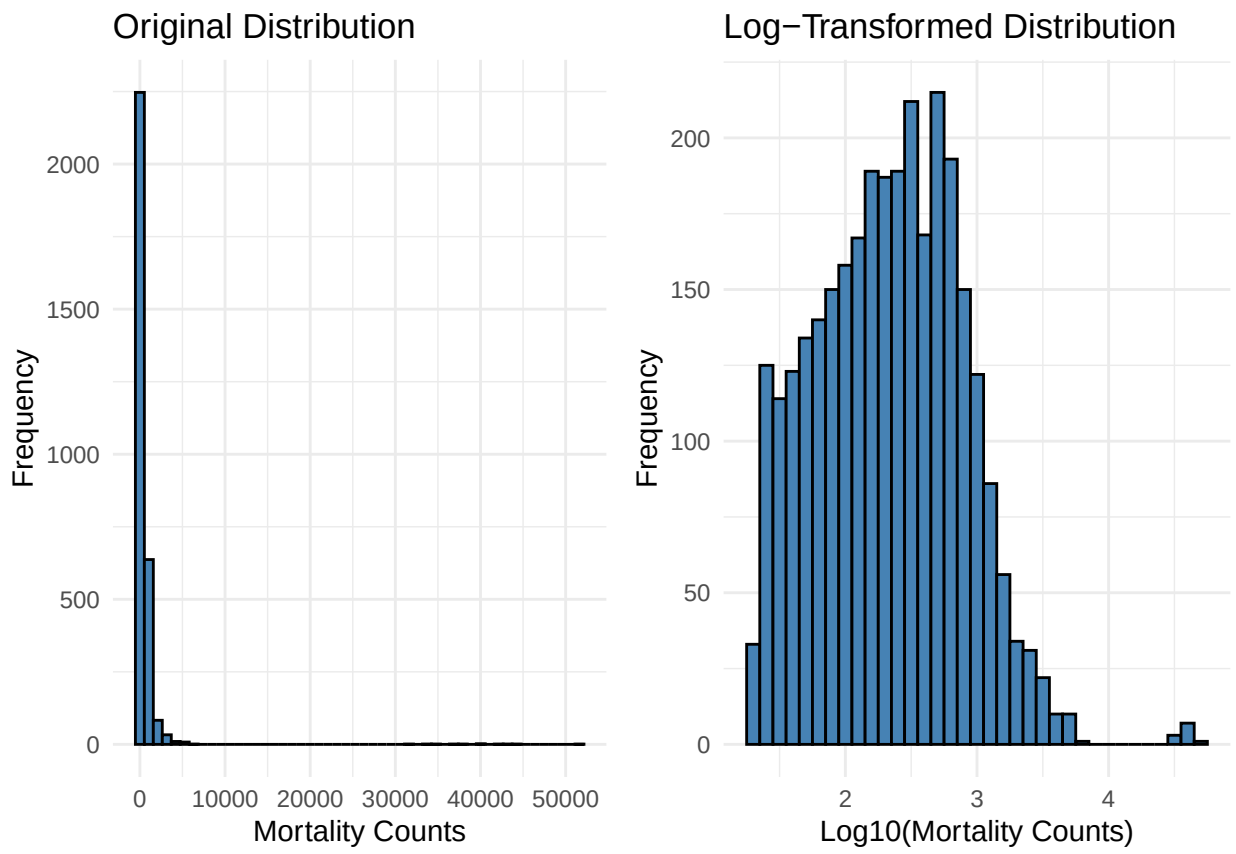


Figure 3: Camparison of the Two Plots

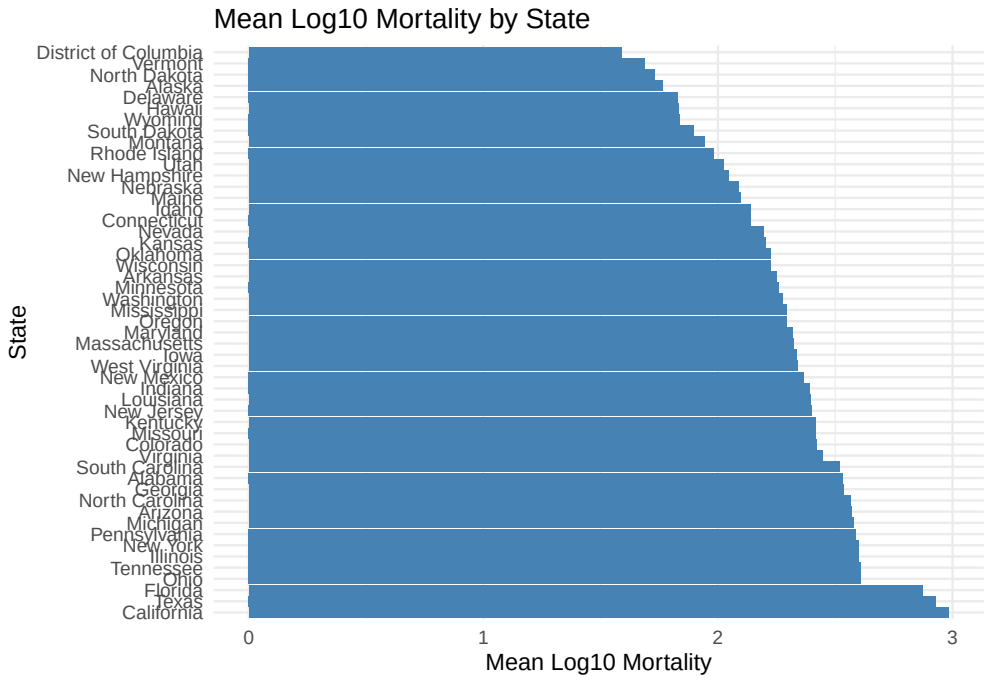


Figure 4: Mean Log10 Mortality Count Rank by State

1. States with High Mean Mortality (Log10 Scale):

States like California, Florida, and Texas have the highest mean log-transformed mortality values, as seen from their positions at the bottom of the chart with the largest bars. This suggests that these states, on average, experience higher mortality counts (in original scale) compared to others.

2. States with Low Mean Mortality (Log10 Scale):

States like District of Columbia, Vermont, and North Dakota have the lowest mean log-transformed mortality values, indicating relatively lower average mortality counts.

3. Possible Drivers of Variation:

Higher mortality counts in populous states like California and Texas might reflect their larger populations or other factors such as age distribution, healthcare access, or environmental factors and low mortality counts in smaller states like Vermont and North Dakota might be due to their smaller populations or other demographic characteristics.

Gender and Race Analysis

In the next step, I divided the "Stratification1" and "StratificationCategory1" columns to extract additional predictors, such as gender and race, for further analysis.

Gender	Mean Value	Max Value	Count
Female	284	2221	564
Male	498	3959	560

Race	Mean Value	Max Value	Count
American Indian or Alaska Native	67.9	263	132
Asian or Pacific Islander	66.0	345	75
Black, non-Hispanic	106	357	314
Hispanic	230	2489	273
White, non-Hispanic	577	2928	550

According to the figure 5, we can defer that:

1. Gender-Based Mortality Differences:

Males have a higher mean mortality rate compared to females. This suggests that mortality is disproportionately higher for males in the population under study.

2. Race/Ethnicity-Based Mortality Differences:

The highest mean mortality rate is observed for the White, non-Hispanic group, which significantly exceeds other racial/ethnic groups.

