Title and Authors

Geographic Trends in Age-Adjusted Heart Disease Mortality in the United States

By Abigail Daniel, Jade Cruz, Yinan Zhu

PHB 221-Final Report

Abstract

Heart disease continues to be the most common cause of death in the United States, yet states do not experience its impact equally. Differences in heart disease mortality often mirror deeper social and structural conditions, including variations in medical care, environmental exposures, and population-level health behaviors. To explore these patterns, we examined state-level, age-adjusted mortality data from the National Center for Health Statistics' Leading Causes of Death dataset, focusing on 1999–2017 heart disease outcomes. After preparing the dataset and limiting it to deaths attributed to heart disease, we produced descriptive graphics and estimated a two-way ANOVA, a linear multiple regression model, spline adjustments to address nonlinear trends, and a random forest model to evaluate how well different methods captured state-specific trajectories. The regression model captured broad national patterns but often missed the curved declines seen in many states, whereas spline models and the random forest approach provided more accurate fits. All analyses pointed to clear and enduring regional patterns: mortality was consistently highest across the South and notably lower in the West and Northeast. These results reinforce that meaningful differences in heart disease outcomes persist across the country and that state-level context plays a substantial role in shaping risk.

Keywords: heart disease mortality, ANOVA, multiple regression model, spline adjustments, random forest model, age-adjusted mortality, state-level variation

---

Background

Despite national declines in heart disease mortality, large geographic disparities persist, with Southern states consistently exhibiting higher death rates compared with states in the West and Northeast. These patterns reflect substantial regional variation in established

heart disease risk factors, including obesity, hypertension, diabetes, smoking, and dietary habits, as well as deeper social and environmental conditions that shape risk across the life course (Diez Roux & Mair, 2010). Such conditions include neighborhood socioeconomic disadvantage, unequal access to preventive care, and differences in opportunities for health-promoting behaviors, all of which contribute to persistent spatial inequities.

Accurately characterizing these disparities requires measures that allow comparisons across populations with differing demographic structures. Age-adjusted mortality rates achieve this by standardizing outcomes to a reference population, enabling researchers to distinguish true differences in disease burden from differences due solely to age. Examining trends in age-adjusted heart disease mortality provides insight into how social and structural determinants influence cardiovascular outcomes over time. Although national declines in coronary disease mortality have been well documented, driven by advances in medical treatment, reductions in key risk factors, and changes in population health behaviors (Ford et al., 2007). The benefits of these improvements have not been distributed evenly across states or regions. Monitoring these geographic patterns remains essential for identifying where the burden is most concentrated, guiding targeted prevention efforts, and informing policies to reduce persistent inequities in cardiovascular health.

Significance of the Problem

Geographic variation in heart disease mortality points to far more than regional differences in numbers—it reflects long-standing inequities in the social and environmental conditions that shape heart disease health. States with persistently high mortality often face structural disadvantages, including limited access to preventive care, fewer opportunities for healthy behaviors, and socioeconomic constraints that elevate chronic disease risk. These upstream influences shape community health over time and help explain why certain regions, particularly in the South, continue to carry a disproportionate burden of heart disease.

Understanding whether these disparities change or remain stable is equally important. Because mortality data are collected repeatedly for each state over many years, analyses must account for both geographic context and temporal trends. Linear and multiple regression, together with interaction terms, allow us to examine whether states follow

similar or divergent patterns over time and whether the regional gaps are narrowing, widening, or largely unchanged. Detecting these patterns is critical for directing prevention efforts, informing resource allocation, and identifying where additional public health investment may be most necessary.

Ultimately, examining the geographic significance of heart disease mortality helps reveal where structural challenges persist and highlights the areas where targeted interventions could yield the greatest improvements in cardiovascular health.

Data

Analysis was conducted on data from the National Center for Health Statistics's Leading Causes of Death dataset, which compiles annual mortality for major causes of death across the United States geographic units. The raw dataset contains 10,868 observations spanning 1999-2017, with information for 52 geographic units: the 50 states, the District of Columbia, and a national "United States" total.

Each row in the dataset contains a combination of Year, State, and Cause of death. Causes are recorded in two related ways. The "Cause Name" variable provides a short name for one of the 11 broad causes of death categories: All causes, Heart Disease, Cancer, Stroke, Unintentional injuries, Chronic lower respiratory disease (CLRD), Diabetes, Alzheimer's disease, Influenza and pneumonia, Suicide, and Kidney disease. The "113 Cause Name" variable gives a more detailed description that corresponds with the ICD-10 groupings. The "Deaths" variable reports the count of deaths for each unique combination of state, year, and cause of death, and the "Age-adjusted Death Rate" variable gives the corresponding age-adjusted mortality rate per 100,000, standardized to the 2000 U.S. population to account for differences in age structure.

For this project, the focus was specifically on observations where the cause name is Heart Disease. The heart disease subset contains 988 rows, corresponding to years 1999-2017 for each of the 52 geographic units. Within this data subset, the number of deaths per year and state ranges from hundreds of deaths in smaller states to hundreds of thousands of deaths for the United States aggregate. The age-adjusted rates span from 115 to 350 deaths per 100,000.

Before analysis, preprocessing was performed to prepare the data for regression modeling. First, the United States aggregate was removed so that the analyses would focus on geographic variation across individual states rather than mixing state-level and national summaries. Then the key variables used in the analysis were checked for missing or clearly invalid values; none were identified, so no additional cleaning was required. Finally, a categorical regional variable was created based on the U.S. Census Bureau classification by grouping states into four regions, Northwest, Midwest, South, and West, to allow for comparisons of heart disease mortality rate across broader areas in addition to state-level results.

Goals

The goal of this study is to examine how age-adjusted heart disease mortality rate varies across states, regions, and time in the United States. To answer this question, our analysis was structured around three core objectives. First, we quantified the differences in mortality levels between states to assess any geographic disparities. Second, we summarized broader regional patterns to determine whether low or high morality states cluster geographically. Third, we evaluated whether past state-level trends can be used to predict future mortality. By understanding these patterns, we can identify where the burden of heart disease remains high, evaluate whether existing interventions are working, how progress has differed across the country, and better anticipate future trends.

Analysis

Our analytical approach focuses on assessing how statistical models can be used to predict future mortality trends and describe any temporal and geographic differences in the morality of heart disease. The analysis contains four main sections, which are data cleaning and preparation, exploratory analysis, statistical testing, and predictive modeling. These four sections allowed us to see broad national patterns while also seeing state level differences in both mortality levels and rates of improvements.

Exploratory Analysis

For our exploratory analysis, we wanted to get a simple overview of how heart disease mortality has changed over time in the U.S. Since the national average of the United States for each year was already given, we plotted its national average heart disease death rate for each year to get a general sense of how mortality has changed over time in the United States. Then, to see trends across the country, we grouped states into four regions (Northeast, Midwest, South, and West) and created regional trend plots. Finally, we visualized all 50 states individually so we could look more closely at how mortality levels change over time and how they vary from state to state.

Statistical Testing

To evaluate how heart disease mortality varies across both geography and time, we performed a two-way ANOVA with State and Year as the two factors, including their interaction. By doing this test, it allowed us to test three main questions at once, which are whether states differ in their overall average mortality, whether mortality changes over time, and whether states improve at the same or different rates.

In order to do a two-way ANOVA test, we state that our null hypothesis assumes that there are no differences in the mean age-adjusted mortality between states, no changes over time, and no interaction between state and year. The alternative hypothesis is that at least one of these factors plays a significant role in the variation in the data. After the two-way ANOVA test, we checked standard model assumptions using residuals vs fitted, Q-Q residuals, scale location, and residual vs leverage plots.

Regression Analysis

For our regression analysis, we compared several regression approaches to model how heart disease mortality changes over time and how those changes vary across states. We first tested with a simple linear where the parameter only included Year, this forced all states to follow the same national slope. For the multiple regression model it added State as a categorical predictor so states could have different baseline levels (intercept), but the rate of which it declined was still the same for all states. Our best case would be an interaction model with Year x State to allow each state to have its own trend over time.

Since our dataset contains multiple years for each state, we also attempted a mixed effects model with a random intercept and random slope for Year within State. Conceptually, this model would be the go to choice because it is designed for repeated measurements and would allow states to vary both in their baseline and slopes. However, we did run into a problem where the model did not converge. This meant that with just our predictor of Year and similar decline across states, the model did not have enough variation to estimate state specific random slopes. Due to the convergence issue, we did not use the mixed effect model as our main model for interface or prediction.

After looking at our exploratory plots and running the residual vs. fitted diagnostics, it made it clear that many state level trends weren't perfectly linear. To account for this model violation assumption, we first tested a quadratic model. The quadratic term helped a bit, but still didn't fully capture the bending patterns over time. Our best fix was a spline model using cubic splits for Year interacted with State. Splines allowed each state to have more flexible room to curve over time. The curved pattern in the residuals vs fitted largely disappeared under the spline modeling, meaning it provided a better fit than linear or quadratic. Overall, our main models are multiple fixed interaction and spline models, which later will be used to check its predictive performance.

Predictive Modeling

To assess how well our models are able to predict future years, we used an 80/20 train test split based on time, where data from 1999-2013 were used for the training set and 2014-2017 were used for the test set. Since we wanted to assess its ability to predict, it was important that the test set consist of later years rather than a random subset of the data. Preferably, we would want to validate the models on morality data beyond 2017, but since our dataset ends there, we planned on using the majority of the dataset to train. We planned on using the majority of the dataset to train while holding out the most recent years so we could show at least some performance sample, then later on in the future use a more updated dataset to test.

We focused on our top two best models, which are multiple regression models with State x Interaction and the spline model. The interaction model allowed each state to have its own trends over time, while the spline model allowed those trends to curve which is much better for capturing non linear changes. After fitting the models, we generated predictions for the test years and compared its accuracies using RMSE and MAE.

We additionally fit a random forest model as a flexible, nonparametric alternative to the regression-based prediction models. For this model, the outcome was the state-level age-adjusted heart disease mortality rate, and the predictors were Year and State. Using the same 80/20 split described above, the random forest model was trained on data from 1999-2013 and evaluated on data from 2014-2017, allowing us to compare its performance directly with the regression models. For our analysis, we grew a forest of 1,000 trees, and at each split, the algorithm randomly selected one of the two available predictors to split on. After fitting the forest, we generated predictions on the test data and measured their accuracy using the same metrics as for the regression models. Variable importance measures were also computed to assess the importance of Year and State.

Results and Interpretation

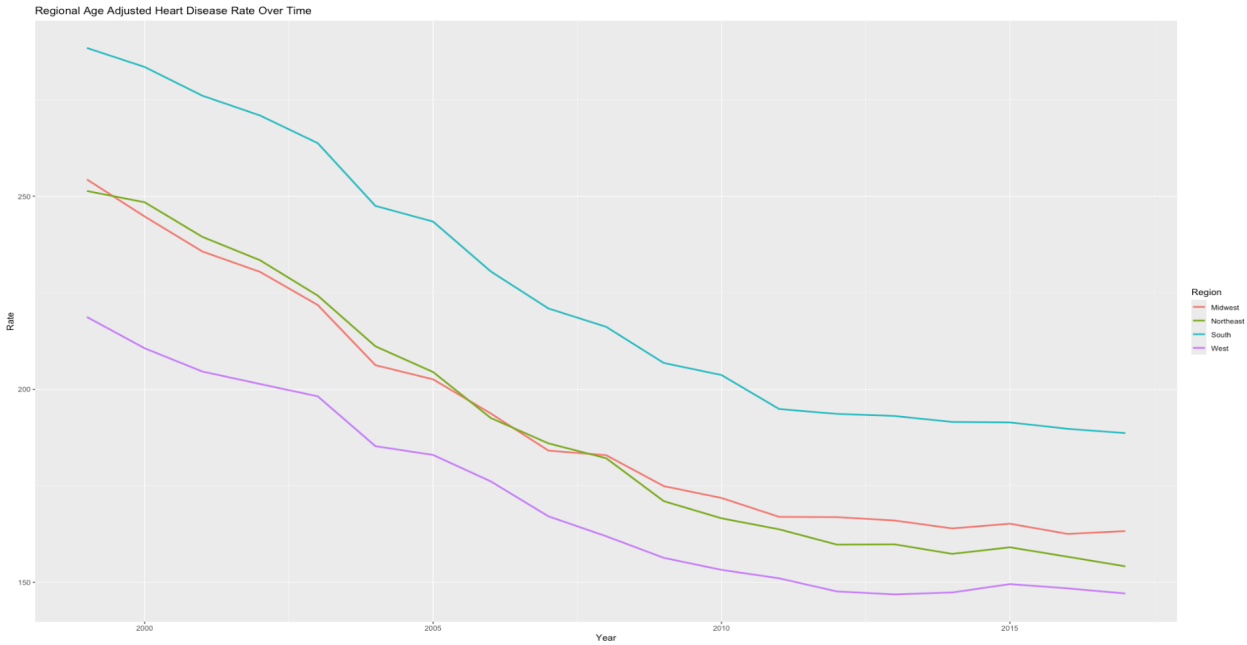


Fig 1. Age adjusted heart disease mortality trends for the four U.S. Census regions from 1999–2017. All of the regions show decline over time with the South consistently highest and West lowest.

When we break the national trend down to regions the differences become much clearer. As shown in Figure 1 all four regions (Northeast, Midwest, South, and West) from 1999-2017 show a steady decline in heart disease mortality, but the levels and pace of

decline are not the same. The South consistently has the highest mortality rates throughout the entire period. While we do see improvements in the South over time, the gap between the South and the rest of the regions never closes. This trend aligns well with already known regional patterns in cardiovascular risk factors such as higher rates of hypertension, smoking, obesity, socioeconomic resources, and access to healthcare. However, the West showed the lowest mortality rates every year. The Northeast and Midwest showed very similar declines, but still show a bit of different trajectories. The Midwest improves a bit more gradually and remains slightly higher by the end of the dataset period. Overall, the regional trends show that while the United States has made progress in reducing heart disease mortality, the improvements aren't evenly shared. The south and west gap show the health inequalities between regions and the need for region specific cardiovascular prevention.

