THE HEALTH IMPACT OF COAL MINING: A REGRESSION ANALYSIS

A Thesis
by
HENNING TOVAR

Submitted to the Graduate School
Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN APPLIED DATA ANALYTICS

May 2020
Walker College of Business

THE HEALTH IMPACT OF COAL MINING: A REGRESSION ANALYSIS

A Thesis
by
HENNING TOVAR
May 2020

APPROVED BY:

Dr. John Whitehead
Chairperson, Thesis Committee

Dr. William Hicks
Member, Thesis Committee

Dr. Lakshmi Iyer
Member, Thesis Committee

Dr. Tim Forsyth
Interim Associate Dean of Graduate Programs and Research, Walker College of Business

Dr. Mike McKenzie
Dean, Cratis D. Williams School of Graduate Studies

ABSTRACT

THE HEALTH IMPACT OF COAL MINING: A REGRESSION ANALYSIS

Henning Tovar
B.A., Friedrich-Alexander-Universität Erlangen-Nürnberg

Chairperson: Dr. John Whitehead

Coal mining has a well-established detrimental effect on the health of coal miners who work in the mines or in direct vicinity to the coal mines. However, it is less clear how mining affects the communities living around the mining area. Recent research has pointed towards increased mortality rates in coal mining areas, but the statistical analyses used to produce this association are flawed. In my research I point out the methodological problems with prior research and argue for the use of hierarchical linear regression instead.

My analysis investigates the relationship between county level mortality rates and coal mining across the entire United States. The multilevel regression model incorporates the effect of time over 8 years and includes county level data for all counties over the time span. Further, the model accounts for control variables used in the literature including economic factors and demographics, county level health indicators, and educational data. Holding all other factors constant, I find that mining does not statistically significantly affect mortality rates for the entirety of the United States.

However, at the state level the effect of coal mining varies considerably, indicating different effects for coal mining states.

ACKNOWLEDGMENTS

Before presenting my thesis work, I would like to extend my gratitude to the people that have helped me succeed up to this point. In particular, I would like to thank my thesis committee: Dr. John Whitehead who helped me greatly with reading through drafts of my thesis and giving me insightful feedback, Dr. William Hicks for answering all of the questions that came to my mind and guiding me through the methodological process, and Dr. Lakshmi Iyer for her flexibility and guidance through the administrative process. Furthermore, I would like to thank my parents and all the friends that gave me support when I realized how much endurance it needs to write a thesis.

DEDICATION

I dedicate my work to the people and mountains of the Appalachians.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Growing up in direct vicinity to the region that for decades was the main coal mining region in Germany, I have always been interested in the politics of coal. In particular, I wanted to study the cultural conflicts that developed along the frontiers of mining. Many political conflicts pivot the economic opportunities associated with coal mining and coal-based power production against renewable energy and less traditional industries. The culture of coal embraces the coal industry as not just another industry but as a lifestyle (Lewin, 2017). However, health is a facet to the political conflicts surrounding coal mining that goes beyond reasoning of economic impact. Coal mining deeply affects the health of miners and potentially the health of mining communities. The health risks for coal miners, like the black lung disease and fatal mining accidents, are well documented and have a place in the public debate. Community health risks, on the other hand, are a less salient issue and are also less researched (Moffatt & Pless-Mulloli, 2003). Consequently, the research question motivating this thesis attempts to illuminate the relationship between coal mining and the health of coal mining communities.

- **RQ**: How does coal mining affect the health of communities living in the vicinity of coal mines.

My research aims to contribute to the body of research by filling in parts of this research gap and explore the relationship between coal mining and county-level mortality rates, as a proxy of community health. In my research I am heavily leaning on a series of articles published by researchers at West Virginia University (Hendryx & Ahern, 2009; Hendryx, Fedorko, & Halverson, 2010). Researchers like Michael Hendryx and colleagues have made numerous efforts to establish a statistical relationship between coal mining and elevated mortality rates on the county level. However, their research is lacking in two key dimensions. First, the published research has a strong focus on the Appalachian region. While coal mining plays a central role in Appalachia, the region can hardly be used to generalize the effect of coal mining on county health. Appalachian counties, on average, suffer from poorer socio-economic conditions in many dimensions and cannot be compared to counties outside the Appalachian region. While Hendryx and colleagues are not claiming such a generalization, my research is driven by an interest to highlight the general connection between coal mining and mortality. The scope of my thesis, thus, goes beyond the Appalachian region and includes counties in the entire United States.

The second gap in the existing body of research is the use of statistical models that suite the temporal structure of the data. Mortality and coal mining are subjects that are inherently time related. Mortality rates are measured in annual deaths per population, while coal production is measured in short tons of coal production over time (quarters, years, etc.). Mortality rates and coal production fluctuate from year to year and statistical models that aim to explain mortality rates need to account for fluctuation over time. However, the current body of research does not make use of statistical models that incorporate time as a variable. Instead, researchers take the average of the variables of interest over the observed time frame and then fit ordinary least squares regression to these averages. This procedure of averaging years of data reduces the variability of the observed effects greatly and weakens statistical models derived from the data. Consequently, the second contribution of my thesis to the research body is to incorporate a time-sensitive statistical approach, namely multilevel linear regression.

CHAPTER 2. LITERATURE REVIEW

Coal mining and coal related industries play an important role in the economy of the United States. While total levels of coal production have decreased over the past decade, the reserves left for exploitation are expected to last for the foreseeable future (Kecojevic & Grayson, 2008). Against the backdrop of these potentially long-lasting coal reserves, the question of whether coal reserves should be exploited becomes a critical one. The answer to this question is linked to the balance of costs and benefits associated with coal mining. Economic output associated with the mining industry can be clearly identified as the benefits of mining. However, the costs of coal mining are less tangible, especially when it comes to the risk coal mining poses to the health of miners and the public at large.

Coal mining has long been identified as an occupation with an increased risk of potentially fatal accidents. A body of literature has focused on the relationship between coal mining and increased mortality in communities surrounding coal mines and activities related to mining (Cortes-Ramirez, Naish, Sly, & Jagals, 2018; Hendryx, 2015). Instead of focusing on the risk that individual miners bear, these studies seek to provide evidence for an increase in risk for the entire county in which the mine is located. Solidifying this relationship could prompt a reconsideration of the impact coal mining has on miners, the environment, and the health of mining communities. As Hendryx and Ahern (2009) claim, including the detrimental effect of coal mining on public health into a benefit-cost analysis of coal mining would outweigh the economic benefits of mining by far. In the following sections, I will discuss the theoretical reasons for linking coal mining and increased mortality and will then go on to elaborate on the empirical findings of this relationship.

2.1 The Theoretical Relationship Between Coal Mining and Mortality

The assumed association between coal mining and increased mortality is based on an increased level of local environmental pollution because of coal mining. The increased pollution then results in an increase in pollution related diseases that increase mortality rates overall. Environmental pollution linked to coal mining can be separated into toxic agents and particular matter, both of which have been associated with increases in illnesses (Esch & Hendryx, 2011; Hendryx, 2009).

The literature on toxic agents such as lead, arsenic, mercury, and cadmium has established a reliable association between these toxic agents and cardiovascular diseases and ischemic and coronary heart diseases (Menke, Muntner, Batuman, Silbergeld, & Guallar, 2006; Menke, Muntner, Silbergeld, Platz, & Guallar, 2009). Furthermore, exposure to lead has been associated with hypertension and kidney diseases (Jain et al., 2007; Lin, Lin-Tan, Li, Chen, & Huang, 2006; Navas-Acien, Guallar, Silbergeld, & Rothenberg, 2007). The process of mining and cleaning coal releases toxic agents into the atmosphere and water bodies around the coal mine, thus increasing the exposure of individuals living in proximity of the coal mine (Hendryx & Ahern, 2008; Hendryx, Yonts, Li, & Luo, 2019).

Furthermore, the mining process creates an increased ambient level of particulate matter (PM) in the area surrounding the mine (Hendryx, 2009). Increased levels of PM can lead to a variety of adverse health effects. Moreover, certain levels of PM have been associated with fatal coronary heart diseases and atherosclerosis. Lung diseases like pulmonary inflammation and general oxidative stress are also associated with increased PM levels (Donaldson et al., 2002). Air pollution, in general, has been associated with increased admissions to the emergency room for a variety of heart diseases (Behringer & Friedell, 2006; Mastin, 2005).

Some of the chemicals used during the mining process are considered carcinogenic and are inadvertently emitted into the ecosystem around the mine. Eventually these substances reach communities living close to mines and negatively impact public health. Elevated cancer rates in areas where coal mining is present have been established by several studies (Christian, Huang, Rinehart, & Hopenhayn, 2011; Hendryx, Fedorko, & Anesetti-Rothermel, 2010; Hendryx, O'Donnell, & Horn, 2008).

The presence of coal mining increases the level of environmental pollutants a community is exposed to and consequently is expected to increase hospitalization and mortality rates (Hendryx, Ahern, & Nurkiewicz, 2007; Hendryx et al., 2019). In the following section, I will discuss the empirical findings of the assumed relationship between coal mining and mortality rates.

2.2. Empirical Findings on the Association of Coal Mining and Mortality Rates

A series of articles published by researchers at the University of West Virginia has resulted in a stream of publications that focus on the association of mortality rates and coal mining (Hendryx & Ahern, 2008, 2009; Hendryx et al., 2008). Most of these studies arrive at similar results and are very similar in terms of data collected and methodology. However, there is some debate in the literature about the validity and reliability of the findings.

Articles first evaluated the relationship between coal mining and community health on the individual level. Individuals' self-reported health indicators were found to be significantly worse in coal mining areas (Hendryx & Ahern, 2008), while hospitalization rates for hypertension and chronic obstructive pulmonary disease were significantly related to quantity of coal mined in states with coal mining (Hendryx et al., 2007). These findings generated a general association between poor health outcomes and coal mining and gave rise to a series of articles that focused on the statistical connection between coal mining and county level mortality rates.

When comparing mean differences between mining counties and non-mining counties, Hendryx (2009) found statistically significantly elevated mortality rates in coal mining counties. Holding other factors constant, coal mining has been associated with an increase in mortality rates of 17 deaths per 100,000 population with a standard error of 7.5 (Esch & Hendryx, 2011; Hendryx et al., 2008). Mountain top removal mining, a particularly intrusive mining technique, has been associated with an increase in mortality rates of about 25 and a standard error of 9.3 (Esch & Hendryx, 2011; Hendryx, 2009). It should be noted that these studies are heavily focused on Southern United States and the Appalachian region in particular.

In a direct response to the series of articles published by Hendryx and colleagues, Borak, Salipante-Zaidel, Slade, and Fields (2012) point towards methodological issues with previously conducted studies. In particular, they highlight that the Appalachian region suffers from overall poor socioeconomic health conditions which lead to increased mortality rates. In a similar article, Buchanich, Balmert, Youk, Woolley, and Talbott (2014) claim that while coal mining is statistically significantly related to cancer related deaths, a similar relationship cannot be inferred for all-cause mortality rates. Furthermore, studies that have found a significant relationship between coal mining and mortality rates employ different operationalizations of key measures and use data from different time frames (Woolley, Meacham, Balmert, Talbott, & Buchanich, 2015).

A meta-study of the literature in 2018 concluded that the evidence for an association between elevated mortality rates and coal mining is outweighing arguments leveled against this relationship (Cortes-Ramirez et al., 2018). Methodological concerns about the conducted studies should be, nonetheless, taken seriously. These methodological issues concern all studies conducted regardless of the results. In the following section I will elaborate upon these concerns.

2.3. Methodological Inconsistencies in the Empirical Findings

Methodological consistency is necessary for reliable and valid conclusions from statistical analysis and ensures comparability between studies. However, the published research on the association between coal mining and mortality rates suffers from several inconsistencies that remain largely unexplained in the literature. These issues concern two areas in particular:

- Operationalization of coal mining
- Treatment of time and trends

While the focus of the previously mentioned studies is the link between coal mining and mortality rates, different studies by the same authors employ different operationalizations of coal mining, and in some instances change the operationalization within one article. In most of the articles, the reasons for a particular operationalization remain unexplained (Borak et al., 2012). Generally, there are three different ways to operationalize the presence of coal mining: (i) the quantity of coal mined per year, (ii) an indicator variable showing the presence of mining, (iii) an ordinal variable reflecting levels of mining.

Several studies operationalize coal mining in terms of the numeric quantity of coal mined per year (Borak et al., 2012; Hendryx & Ahern, 2009). However, Esch and Hendryx (2011) point out that tons of coal mined per year is not-normally distributed. Consequently, Esch and Hendryx (2011) transform the data by taking the natural logarithm of the series. Their article finds a statistically significant relationship between coal mining and mortality rates, while Borak et al. (2012) claim to find no statistically significant relationship without performing a logarithmic transformation.

Another commonly applied operationalization of coal mining is the use of an indicator variable that takes on a value of 1 for counties with any coal mining and a value of 0 for counties with no coal mining (Hendryx, 2009; Hendryx & Ahern, 2009). A variation of this operationalization accounts for the quantity of coal mined and introduces an ordinal variable. The variable distinguishes between counties with no coal mining, low levels of coal mining., and high levels of mining (Borak et al., 2012). However, the value at which this distinction is made changes between studies and there is no explanation why different values are selected. Hendryx et al. (2008) splits counties with coal mining at a production level of 3 million tons of coal per year, while the same author in Hendryx (2009) uses a production level of 4 million tons of coal per year as the dividing value, and finally, Hendryx and Ahern (2009) use the median production level as the dividing value. All three studies claim that the selected value splits coal mining counties into two equally sized groups but do not discuss differing values from previously conducted studies.

Furthermore, published studies employ a problematic treatment of time. The first aspect of this problem concerns data collection. Most studies collect data on key variables from a variety of different time frames. Hendryx, Fedorko, and Halverson (2010), for example, collect mortality data for the years 1997 – 2005, but collect covariate data only for the year 2008. This pattern of differing time frames that do not overlap is common among all the studies cited above. Researchers generally do not discuss a rationale for either collecting data for certain time periods or points in time. This problematic pattern was first pointed out by Borak et al. (2012) in a critique of research authored by Hendryx and colleagues. Nonetheless, Borak et al. (2012) still went on to use data from a different time frame than Hendryx and colleagues.

Secondly, the collected data is aggregated for analysis in a way that excludes the effect of time on mortality rates. Almost all studies cited above collect data over a variety of time periods and then compute the average for variables of interest of this period. For example, Hendryx, Fedorko, and Halverson (2010) collect data on coal mining for 1996 – 2005 and then compute the mean level of coal production per county. The resulting data set contains all variables of interest either at the mean value over a certain period or at a single year value. This data is then used to fit an ordinary least squares regression (OLS) model to predict the association between coal mining and mortality rates. This methodological procedure of collecting data over seemingly arbitrary time periods or single years and then aggregating the data to a cross-sectional-type data set that can be used for OLS was used in the initial studies by Hendryx and Ahern (2009). Almost all studies that followed this research have used the same methodology without explaining methodological choices. A notable exception from this pattern is Hendryx and Holland (2016) which conducts a hierarchical regression model that explicitly includes time as a variable. However, while the study looks at mortality rates from 1968 – 2014, the authors assume all covariates to be constant over this period of 46 years. The covariate data is collected for the year 2015, which is just outside the scope of the analysis.

Treating time this way introduces several problems. First, it is not clear on theoretical grounds, how observations for a single year are related to the averaged observations for several decades. Furthermore, averaging observations over time excludes the effect of time from the analysis. Instead of panel data that includes trends in variables over time, the data are aggregated to a single snapshot of the variables of interest. Consequently, time as a variable to model potential trends is excluded from the data. Furthermore, this procedure reduces the overall variance in the data and eliminates variation for each county that is observed over several years.[1] The number of observations is reduced considerably which results in less reliable statistical models.

[1] When conducting the averaging procedure on my own data to illustrate the variance reduction, the variance of mortality rates is reduced by almost 20% from 23048.43 in the data set containing all observations to 18764.74 in the data set with averaged observations. Furthermore, there is considerable variance in mortality rates within each county over the 8 years of collected data.

Overall, the methodology employed in most studies ignores known trends in mortality rates and coal mining over time. As Hoyert (2012) points out, mortality rates are decreasing over time, which is a finding that a statistical model of mortality rates should include. Findings by Hendryx and Holland (2016) support this conclusion, as the time variable in their hierarchical regression analysis does have a statistically significant effect. An appropriate statistical model of the relationship between coal mining and mortality rates should include panel data of county level mortality rates, coal production levels, and covariates used for the analysis.

2.4. Hypotheses and Expectations

Based on conclusions drawn from the literature, I am developing a statistical model to test for several hypotheses. Firstly, I am interested in the relationship between coal mining and mortality rates. The literature establishes a clear theoretical and empirical association between elevated mortality rates and the presence of coal mining (Hendryx, 2015). Thus, one of the main objectives of my thesis is to examine the evidence for this relationship based on an appropriate statistical model. I am consequently formulating hypothesis H_1, which tests for elevated mortality rates in coal mining counties:

- H_1 = Holding all other variables constant, counties with coal mining have higher mortality rates compared to counties with no coal mining.