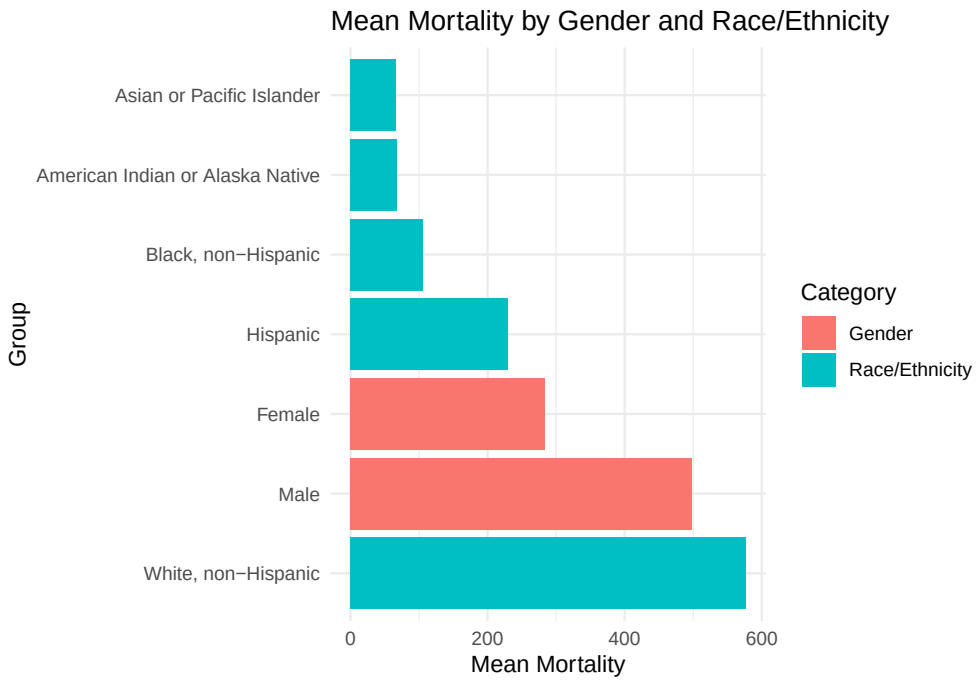


Figure 5: Mean Mortality by Gender and Race

Model Part

Data Preparation

Complete Pooling The complete pooling model assumes that all groups share the same underlying structure, with variations captured through location, demographic stratifications, and binary gender variables. Here is the detailed interpretation:

1. Overall Model Fit Residual Standard Error (RSE): 471.3 The typical deviation of the observed mortality counts from the predicted counts is approximately 471.3 units.

R^2=0.9641, Adjusted R^2=0.9634 which means the model explains 96.41% of the variance in mortality counts, indicating a strong fit. The adjusted R^2 suggests the model remains robust despite having many predictors.

2. Coeffcients

Each coefficient represents the expected change in DataValueAlt (mortality count) for a one-unit increase in the predictor, holding all other predictors constant.

2.1 Intercept (23.389):

The baseline mortality count for the reference group (LocationDesc = Alabama and Stratification1 = Asian or Pacific Islander) when gender is not considered.

2.2 LocationDesc (e.g., California = 1730.417):

The estimated increase or decrease in mortality count for each state compared to the reference state (Alabama). There are some significant States ($p < 0.05$): Examples include California (+1730.417), Texas (+1398.532), and New York(+454.623).

California shows a large positive impact, suggesting higher mortality counts.

2.3 Stratification1 (e.g., Female = 273.802):

The difference in mortality count based on demographic stratifications compared to the reference group (Asian or Pacific Islander): Female (+273.802): Being female is associated with an increase in mortality count compared to the reference group.

Male (+492.281): Being male is associated with a larger increase in mortality count compared to the reference.

White, non-Hispanic (+564.773): Shows significantly higher mortality compared to the reference demographic.

Partial Pooling The partial pooling model accounts for group-level variations in LocationDesc (locations) and Stratification1 (demographics), while estimating global fixed effects for IsFemale and IsMale. Here is the detailed explanation:

1. Overall Model Fit

Residual Standard Error (RSE): 471.3: The typical deviation of the observed mortality counts from the predicted counts is approximately 471.3 units.

R^2=0.9641, Adjusted R^2=0.9634: The model explains 96.41% of the variance in mortality counts, indicating a strong fit. The adjusted R^2 suggests the model remains robust despite having many predictors.

2. Coefficients

Each coefficient represents the expected change in DataValueAlt (mortality count) for a one-unit increase in the predictor, holding all other predictors constant.

2.1 Intercept (23.389):

The baseline mortality count for the reference group (LocationDesc = Alabama and Stratification1 = Asian or Pacific Islander)

2.2 LocationDesc

The estimated increase or decrease in mortality count for each state compared to the reference state (Alabama). Significant States ($p < 0.05$): Examples include California (+1730.417), Texas (+1398.532), and New York (+454.623).

2.3 Stratification

The difference in mortality count based on demographic stratifications compared to the reference group (Asian or Pacific Islander): Female (+273.802): Being female is associated with an increase in mortality count compared to the reference group.