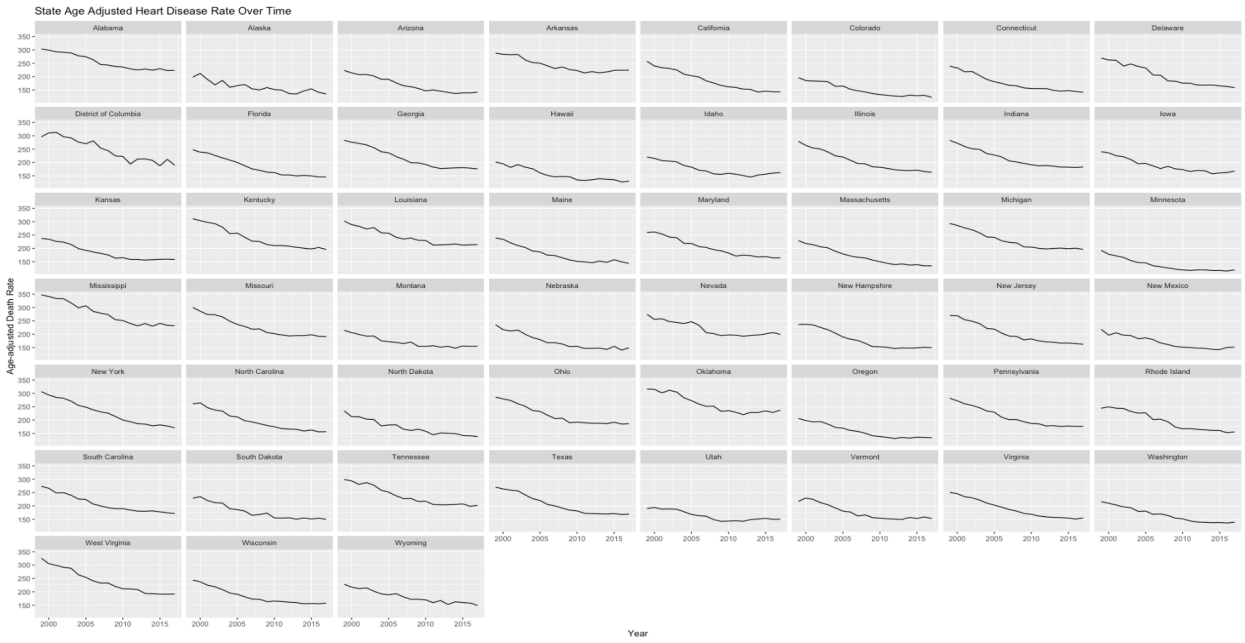


Fig 2. State level age adjusted heart disease mortality trends from 1999-2017. Each state showed an overall decline, but its pace varied in terms of slowdown and improvement.

When we look at each state level trend, we can see a similar non linear decline between them, but the overall levels and pace of improvement are different. Looking at Colorado, Hawaii, and Minnesota they stay at the lower end through the entire period. States such as Mississippi, Arkansas, Alabama, and Oklahoma still remain high. As stated before states improve over time, but the gap between the lowest and highest mortality states never close, which showcases geographic inequalities throughout the study period. Another important pattern that many of the states show is a plateau after around 2011. We

can see that some states flatten more than others, while some show slight increase towards the end. Even within the same region, some states do not follow the same exact trajectory meaning that the regional averages can hide a lot within region variation. Overall, this shows that location still plays a major role in heart disease or cardiovascular outcomes.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
State	50	957770	19155	175.520	<2e-16	***
Year	1	842449	842449	7719.314	<2e-16	***
State:Year	50	37132	743	6.805	<2e-16	***
Residuals	867	94620	109			

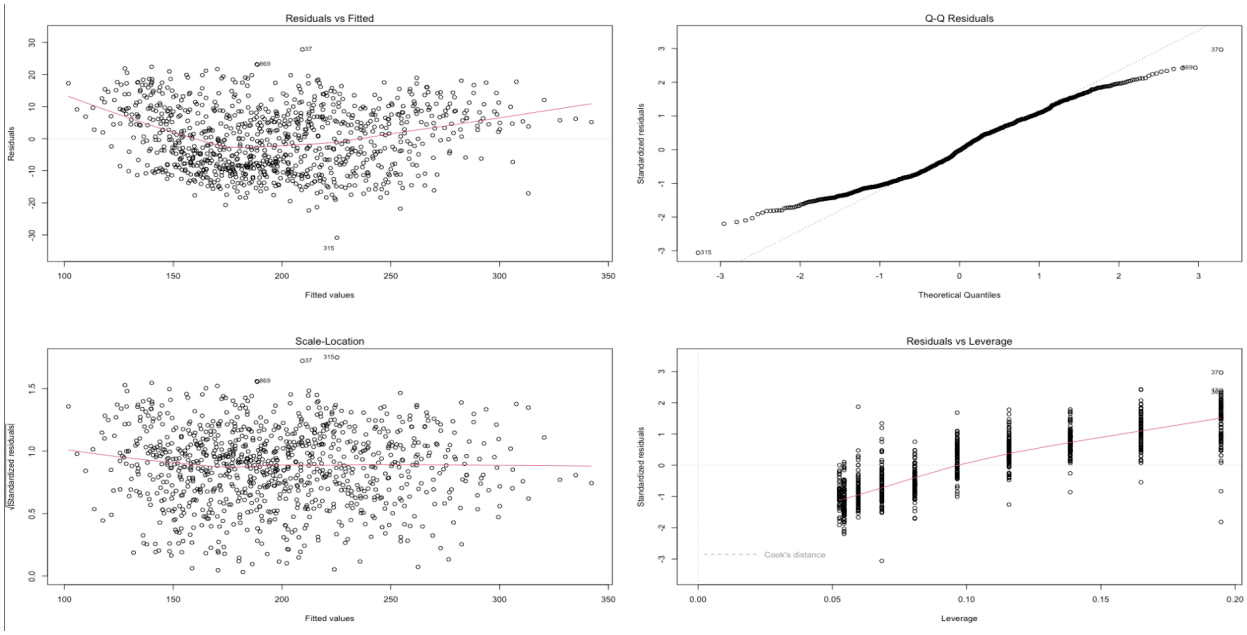


Fig 3. Two way ANOVA for State, Year, and its interaction, as well as its diagnostic check. The curved residual pattern means that a linear model does not capture the non linearity of the data.

The two way ANOVA results show that State, Year, and their interaction are significant predictors of age adjusted heart disease mortality ($p < 2e\text{-}16$). This confirms that mortality levels differ across states, mortality changes over time, and the amount of change is not the same for every state. What was most important out of this was the interaction because it shows that we cannot rely on a single national trend and states improve at different rates. While the model shows that it is statistically significant, the diagnostic plot in Figure 3 shows that there is a clear curved pattern in the residuals. It means that the linear model can't handle the non linear structure in how primary evolves over time. This basically means that the model can detect that states are different and the

rates are declining, but a strict linear form can't fully capture the shape of those declines. To improve the model we went for more flexible approaches, such as using spline regression and random first which allowed morality trends to bend over time where the linear ANOVA model could not.

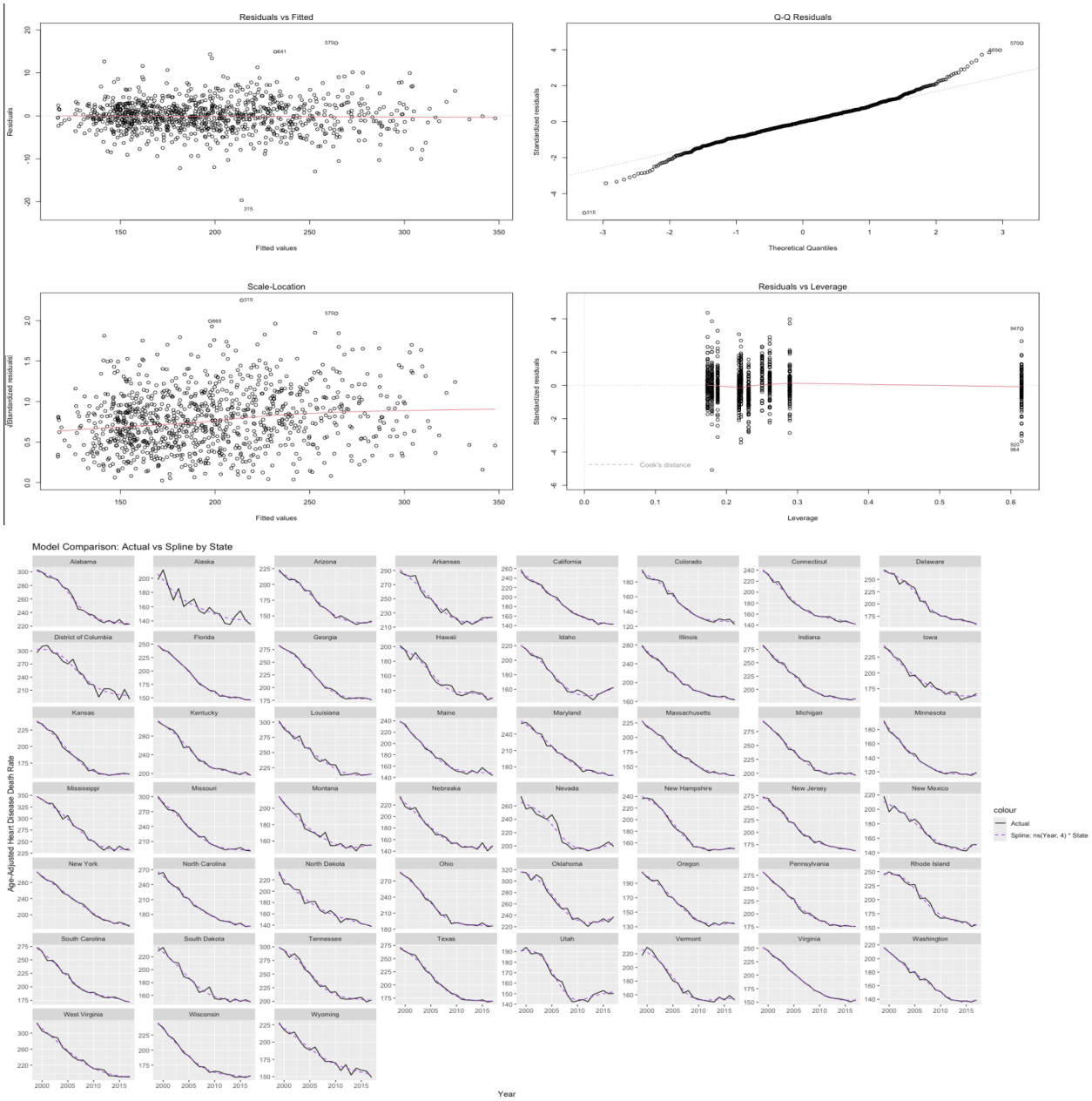
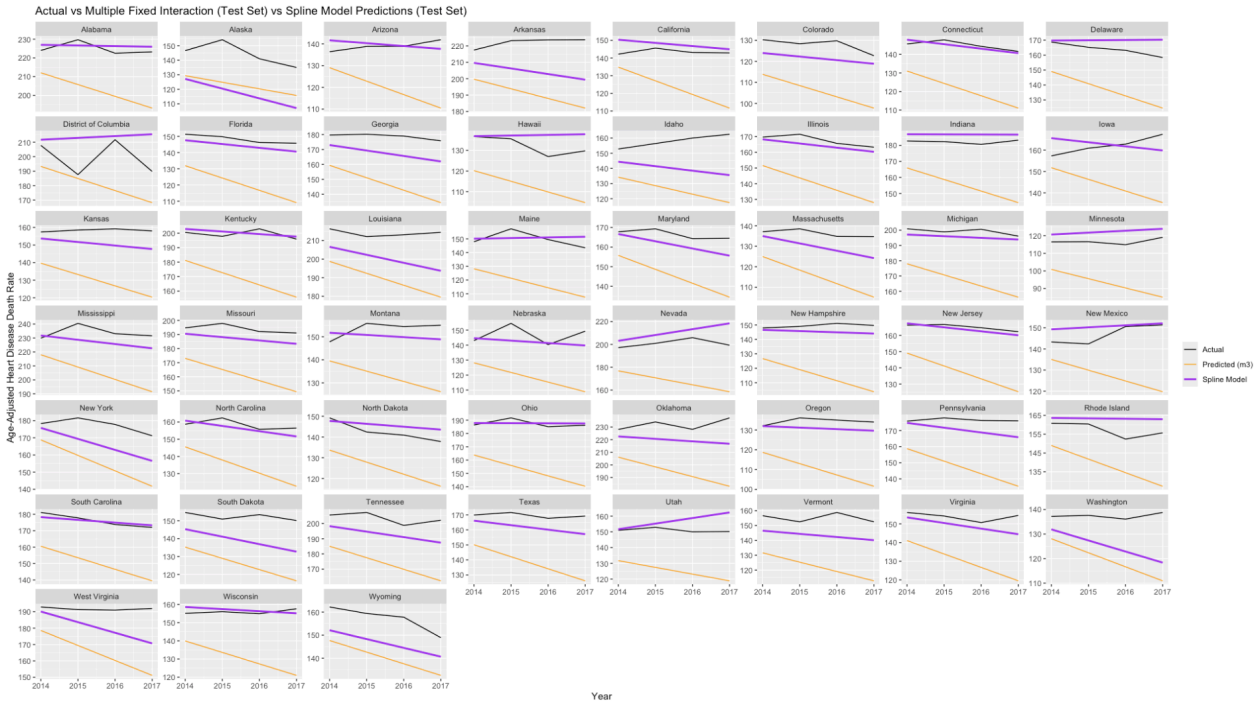


Fig 4. Diagnostic check for spline regression model. The disappearance of the curved pattern means that the spline successfully captures nonlinear decline in heart disease mortality over time and effectively models non linear trends.