Furthermore, according to empirical findings, coal mining disproportionately affects mortality rates in counties with production above the median level of coal mining (Borak et al., 2012; Hendryx, 2009; Hendryx & Ahern, 2009). The second objective of my thesis is to test for the reliability of this additional increase in mortality rates for counties with high levels of coal mining (hypothesis H_2).

- H_2 = Holding all other variables constant, counties with above median levels of coal mining have higher mortality rates compared to counties with below median levels of coal mining.

Thirdly, I am interested in including time as an explicit effect on the relationship between coal mining and mortality rates. Findings by Hendryx and Holland (2016) and Hoyert (2012) demonstrate a downward sloping trend in mortality rates over time. Consequently, I am expecting a negative association between time and mortality rates as summarized in hypothesis H_3:

- H_3 = Holding all other variables constant, mortality rates decrease over the period of study.

The objective of this hypothesis is of methodological rather than substantive interest. As I argued in section 2.3, excluding time as a potential factor from the data poses methodological problems. Thus, H_3 in a broader sense aims at providing statistical reasons for including time into research designs that model the relationship between mortality rates and coal mining.

Findings from several studies have found statistically significant differences in mortality rates between Appalachian counties and non-Appalachian counties (Behringer & Friedell, 2006; Christian et al., 2011; Woolley et al., 2015). Furthermore, scholars found differences in the effect of coal mining based on the geographic location of the counties (Woolley et al., 2015). A fourth objective of my study is to further explore these regional differences summarized in Hypothesis H_4. As laws and policies surrounding mining and mining procedures differ substantially between states (Hendryx & Holland, 2016), regional effects are grouped by state.

- H_4 = Holding all other variables constant, the effect of coal mining on mortality rates differ by state.

The following chapter presents the analytical strategy I apply to test the hypotheses indicated above and discusses the assumptions involved in the modeling approach. Further, I will elaborate on the data I collected for my study and transformations made to the raw data.

CHAPTER 3. METHODS

As discussed in the previous section data on the relationship between mortality rates and coal mining is generally longitudinal data. However, longitudinal and otherwise grouped data pose substantial problems for ordinary least squares regression (OLS). Grouped data violate the assumption of OLS that observations are independent from each other (Gelman & Hill, 2007). In the case of longitudinal data, there is good reason to assume autocorrelation between observations from different years. Countywide coal production levels for the year 2012, for example, are very likely to be highly correlated with production levels for the year 2011. Furthermore, clustered data, for example students clustered in classes, also violate the assumption of independent observations. When comparing students within one class, they will be more similar to each other in contrast to comparisons to students from other classes. Thus, the variance of the observed values is conditional on the group to which the observation belongs (Singer & Willett, 2003).

An OLS model that is fit to clustered (or longitudinal data as a special case of clustered data) may suffer from heteroskedasticity due to violation of independent observations. A possible strategy to deal with this issue is to aggregate the clustered data and fit a regression model to the aggregate. While this is the strategy commonly employed in the literature on coal mining and mortality rates, aggregating data over time reduces the variance in the data considerably and fully erases with-in group variation (Bryk & Raudenbush, 1988). Thus, applying a statistical method that incorporates clustered data into the model improves the modeling approach and allows for more reliable statistical inference (Gelman & Hill, 2007; Raudenbush & Bryk, 2002).

Mixed-effect or hierarchical linear regression models allow to leverage information contained in clustered data. By introducing additional variance components to the regression model, mixed effect models allow for group-level observations to vary around the population average effect of a variable. The fixed effect term of the model represents the population average effect and can be interpreted in a similar fashion to conventional regression coefficients (Gelman & Hill, 2007). The random effect part of the model expresses the variation of groups around the fixed effect. While the variance is assumed to be zero on average, the model relaxes the assumption that the effect of a variable is the exact same for each group. Thus, mixed effect models allow for (i) an analysis of clustered data without aggregation, and (ii) analysis of the group-level variance component.

These advantages of multilevel modeling suit the hypotheses I am testing particularly well, as I am interested in analyzing the effect of coal mining on mortality rates without aggregating the data. Furthermore, I am interested in the variation of the effect of coal mining on mortality rates between states. This variation is represented by the additional variance components introduced in the multilevel model. In the following section I will go over the model selection process and formalize the multi-level regression model.

3.1 Model Selection

The purpose of this study is to replicate conceptualizations and measurements from research designs employed in the literature and use a statistical method that allows for non-aggregated data. Thus, the variable selection and measurement choices are generally a replication of the approaches pursued in studies found in the literature. In the following sections, I will elaborate on the variable selection and measurement of the selected variables. The statistical model used in this study includes several interaction terms that are generally not included in the literature. The theoretical reasons for these interaction terms will be given in a separate section along with the full formalization of the statistical model.

3.1.1 Dependent Variable: Mortality

Virtually all previous studies operationalize mortality in terms of county level mortality rates. The National Center for Health Statistics (NCHS) computes annual county level mortality rates and makes these accessible to the public (NCHS, 2017). Mortality rates are calculated as the proportion of raw death counts to the county population and are then multiplied by 100,000. The final crude mortality rate reflects the mortality per 100,000 population. However, death counts are affected by the underlying age distribution of the county population. Cancer cases and consequently cancer-related mortality, for example, increase with increasing median age of a county. Thus, crude mortality rates are transformed to age-adjusted mortality rates by calculating the weighted average of age-specific death rates. To ensure comparability, all crude mortality rates are adjusted to the 2000-census population (Anderson &

Rosenberg, 1998). Consequently, the dependent variable for the statistical model is the county level age-adjusted mortality rate.

Figure 1 shows the total distribution of mortality rates and state-level mortality rates over time. While the distribution is approximately normal, it becomes obvious that states vary substantially in regard to their average mortality rate. Mortality rates in the United States have decreased significantly over the past several decades (Hoyert, 2012). Measured over the relatively short period of time included in this study, however, the mortality rates remain relatively stable. However, different states range widely in their respective mortality rates. According to Figure 1, mortality rates spread over a range of about 400 deaths per 100,000 population. These findings indicate that a statistical model that measures county-level mortality, should include variation in state-level baseline mortality. Thus, I will include a state-level random intercept variance component into the multilevel regression model.

3.1.2 Independent Variable: Coal Mining

As this study investigates the effect of coal mining on county-level mortality rates, the key independent variable is county-level coal mining. Previous studies have generally approached the measurement of coal mining in two different ways. Intuitively, since the effect of coal mining is the research focus, many studies operationalize mining in quantitative terms of tons of coal mined per year (Borak et al., 2012; Esch & Hendryx, 2011; Hendryx, 2009; Hendryx & Ahern, 2009). However, coal production is not normally distributed and thus including the untransformed variable into a regression poses a problem in the modeling process. Figure 2 shows the logged coal production levels across coal producing counties in the United States. It should be noted that counties without coal production are excluded from this graph. As most counties have no coal mining, the distribution would have a high peak around zero with a large number of outliers when including counties that do not produce coal.

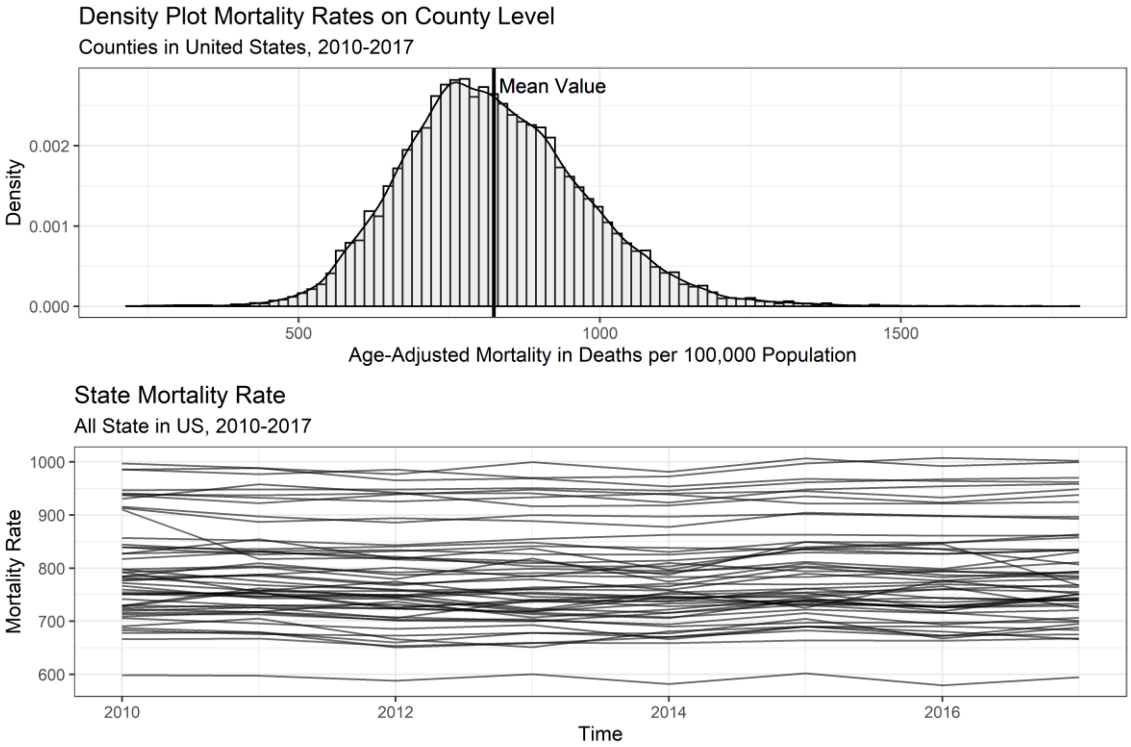


Figure 1. Mortality Rates. Total Distribution and Development Over Time

In order to account for the problematic distribution of coal mining, other studies have operationalized coal mining in terms of the presence of any coal mining (Borak et al., 2012; Hendryx, 2009; Hendryx & Ahern, 2009). The measurement is operationalized as an indicator variable that takes on a value of 1 for counties that show any level of coal mining during the study period. The research designs also include an indicator variable for counties with coal production levels above the median value. As research has shown, these counties suffer from an increased effect of coal mining on mortality rates (Hendryx & Ahern, 2009). Consequently, for this research project, coal mining is operationalized in terms of the presence of any coal mining combined with an additional indicator variable for above median levels of coal production.

While different states appear to have different baseline mortality rate levels, it seems that for coal producing states, the effect of coal mining on mortality (i.e. the slope) varies. Figure 3 shows the state-level trend line of the relationship between mortality rates and coal production. This variation is also supported on theoretical grounds. The type of coal that is produced and extraction techniques used for coal production differ across states. While, for example, Pennsylvania produces anthracite coal, states in the South produce coal with higher levels of sulfur (Hendryx, 2015; Hendryx & Holland, 2016). Furthermore, states in the Appalachians are more likely to employ mountaintop removal procedures that have been linked to higher mortality rates (Hendryx, Fedorko, & Anesetti-Rothermel, 2010; Hendryx & Holland, 2016; Hendryx et al., 2019). The multi-level regression model attempts to capture this varying effect of coal mining by including an additional state-level variance component that allows the slope of coal mining to vary around the population average effect.

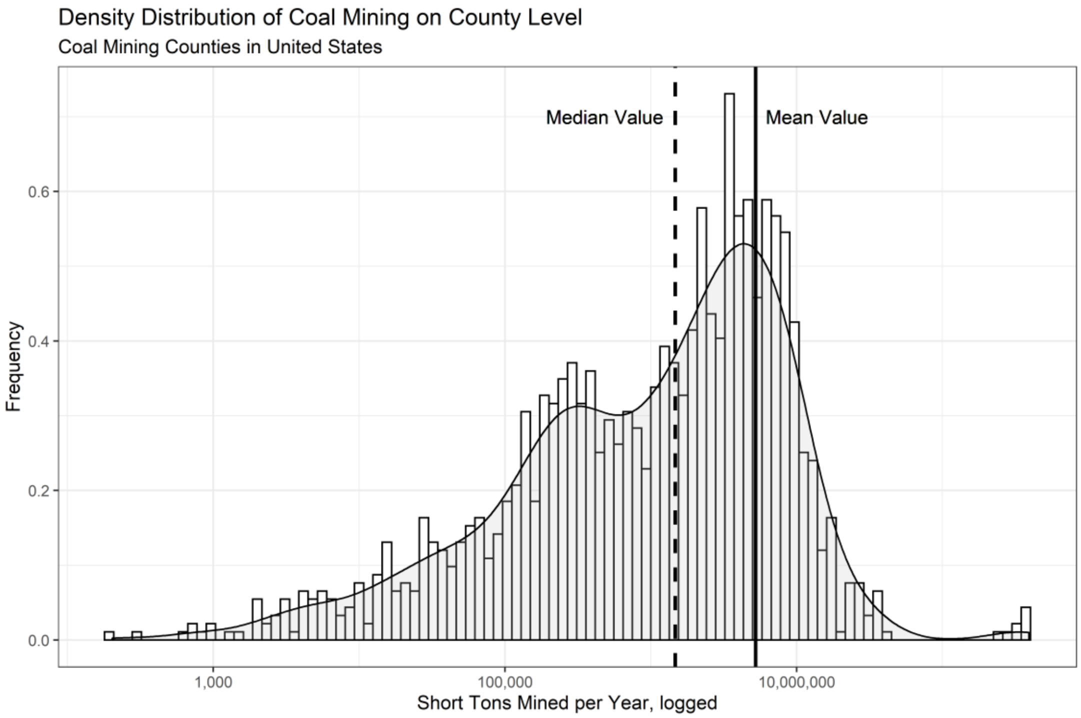


Figure 2. Distribution of County-Level Coal Production

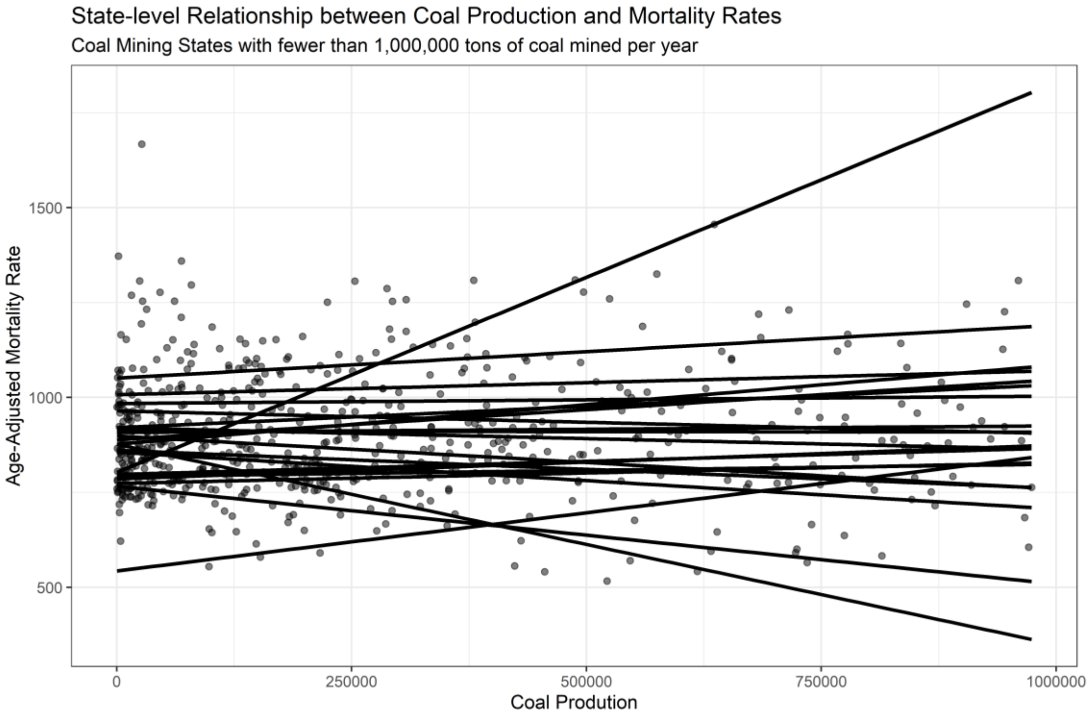


Figure 3. State-Level Relationship Between Coal Mining and Mortality Rates.

3.1.3 Control Variables at the County Level

While the influence of coal mining on mortality rates is the key focus of this study, there are other variables for which the model has to control. Presumably, the impact of these variables is more substantial than coal mining. The following section goes over control variables included in the multilevel model and variable operationalization.

Economic circumstances influence many life choices and the longevity of individuals. Thus, favorable economic circumstances are presumably tied to lower mortality rates (Adler & Ostrove, 1999; DeNavas-Walt, Proctor, & Smith, 2010). The literature generally includes income, unemployment rate, and poverty rate as covariates. Unemployment is measured as the county-level unemployment rate. Poverty is measured as the percent of households below the federal poverty line. Income is measured in terms of median household income (Borak et al., 2012; Hendryx & Ahern, 2009; Woolley et al., 2015).