Male (+492.281): Being male is associated with a larger increase in mortality count compared to the reference.

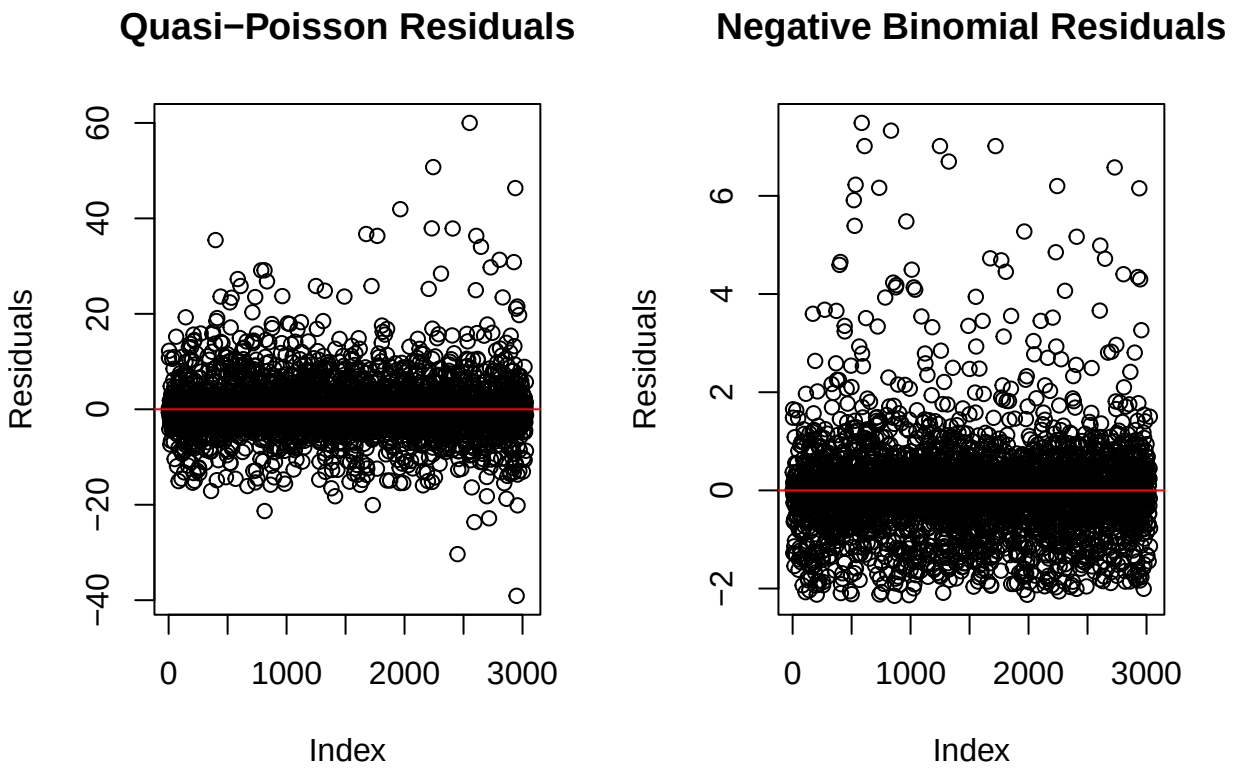
White, non-Hispanic (+564.773): Shows significantly higher mortality compared to the reference demographic.

No Pooling

Poisson Model A dispersion statistic of 38 is a clear indication of severe overdispersion in the Poisson regression model. This means that the variance of the dependent variable (DataValueAlt) is much larger than the mean, violating the Poisson assumption that variance equals the mean. So it is necessary for a change in model.

Model	Degrees of Freedom (df)	AIC
Poisson Model	59	134969.56
Negative Binomial Model	60	36471.35

The result indicates that using negative binomial model is much better than original poisson model.



According to the residuals plots, the negative binomial model is better than quasi-poisson model for it has fewer outliers and a more even distribution around 0 (negative binomial: -2~6, quasi-poisson: -40~60)

```
AIC(complete_pooling_model, partial_pooling_model, no_pooling_model, neg_binomial_model)
```

Comparison of the models

```
##                         df      AIC
## complete_pooling_model  60 45961.07
## partial_pooling_model    4 46394.58
## no_pooling_model       300 44426.70
## neg_binomial_model      60 36471.35
```

1. Interpretation

1.1 Complete Pooling (AIC = 45961.07):

The AIC is relatively high, suggesting this model has the weakest fit compared to the others. This is expected because the complete pooling model oversimplifies by assuming all groups have the same structure, ignoring group-level differences.