Since the ANOVA, regression models, and diagnostic checks show a clear curvature in the residuals, we fit a spline regression model to better capture the non linear decline in

heart disease mortality over time. In figure 4, we can see that the residual vs. fitted curve disappears, which means that the spline successfully captured non linear structure that the linear models were not able to get. We used a natural cubic spline with four degrees of freedom which gave enough flexibility for any beginning trend over time. In terms of real world interpretation, this matters because the U.S. heart disease mortality does not improve at a constant linear rate. In a public health setting, its process usually accelerates during certain periods and something slows down during others, which really shows that a linear model can not reflect these changes. So by allowing each state's curve to bend naturally the spline provides a more realistic picture of how heart disease risk has changed over time and where some states may be stalling in terms of improvement.

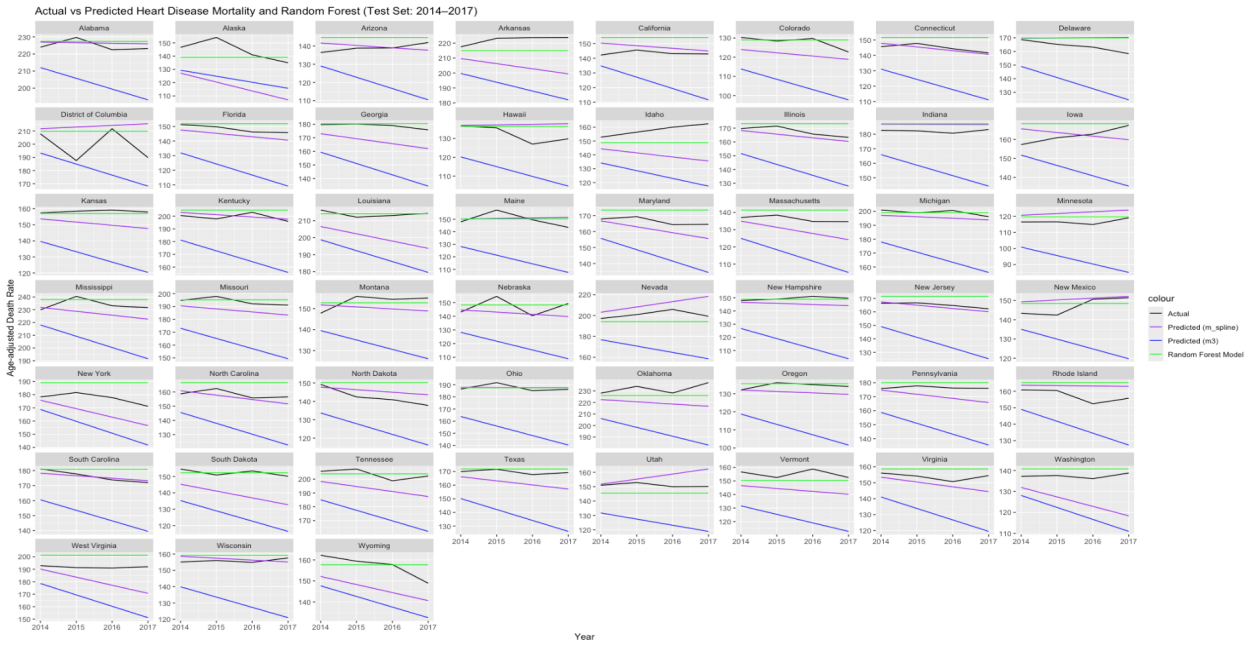


Model <chr>	RMSE <dbl>	MAE <dbl>
m3 (Year*State)	27.514214	25.926974
m_spline	9.525999	7.259065

Fig 5. Actual vs. predicted mortality rates for each state using a spline model. Spline model captures general direction and curvature of the trends, much better in modeling more irregular patterns across states.

When we evaluated the spline model on the test set (2014-2017), we saw that it captures the overall direction and curvature of the mortality trends for most states, but the fit is not perfect. Many states show noticeable gaps between the predicted and actual values, which

tells us that even though the spline model is more flexible than the linear models., it still struggles with states that have less consistent patterns. The spline model got an RMSE of about 9.5 and MAE of 7.3 deaths per 100,000, which is about 3-5% prediction error. This is reasonably small for a model only using Year and State as its predictors, but it is not the best compared to adding additional predictors. But within our dataset, where states follow similar trends, it shows that the spline captures broad non linear trends without fully nailing the in details of every state. If we used this same model structure to a global dataset while still only using Year and State/Country as predicted, the performance would most likely be much worse because internal morality varies a lot more than the differences we see in U.S states.



Model <chr>	RMSE <dbl>	MAE <dbl>
m3 (Year*State)	27.514214	25.926974
m_spline	9.525999	7.259065
RF (Year * State)	6.281032	5.000903

Fig 6. Actual vs Predicted Disease Mortality using random forest model. Captures the general direction much better than all the other models.

On the test set, the random forest model produced the lowest prediction error among all models (RMSE = 6.28, MAE = 5). In comparison, the spline regression models had slightly higher prediction errors, while the linear models had substantially larger ones. The panel plot of actual versus predicted heart disease mortality rates shows that the random forest model tracked the observed age-adjusted death rates more closely for most states, especially those that exhibited trends with noticeable curvature or diverged from a

simple linear pattern. The random forest model also explained approximately 95% of the variance in age-adjusted heart disease mortality in the test set R^2.

These results indicate that allowing for a flexible, nonlinear relationship between year and state improves the prediction of heart disease mortality rates. While the spline model already captured most of the nonlinear patterns, the random forest reduced both the RMSE and MAE the most and better matched the observed trends in states with more complex patterns. The high R^2 also suggests that year and state together explain a large share of the variation in the age-adjusted mortality rate on a state-level, although additional predictors are likely needed to better capture short-term variability.

Conclusion

This project examined how age-adjusted heart disease mortality varies across U.S. states, whether there are regional patterns, and whether past trends could be used to predict future state-level trends. Using data from 1999-2017, the analysis showed that heart disease has decreased over time, but persistent geographic differences remain. South states tended to have higher age-adjusted death rates than those in other regions, while states in the West had the lowest rates, showing a geographically concentrated problem.

Modeling helped deepen the understanding of these patterns. A two-way ANOVA and regression models showed that both Year and State, and their interactions, are essential for explaining differences in mortality rates. Simple models, with only a linear trend, could not capture how differently states changed over time. Diagnostic checks further motivated the use of spline models, which allowed the year effect to be nonlinear. This model matched the observed state patterns over time more efficiently.

Predictive modeling trained models on data from 1999-2013 and evaluated them on a test set of years from 2014-2017 using RMSE and MAE. The spline model outperformed linear models, but the random forest had the best predictive performance.

Overall, the project shows that heart disease mortality in the United States has decreased nationwide. However, geographic disparities are still prevalent, and simple information about time and location can be used to make reasonable short-term predictions of state-level heart disease mortality. Additional predictors, such as patient-level and socioeconomic factors, are likely needed to capture short-term within-state variation better

---

Discussion

This study explored geographic differences in heart disease mortality across the United States using age-adjusted death rates from the NCHS. Using simple and multiple linear regression, we examined how mortality levels varied across Census regions and assessed whether regions accounted for meaningful variation in outcomes. Across all models, we found clear and consistent disparities: Southern states had the highest age-adjusted mortality, while states in the West and Northeast showed much lower rates. These results highlight substantial geographic inequalities in cardiovascular health and demonstrate that where people live can strongly influence their overall risk.

Several limitations should be acknowledged when interpreting these findings. Because our analysis used state-level mortality data, we were not able to include individual-level risk factors such as smoking, obesity, hypertension, diet, physical activity, or treatment adherence. These variables play a major role in cardiovascular disease, and without them, we cannot assess their contribution to the differences observed across states. As a result, our models identify patterns but cannot make strong causal claims about the specific drivers of regional disparities.

Although age-adjusted rates help standardize states by accounting for a common age structure, they account for only one demographic dimension. States vary widely in racial and ethnic composition, socioeconomic status, and the rural–urban mix—all of which influence cardiovascular risk independently of age. These unmeasured factors may lead to residual confounding, even when age-standardized mortality rates are used.

Another limitation comes from relying on broad Census regions. While these categories help summarize large-scale patterns, they may hide important differences within regions, especially in the South, where socioeconomic and healthcare conditions vary

substantially. Such within-region variation reflects local policy environments and community-level contexts that our models cannot fully capture.

Despite these limitations, our regression approach provides a useful descriptive overview of geographic inequalities in heart disease mortality. The persistence of regional differences suggests that place-based factors remain an important influence on cardiovascular outcomes and highlights the need for public health strategies that address structural inequities and improve access to supportive environments nationwide.

Future Directions

Several ways of this work could strengthen the analysis and provide a deeper understanding of the drivers of geographic disparities in heart disease mortality. A central next step is to incorporate additional state-level covariates, such as poverty rates, educational attainment, health insurance coverage, and indicators of healthcare access, into the regression framework. Including these variables would help explain a larger share of the variation in mortality and offer clearer insight into the structural and policy-related factors that shape cardiovascular outcomes across states.

Another important direction is to expand the analysis to examine trends over time rather than relying solely on cross-sectional comparisons. Modeling changes across multiple years would help reveal whether states are improving at similar or diverging rates, and whether regional gaps in mortality are narrowing or widening. In parallel, applying spatial analytic techniques could capture geographic clustering that extends beyond state borders and may reflect shared environmental or socioeconomic conditions.

Finally, future studies would benefit from analyses that move beyond state-level averages to examine variation within population subgroups. Stratifying mortality outcomes by race and ethnicity, sex, or age group, where data permit could uncover disparities that are obscured at more aggregated levels. Such subgroup analyses would provide a more detailed account of who is most affected by cardiovascular mortality and offer a stronger foundation for equity-focused public health initiatives.

Contributions

For this project, Yinan made contributions to the conceptual and written components of the report. She drafted the Abstract, Background, Significance of the Problem, and Discussion and Future Directions, which together established the public health relevance of geographic disparities in heart disease mortality and articulated the rationale for the analytical approach. In the Abstract, she synthesized the study's aims, methods, and principal findings into a clear and concise summary. Her work on the Discussion and Future Directions offered a critical interpretation of the mixed regression model results, highlighted key limitations, and proposed meaningful directions for subsequent research. Collectively, these contributions shaped the coherence of the report and strengthened its broader implications for cardiovascular population health.

For this project, Jade made contributions to the computational and analytical components of the report, which included most of the coding pipeline from data cleaning, visualization, and statistical modeling. Contributions included exploratory analysis that showcased state, regional, and national patterns in heart disease mortality, statistical methods which included ANOVA, linear regression, multiple linear regression, random intercept mixed regression model, and predictive modeling using training/test splits. Model performance was evaluated using RMSE, MAE, and fitted vs. actual plotted comparisons. Written components include goals, analysis, results and interpretation sections, showcasing that relying only on Year and State is not enough to provide an accurate predictive model, specifically for short term variation, emphasizing the need for additional predictors such as patient level data and socioeconomic data to improve future model accuracies.

For this project, Abigail made contributions to the descriptive, diagnostic, and advanced modeling components of the report. Her work included developing the data description, summarizing key variables from the national causes of death dataset, and explaining how variables were defined and prepared for analysis. She implemented and evaluated the random forest model as a flexible alternative to the regression models, including tuning model parameters, computing variable importance measures, and comparing the predictive performance with spline models. She also conducted diagnostic tests for the regression model to evaluate its adequacy and guide interpretation. Written components also included analysis, results, and interpretation sections as well as the conclusion, which integrated results across models.