Aside from economic factors, a statistical model of mortality rates should control for the demographic characteristics of the county. Increases in educational attainment on the aggregate level are generally associated with lower mortality rates. However, there are distinct differences in the effect of different levels of educational attainment (Desjardins & Schuller, 2006). Thus, education levels are included in the model in terms of the county-level rate of high school graduates and the rate of individuals within the county that hold a bachelor's or graduate degree.

Furthermore, age, gender, and racial make-up influence the general mortality of a county and are included as control variables. Higher average age, measured in county-level median age, is generally associated with high mortality rates (Anderson & Rosenberg, 1998). Since males generally experience shorter life expectations, the percentage of the county population that is male is commonly included in statistical modeling of mortality rates (Borak et al., 2012; Christian et al., 2011). Lastly, racial minorities suffer from shorter life-expectations on average and thus, the racial make-up of the county is included as a control. Race is measured in three separate variables as the percentage of the population that is Black, American Indian, or Hispanic.

County-level mortality rates are generally linked to the health characteristics of the respective county. Similar to other research designs (e.g., Christian et al. (2011); Esch and Hendryx (2011); Hendryx (2009)), several county level health indicators are included in the model. These include the county level smoking rate, obesity rate, and alcoholism rate. The measurement strategy is taken from the County Health Ranking.[2] Furthermore, access to healthcare is included as a control variable as two separate variables that measure the percentage of the total population that is uninsured and the proportion of the county population to the number of primary care physicians.

The general geographic characteristics of a county also influence mortality rates and hence are included as controls. Rural counties and counties in Southern states generally suffer from increased mortality rates (Hendryx, Fedorko, & Halverson, 2010). Thus, rurality is included into the modeling approach and is measured based on the Urban-Rural Continuum Code of the US Department of Agriculture (USDA, 2004). While USDA distinguishes between 9 different steps from urban to rural, previous studies

[2] See https://www.countyhealthrankings.org/. Obesity is measured as individuals with a BMI above 30. Smoking as individuals who report as current smokers. Alcoholism is measured as the percentage of adults who report heavy drinking. Accessed: 3/15/2020.

have reduced these to a dichotomous variable that takes on the value of 1 for counties with that are classified as nonmetropolitan (codes 4-9) and the value 0 for metropolitan areas (codes 1-3). Rurality is included into the study design based on this operationalization.

Counties that are part of a Southern state are measured by an indicator variable with Southern states being identified in accordance to previous research (Hendryx, 2009). Furthermore, research has indicated that after controlling for economic and demographic factors, Appalachian counties are disproportionally affected by the effect of coal mining on mortality rates (Borak et al., 2012; Hendryx, 2009; Woolley et al., 2015). Therefore, the modeling approach in this study includes an indicator variable that takes on the value of 1 for counties that are indicated as part of the Appalachian region by the Appalachian Regional Commission (ARC, 2020). As the effect of mining on mortality rates is theoretically related to environmental pollutants as well as particulate matter, the geographic size of a county influences the distance of individuals to coal mines and consequently the level of exposure (Hendryx & Ahern, 2008). On the aggregate level this variable is measured in county size in square miles as indicated by the Census Bureau (U.S. Census Bureau, 2010).

As the research design for this project includes observations of counties in the United States from several consecutive years, it is necessary to incorporate the effect of time into the modeling strategy. Time is measured as a counter variable that starts with a value of 0 for the first recorded year and increases in increments of 1 for each following year. As the effect of time does not necessarily have to be linear, the model also includes a quadratic term for the effect of time.

3.1.4 Interaction Terms and Formalized Model

Presumably, the effects of several control variables interact with coal mining or across covariates. Thus, the model includes five interaction terms to account for these interdependencies. The Appalachian region is one of the major coal mining areas in the United States. Several studies have found that counties in the Appalachian suffer at an elevated rate from increased mortality and diseases that can be associated with coal mining (Esch & Hendryx, 2011; Hendryx, 2009, 2015; Hendryx & Ahern, 2009). Furthermore, counties that produce above median levels of coal per year and are located within Appalachia suffer at an even higher rate than counties producing above the median level of coal per year outside of Appalachia (Borak et al., 2012; Hendryx, 2015). Consequently, the model includes a variable that accounts for an interaction between coal mining and Appalachian location and a variable that accounts for the interaction between above median mining levels and Appalachian location.

Furthermore, demographic covariates interact with the effect of coal mining on mortality rates. Coal mining facilities and the operation of these facilities generally induce an influx of qualified labor (Que, Awuah-Offei, & Samaranayake, 2015). The presence of coal mining thus might interact with the variables measuring county level educational characteristics. Thus, the model includes two additional terms accounting for interactions between coal mining and the county rate of high school graduate as well as the rate of individuals with a bachelor's or graduate degree. Lastly, research has shown elevated mortality rates for Hispanic males (Hoyert, 2012). Thus, the model includes a fifth interaction term that focuses on the interaction between the percentage of males of the county population and the percentage of the population that belongs to the Hispanic community.

The equation below shows the fully formalized multilevel regression model that predicts mortality rates (y) for county *j* in state *i*. The intercept for each state is indicated by the α_i term in the regression equation. Further, the δ_n coefficients indicate the interaction terms. As the state-level intercept is allowed to vary around the population intercept, ς_{1i} indicates the variance component for the random intercept. The second variance component, ς_{2i}, allows for the slope of coal mining to vary at the state-level around the population average effect of coal mining on mortality rates.

$$
\begin{aligned}
y_{ij} = \ \alpha_i + \ & \beta_1 \cdot medianmining_{ij} + \ \beta_2 \cdot appalachia_{ij} + \beta_3 \cdot rural_{ij} \\
& + \beta_4 \cdot unemployment_{ij} + \beta_5 \cdot income_{ij} + \beta_6 \cdot poverty_{ij} \\
& + \beta_7 \cdot age_{ij} + \beta_8 \cdot hsgrad_{ij} + \ \beta_9 \cdot bagrad_{ij} + \beta_{10} \cdot landarea_{ij} \\
& + \beta_{11} \cdot malepop_{ij} + \beta_{12} \cdot blackpop_{ij} + \ \beta_{13} \cdot amerindpop_{ij} \\
& + \beta_{14} \cdot hispanicpop_{ij} + \beta_{15} \cdot phyaccess_{ij} + \beta_{16} \cdot alcoholism_{ij} \\
& + \beta_{17} \cdot obesity_{ij} + \beta_{18} smoke + \beta_{19} time + \ \beta_{20} time^2 \\
& + \beta_{21} southernstate + \beta_{21} uninsuredpop \\
& + \ \gamma_{1i} \cdot coalmining \\
& + \delta_1 \cdot [coalmining \times appalachia]_{it} \\
& + \delta_2 \cdot [medianmining \times appalchia]_{it} \\
& + \delta_3 \cdot [coalmining \times hsgrad]_{it} \\
& + \delta_4 \cdot [coalmining \times bagrad]_{it} + \ \varsigma_{1i} \ + \varsigma_{2i} \cdot coalmining + \varepsilon_{ij}
\end{aligned}
$$

3.1.5. Variable Transformation

As the key independent variables are binary indicator variables, I am rescaling all continuous variables in the model by first centering the variables and then dividing them by two standard deviations. This technique has been argued for by Gelman (2008) in order to improve interpretability of regression coefficients in case of a mixture of continuous and binary input variables. Dividing by two standard deviations rather than one standard deviation, as is the usual process of variable standardization, allows for an easier comparison between regression coefficients of indicator and continuous variables. As Gelman argues, binary variables that are evenly distributed vary with a standard deviation of 0.5. Thus, when comparing the coefficient of a binary variable to the coefficient of a traditionally standardized continuous variable, the comparison overstates the effect of the binary variable. Consequently, all continuous input variables are standardized by two standard deviations.

3.2 Data

The study design combines data sets over a period from 2010 – 2017. One of the main objectives of the research design is to avoid aggregation over time, while preserving the county-level as the level of analysis. This focus comes with the trade-off of data availability. Reliable county-level information about the covariates included in the statistical model are not readily available pre-2010. This is mainly due to the fact that prior to 2010 the American Community Survey (ACS), the main source for covariate information, was based on a sample of all counties. Only after 2010 did the ACS produce estimates about the economic and demographic factors present in all counties of the United States. Rather than imputing values for counties not included prior to 2010, I am limiting the time frame for my research design to 2010 – 2017.

Sampling of the observed counties is not necessary, as the unit of analysis are counties within the United States, and it is reasonably possible to collect information on every single county. However, the CDC does not provide information on mortality rates for counties with less than 20 deaths. A further discussion of this issue and a list of all counties that are excluded from the research design can be found in the Appendix A. In total, the number of counties included in the research design are 2994, which sums up to 23,952 observations over the period of study of 8 years (2010-2017).

Information on the variables included in the statistical model were collected from a variety of publicly available sources. Table 1 shows a list of institutions and data sources included in the research design. Mortality rates were retrieved from the Compressed Mortality File made available through the National Center for Health Statistics (NCHS, 2017). The data file contains information on age group mortality, crude and age-adjusted mortality rates, and causes of death. Further, information on annual county-level coal production was retrieved from the Energy Information Administration (EIA, 2020). Information on control variables, including economic and demographic, and geographic characteristics, were collected from the ACS conducted by the U.S. Census Bureau (U.S. Census Bureau, 2018). Lastly, health indicators for each county were collected from County Health Rankings & Roadmaps. This institution compiles information from the CDC's Behavioral Risk Factor Surveillance System survey and makes health indicators for the county level publicly available. Information curated by County Health Rankings has been used in the literature (Hendryx & Holland, 2016). Missing values for health indicators on the county-level were imputed as the respective state averages for the respective year.

Table 1. Data Sources

Institution	Information	Data Source
National Center for Health Statistics	Mortality Rates	Compressed Mortality File, 1999-2017
Energy Information Administration	Coal Production	Coal Data Browser, 2010-2017
U.S. Census Bureau	Economic and Demographic Control Variables	American Community Survey, 2010-2017
Appalachian Regional Commission	Appalachian Counties	Counties in Appalachia 2020.
County Health Rankings	County Health Indicators	County Health Rankings, 2010-2017
U.S. Department of Agriculture	Rural Counties	Rural-Urban Continuum Code 2013.

3.3. Descriptive Statistics

Table 2 shows descriptive statistics for all continuous variables. State-level descriptive statistics for indicator variables can be found in Appendix B. The skewed distribution of annual coal production becomes obvious from the data presented in Table 2. Annual production ranges from 0 to almost 400,000,000 with a median of 0 and a mean of 315,058.5 tons of annual coal production. Furthermore, the standard deviation of 6,363,854 indicates strong variation in the data. The clearly non-normal distribution points towards the rationale of measuring coal mining in terms of a binary variable. Mortality rates also show a wide spread from 227 to 1,793.6 annual deaths per 100,000 population and considerable variation in the data with a standard deviation of 151. However, the mean and median value are very close together, especially when considering the wide spread of the distribution.

The economic covariates included in the model show a wide range with a concentration around the measures of centrality. The median and mean values for household income differ by about $2,000 around $45,000. While the distribution shows a wide range, the distribution is narrowly spread around the mean value with a standard deviation of $12,276. Similarly, median and mean values for educational attainment and poverty rate differ by 1-2 percentage points from each other. In both cases the standard deviation indicates a somewhat narrow spread around the mean value. In case of median county age, the distribution is even closer, with median and average age being 0.1 years apart from each other and a narrow spread with a standard deviation of 5 years. Furthermore, the median age of 40.5 roughly splits the distribution ranging from 21.4 – 66.4 in half.

Table 2. Descriptive Analysis Continuous Variables

Variable	Minimum	Maximum	Median	Mean	SD
Mortality	227	1,793.6	812.1	823.831	151.817
Coal Production	0	392,528,314	0	315,058.5	6,363,854
Unemployment	0	29.9	7.6	8.01075	3.481
Income	18,972	129,588	44,567.5	46,488.87	12,276.01
HS Graduates	6	73.9	35.2	34.997	7.505
BA or Higher	0	80.2	18.1	20.322	9.65817
Poverty Rate	1.1	52	15.7	16.374	6.3776
Age	21.4	66.4	40.5	40.4	5.00484
Male Population	37.4	80.8	49.5	49.957	2.26205
American Indian	0	87	0.3	1.693	6.6922
Black	0	86.9	2.4	9.286	14.632
Hispanic	0	99.2	3.5	8.416	13.17435
Alcoholism	0	42.3	15.46324	15.309	4.53545
Obesity	10.7	48.1	30.3	30.269	4.28535
Smoking	0	51.1	19.787	20.234	5.16706
Uninsured	2.721	48.4	16.7	17.118	5.73218
Prim Care Access	158.243	24,939	1,971.25	2,573.647	2,136.795
County Size	2	145,504.79	602.76	1,094.38	3,647.492

Figure 4 shows the distribution of population measurements. While the distributions are generally wide, they are narrowly distributed around the median value. Standard deviations for population measurements range between 6-14 indicating a narrow to moderate spread relative to the wide range of the distribution. The long tail of the distribution skews the mean value of the distribution towards the outliers of the distribution. An example of a county on the right of the distribution is Starr County, Texas with a Hispanic population of about 99% of the total population for the years 2016 and 2017. However, the mean value is within one standard deviation from the median value. Thus, further transformation is not necessary before fitting the data.[3]

The distribution of county size in square miles displays a similar spread with a number of very large counties in Alaska that are between 20 and 40 standard deviations bigger than the mean value. Furthermore, independent cities in Virginia introduce very small values into the data (about 2 square miles for the smallest city). This wide spread is reflected in the substantial difference between the median and mean value as well as the substantial standard deviation. However, as the outlier and influence analysis in the next chapter shows, transforming county size by a logarithmic transformation did not deal with this problem appropriately.

[3] Outlier identification and treatment after fitting a model to the data will be discussed in the results section.

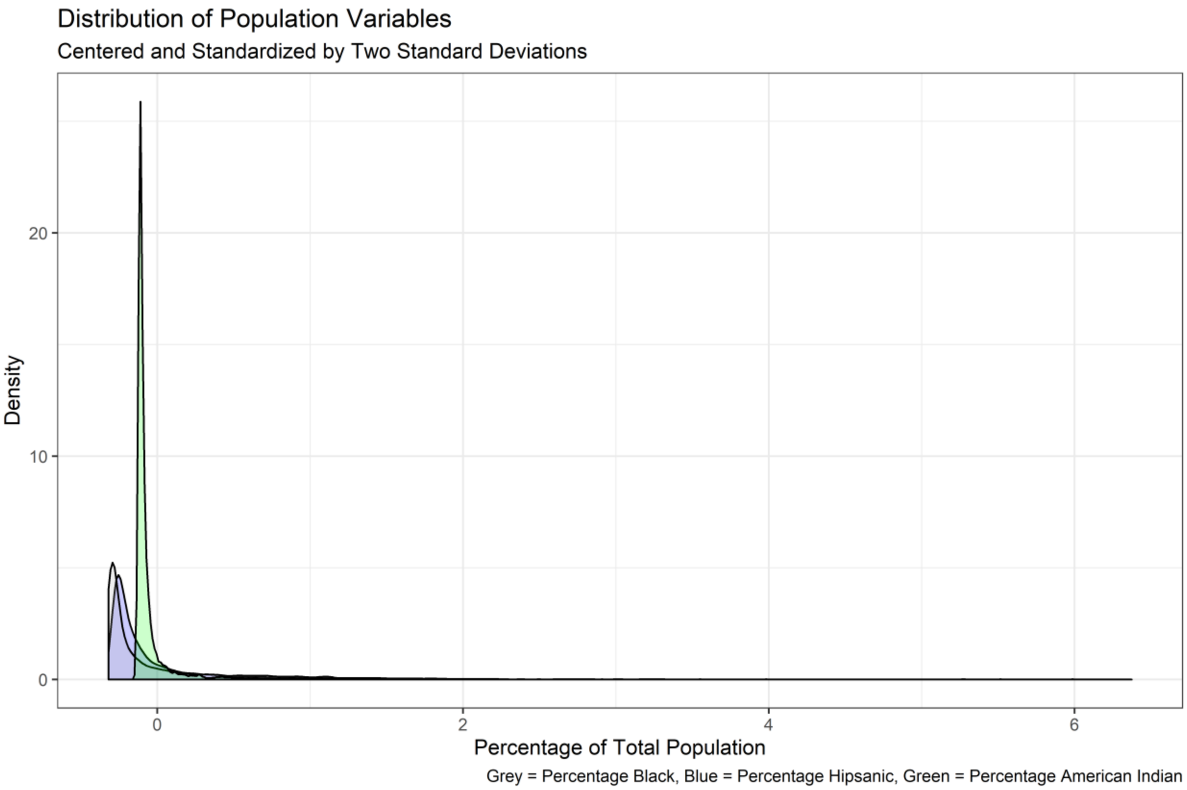


Figure 4. Population Distribution - Control Variables

CHAPTER 4. RESULTS

In the following section, I will present the results of the multilevel regression model. The model was fit to the data in R using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015; Doran, Bates, Bliese, & Dowling, 2007). While the lme4 package has some limitations in fitting multilevel regression models, especially in regards to the selection of a variance-covariance matrix, the package is widely used and is supported by a variety of additional packages (Gelman & Hill, 2007; Nieuwenhuis, Grotenhuis, & Pelzer, 2012). As the calculation of P-Values for multilevel regression models is subject to academic debate, the lme4 package does not provide P-Values (Bates et al., 2015). In the presentation of my work, I will report P-Values but, following conventions suggested by Wasserstein and Lazar (2016), I will be focusing on statistical uncertainty in terms of confidence intervals. P-Values are calculated based on the methods discussed in Kuznetsova, Brockhoff, and Christensen (2017).