1.2 Partial Pooling (AIC = 46394.58):

The AIC is higher than the complete pooling model, indicating a slightly worse trade-off between fit and complexity. However, the partial pooling model may still be preferred in hierarchical data when overfitting is a concern or when interpretability and generalizability are priorities.

1.3 No Pooling (AIC = 44426.70):

This model has the lowest AIC, suggesting the best fit to the data among the three models. However, no pooling models can overfit because they do not share information between groups, and their complexity is reflected in the high degrees of freedom (300).

1.4 Negative Binomial Model (AIC = 36471.35)

This is the best model due to its smallest AIC and an appropriate degrees of freedom.

2. Model Selection

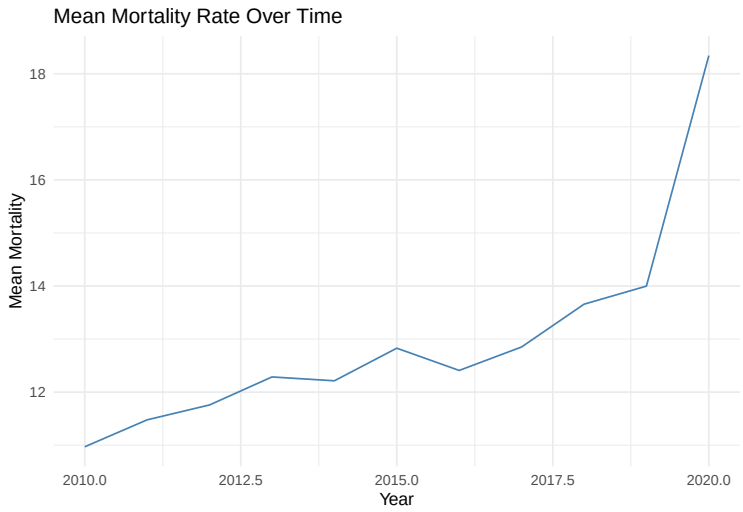
The Negative Binomial model is the best choice due to its:

2.1 Lowest AIC, indicating the best balance between fit and complexity.

2.2 Ability to explicitly model overdispersion, which is a key issue in the dataset.

Rate Data Analysis

Basic Visualization



According to figure 6, South Dakota has the highest mean mortality rate among the top 10 states(Over 30). New Mexico and Montana follow closely behind, indicating significant health-related challenges in these states.

According to figure 7, Virginia, Connecticut, and Vermont have the lowest mean mortality rates among all states in the dataset, suggesting relatively better health outcomes or lower incidence of the measured event.