References

Diez Roux, A. V., & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences, 1186*(1), 125–145. https://doi.org/10.1111/j.1749-6632.2009.05333.x

Ford, E. S., Ajani, U. A., Croft, J. B., Critchley, J. A., Labarthe, D. R., Kottke, T., Giles, W. H., & Capewell, S. (2007). Explaining the decrease in U.S. deaths from coronary disease, 1980–2000. *New England Journal of Medicine, 356*(23), 2388–2398. https://doi.org/10.1056/NEJMsa053935

National Center for Health Statistics. (2024). *NCHS Leading Causes of Death, United States*. U.S. Department of Health and Human Services. https://catalog.data.gov/dataset/nchs-leading-causes-of-death-united-states

PHB_221_Final_Project

Abigail, Jade, Yinan

2025-11-13

Dataset and Libraries

```
# libraries
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.3.3
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 4.3.3
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.3
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(alr4)
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
##
## Loading required package: effects
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(splines)

set.seed(54)

# dataset
data <- read_csv("NCHS_-_Leading_Causes_of_Death__United_States.csv")
```

```
## Rows: 10868 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): 113 Cause Name, Cause Name, State
## dbl (3): Year, Deaths, Age-adjusted Death Rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Collection and Cleaning

```
# view first few rows of data
head(data)
```

```
## # A tibble: 6 x 6
##    Year `113 Cause Name`        `Cause Name` State Deaths Age-adjusted Death R~1
##   <dbl> <chr>                  <chr>        <chr>  <dbl>                  <dbl>
```

```
## 1  2017 Accidents (unintention~ Unintention~ Unit~ 169936                  49.4
## 2  2017 Accidents (unintention~ Unintention~ Alab~   2703                  53.8
## 3  2017 Accidents (unintention~ Unintention~ Alas~    436                  63.7
## 4  2017 Accidents (unintention~ Unintention~ Ariz~   4184                  56.2
## 5  2017 Accidents (unintention~ Unintention~ Arka~   1625                  51.8
## 6  2017 Accidents (unintention~ Unintention~ Cali~  13840                  33.2
## # i abbreviated name: 1: `Age-adjusted Death Rate`
```

```
# check the number of rows and cols
dim(data)
```

```
## [1] 10868     6
```

```
# filter data
filtered_data <- data[data$`Cause Name` == "Heart disease", ]

head(filtered_data)
```

```
## # A tibble: 6 x 6
##    Year `113 Cause Name`        `Cause Name` State Deaths Age-adjusted Death R~1
##   <dbl> <chr>                   <chr>        <chr>  <dbl>                   <dbl>
## 1  2017 Diseases of heart (I00~ Heart disea~ Unit~ 647457                   165
## 2  2017 Diseases of heart (I00~ Heart disea~ Alab~  13110                   223.
## 3  2017 Diseases of heart (I00~ Heart disea~ Alas~    814                   135
## 4  2017 Diseases of heart (I00~ Heart disea~ Ariz~  12398                   142.
## 5  2017 Diseases of heart (I00~ Heart disea~ Arka~   8270                   224.
## 6  2017 Diseases of heart (I00~ Heart disea~ Cali~  62797                   143.
## # i abbreviated name: 1: `Age-adjusted Death Rate`
```

```
dim(filtered_data)
```

```
## [1] 988   6
```

```
# missing data
colSums(is.na(filtered_data))
```

```
##                    Year          113 Cause Name              Cause Name
##                       0                       0                       0
##                   State                  Deaths Age-adjusted Death Rate
##                       0                       0                       0
```

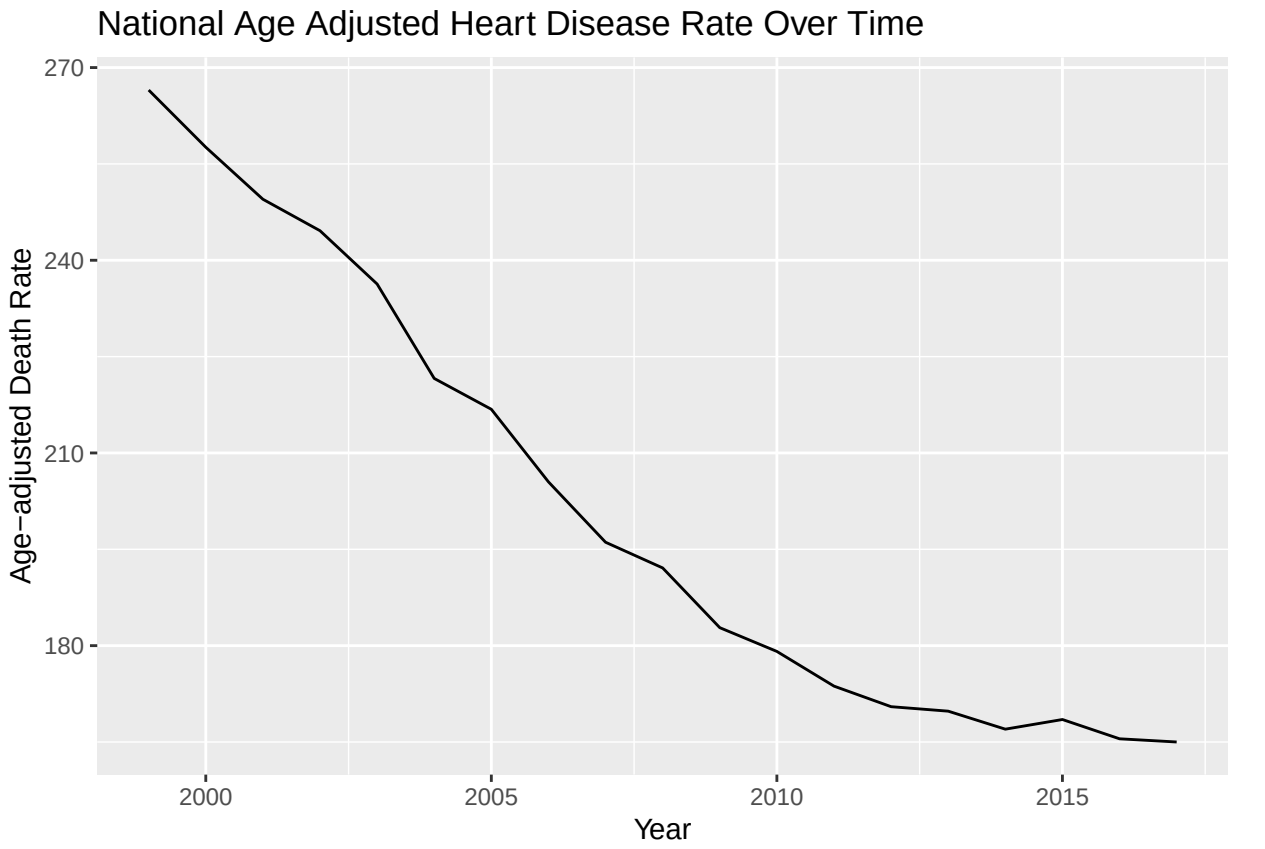
Explantory Analysis

```
# national level
national_data <- data |>
  filter(`Cause Name` == "Heart disease", State == "United States")

head(national_data)
```

```
## # A tibble: 6 x 6
##    Year `113 Cause Name`        `Cause Name` State Deaths Age-adjusted Death R~1
##   <dbl> <chr>                   <chr>        <chr>  <dbl>                   <dbl>
## 1  2017 Diseases of heart (I00~ Heart disea~ Unit~ 647457                   165
## 2  2016 Diseases of heart (I00~ Heart disea~ Unit~ 635260                   166.
## 3  2015 Diseases of heart (I00~ Heart disea~ Unit~ 633842                   168.
## 4  2014 Diseases of heart (I00~ Heart disea~ Unit~ 614348                   167
## 5  2013 Diseases of heart (I00~ Heart disea~ Unit~ 611105                   170.
```

```
## 6  2012 Diseases of heart (I00~ Heart disea~ Unit~ 599711                    170.
## # i abbreviated name: 1: `Age-adjusted Death Rate`
```

```r
ggplot(data = national_data, mapping = aes(x = Year, y = `Age-adjusted Death Rate`)) +
  geom_line() +
  labs(title = "National Age Adjusted Heart Disease Rate Over Time")
```

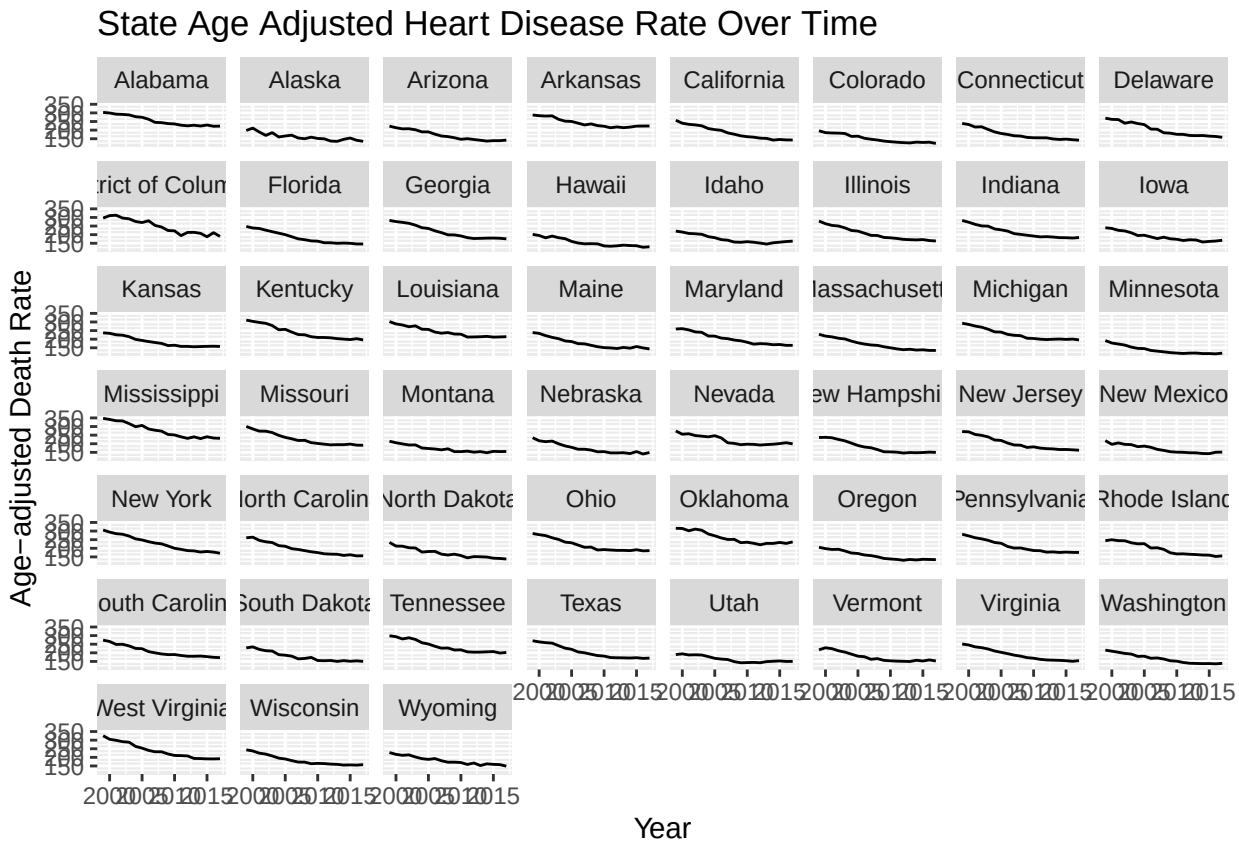



```r
# state level

state_only <- filtered_data |>
  filter(State != "United States") |>
  mutate(
    Region = case_when(
      State %in% c("Maine", "New Hampshire", "Vermont", "Massachusetts",
                   "Rhode Island", "Connecticut", "New York", "New Jersey",
                   "Pennsylvania") ~ "Northeast",
      State %in% c("Ohio", "Indiana", "Illinois", "Michigan", "Wisconsin",
                   "Minnesota", "Iowa", "Missouri", "North Dakota", "South Dakota",
                   "Nebraska", "Kansas") ~ "Midwest",
      State %in% c("Delaware", "Maryland", "District of Columbia", "Virginia",
                   "West Virginia", "North Carolina", "South Carolina", "Georgia",
                   "Florida", "Kentucky", "Tennessee", "Alabama", "Mississippi",
                   "Arkansas", "Louisiana", "Oklahoma", "Texas") ~ "South",
      State %in% c("Montana", "Idaho", "Wyoming", "Colorado", "New Mexico",
                   "Arizona", "Utah", "Nevada", "Washington", "Oregon", "California",
                   "Alaska", "Hawaii") ~ "West"
    )
  )
```