4.1. Influence Analysis

As indicated by the descriptive analysis, there are a number of observations that have the potential to be outliers. Observations that lie on the far-right end of the racial make-up distribution, and counties in Alaska that are several magnitudes larger than the mean value for county size can overly influence the regression analysis.[4] In order to

[4] Logarithmic transformations are a common tool to handle outlier influence before the regression is fit. However, when using logged values for potentially influential variables, the influence of outlier groups

produce reliable and accurate statistical estimates, overly influential data should be removed from the data. The influence analysis is conducted before the final model is fit to the data, as residual diagnostics cannot fully account for the influence of outliers (Nieuwenhuis et al., 2012; Van der Meer, Te Grotenhuis, & Pelzer, 2010). Statistical influence is understood as the power a single observation has on the estimated regression parameters. Highly influential observations can skew the parameters and consequently negatively influence the accuracy of coefficients, confidence intervals, and the generalizability of the results (Imai, 2017; Van der Meer et al., 2010).

In general, statistical influence can be estimated by iteratively excluding observations and fitting the regression model without the respective observation. In the next step regression parameters are compared between the model that was fit to the full data and the model fitted to the data excluding the *i*-th observation. The changes in regression parameters indicate the influence the respective observation has over the model estimates (Imai, 2017). In the case of multilevel regression, however, Van der Meer et al. (2010) argue that the grouping variable, as an essential part of the regression model, should also be considered as a potential source of influence. The statistical method developed to identify influential groups applies the same logic of iteratively deleting cases, to the grouping variable. Consequently, influence measures are calculated based on iteratively fitting models to data that exclude all observations belonging to the *i*-th group and comparing the parameters to the model that is based on the full data (Nieuwenhuis et al., 2012; Van der Meer et al., 2010).

and variables increased rather than decreased. Thus, linear values are used for a full outlier analysis and treatment.

The left graph in Figure 5 shows the influence of the state-level groups on the entire model. Influence has been calculated as Cook's distance which is an influence measure that considers the influence of data points on all model parameters. Figure 15 through Figure 16 in Appendix C take a closer look at the influence of state-level groups on specific variable coefficients. Overall, Figure 5 shows a clear outlier influence of variables grouped under Texas. The solid line indicates the mean value of Cook's distance, while the dashed line shows the rule of thumb cut-off value of three times the mean value. While there are several groups that go beyond this cut-off value, Texas clearly outweighs all other groups by a factor of two to three.[5] Thus, some form of outlier treatment should be undertaken.

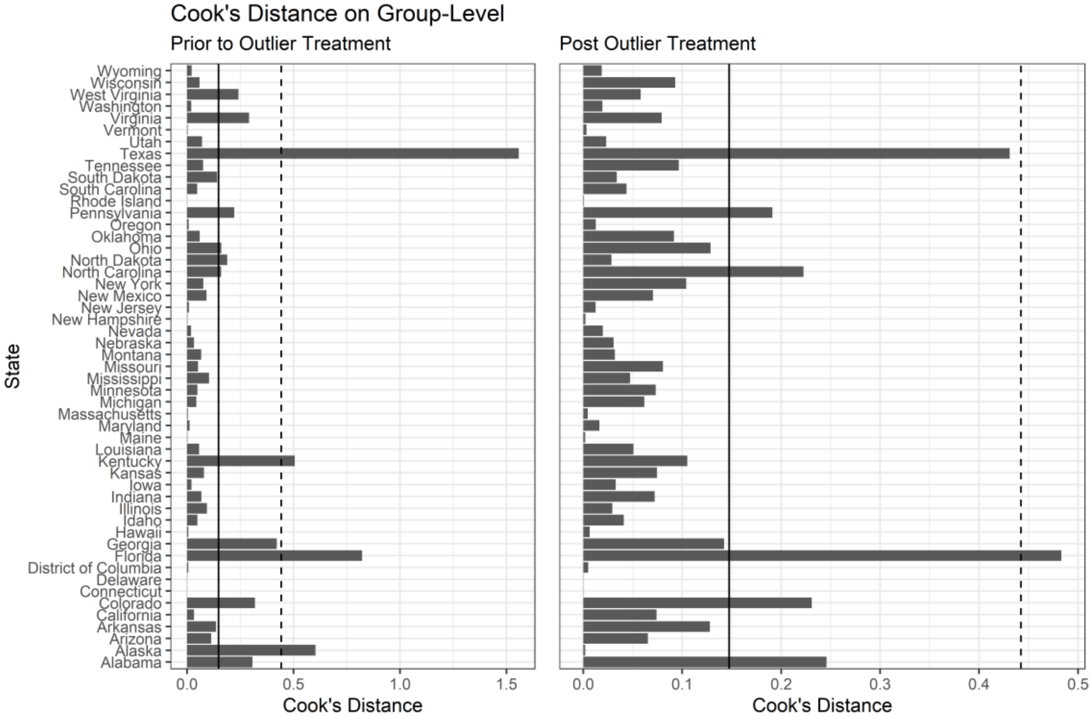


Figure 5. State-Level Influence Measures

[5] As the analysis discussed in the appendix shows, observations that are part of Texas have a strong influence on the beta-coefficient of the percentage of the population that is Hispanic.

However, excluding Texas entirely from the regression model would eliminate 223 counties per year which adds up to 7.5% of the entire data. Hence, influence statistics were calculated for every single observation in the data. Table 3 shows the summary statistics for Cook's distance on the observation-level. The values are widely spread with the with maximum observation being 68.35 standard deviation away from the smallest observation. Furthermore, the distribution has a long right tail with mean and median observation being much closer to the minimum than to the maximum observation. This distribution is not surprising as only a few observations should have a high influence value. While the number of overly influential observations is still large, it is more widely spread across all states. Table 10 and Table 11 in the Appendix present summary statistics by state and year for observations that are excluded from the regression analysis.

Table 3. Summary Statistics,
Observation-Level Influence Measures

Measure	Value
Minimum	5E-11
Maximum	0.0175
Mean	5E-05
Median	7E-06
SD	0.0003
Sum of Outliers	1521 (6.35%)

Excluding overly influential observations from the data reduces the total amount of observations by 6.35%. However, these are somewhat similarly distributed across states and years with an average reduction of 6.8% per state and 6.4% per year. The only exception to this is Alaska with a reduction of 35% of all observations. However, this is hardly surprising, as Alaska is a state that does not compare to any other state in the US. Utilizing observation-level influence treatment rather than group level treatment, thus, reduces the number of excluded observations while preserving all groups. In the case of Texas, the elimination rate is reduced from 100% to 11.8% and the influence is strongly reduced as the graph on the right-hand side of Figure 5 shows.

4.2. Regression Results

Table 4 shows the regression output of the multilevel regression model and Table 5 shows goodness of fit measures for the model. As all continuous input variables are standardized, the intercept α_i indicates the average state-level intercept when the continuous variables are held at their averages and the indicator variables are 0. The variance component σ_{ς_1}indicates the standard deviation of the variance at which states vary around the intercept. Thus, on average, states vary around the intercept α_i by about 41 deaths per 100,000 population. The effect of time indicates that while mortality rates decrease over time, this reduction is moderated by the quadratic effect of time.

Figure 6 shows the estimated variable parameters as well as the respective 95%-confidence intervals. The x-axis indicates the size of the estimated parameters and red line highlights zero. As all continuous variables are standardized, the magnitude of the effects can be compared in terms of change in mortality rates for a two standard deviation change in the independent variable. Point estimates that are marked as blue are statistically significant, while for red points the associated confidence intervals include zero. The effects are ordered by the magnitude of the estimated parameter effect from the most negative to the most positive effect size.

Table 4. Regression Output

Parameter	Estimate	SE
Intercept (α_i)	794.622*	6.851
Coal Mining	-17.734	16.027
Above Median Mining	9.001	9.356
Appalachia	4.837*	2.238
HS Grad Rate	12.746*	1.591
BA Grad Rate	-69.241*	1.962
Male Population	-21.879*	1.323
Hispanic Population	-58.921*	1.944
Coal Mining x Appalachia	44.394*	15.130
Above Median Mining x Appalachia	35.516*	11.619
Male Population x Hispanic Population	15.334*	2.169
Coal Mining x HS Grad Rate	-29.803*	7.451
Coal Mining x BA Grad Rate	-26.317*	8.089
Poverty Rate	46.047*	2.640
Median Age	-32.396*	1.528
Black Population	-8.039*	1.873
Southern State	54.314*	13.212
Rural County	-2.803*	1.368
Unemployment Rate	26.824*	1.855
American Indian Population	22.160*	1.916
Median Income	-40.337*	2.620
Physician Access	-10.248*	1.172
Uninsured Population	-31.009*	1.900
Alcoholism Rate	-10.864*	1.641
Obesity Rate	12.055*	1.729
Smoking Rate	35.002*	1.534
Time	-8.251*	0.903
Time squared	1.839*	0.121
County Size	-3.474	2.598
σ_{ς_1}	41.098	
σ_{ς_2}	69.828	
$\rho_{\varsigma_1 \varsigma_2}$	-0.367	
σ_{ε}	76.906	
Dependent Variable = County-Level Mortality Rates, * = statistically significant at $p < 0.05$		

Table 5. Regression Output. Goodness of Fit Measures

ICC	0.526
AICc	25,8709
BIC	25,8974
N	22,431
Groups	51
Multilevel Regression Model, with State as Grouping-Variable. Dependent Variable = County-Level Mortality Rates.	

The presence of any coal mining as well as coal mining above the median production level are not significantly associated with changes in county-level mortality rates. Furthermore, the associated 95%-confidence intervals are widespread indicating a large amount of statistical uncertainty. However, the effect of coal mining and mining above the median production level is included in several interaction terms. While the variable is significantly associated with increased mortality rates as a main effect, the interaction terms show statistically significant changes in mortality rates. Thus, for coal mining counties in Appalachia, mortality rates are increased by about 44 deaths per 100,00 population. Furthermore, for coal mining counties in Appalachia that produce above the median coal production level, mortality rates are increased by about 80 deaths per 100,000 population.[6]

The presence of coal mining moderates the effect of higher average levels of education significantly. In the case of high school graduate rates, a cross-over interaction is occurring. For coal mining counties, a two standard deviation increase in

[6] While the interaction effects are listed separately in the regression output, it is logically impossible for a county in Appalachia to produce above the median coal mining level but not be a coal mining county. Thus, the effect can be combined.

high school graduates is associated with a decrease in mortality rates by 17 deaths per 100,000 population. For non-coal mining counties, a two standard deviation increase in high school graduate rates is associated with an increase in mortality rates of approximately 13 deaths per 100,000 population. The effect of increases in the rate of individuals with a college or graduate degree, is magnified for coal mining counties. Holding other variables constant, a two standard deviation increase is associated with a decrease in mortality rates by about 95 deaths per 100,000 population. For non-coal mining counties, a two standard deviation increase in college and graduate degree holders is associated with a decrease of mortality rates by 69.

The variance component σ_{ς_2}indicates the state-level variation in the slope of coal mining. On average, the effect of coal mining on county-level mortality rates varies by about 70 deaths per 100,000 population depending on the state the county is located in. This wide variation also explains the wide 95%-confidence intervals for the effect of coal mining. Furthermore, $\rho_{\varsigma_1\varsigma_2}$ indicates the correlation between random slopes and random intercepts. The negative correlation coefficient indicates that states with higher intercept values tend to have smaller values for the slope of coal mining on mortality rates. This relationship can also be seen in Figure 7. The values for random slopes and intercepts are points estimates at the end of a large number of iterations. Variance components are estimated by a stochastic procedure, and it should be noted that there is uncertainty associated with point estimates. The overall relationship expressed in the correlation coefficient can nonetheless be established.

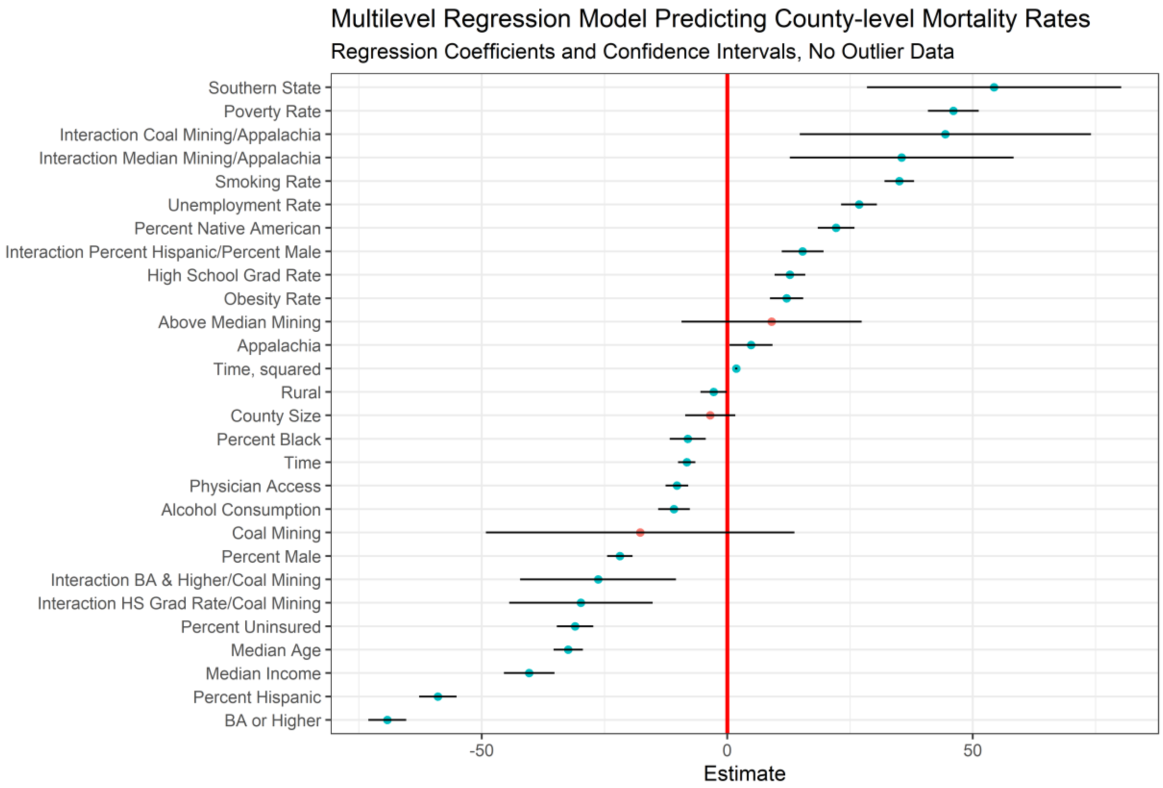


Figure 6. Variable Coefficients

The effect of geographic covariates is somewhat mixed. While the effect for the size of counties in square miles is not statistically significantly different from zero, mortality rates for rural counties are on average 3 deaths per 100,000 population lower. However, in both cases the effect is relatively small in substantial terms. Thus, rurality as well as county size are not substantively important for the prediction of county-level mortality rates. However, for counties in Southern states mortality rates are significantly and substantially different from zero. On average, mortality rates in Southern states are increased by 54 deaths per 100,000 population. In relative terms, the increase associated with being located in a Southern state is the strongest increasing factor on mortality rates.