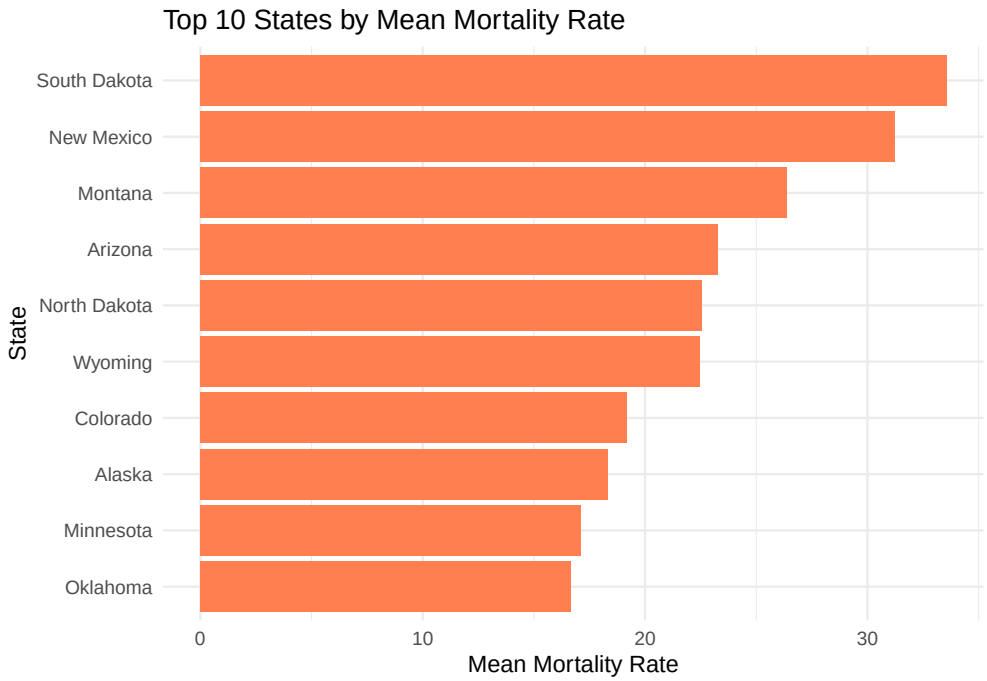


Figure 6: Top 10 States by Mean Mortality Rate

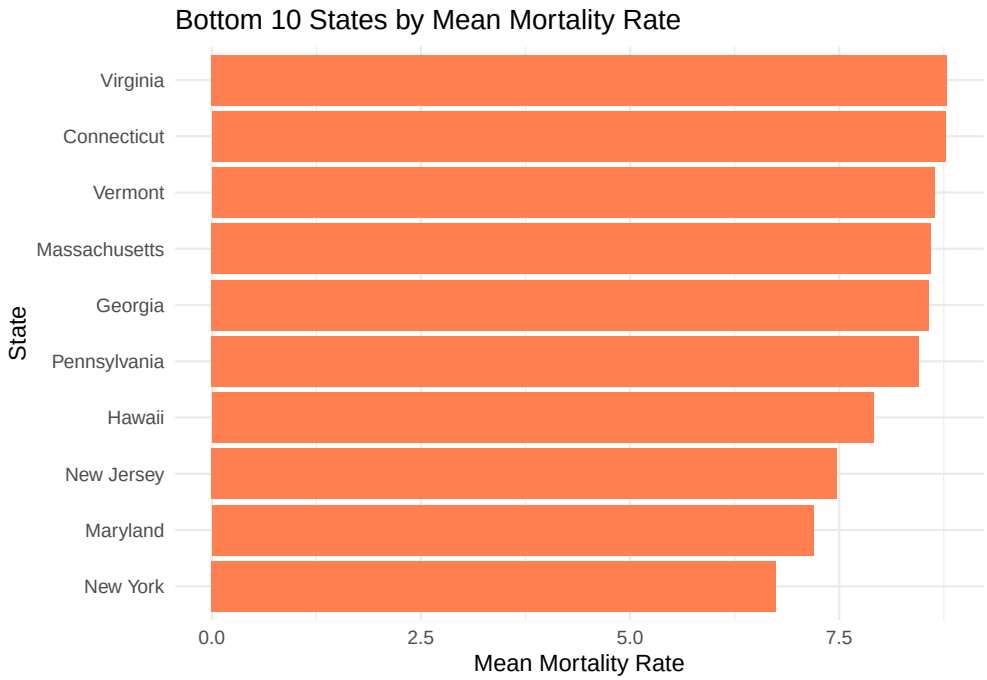


Figure 7: Bottom 10 States by Mean Mortality

Model Part

Data Preparation

Linear Model

Logistic Model Logistic regression is only applicable if the outcome have a binary version of the mortality rate (e.g., "high" vs. "low"). So I choose the median of the mortality rate as the threshhold.

Interpretation of the Confusion Matrix:

Prediction	Reference: 0	Reference: 1
Predicted 0	252	48
Predicted 1	47	259

The model achieves high accuracy (84.32%) with balanced sensitivity (84.28%) and specificity (84.36%).

High precision values for both classes indicate that the model makes reliable predictions for both class "0" and class "1."

Poisson Model The overdispersion coefficient is 0.8883045, indicating no need for change of poisson model.

Comparison of the Models

Model	Degrees of Freedom (df)	AIC
Complete Pooling Model	60	45961.07
Partial Pooling Model	4	46394.58
No Pooling Model	300	44426.70
Negative Binomial Model	60	36471.35

According to the table, the logistic model is the best of the three models.