```
head(state_only)
```

```
## # A tibble: 6 x 7
##    Year `113 Cause Name` `Cause Name` State Deaths Age-adjusted Death R~1 Region
##   <dbl> <chr>            <chr>        <chr>  <dbl>                   <dbl> <chr>
## 1  2017 Diseases of hea~ Heart disea~ Alab~  13110                    223. South
## 2  2017 Diseases of hea~ Heart disea~ Alas~    814                    135  West
## 3  2017 Diseases of hea~ Heart disea~ Ariz~  12398                    142. West
## 4  2017 Diseases of hea~ Heart disea~ Arka~   8270                    224. South
## 5  2017 Diseases of hea~ Heart disea~ Cali~  62797                    143. West
## 6  2017 Diseases of hea~ Heart disea~ Colo~   7060                    123. West
## # i abbreviated name: 1: `Age-adjusted Death Rate`
```

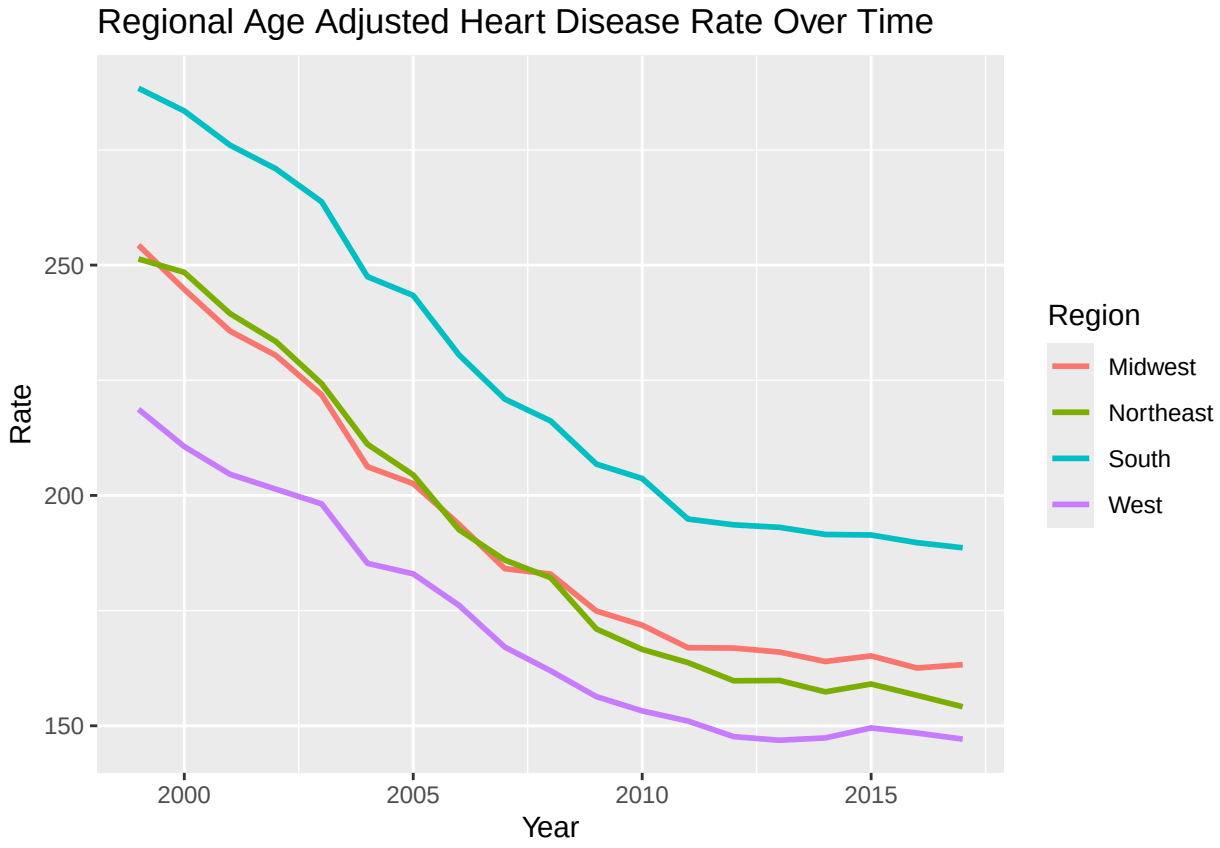
```
ggplot(data = state_only, mapping = aes(x = Year, y = `Age-adjusted Death Rate`)) +
  geom_line() +
  facet_wrap(~ State) +
  labs(title = "State Age Adjusted Heart Disease Rate Over Time")
```



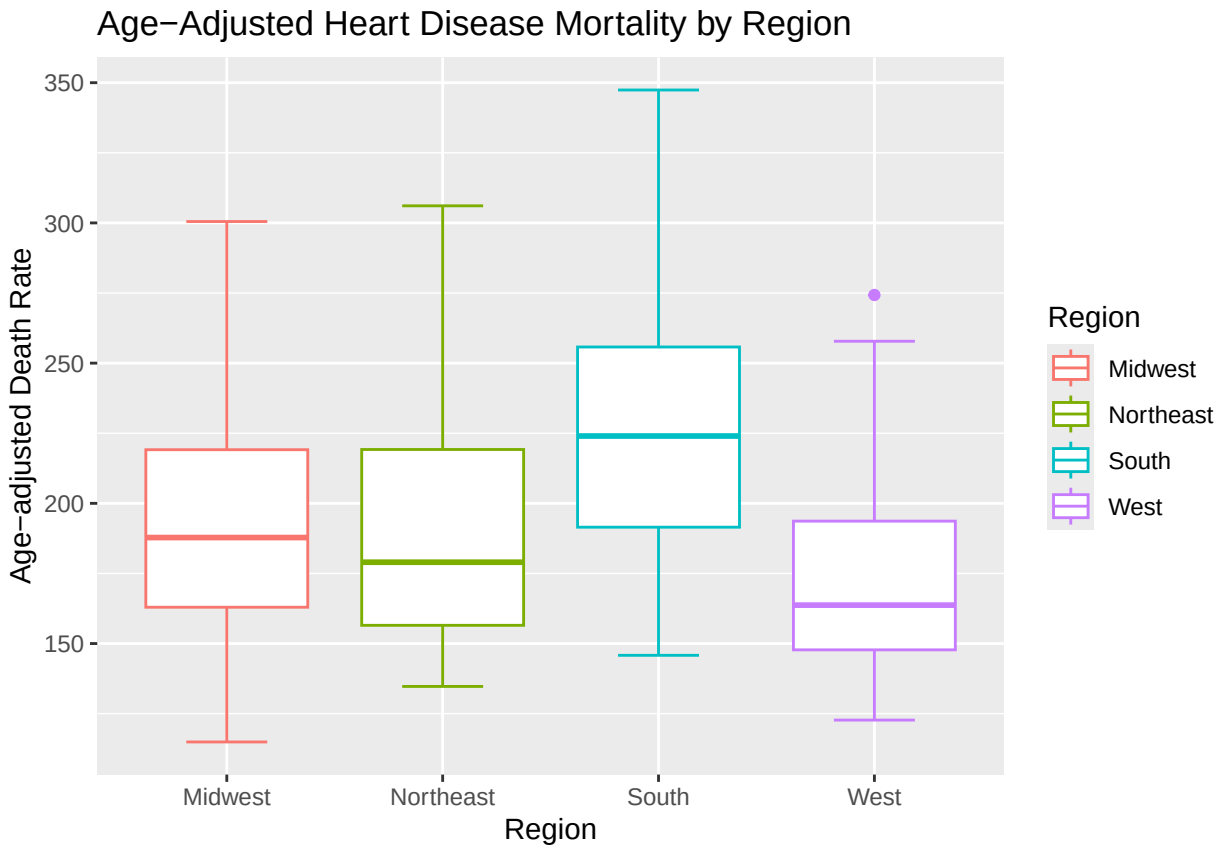

```
# regional level
regional_summary <- state_only |>
  group_by(Region, Year) |>
  summarise(
    Rate = mean(`Age-adjusted Death Rate`)
  )
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
ggplot(data = regional_summary, mapping = aes(x = Year, y = Rate, color = Region)) +
  geom_line(linewidth = 1) +
  labs(title = "Regional Age Adjusted Heart Disease Rate Over Time")
```




```
ggplot(data = state_only, mapping = aes(x = Region, y = `Age-adjusted Death Rate`, color = Region)) +
  geom_boxplot(staplewidth = 0.5) +
  labs(title = "Age-Adjusted Heart Disease Mortality by Region")
```

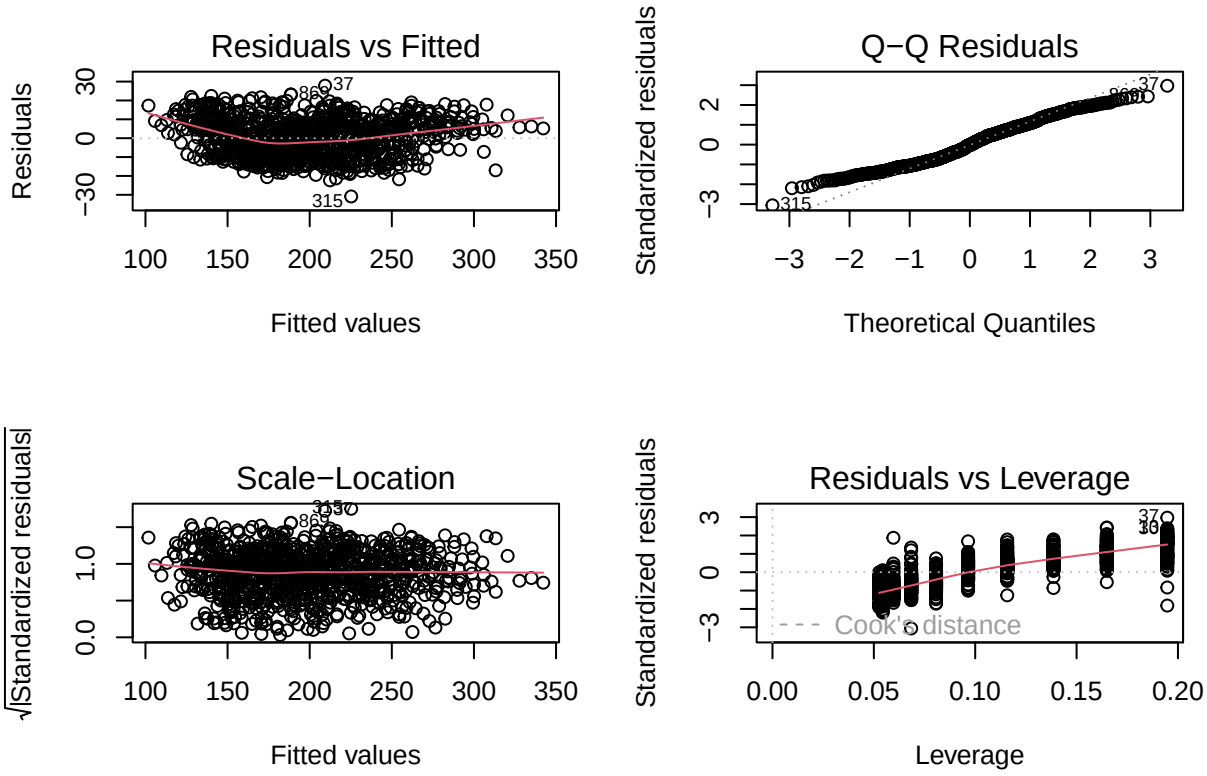


Statistical Testing

```
# two way anova testing
# * (interaction term) bc each state has their own slope, if + was used then slope for Year is the same
two_way <- aov(`Age-adjusted Death Rate` ~ State * Year, data = state_only)
summary(two_way)
```

```
##              Df Sum Sq Mean Sq  F value Pr(>F)
## State        50 957770   19155  175.520 <2e-16 ***
## Year          1 842449  842449 7719.314 <2e-16 ***
## State:Year   50  37132     743    6.805 <2e-16 ***
## Residuals   867  94620     109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# check assumptions: homoscedasticity, normality of residuals, linearity, and influential pts
par(mfrow = c(2,2))
plot(two_way)
```




```
D <- cooks.distance(two_way)
sum(D > 1)
```

```
## [1] 0
```

Regression Analysis

```
# simple linear regression
m1 <- lm(`Age-adjusted Death Rate` ~ Year, data = state_only)
#summary(m1)


# multiple regression +
m2 <- m_additive <- lm(`Age-adjusted Death Rate` ~ Year + State, data = state_only)
#summary(m2)

# multiple regression *
m3 <- lm(`Age-adjusted Death Rate` ~ Year * State, data = state_only)
#summary(m3)

# mixed model does not work as slopes between states are too simila, receives convergence error

#mixed1 <- lmer(`Age-adjusted Death Rate` ~ Year + (1 | State), data = state_only)
#summary(mixed1)

mixed2 <- lmer(`Age-adjusted Death Rate` ~ Year + (Year | State), data = state_only)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

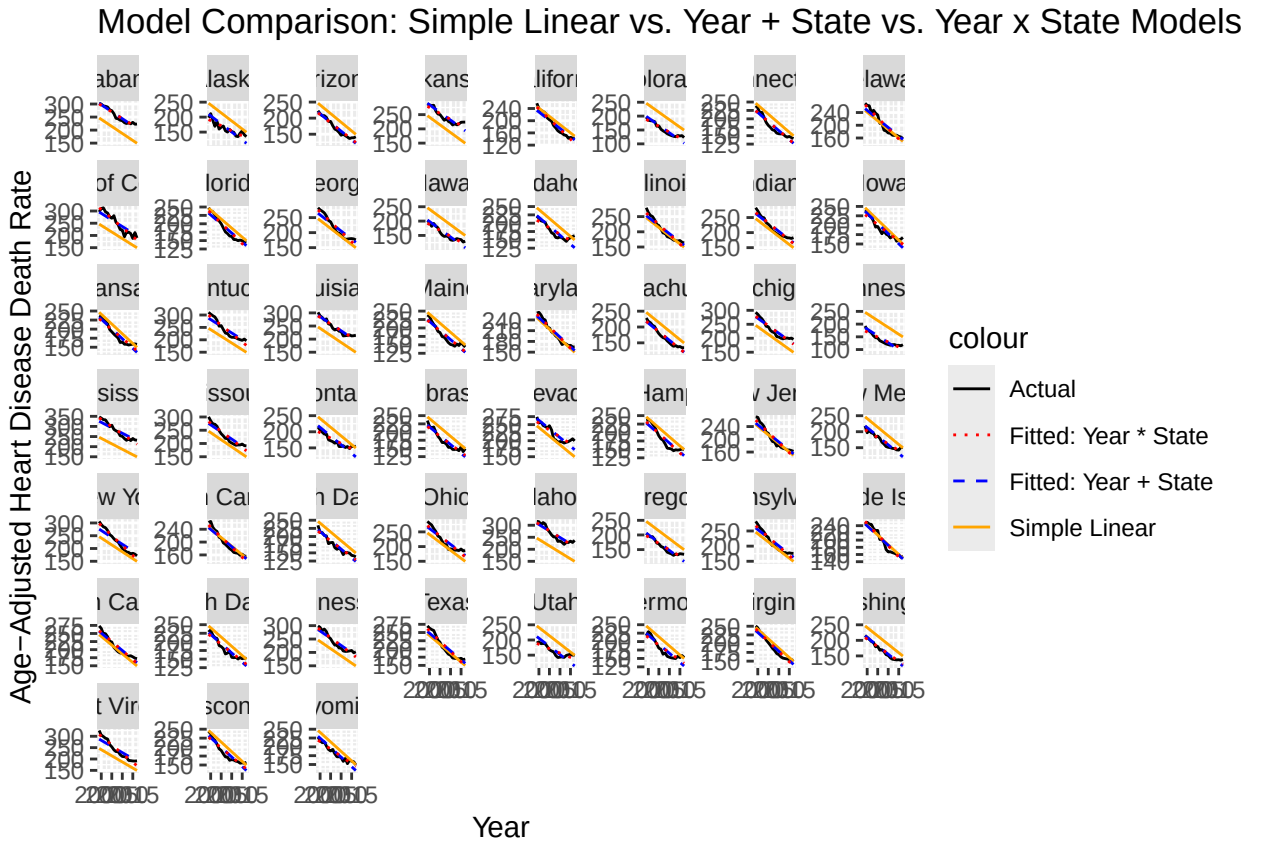
```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 2 negative eigenvalues
```

```
#summary(mixed2)


# plot comparison

state_only <- state_only |>
  mutate(
    fitted_m1 = fitted(m1),
    fitted_m2 = fitted(m2),
    fitted_m3 = fitted(m3),
    #fitted_mixed1 = fitted(mixed1),
    fitted_mixed2 = fitted(mixed2)
  )

head(state_only)
```

```
## # A tibble: 6 x 11
##    Year `113 Cause Name` `Cause Name` State Deaths Age-adjusted Death R~1 Region
##   <dbl> <chr>            <chr>        <chr>  <dbl>                  <dbl> <chr>
## 1  2017 Diseases of hea~ Heart disea~ Alab~  13110                   223. South
## 2  2017 Diseases of hea~ Heart disea~ Alas~    814                   135  West
## 3  2017 Diseases of hea~ Heart disea~ Ariz~  12398                   142. West
## 4  2017 Diseases of hea~ Heart disea~ Arka~   8270                   224. South
## 5  2017 Diseases of hea~ Heart disea~ Cali~  62797                   143. West
## 6  2017 Diseases of hea~ Heart disea~ Colo~   7060                   123. West
## # i abbreviated name: 1: `Age-adjusted Death Rate`
## # i 4 more variables: fitted_m1 <dbl>, fitted_m2 <dbl>, fitted_m3 <dbl>,
## #   fitted_mixed2 <dbl>
```