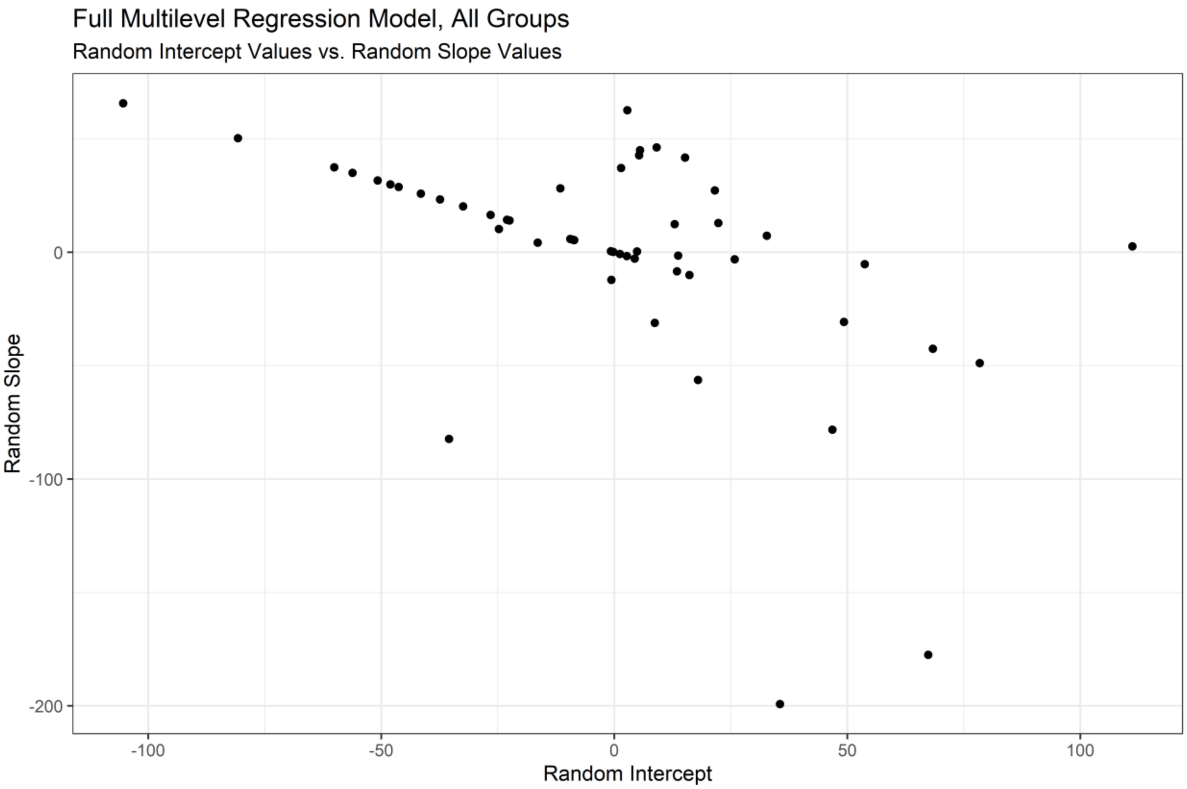


Figure 7. Correlation Random Slope and Intercept

The estimated effects for economic covariates confirm the expected results. On average an increase in poverty rates by two standard deviations is associated with an increase in mortality rates by 46. Similarly, a two standard deviation increase in unemployment while holding other variables constant, is reflected in an increase in mortality rates by about 27 deaths per 100,000 population. An increase by two standard deviations in the median income level, on the other hand, is associated with a decrease in the predicted mortality rate by 40. In relative terms to other variables, the effects associated with economic covariates have substantial influence on county-level mortality rates.

For demographic covariates, the effect sizes differ considerably. Increases in the black population of a county are on average associated with a mild decrease in mortality rates (8 deaths per 2 standard deviations). However, increases in the American Indian population are on average associated with a more substantial increase in mortality rates (22 deaths per 2 standard deviations). The change in mortality rates associated with the Hispanic population of a county is substantially larger. On average and while holding other variables constant at their mean values, a two standard deviation increase in the Hispanic population of a county is associated with a decrease in mortality rates of 59 deaths per 100,000 population. However, this effect is moderated by changes in the percentage of the male population. An increase in male population by two standard deviations on average reduces mortality rates by 22. The interaction between age and Hispanic population indicates that the effect of an increase in both male population and Hispanic population by two standard deviations each is reduced by 15 to a reduction of 66 deaths per 100,000 population. Without the moderating effect of the interaction, a reduction of 81 deaths per 100,000 population would be expected from the coefficient of each main effect.

Health indicators, furthermore, have a substantial influence on mortality rates. Variables measuring access to healthcare indicate that an increase in the ratio of primary care physicians to the total population by two standard deviations is associated with a decrease in mortality rates by 10 deaths per 100,000 population. Further, increases in the rate of smoking and obesity are associated with an increase in mortality rates. In contrast to the expected effects, increases in alcoholism and the percentage of the population that is uninsured are associated with reduced mortality rates.

4.3. Regression Diagnostics

Similar to conventional OLS regression, it is necessary to perform a diagnostic analysis on a multilevel regression model. In particular, regression diagnostics should check the produced residuals on the observation-level as well as the group-level. The Figure below shows the total residual distribution and the plot of fitted values against the actual values. The residuals appear to be normally distributed with a mean value close to zero. Further, when plotting fitted values against the actual values in the data, Figure 8 shows a clear diagonal trend that indicates the ability of the model to predict the dependent variable. There appear to be a few cases of outliers on the left and right of the plot.

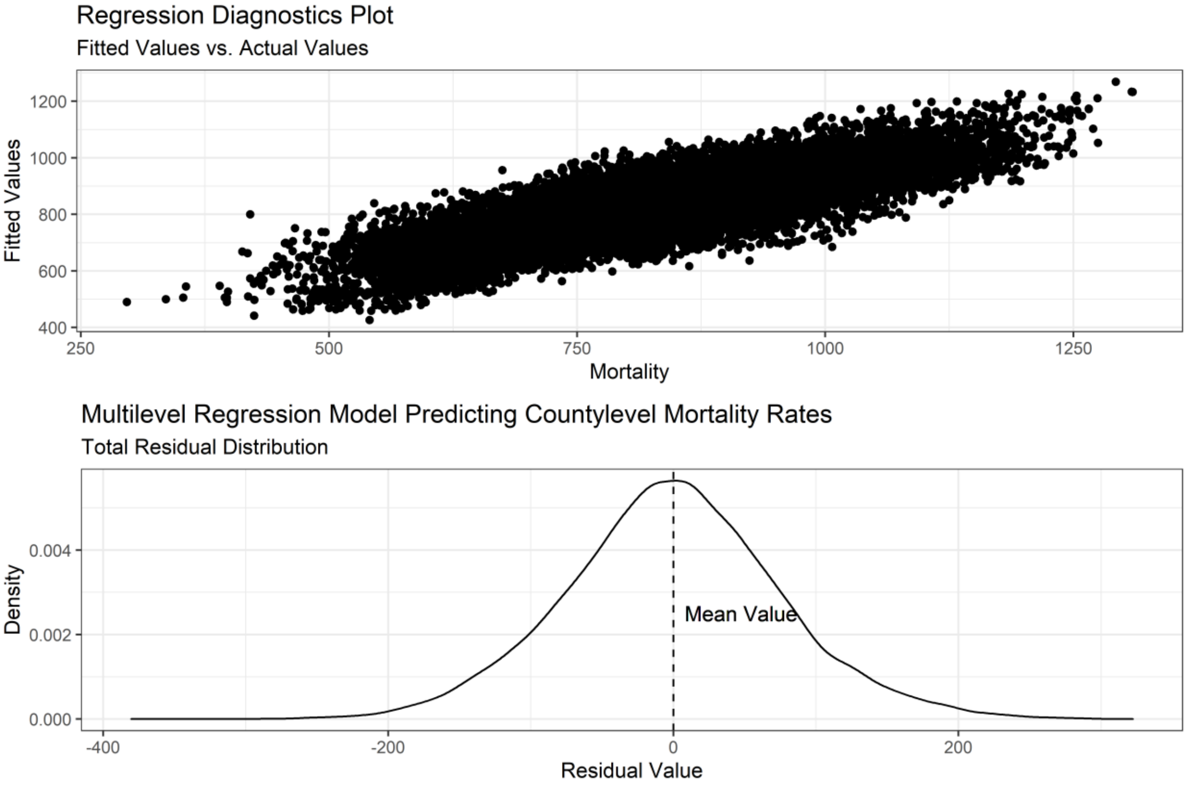


Figure 8. Regression Diagnostics, Residual Distribution

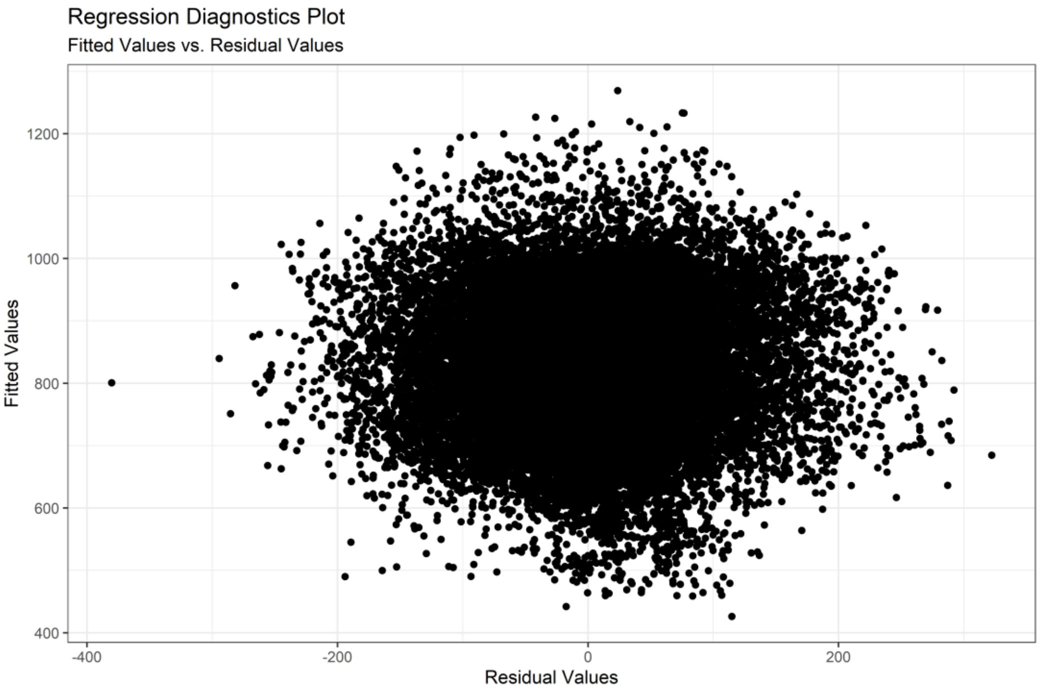


Figure 9. Residual Diagnostics, Fitted Values vs. Residual Values

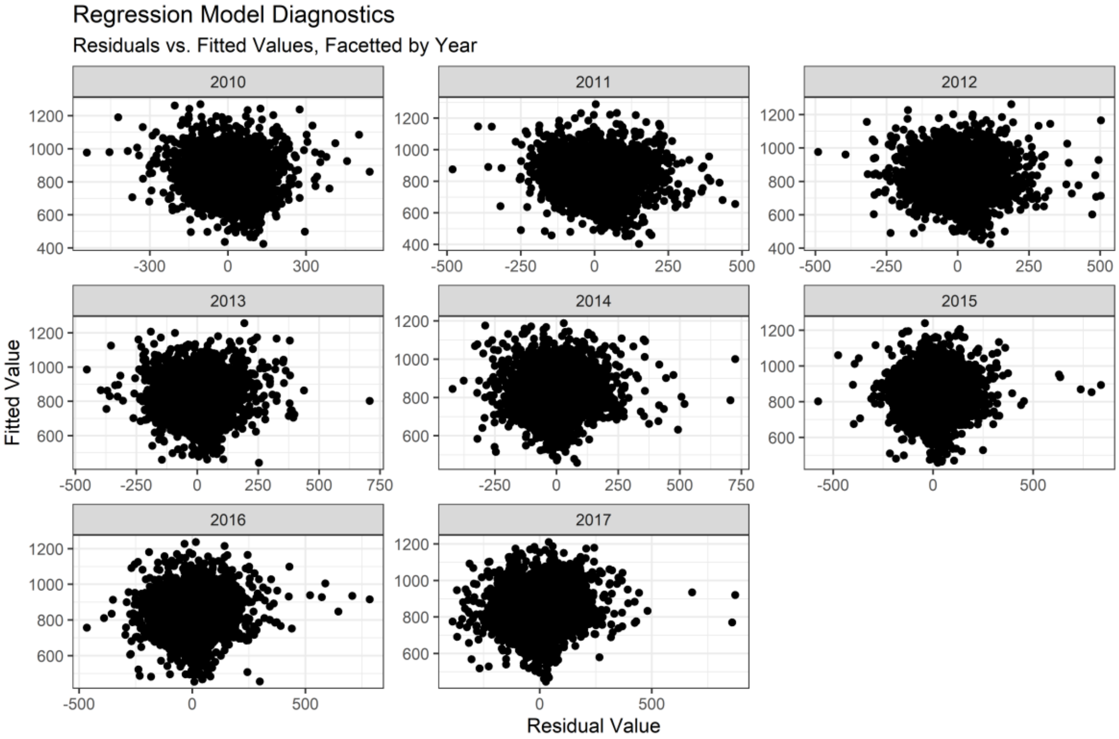


Figure 10. Residual Diagnostics, Fitted Values vs. Residual Values facetted by Year

Furthermore, Figure 9 plots fitted values against residual values. As becomes clear from the plot, there is no obvious pattern in the data that would indicate heteroskedasticity in the residuals. However, as mentioned above, in the case of multilevel regression it is essential to check the residual distributions by grouping variables. Thus, the figures included in Appendix E show univariate residual distribution as well as scatterplots for the relationship between residual values and fitted values broken down by state.[7] Figure 10 shows the relationship of residuals to fitted values broken down by year.

The regression model does not include year as a grouping variable, but the longitudinal character of the research design prompts the necessity of checking annual residuals. The yearly plots, overall, show a similar distribution as Figure 9. However, while there are fewer outlier combinations of residual and fitted values, there seems to be a slight dent in residual distribution on the lower right of the scatterplots for 2014-2017. In total, this does not sum up to a clear pattern of the yearly plots and there does not appear to be a trend present in the residuals.

Apart from the residual values, the variance components of multilevel regression models also require to be inspected. Figure 11 shows the estimated random intercepts and random slopes and includes the respective confidence intervals. The confidence intervals are centered around the median value of the iteratively produced random

[7] As there are 51 states/groups included in the analysis, these plots take up substantial space. They are, thus, included in the Appendix together with a discussion of the results. Furthermore, the single figures are excluded from the list of figures as the value of the plots stems from the combination of all plots rather than each single plot by itself.

effects. A central assumption of multilevel regression models it that while effects are allowed to vary around the population average, this variation will on average be equal to zero. Figure 11 shows that while for some states the parameter effect varies significantly from the population average, for most states the confidence interval includes zero. Thus, overall the random effects can be expected to average zero.

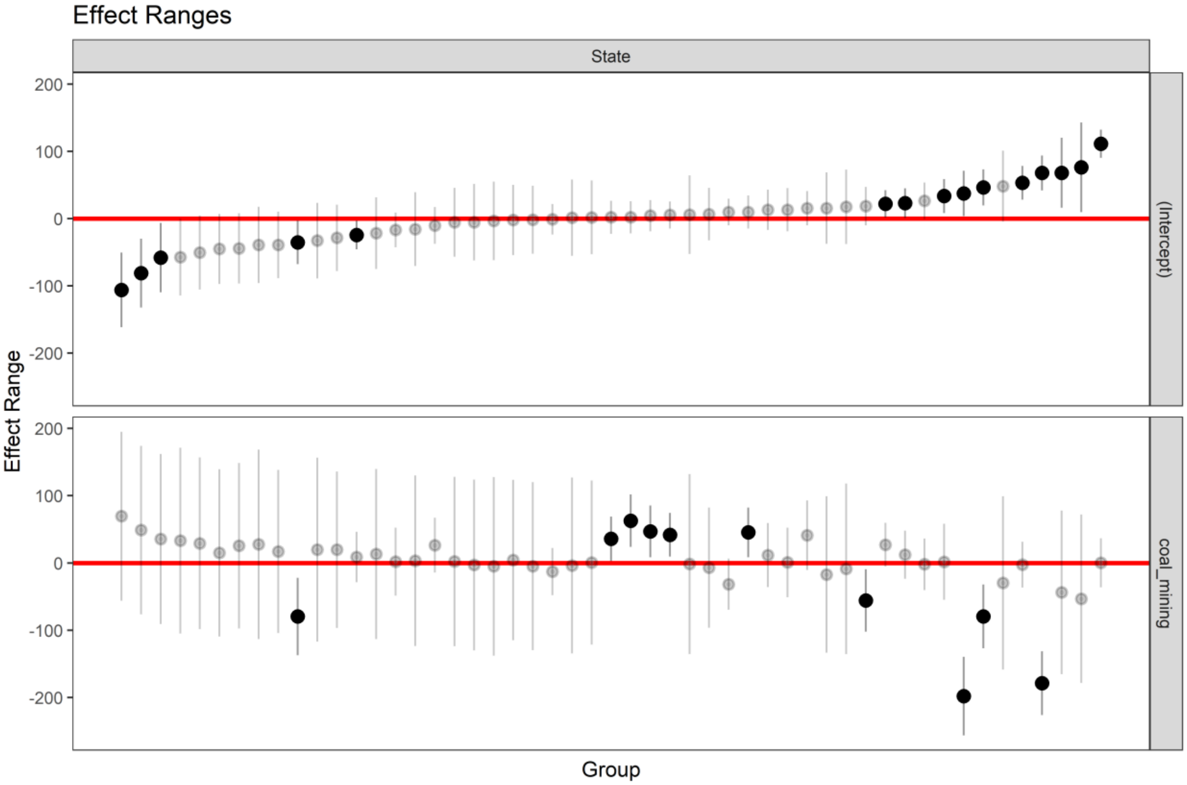


Figure 11. Regression Diagnostics, Variance Components

The residuals of the regression model are randomly distributed with a mean of zero. Furthermore, the distribution broken down by year does not show a pattern in the relationship between fitted values and residuals. Broken down by state the residuals appear to be normally distributed most of the time. However, as the number of observations per state varies greatly, the residuals are not normally distributed for all states.