Conclusion

Results Summary

The analysis revealed significant geographic and demographic disparities in chronic liver disease mortality across the United States. States such as South Dakota, **New Mexico, and Montana exhibited the highest mean mortality rates, while Vermont, the District of Columbia, and Hawaii consistently reported the lowest. Demographic analysis highlighted that males and White, non-Hispanic populations had notably higher mortality rates compared to other groups.

Models Summary

Modeling efforts demonstrated that negative binomial regression effectively addressed overdispersion in the data, with an AIC of 36,471.35, outperforming the Poisson model. Logistic regression achieved an accuracy of 84.32% in classifying high and low mortality rates, with balanced sensitivity and specificity. These findings underscore the importance of tailored interventions to address disparities, particularly in high-risk states and demographic groups, emphasizing the need for public health programs targeting alcohol-related behaviors and chronic disease prevention.

Validation

1. Geographic Disparities in Liver Disease Mortality

1.1 "Geographic Variability in Liver Disease-Related Mortality Rates in the United States"

"While chronic liver disease is the 12th leading cause of death in all Americans, it is the fourth leading cause of death in those 45-54 years of age and the sixth leading cause of death in Hispanic Americans. In the United States, liver disease mortality has been attributed to individual characteristics such as ethnicity, race, obesity, and alcohol consumption."

"Figure 1 and Table 1 show significant variability in age-adjusted liver disease mortality at a state level. Age-adjusted liver disease mortality ranges from 6.4 to 17.0 per 100,000. In the northeastern United States, rates of age-adjusted liver disease mortality are the lowest in the country. New Hampshire and New York have the lowest rates in the United States (6.4 and 6.6/100,000, respectively). In contradiction to this general assumption, West Virginia's rate is in the highest quartile, with a rate of 10.7/100,000. States in the west and central southwest carry some of the highest liver disease mortality rates, with New Mexico reporting the highest liver disease mortality, 17.0/100,000. The southern state of Georgia does not fit this assumption and falls into the lowest quartile, with a rate of 7.5/100,000. State size did not impact the variability in mortality rates."

This study examines interstate variability in liver disease mortality, highlighting significant geographic differences that may inform public health policies. (link Text)

2. Demographic Disparities in Alcohol-Related Mortality

2.1 "Racial and Ethnic Disparities in Alcohol-Attributed Deaths in the United States, 1999–2020"

"Between 1999 and 2020, a total of 605,948 individuals died from alcohol-related causes in the US. The highest AAMR was observed among American Indian/Alaska Natives, who were 3.6 times as likely (95% CI: 3.57, 3.67) to die from alcohol-related causes compared to Non-Hispanic Whites. Non-Hispanic Blacks, Asians/Pacific Islanders, and Hispanics showed lower AAMRs compared to Non-Hispanic Whites. These results were similar when stratified by sex, with American Indian/Alaska Native males being 3.2 times as likely (95% CI: 3.09, 3.21) and females being 4.8 times as likely (95% CI: 4.73, 4.96) to die from alcohol compared to Non-Hispanic Whites."

This study examines the burden and trends in alcohol-attributed mortality rates by race and ethnicity, revealing significant disparities among different groups. (Link Text)

3. Trends in Alcohol-Related Mortality

3.1 "Alcohol-Related Deaths in the U.S. More Than Double from 1999 to 2020"

"The 85 and older age group saw a possible but nonsignificant increase. Additionally, individuals aged 55-64 had both the steepest rise in mortality and the highest absolute rates in both 1999 and 2020. Both men and women experienced significant increases in alcohol-related deaths, but men had the highest rates in both

years and saw the steepest increase overall. Women, however, saw the largest proportional rise, with deaths increasing from 4.8 per 100,000 in 1999 to 12 in 2020."

"Deaths in women increased two-and-a-half times, while Asian and Pacific Islander communities experienced the steepest rise of 2.4 times. Regionally, the Midwest experienced the greatest jump, with an increase of 2.5 times in alcohol-related mortality, followed by the Northeast, West and South."

This study reports a significant increase in alcohol-related deaths over two decades, with notable rises among younger adults and women.(Link Text)