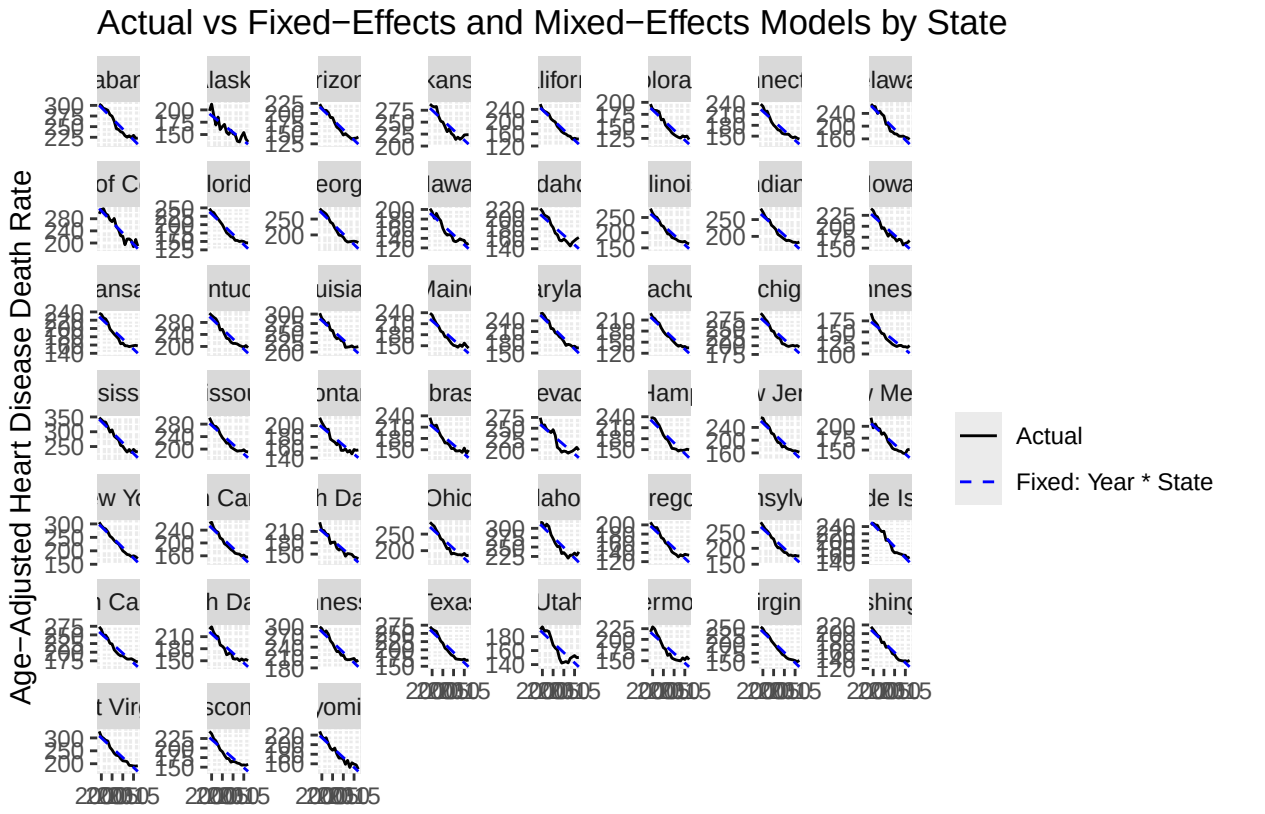
```
ggplot(data = state_only, mapping = aes(x = Year)) +

  # actual data
  geom_line(mapping = aes(y = `Age-adjusted Death Rate`, color = "Actual")) +

  # linear
  geom_line(mapping = aes(y = fitted_m1, color = "Simple Linear")) +

  # year + state
  geom_line(mapping = aes(y = fitted_m2, color = "Fitted: Year + State"), linetype = "dashed") +

  # year * state
  geom_line(mapping = aes(y = fitted_m3, color = "Fitted: Year * State"), linetype = "dotted") +

  facet_wrap(~ State, scales = "free_y") +

  labs(
    title = "Model Comparison: Simple Linear vs. Year + State vs. Year x State Models",
    y = "Age-Adjusted Heart Disease Death Rate"
  ) +

  scale_color_manual(values = c(
    "Actual" = "black",
```

```
    "Simple Linear" = "orange",
    "Fitted: Year + State" = "blue",
    "Fitted: Year * State" = "red"
  ))
```




```
ggplot(data = state_only, mapping = aes(x = Year)) +
  # actual data
  geom_line(mapping = aes(y = `Age-adjusted Death Rate`, color = "Actual")) +

  # m3
  geom_line(aes(y = fitted_m3, color = "Fixed: Year * State"), linetype = "dashed") +

  # mixed1
  #geom_line(aes(y = fitted_mixed1, color = "Mixed: (1 | State)"), linetype = "dashed") +

  # mixed2
  # geom_line(mapping = aes(y = fitted_mixed2, color = "Mixed: (Year | State)"), linetype = "dotted") +

  facet_wrap(~ State, scales = "free_y") +

  labs(
    title = "Actual vs Fixed-Effects and Mixed-Effects Models by State",
    y = "Age-Adjusted Heart Disease Death Rate",
    color = ""
  ) +

  scale_color_manual(values = c(
```

```
    "Actual" = "black",
    "Fixed: Year * State" = "blue"
    #"Mixed: (1 | State)"  = "darkgreen",
    #"Mixed: (Year | State)" = "red"
  ))
```



Why mixed effects doesn't work? - negative values

```
# Random effects covariance matrix for (Intercept, Year) by State
vc <- VarCorr(mixed2)$State
Sigma <- as.matrix(vc)
Sigma
```

```
##             (Intercept)          Year
## (Intercept) 331.1497371 -0.4385832624
## Year         -0.4385833  0.0006104929
## attr(,"stddev")
## (Intercept)        Year
## 18.19752008  0.02470815
## attr(,"correlation")
##             (Intercept)       Year
## (Intercept)   1.0000000 -0.9754377
## Year         -0.9754377  1.0000000
```

```
eigen(Sigma)$values
```

```
## [1] 3.311503e+02 2.962181e-05
```

```
# Check singularity
isSingular(mixed2, tol = 1e-4)
```

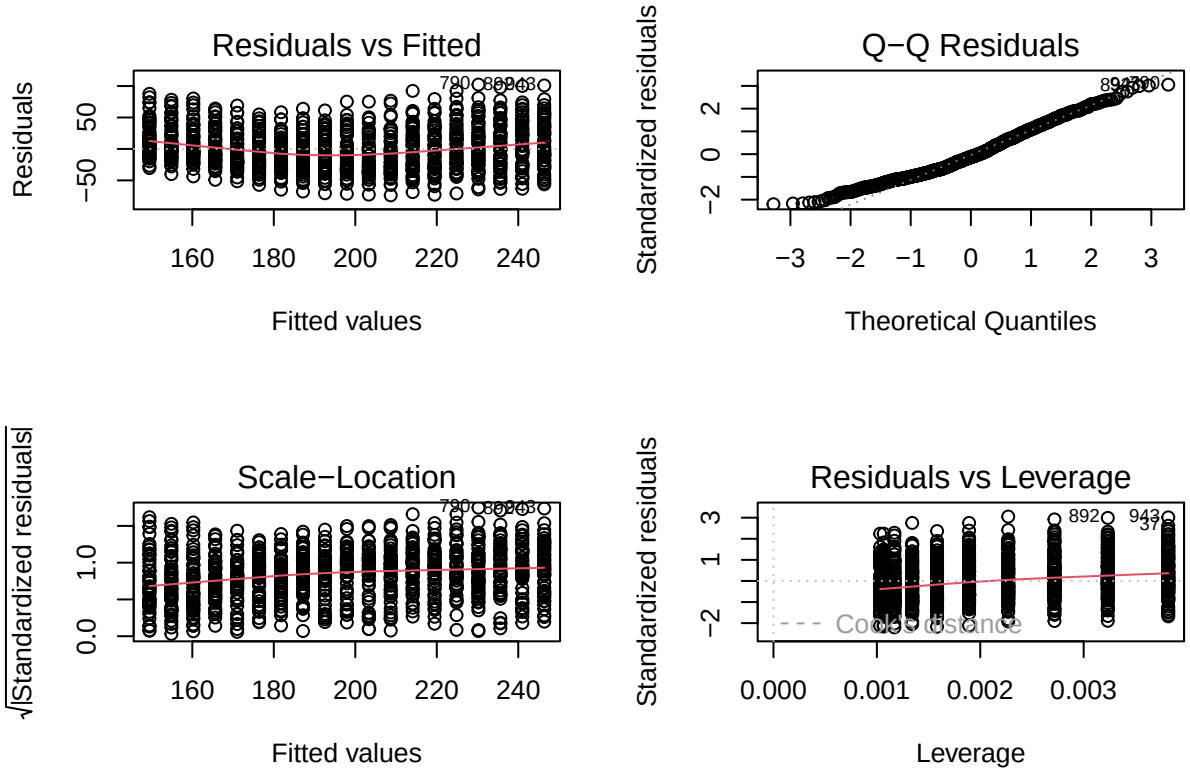
```
## [1] FALSE
```

```
# Hessian eigenvalues
H <- mixed2@optinfo$derivs$Hessian
eigen(H)$values
```

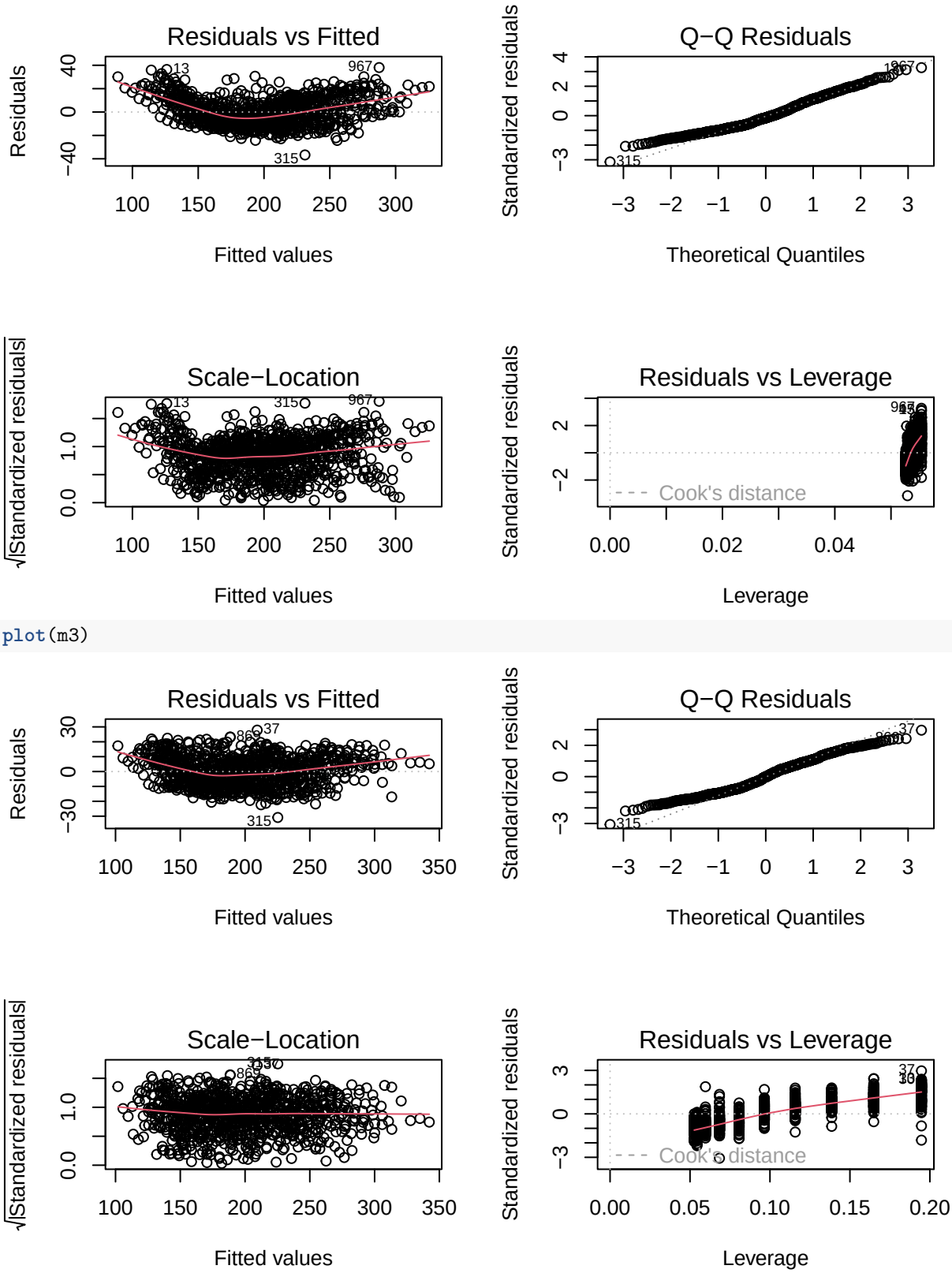
```
## [1]  1.042869e+08 -6.804950e-01 -1.463193e+06
```