4.4. Model Comparisons

The predominantly applied methodological strategy in the literature, as discussed in section 2.3. computes the averages of county-level observations and fits an OLS regression model to the averaged data. The purpose of this section is to compare the model discussed above to a model based on the averaging procedure commonly applied in the literature. Furthermore, to illustrate the reasoning behind multilevel regression, the multilevel model is also compared to an OLS regression model that is fit to completely pooled data. Pooled data disregards the clustered character of longitudinal data and treats all observations as independent from each other.

Table 6 shows regression outputs and summary statistics for the three models. The comparison between the multilevel regression model and the model that was fit to the averaged data, the difference in estimated standard errors becomes obvious. It should be noted that the summary statistics cannot be compared, since the models were fit to substantially different data. However, the statistical uncertainty of the model predictions can be compared. Standard errors for the averaged model are often several times larger than the standard errors for the multilevel regression model that was fit to a data set with much higher levels of variance. Table 6 also highlights the substantial difference in the number of observations between the two models. The results in Table 6 show that while the OLS regression model is fit to a smaller data set with fewer variation in the data, the model estimates are associated with more statistical uncertainty.

When comparing the multilevel regression model to the OLS model that was fit to pooled data, the advantages of the multilevel regression model become clear. As both models are fit to generally the same data set, the summary statistics in Table 7 can be compared. The substantial differences in the Bayesian Information Criteria (BIC) and the corrected Akaike

Table 6. Model Comparison

	Multilevel Regression		Averaged Data OLS		Pooled Data OLS	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	794.622*	6.851	808.529*	3.266	799.212*	1.550
Coal Mining	-17.734	16.027	8.521	12.484	18.352*	7.098
Above Median Mining	9.001	9.356	15.491	15.942	2.070	8.746
Appalachia	4.837*	2.238	-11.681*	5.102	2.303	2.401
HS Grad Rate	12.746*	1.591	11.221*	5.074	19.326*	1.860
BA Grad Rate	-69.241*	1.962	-71.082*	6.261	-68.964*	2.349
Male Population	-21.879*	1.323	-21.645*	3.087	-22.283*	1.490
Hispanic Population	-58.921*	1.944	-36.556*	4.260	-41.982*	1.872
Coal Mining x Appalachia	44.394*	15.130	20.098	16.680	11.722	8.916
Median Mining x App.	35.516*	11.619	30.672	20.727	48.926*	11.426
Male Pop. x Hisp. Pop	15.334*	2.169	15.342*	4.256	15.629*	2.063
Coal Mining x HS Grad Rate	-29.803*	7.451	-52.758*	14.275	-28.270*	6.733
Coal Mining x BA Grad Rate	-26.317*	8.089	-41.698*	15.011	-32.245*	6.993
Poverty Rate	46.047*	2.640	57.473*	7.145	67.125*	2.906
Median Age	-32.396*	1.528	-28.901*	3.716	-29.284*	1.696
Black Population	-8.039*	1.873	-8.989*	4.168	-10.627*	1.917
Southern State	54.314*	13.212	34.616*	4.300	52.745*	1.945
Rural County	-2.803*	1.368	-0.147	3.513	-0.147	1.696
Unemployment Rate	26.824*	1.855	6.800	4.200	2.016	1.825
American Indian Pop.	22.160*	1.916	18.543*	3.464	24.726*	1.647
Median Income	-40.337*	2.620	-8.019	6.782	-18.859*	2.903
Physician Access	-10.248*	1.172	-16.239*	3.098	-7.886*	1.421
Uninsured Population	-31.009*	1.900	-10.107*	4.504	-17.133*	1.859
Alcoholism Rate	-10.864*	1.641	-45.015*	3.785	-30.205*	1.666
Obesity Rate	12.055*	1.729	24.019*	4.418	28.010*	1.728
Smoking Rate	35.002*	1.534	63.780*	4.065	39.114*	1.658
Time	-8.251*	0.903				
Time squared	1.839*	0.121				
County Size	-3.474	2.598	-6.592*	3.104	-3.094*	1.501
σ_{ς_1}	41.098					
σ_{ς_2}	69.828					
$\rho_{\varsigma_1\varsigma_2}$	-0.367					
σ_{ε}	76.906					
N	22,431		2,994		23,952	
AICc	258,709		34,268		289,883	
BIC	258,974		34,436		290,109	

Information Criteria (AICc) for each model show that incorporating the clustered structure of the data into the regression analysis greatly improves the fit of the model. Furthermore, the residual diagnostics shown in Figure 12 show the improved fit of the multilevel regression model. The bottom left plot in Figure 12 shows a more pronounced pattern in the distribution of residuals compared to the distribution of residuals of the multilevel regression model (also discussed in section 4.3.). The top half plots of observed against fitted values reveal that the multilevel regression model achieves a tighter fit of the predicted values to the data. The model estimated on the pooled data over and underpredicts mortality rates and consequently, provides less reliable results.

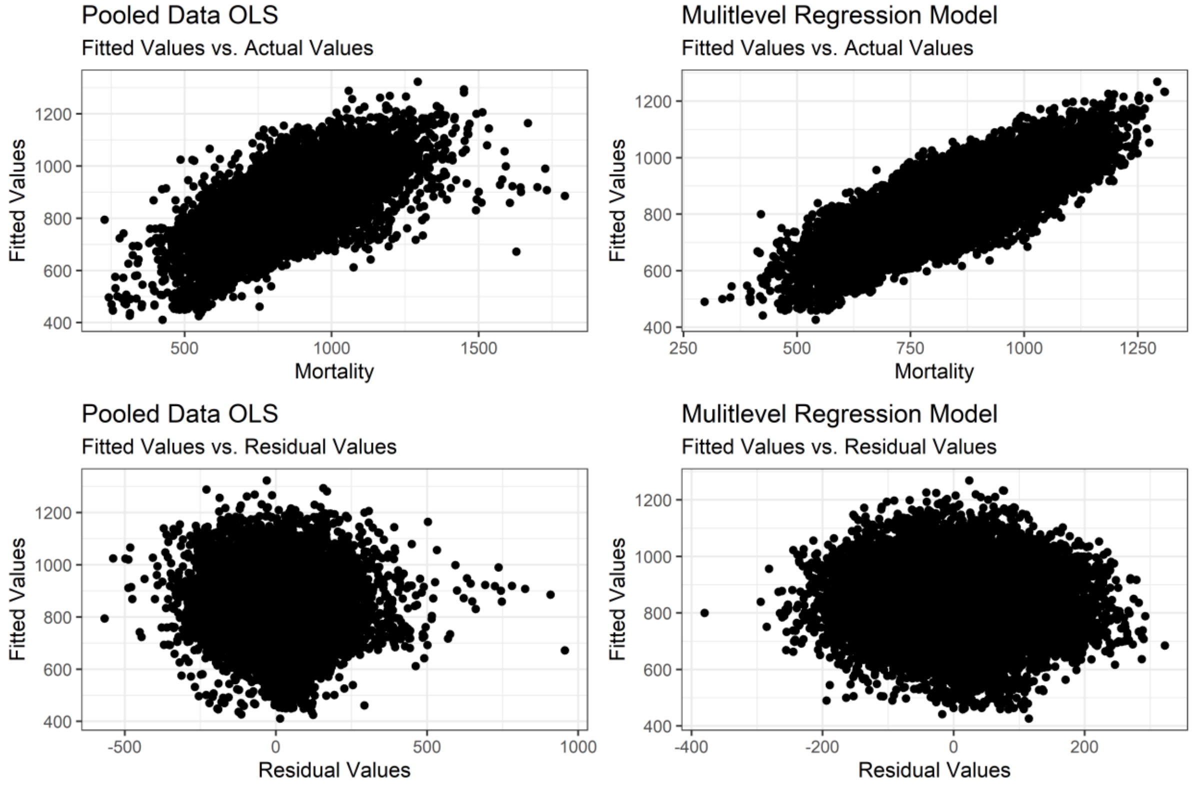


Figure 12. Model Comparison: Residual Diagnostics

CHAPTER 5. DISCUSSION

In this section I will go over the implications of the regression results for my hypotheses and then elaborate on the consequences of my results for the broader literature. According to H_1, counties with coal mining are assumed to have higher mortality rates compared to counties without coal mining. The regression analysis discussed above shows mixed results for this hypothesis. The presence of coal mining by itself does not have a statistically significant influence on county-level mortality rates. However, for counties in the Appalachians the effect of coal mining is associated with a significant increase in mortality rates. Furthermore, as Figure 6 shows, the increase in mortality rates for coal mining counties in the Appalachians is one of the largest increases compared to other variable coefficients. The variance component of the regression model reflects the mixed results. On average, the state-level effect of coal mining varies by almost 70 deaths per 100,000 population. Thus, while coal mining cannot be concluded to have an effect on mortality rates for every state, it does appear to influence mortality rates for counties in some states and especially for counties in Appalachia. Based on the evidence from the regression analysis, the null hypothesis cannot be conclusively rejected, as for most counties the effect of coal mining is not significantly different from zero. However, the analysis also provides evidence that for some counties the effect of coal mining is significantly and very substantially different from zero. Consequently, the hypothesis cannot be conclusively confirmed or rejected.

Similarly, mixed results are found for hypothesis H_2. The hypothesis assumes that coal mining above median production levels additionally increases mortality rates for coal mining counties. However, the model output indicates that the influence of above median level coal production is not statistically significant. While the main effect does not appear to be significant, the interaction term for Appalachian counties is significant in a statistical and substantial sense. For counties in Appalachia, mortality rates are increased by approximately 36 deaths per 100,000 population. Thus, the analysis does not allow to reject the null hypothesis in favor of hypothesis H_2. However, there is evidence that for some counties the null hypothesis can be rejected in favor of the alternative hypothesis.

Despite mixed results for hypotheses H_1 and H_2, the regression analysis finds clear evidence for hypothesis H_3. Over time, mortality decreases on average by 8.251 deaths per 100,000 population, which is moderated by a squared effect of time with a coefficient of 1.839. It should be noted that while the effects are clearly statistically significant, they are based on a somewhat small sample of 8 years. Thus, the effect sizes of the coefficients are not of primary interest. As the literature review points out, most other studies conducted on the relationship between of mortality rates and coal mining deliberately eliminate the effect of time. Thus, hypothesis H_3 is aimed at providing evidence that when included into the analysis, time has a significant influence. The regression results allow to reject the null hypothesis that time has no influence on mortality rates with a high degree of statistical certainty.

In regard to hypothesis H_4, the analysis provides confirmatory evidence. Table 7 shows the confidence intervals for the variance components of the multilevel regression model. Following these results, states vary significantly from zero in their average mortality rates. The effect of coal mining for coal mining states varies significantly on the state level as well. Consequently, there is sufficient statistical evidence to reject the null hypothesis that the effect of coal mining does not vary on the state level. Similarly to hypothesis H_3, the focus of hypothesis H_4 is the confirmatory evidence for group-level variation rather than the specific value of the variance component.

Table 7. Confidence Intervals, Variance Components

Variance Component	Profile Likelihood Based 95%-Confidence Interval
Random Interval (σ_{ς_1})	[33.122; 49.980]
Random Slope (σ_{ς_2})	[49.615; 98.128]

The results of my thesis add to the literature in a substantial and methodological way. Firstly, the study results explore the relationship between coal mining and mortality rates on a broader scale than previous studies. The scope of the research design includes all states in the United States and collects annual data over eight years. Furthermore, the data is not aggregated and thus the effect of time is included into the model. The mixed results for the effect of coal mining and mining above the median production level continue a substantial debate in the literature. Numerous studies by Michael Hendryx and colleagues found evidence for a statistically significant positive relationship between coal mining and mortality rates (Hendryx, 2009, 2015; Hendryx et al., 2007; Hendryx & Ahern, 2008).However, Buchanich et al. (2014) as well as Woolley et al. (2015) found contradicting statistical evidence. According to their results, coal mining is not generally associated with increased mortality rates. However, they also find statistical evidence for a significant relationship between coal mining and mortality rates for some counties in Appalachia. Thus, my study results extend the collection of evidence finding no sufficient statistical evidence to support a generally positive relationship between coal mining and mortality rates.

For counties within Appalachia, however, the effect of coal mining on mortality rates is statistically significant and the effects sizes are very substantial. These findings, in fact, support the findings of studies by Hendryx and colleagues that were focused exclusively on the Appalachian region (Esch & Hendryx, 2011; Hendryx, 2009; Hendryx & Ahern, 2009). However, for more reliable statistical inference, the focus of these studies should be extended beyond Appalachia, which is a region that is hardly comparable to other parts of the United States (Behringer & Friedell, 2006). Consequently, the substantial increases in mortality rates associated with coal mining and above median level coal production for Appalachian counties support conclusions made by most previous studies.

However, the results of my study point towards several methodological issues present in the body of literature on the relationship between coal mining and mortality rates. As a consequence of rejecting the null hypothesis for hypotheses H_3 and H_4 state-level grouping and time should be included into statistical modeling approaches. The overwhelming majority of research designs aggregate data over time and then fit an OLS regression model to the data. Aggregating data over time does not only reduces variation of input variables but also eliminates time as an input variable. Following H_4, time has a significant influence on mortality rates. Consequently, statistical models should be fit to annual data rather than aggregate data.

Furthermore, state-level variation should be included into statistical modeling. As findings by other scholars have pointed out, the relationship between coal mining and mortality rates varies regionally (Borak et al., 2012; Buchanich et al., 2014; Woolley et al., 2015). Grouping regional effects by states is convenient as laws and policies surrounding mining and mining procedures differ substantially between states (Hendryx & Holland, 2016). Subsequently, a data analysis of the influence of coal mining on mortality rates should include annual data of county and state level variables.

CHAPTER 6. CONCLUSION

The results of my research give a mixed answer to my initial research question. Based on the collected data and the applied statistical approach, a decisive conclusion about the effect of coal mining on mortality rates cannot be made. However, the results still have methodological and material implications. States vary substantially in their mortality rates and in the effect coal mining has on those mortality rates. This insight should be reflected, when the cost and benefits of coal mining are under consideration. Furthermore, these results point towards the research gap that my thesis attempts to fill in. The application of advanced statistics that incorporate the clustered character of data into the regression analysis improves the modeling of relationship between coal mining and mortality rates.

For some states, the effect of coal mining on public health is much more pronounced than for others. In terms of future research, these differences should be explored. States differ in laws and regulations concerning coal mining. The physical composition of coal also differs between states from different regions. Future research should explore the influence of these differences on mortality rates. Furthermore, it is not clear if coal mining affects different causes of deaths to a different degree. Research by Buchanich et al. (2014) finds a significant relationship between coal mining and cancer-related deaths but not between other causes of death. While the study is limited to the Appalachian region, this insight should be explored in future research on the national scale.

However, the results point towards a clear relationship between elevated mortality rates and coal mining in the Appalachian region. These findings support a general trend found in the research literature. When comparing factors that influence mortality rates, coal mining and especially high levels of coal production are the strongest risk factor for Appalachian counties. For a region that has formed a cultural symbiosis with coal mining this relationship is especially impactful. Coal mining in the Appalachians is not just economic activity but rather "a way of life" (Lewin, 2017). Consequently, the substantial risk coal mining poses to public health should be considered when coal mining is the subject of public debate.

Lastly, the study results point towards the influence time has when modeling mortality rates. As discussed in Chapter 5, time has a significant influence on mortality rates and should be included in an analysis. However, when compared to other studies, the exact coefficients found in this analysis seem to overstate the influence of time (Hendryx & Holland, 2016). This is likely due to the fact that the period of time under investigation is rather short (2010-2017). Consequently, future research should focus on investigating the effect of time for a broader time fare. In order to combine county-level analysis for studies before 2010, county-level covariate data have to be estimated. As the American Community Survey does not collect county-level information for all counties prior to 2010, data collection becomes a more challenging aspect of future research.

If the estimation of covariates is successful, future research could attempt to incorporate further clustering of the data into the modeling approach. At this point, the multilevel regression model assumes that county-level observations are nested within states. However, with a larger number of years, it would be possible to further model annual county observations within the county itself. This third nesting level would reflect the longitudinal character of the data and could improve results. However, this approach would assume a more complicated statistical relationship and it could be necessary to relax assumptions about temporal dependencies. Consequently, future research that includes a third level could incorporate different variance-covariance structures.

REFERENCES

Adler, N. E., & Ostrove, J. M. (1999). Socioeconomic status and health: what we know and what we don't. *Annals of the New York Academy of Sciences, 896*(1), 3-15. doi:10.1111/j.1749-6632.1999.tb08101.x

Anderson, R. N., & Rosenberg, H. M. (1998). *Age standardization of death rates: implementation of the year 2000 standard*. Hyattsville, Maryland: U.S. Dept. of the Health and Human Services, Public Health Service

ARC. (2020). Counties in Appalachia. Retrieved from https://www.arc.gov/appalachian_region/CountiesinAppalachia.asp

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 1*(1). Retrieved from https://www.jstatsoft.org/v067/i01

Behringer, B., & Friedell, G. H. (2006). Appalachia: where place matters in health. *Preventing chronic disease, 3*(4), A113-A113. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1779277/

Borak, J., Salipante-Zaidel, C., Slade, M. D., & Fields, C. A. (2012). Mortality disparities in Appalachia reassessment of major risk factors. *Journal of Occupational and Environmental Medicine, 54*(2), 146-156. doi:10.1097/JOM.0b013e318246f395

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: a challenge to conventional interpretations. *Psychological bulletin*(3), 396.

Buchanich, J. M., Balmert, L. C., Youk, A. O., Woolley, S. M., & Talbott, E. O. (2014). General mortality patterns in Appalachian coal-mining and non-coal-mining counties. *Journal of*

Occupational and Environmental Medicine, 56(11), 1169-1178. doi:10.1097/jom.0000000000000245

Bureau, U. S. C. (2010). *Population, housing units, area, and density: 2010 - United States* (GCT-PH1).

Bureau, U. S. C. (2018). *American Community Survey 1-year Estimates, 2010-2017*. Washington, DC

Christian, W. J., Huang, B., Rinehart, J., & Hopenhayn, C. (2011). Exploring geographic variation in lung cancer incidence in Kentucky using a spatial scan statistic: elevated risk in the Appalachian coal-mining region. *Public health reports (Washington, D.C. : 1974), 126*(6), 789-796. doi:10.1177/003335491112600604

Cortes-Ramirez, J., Naish, S., Sly, P. D., & Jagals, P. (2018). Mortality and morbidity in populations in the vicinity of coal mining: a systematic review. *BMC Public Health, 18*(1), 721. doi:10.1186/s12889-018-5505-7

DeNavas-Walt, C., Proctor, B., & Smith, J. (2010). *Income, poverty, and health insurance coverage in the United States: 2009 (U.S. Census Bureau, Current Population Reports, P60–238)*. Washington, DC: U.S. Census Bureau

Desjardins, R., & Schuller, T. (2006). Measuring the effects of education on health and civic engagement. Retrieved from https://www.oecd.org/education/innovation-education/measuringtheeffectsofeducationonhealthandcivicengagement.htm

Donaldson, K., Brown, D., Clouter, A., Duffin, R., MacNee, W., Renwick, L., . . . Stone, V. (2002). The pulmonary toxicology of ultrafine particles. *Journal of Aerosol Medicine, 15*(2), 213-220. doi:10.1089/089426802320282338

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel rasch model: with the lme4 package. *Journal of Statistical Software, 1*(2). Retrieved from https://www.jstatsoft.org/v020/i02

EIA. (2020). *Coal Production and Preparation Report*. (EIA-7A). Washington, DC Retrieved from https://www.eia.gov/coal/data/browser/

Esch, L., & Hendryx, M. (2011). Chronic Cardiovascular Disease Mortality in Mountaintop Mining Areas of Central Appalachian States. *Journal of Rural Health, 27*(4), 350-357. doi:10.1111/j.1748-0361.2011.00361.x

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865-2873. doi:10.1002/sim.3107

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Hendryx, M. (2009). Mortality from heart, respiratory, and kidney disease in coal mining areas of Appalachia. *International Archives of Occupational and Environmental Health, 82*(2), 243-249. doi:10.1007/s00420-008-0328-y

Hendryx, M. (2015). The public health impacts of surface coal mining. *The Extractive Industries and Society, 2*(4), 820-826. doi:10.1016/j.exis.2015.08.006

Hendryx, M., Ahern, M., & Nurkiewicz, T. (2007). Hospitalization patterns associated with Appalachian coal mining. *Journal of Toxicology and Environmental Health, 70*, 2064-2070. doi:10.1080/15287390701601236

Hendryx, M., & Ahern, M. M. (2008). Relations between health indicators and residential proximity to coal mining in West Virginia. *American Journal of Public Health, 98*(4), 669-671. doi:10.2105/AJPH.2007.113472

Hendryx, M., & Ahern, M. M. (2009). Mortality in Appalachian coal mining regions: the value of statistical life lost. *Public Health Reports, 124*(4), 541-550. doi:10.1177/003335490912400411

Hendryx, M., Fedorko, E., & Anesetti-Rothermel, A. (2010). A geographical information system-based analysis of cancer mortality and population exposure to coal mining activities in

West Virginia, United States of America. *Geospatial Health, 4*(2), 243-256. doi:10.4081/gh.2010.204

Hendryx, M., Fedorko, E., & Halverson, J. A. (2010). Pollution sources and mortality rates across rural-urban areas in the United States. *Journal of Rural Health, 26*(4), 383-391. doi:10.1111/j.1748-0361.2010.00305.x

Hendryx, M., & Holland, B. (2016). Unintended consequences of the Clean Air Act: Mortality rates in Appalachian coal mining communities. *Environmental Science & Policy, 63*, 1-6. doi:10.1016/j.envsci.2016.04.021

Hendryx, M., O'Donnell, K., & Horn, K. (2008). Lung cancer mortality is elevated in coal-mining areas of Appalachia. *Lung Cancer, 62*(1), 1-7. doi:10.1016/j.lungcan.2008.02.004

Hendryx, M., Yonts, S. D., Li, Y., & Luo, J. (2019). Mountaintop removal mining and multiple illness symptoms: A latent class analysis. *Science of The Total Environment, 657*, 764-769. doi:https://doi.org/10.1016/j.scitotenv.2018.12.083

Hoyert, D. L. (2012). *75 years of mortality in the United States, 1935-2010*. (1941-4927

0276-4733). Washington, DC: U.S. Department of Health and Human Services Retrieved from https://purl.fdlp.gov/GPO/gpo114224

Imai, K. (2017). *Quantitative social science : An introduction*. Princeton: Princeton University Press.

Jain, N. B., Potula, V., Schwartz, J., Vokonas, P. S., Sparrow, D., Wright, R. O., . . . Hu, H. (2007). Lead levels and ischemic heart disease in a prospective study of middle-aged and elderly men: the VA normative aging study. *Environmental health perspectives, 115*(6), 871-875. doi:10.1289/ehp.9629

Kecojevic, V., & Grayson, R. (2008). An analysis of the coal mining industry in the United States. *Minerals & Energy - Raw Materials Report, 23*, 74-83. doi:10.1080/14041040802181790

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 1*(13). Retrieved from https://www.jstatsoft.org/v082/i13

Lewin, P. G. (2017). "Coal is not just a job, it's a way of life": The cultural politics of coal production in central Appalachia. *Social Problems, 66*(1), 51-68. doi:10.1093/socpro/spx030

Lin, J.-L., Lin-Tan, D.-T., Li, Y.-J., Chen, K.-H., & Huang, Y.-L. (2006). Low-level environmental exposure to lead and progressive chronic kidney diseases. *The American Journal of Medicine, 119*(8), 707.e701-707.e709. doi:https://doi.org/10.1016/j.amjmed.2006.01.005

Mastin, J. P. (2005). Environmental cardiovascular disease. *Cardiovascular Toxicology, 5*(2), 91-94. doi:10.1385/CT:5:2:091

Menke, A., Muntner, P., Batuman, V., Silbergeld, E. K., & Guallar, E. (2006). Blood lead below 0.48 umol/L (10 ug/dL) and mortality among US adults. *Circulation, 114*(13), 1388-1394. doi:10.1161/CIRCULATIONAHA.106.628321

Menke, A., Muntner, P., Silbergeld, E. K., Platz, E. A., & Guallar, E. (2009). Cadmium levels in urine and mortality among U.S. adults. *Environmental health perspectives, 117*(2), 190-196. doi:10.1289/ehp.11236

Moffatt, S., & Pless-Mulloli, T. (2003). "It wasn't the plague we expected." Parents' perceptions of the health and environmental impact of opencast coal mining. *Social Science & Medicine, 57*(3), 437-451. doi:https://doi.org/10.1016/S0277-9536(02)00369-6

Navas-Acien, A., Guallar, E., Silbergeld, E. K., & Rothenberg, S. J. (2007). Lead exposure and cardiovascular disease--a systematic review. *Environmental health perspectives, 115*(3), 472-482. doi:10.1289/ehp.9785

NCHS. (2017). *Compressed Mortality File, 1999-2016 (machine readable data file and documentation, CD-ROM Series 20, No. 2V) as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.* Retrieved from: https://wonder.cdc.gov/controller/datarequest/D140

Nieuwenhuis, R., Grotenhuis, M. t., & Pelzer, B. (2012). influence.ME: Tools for detecting influential data in mixed effects models. *R Journal, 4*(2), 38-47. doi:10.32614/RJ-2012-011

Que, S., Awuah-Offei, K., & Samaranayake, V. A. (2015). Classifying critical factors that influence community acceptance of mining projects for discrete choice experiments in the United States. *Journal of Cleaner Production, 87*, 489-500. doi:https://doi.org/10.1016/j.jclepro.2014.09.084

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : applications and data analysis methods* Thousand Oaks [etc.]: Sage Publication.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*: Oxford University Press.

USDA. (2004). *Measuring Rurality: Rural-Urban Continuum Codes*. Washington, DC Retrieved from http://www.ers.usda.gov/briefing/rurality/ruralurbcon/

Van der Meer, T., Te Grotenhuis, M., & Pelzer, B. (2010). Influential cases in multilevel modeling: A methodological comment. *American Sociological Review, 75*(1), 173-178. doi:10.1177/0003122409359166

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133. doi:10.1080/00031305.2016.1154108

Woolley, S. M., Meacham, S. L., Balmert, L. C., Talbott, E. O., & Buchanich, J. M. (2015). Comparison of mortality disparities in central Appalachian coal-and non-coal-mining counties. *Journal of Occupational and Environmental Medicine, 57*(6), 687-694.

doi:10.1097/jom.0000000000000435

APPENDIX

Appendix A – Sample Frame

The study design includes counties in the United States between 2010 – 2017. However, the sample frame is limited by the availability of data on mortality rates. The CDC suppresses counties with less than 10 deaths from the sample due to privacy concerns, while counties with less than 20 deaths are marked as unreliable due to data concerns. Below is a list of all counties that are excluded from the sample due to reliability or privacy concerns. Further, the list indicates the presence of coal mining in the excluded county. Overall 140 counties were excluded due to privacy concerns which represent about 4% of the sampling frame.

Table 8. List of All Counties Excluded from Sample

Name of County	Coal Mining
Aleutians East Borough, AK	No
Bristol Bay Borough, AK	No
Denali County, AK	Yes
Haines Borough, AK	No
Hoonah-Angoon Census Area, AK	No
Lake and Peninsula Borough, AK	No
Petersburg Borough/Census Area, AK	No
Prince of Wales-Hyder Census Area, AK	No
Prince of Wales-Outer Ketchikan Census Area, AK	No
Skagway-Hoonah-Angoon Census Area, AK	No
Wrangell City and Borough, AK	No
Wrangell-Petersburg Census Area, AK	No
Alpine County, CA	No
Broomfield County, CO	No

Name of County	Coal Mining
Cheyenne County, CO	No
Custer County, CO	No
Dolores County, CO	No
Gilpin County, CO	No
Jackson County, CO	No
Kiowa County, CO	No
Mineral County, CO	No
Ouray County, CO	No
San Miguel County, CO	No
Chattahoochee County, GA	No
Echols County, GA	No
Quitman County, GA	No
Taliaferro County, GA	No
Webster County, GA	No
Butte County, ID	No
Camas County, ID	No
Clark County, ID	No
Greeley County, KS	No
Hamilton County, KS	No
Haskell County, KS	No
Hodgeman County, KS	No
Kiowa County, KS	No
Lane County, KS	No
Stanton County, KS	No
Wallace County, KS	No
Wichita County, KS	No
Keweenaw County, MI	No
Issaquena County, MS	No
Carter County, MT	No
Daniels County, MT	No
Garfield County, MT	No
Golden Valley County, MT	No
Granite County, MT	No
Judith Basin County, MT	No
Liberty County, MT	No
McCone County, MT	No
Meagher County, MT	No
Powder River County, MT	No
Prairie County, MT	No
Treasure County, MT	No
Wheatland County, MT	No
Wibaux County, MT	No
Banner County, NE	No

Name of County	Coal Mining
Deuel County, NE	No
Dundy County, NE	No
Frontier County, NE	No
Gosper County, NE	No
Grant County, NE	No
Hayes County, NE	No
Hooker County, NE	No
Keya Paha County, NE	No
Logan County, NE	No
Loup County, NE	No
Rock County, NE	No
Sioux County, NE	No
Thomas County, NE	No
Wheeler County, NE	No
Esmeralda County, NV	No
Eureka County, NV	No
Storey County, NV	No
Harding County, NM	No
Burke County, ND	No
Golden Valley County, ND	No
Grant County, ND	No
Kidder County, ND	No
Oliver County, ND	Yes
Renville County, ND	No
Sheridan County, ND	No
Steele County, ND	No
Gilliam County, OR	No
Sherman County, OR	No
Wheeler County, OR	No
Buffalo County, SD	No
Campbell County, SD	No
Faulk County, SD	No
Haakon County, SD	No
Hanson County, SD	No
Harding County, SD	No
Hyde County, SD	No
Jackson County, SD	No
Jerauld County, SD	No
Jones County, SD	No
Mellette County, SD	No
Sanborn County, SD	No
Stanley County, SD	No
Sully County, SD	No

Name of County	Coal Mining
Ziebach County, SD	No
Armstrong County, TX	No
Borden County, TX	No
Briscoe County, TX	No
Cottle County, TX	No
Culberson County, TX	No
Dickens County, TX	No
Edwards County, TX	No
Foard County, TX	No
Glasscock County, TX	No
Hudspeth County, TX	No
Irion County, TX	No
Jeff Davis County, TX	No
Kent County, TX	No
McMullen County, TX	No
Motley County, TX	No
Oldham County, TX	No
Reagan County, TX	No
Roberts County, TX	No
Sherman County, TX	No
Sterling County, TX	No
Stonewall County, TX	No
Throckmorton County, TX	No
Terrell County, TX	No
Daggett County, UT	No
Piute County, UT	No
Rich County, UT	No
Upton County, TX	No
Wayne County, UT	No
Bedford city, VA	No
Clifton Forge city, VA	No
Emporia city, VA	No
Highland County, VA	No
Garfield County, WA	No
Niobrara County, WY	No

Appendix B – State-Level Descriptive Analysis

The table below shows descriptive statistics for state-level indicator variable. The values are given as proportions of counties within a state. For example, 56.7% of the counties in Alabama are considered rural counties and 13.6% of all counties within Alabama produce coal. The total number of counties per state are listed as n. The summary statistics were computed prior to the outlier analysis and treatment.