- Assumption checks for m1, m2,m3

```
par(mfrow=c(2,2))
plot(m1)
```



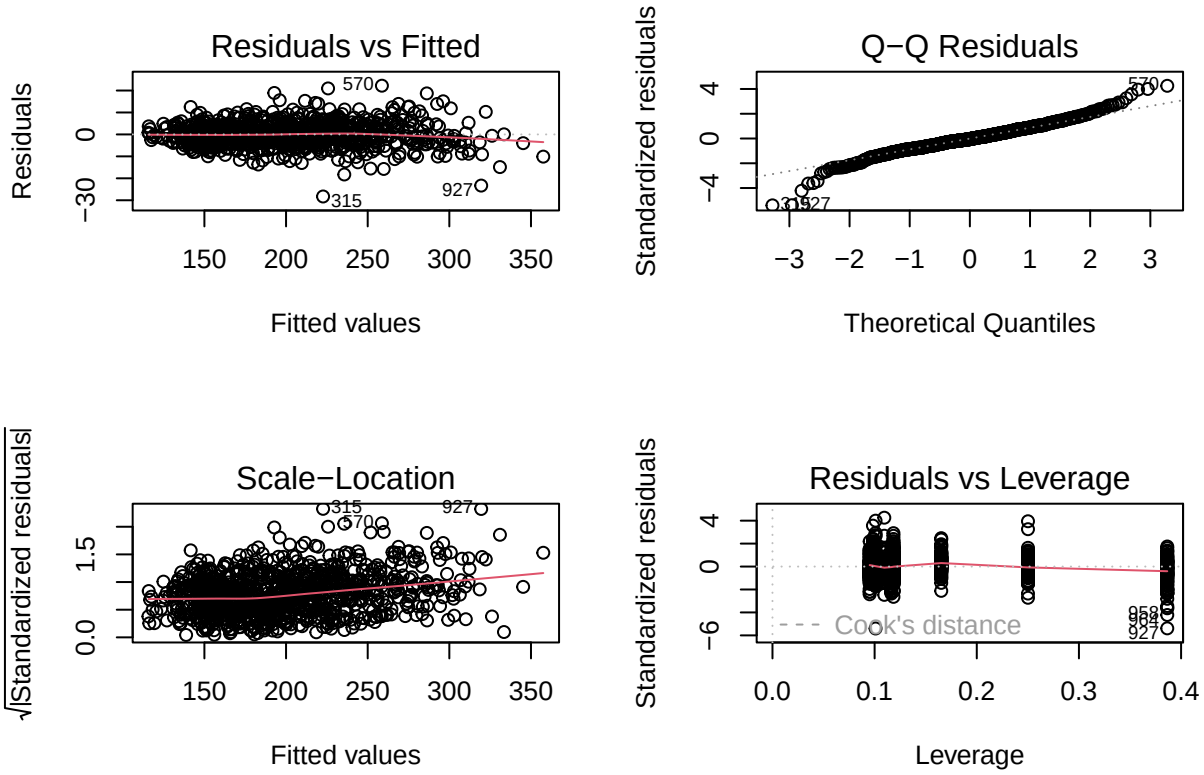

```
plot(m2)
```



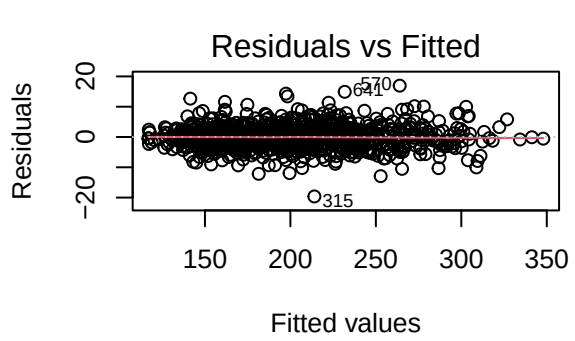

```
plot(m3)
```

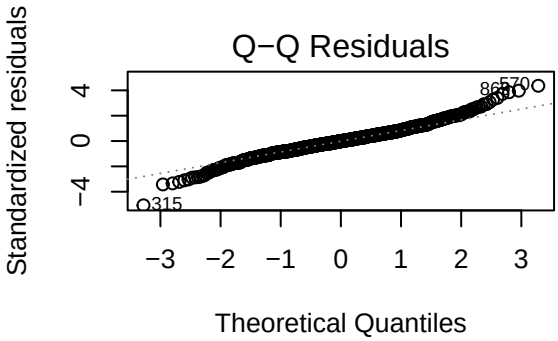

- Main Focus m3 (multiple regression interaction)
  - fix diagnostic check, mainly linearity
  - tested with quadratic and splines

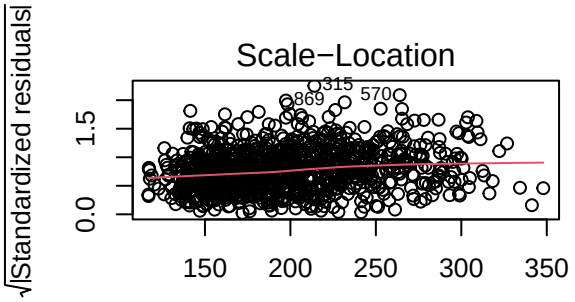
```
m_quad <- lm(`Age-adjusted Death Rate` ~ (Year + I(Year^2)) * State,
             data = state_only)

par(mfrow = c(2,2))
plot(m_quad)
```

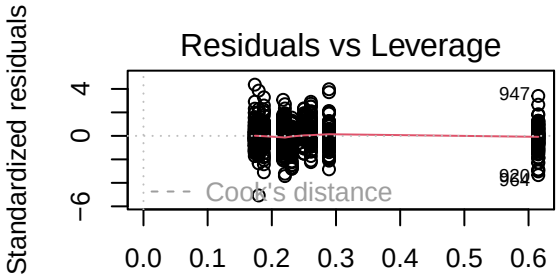



```
m_spline <- lm(`Age-adjusted Death Rate` ~ ns(Year, df = 4) * State,
               data = state_only)

par(mfrow = c(2,2))
plot(m_spline)
```

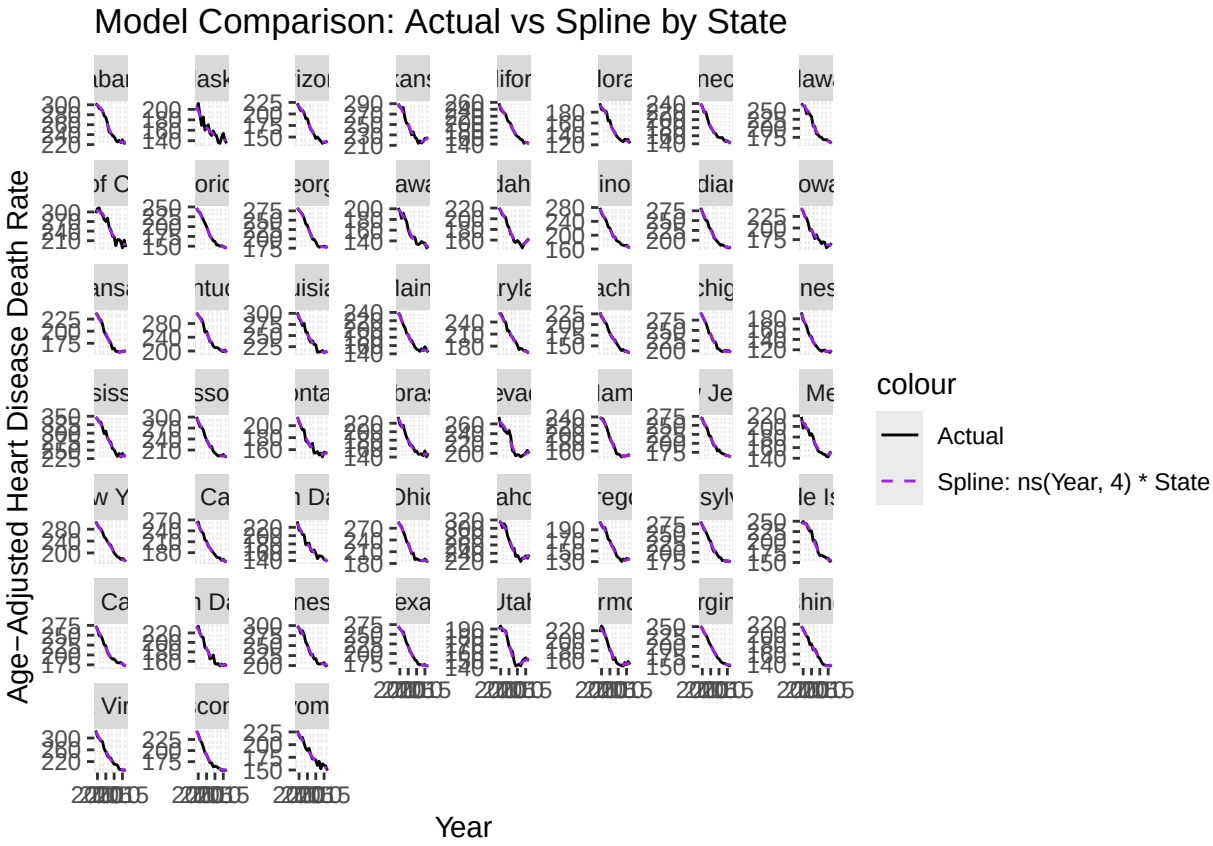









```
# Add fitted values to the data
state_only <- state_only |>
  mutate(
    fitted_spline  = fitted(m_spline)   # spline
  )

# Plot: Actual vs Spline
ggplot(data = state_only, aes(x = Year)) +

  # actual data
  geom_line(mapping = aes(y = `Age-adjusted Death Rate`, color = "Actual")) +

  # spline
  geom_line(mapping = aes(y = fitted_spline, color = "Spline: ns(Year, 4) * State"),
            linetype = "dashed") +

  facet_wrap(~ State, scales = "free_y") +
  labs(
    title = "Model Comparison: Actual vs Spline by State",
    y = "Age-Adjusted Heart Disease Death Rate",
  ) +
  scale_color_manual(values = c(
    "Actual"                       = "black",
    "Spline: ns(Year, 4) * State" = "purple"
  ))
```



#

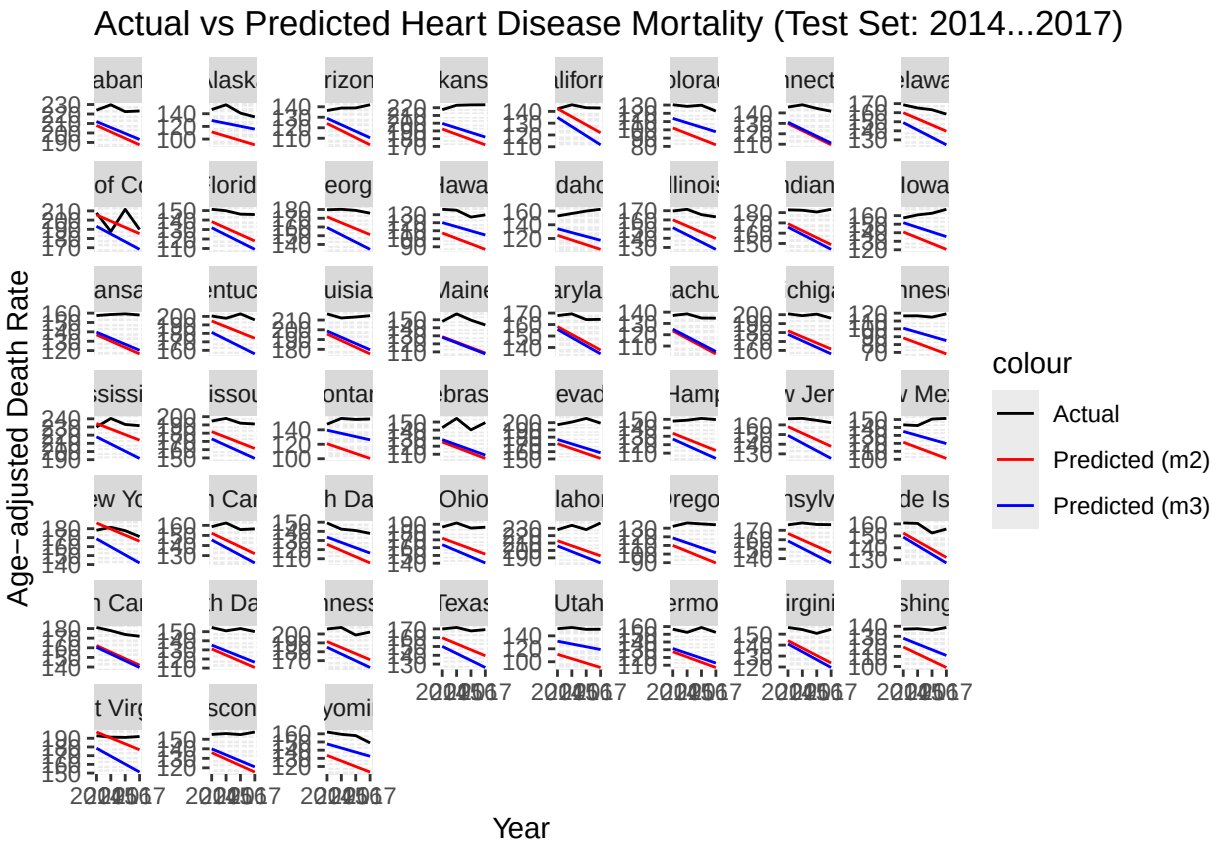
Predictive Modeling

- Train Models:
  - 80/10/10 does not work bc we cant randomly shuffle years
  - plan: use most of data as train (80%), but still want to show something for presentation (20%)
    - possible future direction: use a more updated dataset > 2017

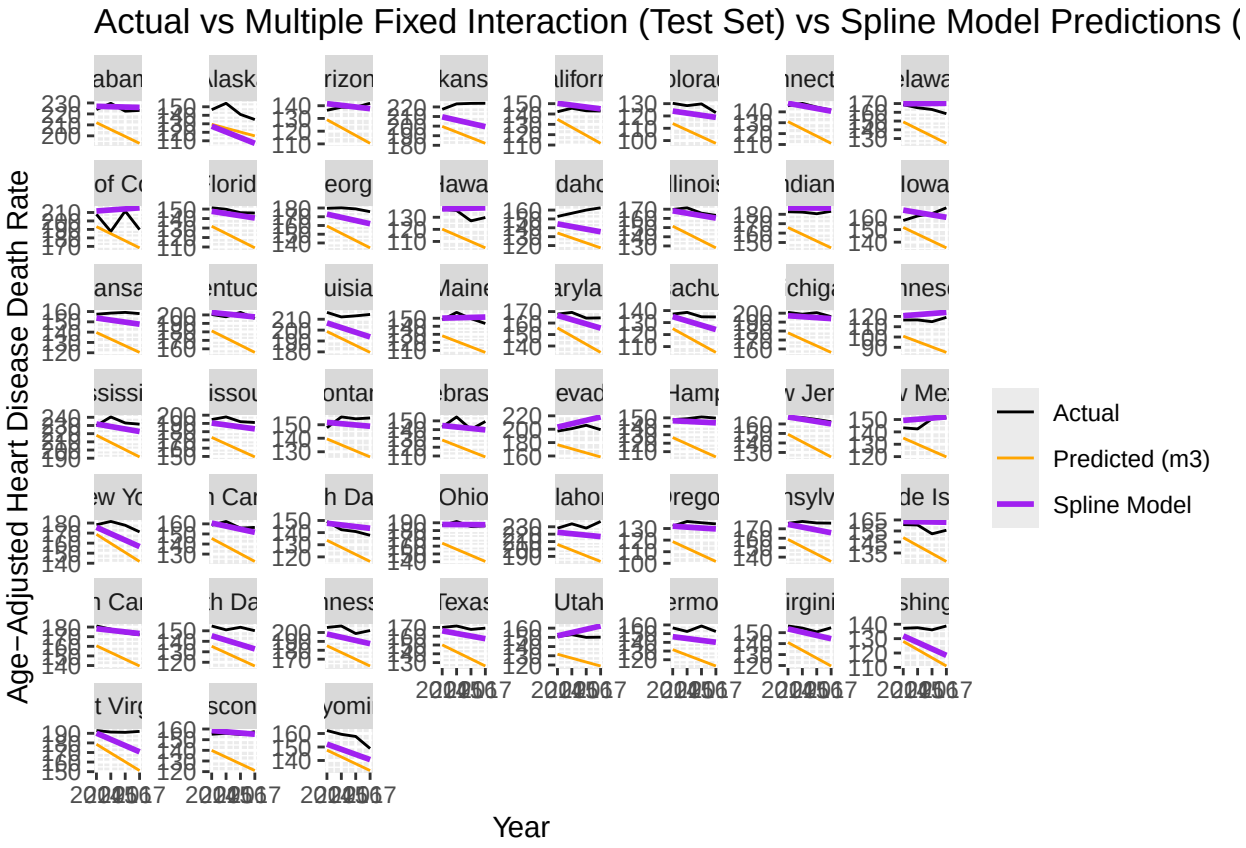
```
train <- state_only |>
  filter(Year <= 2013)

test <- state_only |>
  filter(Year > 2013)

# train models
m2_train <- lm(`Age-adjusted Death Rate` ~ Year + State, data = train)
m3_train <- lm(`Age-adjusted Death Rate` ~ Year * State, data = train)

# predict on test set

test$pred_m2 <- predict(m2_train, newdata = test)
test$pred_m3 <- predict(m3_train, newdata = test)

ggplot(data = test, mapping = aes(x = Year)) +

  geom_line(mapping = aes(y = `Age-adjusted Death Rate`, color = "Actual")) +

  geom_line(mapping = aes(y = pred_m2, color = "Predicted (m2)")) +

  geom_line(mapping = aes(y = pred_m3, color = "Predicted (m3)")) +
```

```
  facet_wrap(~ State, scales = "free_y") +

  labs(title = "Actual vs Predicted Heart Disease Mortality (Test Set: 2014-2017)") +

  scale_color_manual(values = c("Actual" = "black", "Predicted (m3)" = "blue", "Predicted (m2)" = "red"))
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality (Test Set:
## 2014-2017)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality (Test Set:
## 2014-2017)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality (Test Set:
## 2014-2017)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality (Test Set:
## 2014-2017)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality (Test Set:
## 2014-2017)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality (Test Set:
## 2014-2017)' in 'mbcsToSbcs': dot substituted for <93>
```



```r
# train/test spline model

m_spline <- lm(`Age-adjusted Death Rate` ~ ns(Year, 4) * State, data = train)


# spline
test$pred_spline <- predict(m_spline, newdata = test)

rmse_spline  <- rmse(test$`Age-adjusted Death Rate`, test$pred_spline)
mae_spline   <- mae(test$`Age-adjusted Death Rate`, test$pred_spline)

ggplot(data = test, aes(x = Year)) +

  # actual data
  geom_line(aes(y = `Age-adjusted Death Rate`, color = "Actual")) +

  # m3
  geom_line(mapping = aes(y = pred_m3, color = "Predicted (m3)")) +

  # spline predicted
  geom_line(aes(y = pred_spline, color = "Spline Model"), linewidth = 1) +

  facet_wrap(~ State, scales = "free_y") +

  labs(
    title = "Actual vs Multiple Fixed Interaction (Test Set) vs Spline Model Predictions (Test Set)",
    y = "Age-Adjusted Heart Disease Death Rate",
    color = ""
  ) +

  scale_color_manual(values = c(
    "Actual" = "black",
    "Spline Model" = "purple",
    "Predicted (m3)" = "orange"
  ))
```



Evaluation and Interpretation

- look at RMSE and MAE
- for final report go more in detail what these results mean

```
# check models accuracies
# RMSE and MAE
rmse_m2 <- rmse(test$`Age-adjusted Death Rate`, test$pred_m2)
mae_m2  <- mae(test$`Age-adjusted Death Rate`, test$pred_m2)

rmse_m3 <- rmse(test$`Age-adjusted Death Rate`, test$pred_m3)
mae_m3  <- mae(test$`Age-adjusted Death Rate`, test$pred_m3)

rmse_spline <- rmse(test$`Age-adjusted Death Rate`, test$pred_spline)
mae_spline  <- mae(test$`Age-adjusted Death Rate`, test$pred_spline)

# accuracies
data.frame(
  Model = c("m3 (Year*State)", "m_spline"),
  RMSE = c(rmse_m3, rmse_spline),
  MAE  = c(mae_m3, mae_spline)
)
```

```
##             Model      RMSE       MAE
## 1 m3 (Year*State) 27.514214 25.926974
## 2        m_spline  9.525999  7.259065
```

Random Forest

```
train$State = factor(train$State)
test$State = factor(test$State)

rm_model = randomForest(`Age-adjusted Death Rate`~Year*State, data = train, ntree = 1000, mtry = 2, imp

print(rm_model)
```

```
##
## Call:
##  randomForest(formula = `Age-adjusted Death Rate` ~ Year * State,      data = train, ntree = 1000, m
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 38.46033
##                     % Var explained: 98.08
```

```
test$pred_rm = predict(rm_model, newdata=test)
rmse_rfm = rmse(test$`Age-adjusted Death Rate`, test$pred_rm)
mae_rfm = mae(test$`Age-adjusted Death Rate`, test$pred_rm)
```

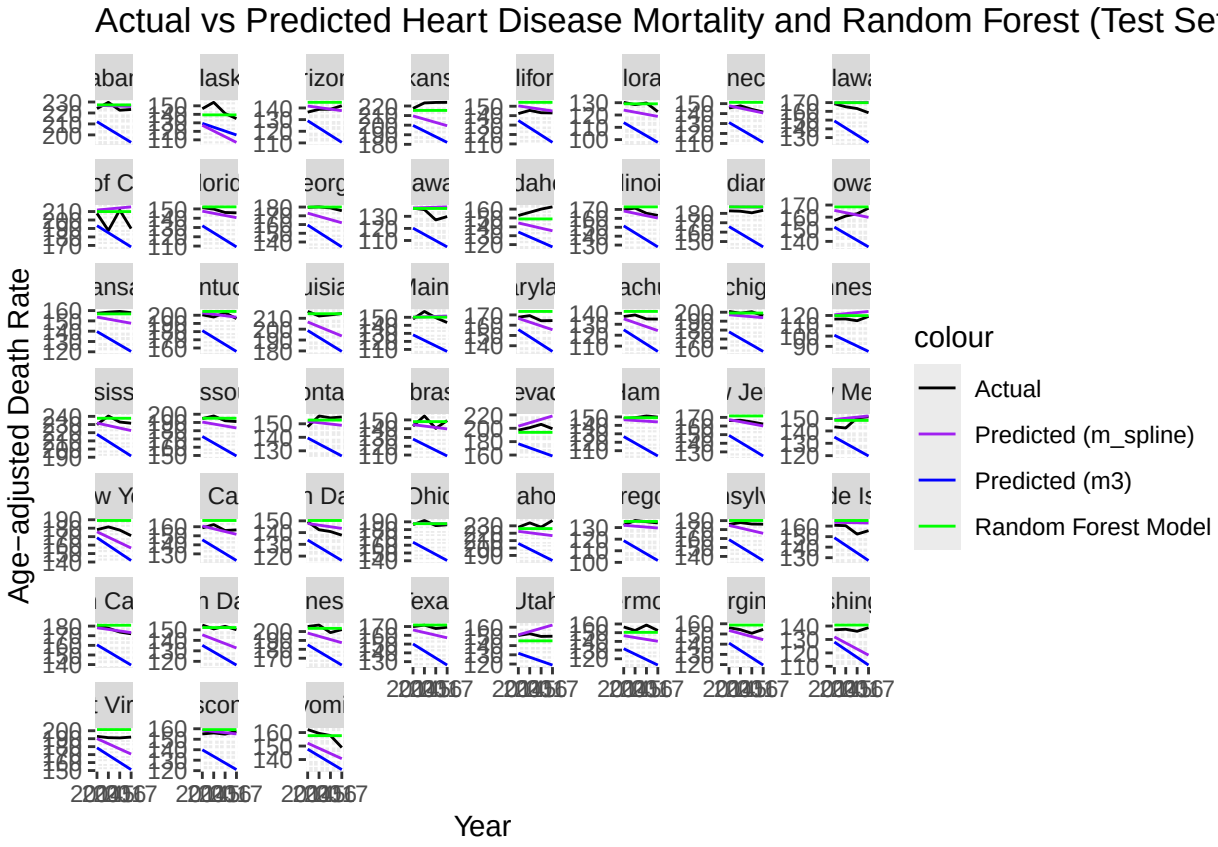
```
# R^2
y_test = test$`Age-adjusted Death Rate`
y_rmf = test$pred_rm

ss_res = sum((y_test - y_rmf)^2)
ss_tot = sum((y_test-mean(y_test))^2)
R2_rmf = 1 - ss_res/ss_tot

R2_rmf
```

```
## [1] 0.9497093
```

So about 95% of the variation in state-level age-adjusted heart-disease mortality(2014-2017) is explained by the random forest model using only Year and State

```
ggplot(data = test, mapping = aes(x = Year)) +

  geom_line(mapping = aes(y = `Age-adjusted Death Rate`, color = "Actual")) +

  geom_line(mapping = aes(y = pred_m3, color = "Predicted (m3)")) +

  #geom_line(mapping = aes(y = pred_mixed1, color = "Predicted (m1)")) +

  geom_line(mapping = aes(y = pred_spline, color = "Predicted (m_spline)")) +

  geom_line(mapping = aes(y = pred_rm, color = "Random Forest Model"))+

  facet_wrap(~ State, scales = "free_y") +

  labs(title = "Actual vs Predicted Heart Disease Mortality and Random Forest (Test Set: 2014-2017)") +

  scale_color_manual(values = c("Actual" = "black", "Predicted (m3)" = "blue", "Random Forest Model" =
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality and Random
## Forest (Test Set: 2014-2017)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality and Random
## Forest (Test Set: 2014-2017)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality and Random
## Forest (Test Set: 2014-2017)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality and Random
## Forest (Test Set: 2014-2017)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality and Random
## Forest (Test Set: 2014-2017)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Actual vs Predicted Heart Disease Mortality and Random
## Forest (Test Set: 2014-2017)' in 'mbcsToSbcs': dot substituted for <93>
```




```
data.frame(
  Model = c("m3 (Year*State)", "m_spline", "RF (Year * State)"),
  RMSE = c(rmse_m3 , rmse_spline, rmse_rfm),
  MAE  = c(mae_m3, mae_spline, mae_rfm)
)
```

```
##               Model      RMSE       MAE
## 1   m3 (Year*State) 27.514214 25.926974
## 2          m_spline  9.525999  7.259065
## 3 RF (Year * State)  6.299230  5.015603
```