Table 9. State-Level Descriptive Statistics for Indicator Variables

State	n	rural	Above Median Mining	Coal Mining	Appalachia
Alabama	536	0.567	0.047	0.136	0.552
Alaska	128	0.813	0.039	0.047	0.000
Arizona	120	0.467	0.067	0.067	0.000
Arkansas	600	0.733	0.000	0.013	0.000
California	456	0.351	0.000	0.000	0.000
Colorado	416	0.712	0.084	0.142	0.000
Connecticut	64	0.125	0.000	0.000	0.000
Delaware	24	0.000	0.000	0.000	0.000
District of Columbia	8	0.000	0.000	0.000	0.000
Florida	536	0.343	0.000	0.000	0.000
Georgia	1232	0.532	0.000	0.000	0.240
Hawaii	32	0.500	0.000	0.000	0.000
Idaho	328	0.732	0.000	0.000	0.000
Illinois	816	0.608	0.098	0.132	0.000
Indiana	736	0.522	0.060	0.094	0.000
Iowa	792	0.788	0.000	0.000	0.000
Kansas	768	0.802	0.000	0.010	0.000
Kentucky	960	0.708	0.105	0.221	0.450
Louisiana	512	0.453	0.016	0.029	0.000
Maine	128	0.688	0.000	0.000	0.000
Maryland	192	0.208	0.016	0.083	0.125
Massachusetts	112	0.214	0.000	0.000	0.000
Michigan	656	0.683	0.000	0.000	0.000
Minnesota	696	0.690	0.000	0.000	0.000
Mississippi	648	0.790	0.012	0.020	0.296
Missouri	920	0.704	0.000	0.009	0.000
Montana	328	0.902	0.073	0.098	0.000
Nebraska	600	0.827	0.000	0.000	0.000

State	n	rural	Above Median Mining	Coal Mining	Appalachia
Nevada	112	0.786	0.000	0.000	0.000
New Hampshire	80	0.700	0.000	0.000	0.000
New Jersey	168	0.000	0.000	0.000	0.000
New Mexico	256	0.781	0.063	0.063	0.000
New York	496	0.387	0.000	0.000	0.226
North Carolina	800	0.540	0.000	0.000	0.290
North Dakota	344	0.884	0.047	0.047	0.000
Ohio	704	0.568	0.047	0.156	0.364
Oklahoma	616	0.766	0.000	0.058	0.000
Oregon	264	0.606	0.000	0.000	0.000
Pennsylvania	536	0.448	0.082	0.382	0.776
Rhode Island	40	0.000	0.000	0.000	0.000
South Carolina	368	0.435	0.000	0.000	0.130
South Dakota	400	0.840	0.000	0.000	0.000
Tennessee	760	0.558	0.000	0.028	0.547
Texas	1808	0.659	0.039	0.048	0.000
Utah	200	0.600	0.105	0.150	0.000
Vermont	112	0.786	0.000	0.000	0.000
Virginia	1048	0.389	0.021	0.047	0.183
Washington	304	0.447	0.000	0.000	0.000
West Virginia	440	0.618	0.286	0.455	1.000
Wisconsin	576	0.639	0.000	0.000	0.000
Wyoming	176	0.909	0.182	0.227	0.000

Appendix C – Outlier and Influence Analysis

The tables and figures below provide more insight into the analysis of overly influential observations as discussed in section 4.1. The tables show the sum of counties excluded from analysis per state. For every year each county is counted as a unique observation. For example, Table 10 shows that for Alaska a total of 48 counties have been excluded from analysis over the full 8 years. It should be noted, that the same county can be counted for each year. In order to put this reduction into perspective, the table shows the reduction of total counties per state. For Alaska, the number of counties has been reduced by 37.5%. Furthermore, the table shows the average mortality of excluded counties and the percentage of counties with coal mining that were excluded. Lastly, the table shows the mean value of the percentage of the population that is Hispanic as well as the mean land area. The outlier analysis discussed in section 4.1. showed that these two variables were particularly affected by influential observations. It should be noted that these quantities are standardized by two standard deviations. Thus, on average counties that were excluded from analysis in Alabama are 0.416 standard deviations ($0.208 \cdot 2$) smaller than the average county size. At the end of the table are the overall mean values for all states. Table 11 shows the same quantities as Table 10, but broken down by year rather than by state.

Table 10. Summary Statistics Overly Influential Observations, By State

State	Number of Counties	Reduction in Observations	Mean Mortality	Percent of Counties with Mining	Mean standardized Percentage Hispanic Population	Mean standardized Land Area
Alabama	42	7.836%	975.619	50.00%	-0.208	-0.040
Alaska	48	37.500%	943.919	8.33%	-0.200	6.694
Arizona	13	10.833%	804.946	7.69%	0.527	1.049
Arkansas	15	2.500%	1007.520	0.00%	-0.207	-0.057
California	7	1.535%	515.314	0.00%	0.764	0.142
Colorado	63	15.144%	651.510	31.75%	0.499	0.114
Florida	31	5.784%	927.348	0.00%	0.203	-0.071
Georgia	116	9.416%	889.034	0.00%	-0.122	-0.103
Idaho	29	8.841%	736.355	0.00%	0.035	0.054
Illinois	38	4.657%	901.937	63.16%	-0.212	-0.096
Indiana	16	2.174%	921.125	31.25%	-0.247	-0.106
Iowa	5	0.631%	960.880	0.00%	-0.251	-0.088
Kansas	46	5.990%	897.848	0.00%	0.086	-0.040
Kentucky	151	15.729%	1129.158	60.93%	-0.260	-0.105
Louisiana	23	4.492%	911.148	4.35%	-0.248	-0.051
Maryland	1	0.521%	1130.500	0.00%	-0.156	-0.139
Massachusetts	1	0.893%	700.600	0.00%	-0.092	-0.144
Michigan	7	1.067%	818.271	0.00%	-0.252	-0.045
Minnesota	6	0.862%	1019.367	0.00%	-0.219	-0.037
Mississippi	61	9.414%	1008.493	6.56%	-0.267	-0.072
Missouri	19	2.065%	925.547	0.00%	-0.259	-0.075
Montana	33	10.061%	1000.436	39.39%	-0.225	0.253
Nebraska	20	3.333%	787.320	0.00%	0.003	-0.035
Nevada	11	9.821%	836.373	0.00%	0.283	0.830
New Mexico	36	14.063%	839.900	0.00%	1.596	0.350
New York	4	0.806%	798.750	0.00%	0.274	-0.062
North Carolina	21	2.625%	782.343	0.00%	-0.145	-0.098
North Dakota	42	12.209%	955.281	4.76%	-0.234	0.022
Ohio	34	4.830%	840.232	79.41%	-0.281	-0.089
Oklahoma	25	4.058%	892.768	0.00%	0.179	0.006
Oregon	4	1.515%	574.450	0.00%	0.116	0.701
Pennsylvania	30	5.597%	789.060	80.00%	-0.195	-0.048
South Carolina	8	2.174%	892.313	0.00%	-0.077	-0.082

State	Number of Counties	Reduction in Observations	Mean Mortality	Percent of Counties with Mining	Mean standardized Percentage Hispanic Population	Mean standardized Land Area
South Dakota	43	10.750%	1015.291	0.00%	-0.246	0.067
Tennessee	24	3.158%	1083.929	0.00%	-0.260	-0.117
Texas	214	11.836%	850.171	3.74%	1.717	0.021
Utah	19	9.500%	847.442	36.84%	-0.073	0.574
Vermont	1	0.893%	492.800	0.00%	-0.278	-0.059
Virginia	117	11.164%	923.758	5.13%	-0.078	-0.128
Washington	4	1.316%	524.750	0.00%	0.423	-0.063
West Virginia	71	16.136%	990.310	90.14%	-0.284	-0.090
Wisconsin	10	1.736%	1002.010	0.00%	-0.193	-0.091
Wyoming	12	6.818%	696.025	50.00%	-0.108	0.303
Mean Values	35.372	6.797%	864.934	15.196%	0.019	0.208

Table 11. Summary Statistics Overly Influential Observations, By Year

Year	Number of Counties	Reduction in Observations	Mean Mortality	Percent of Counties with Mining	Mean Standardized Percentage Hispanic Population	Mean Standardized Land Area
2010	200	6.680%	888.697	20.00%	0.181	0.208
2011	178	5.945%	934.382	24.16%	0.135	0.143
2012	165	5.511%	908.172	26.06%	0.230	0.160
2013	199	6.647%	907.366	24.12%	0.176	0.247
2014	215	7.181%	891.261	25.58%	0.134	0.233
2015	188	6.279%	940.289	19.15%	0.232	0.229
2016	154	5.144%	942.956	16.88%	0.201	0.234
2017	222	7.415%	900.985	17.12%	0.220	0.211
Mean	190.125	6.350%	914.263	0.216	0.188	0.208

Figure 13 through Figure 15 show the influence of excluding overly influential observations from the regression model. The measure is broken down by state and influence prior to outlier treatment is juxtaposed to influence post outlier treatment. The figures show clearly the effect of removing overly influential observations from the data. Figure 12 shows the reduction in the influence of overly influential observations on the beta coefficient of the percentage of the population that is Hispanic. While the influence of Florida and Colorado is still substantial, the influence of Texas is reduced substantial. Furthermore, Figure 14 shows the effect of outlier treatment on the interaction coefficient between coal mining and Appalachia. The figure highlights the substantial reduction of influential observations in Kentucky.

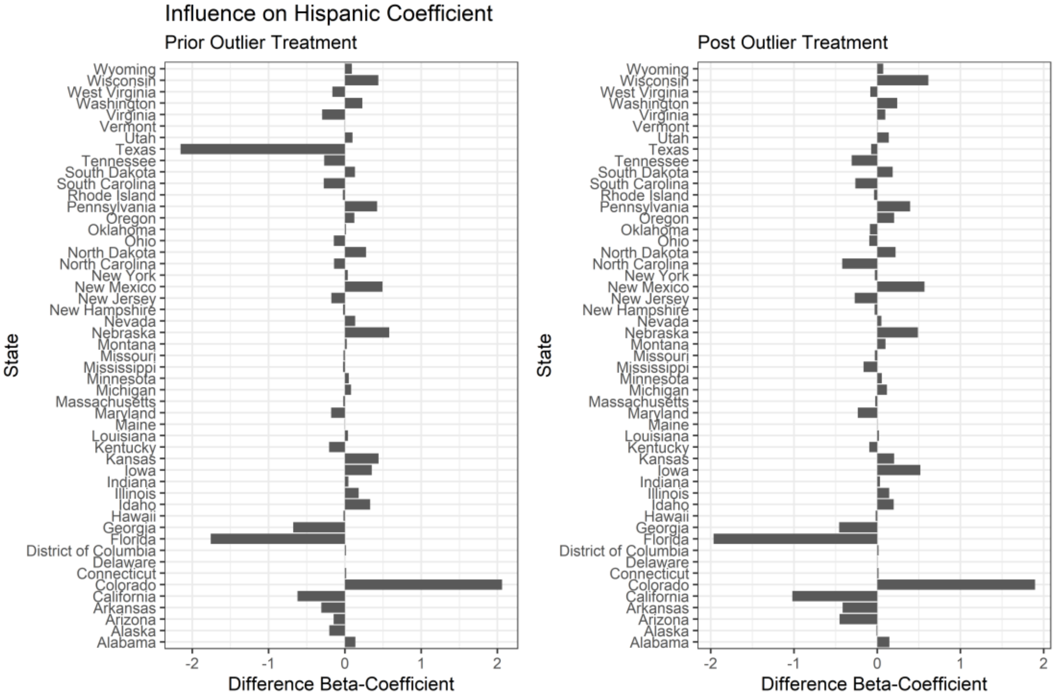


Figure 13. Outlier Influence - Difference in Hispanic Population Beta-Coefficient

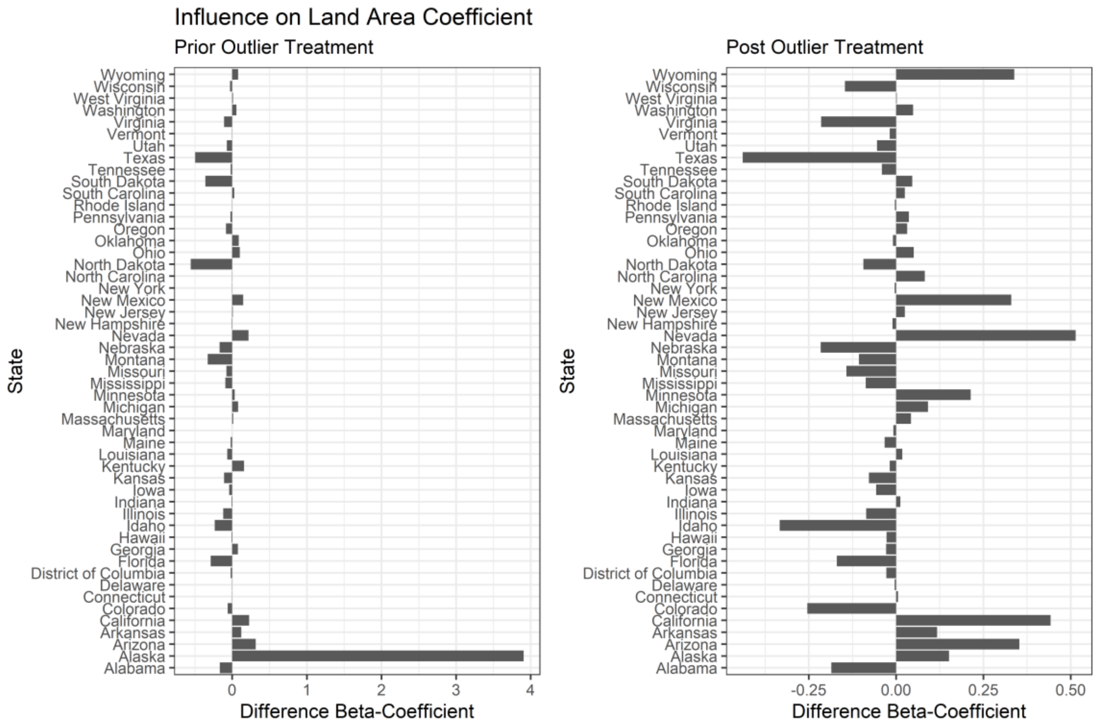


Figure 14. Outlier Influence - Difference in Land Area Beta-Coefficient

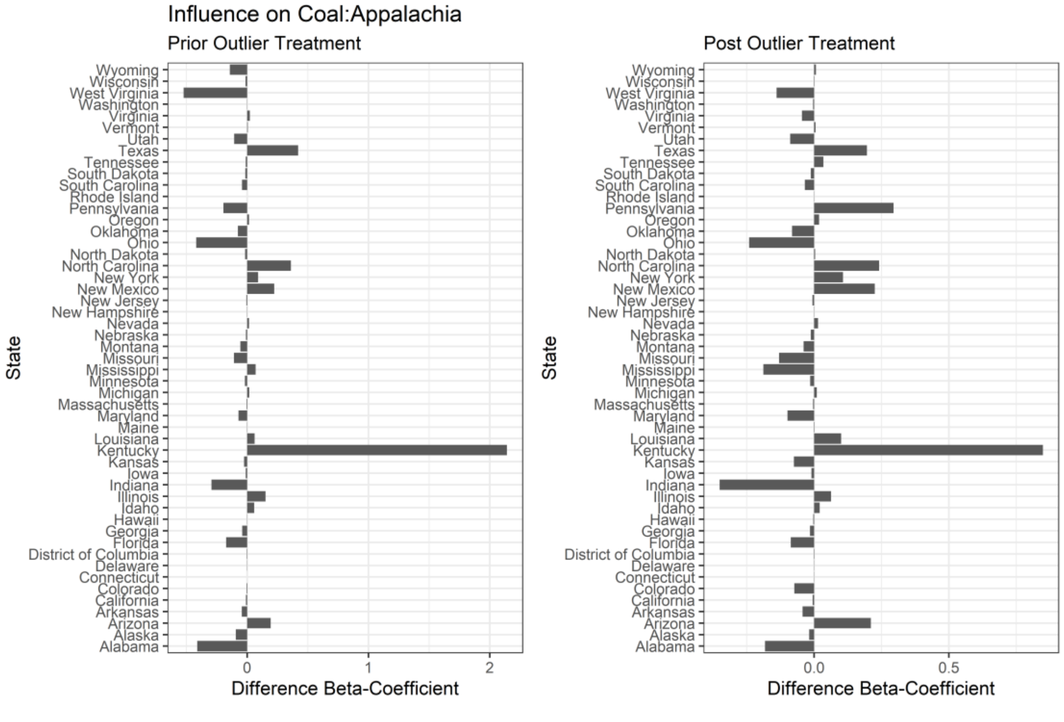


Figure 15. Outlier Influence - Difference in Interaction Coal Mining and Appalachia

Appendix D – Regression Model Summary Table

The regression output for four different multilevel models that were fit to the data is shown in Table 12.[8] The unconditional mean model was fit to the data without predictor variables but a varying intercept at the state-level. This unconditional means model provides the benchmark for all other models. In the next two steps, unconditional growth models were fit that include coal mining as the only predictor variable. The second unconditional growth model further allows for a varying slope of coal mining. Lastly, the full model was fit to the data.[9] The table can be used to observe the increase in explanatory power from the unconditional means to the full model. Furthermore, the intraclass correlation coefficient (ICC) increases substantially between models with only a random intercept and models that allow for a random slope of coal mining. ICC indicates the amount of variance that is contained in clustering. A high ICC indicates that observations within a cluster are very similar to each other. The increasing ICC indicates that additional variation in the data can be explained by allowing for a random slope.

Figure 16 shows the random slope value for each state plotted against the respective random intercept. The value pairs for the full model are gray, while the value pairs for the unconditional growth model are black. Both models show a clear

[8] The model summary shows the multilevel regression models as fit to the entire data (prior to influence analysis and treatment). I am showing the fit to the full data as the progression from the unconditional mean model to the full model is part of the model selection process that takes place before the outlier analysis is undertaken.

[9] The intermediate models that were calculated during the model selection process are not presented in Table 11. The regression outputs and summary statistics are available upon request and on the GitHub repository.

downward trending correlation between random slope and random intercept values. This trend indicates that states with higher random intercept values tend to have lower random slope values for the effect of coal mining on mortality. However, the full model indicates a stronger negative correlation between the slope and intercept values. The regression output in Table 12 reflects this observation with the difference of about 0.16 in the correlation term $\rho_{\varsigma_1 \varsigma_2}$.

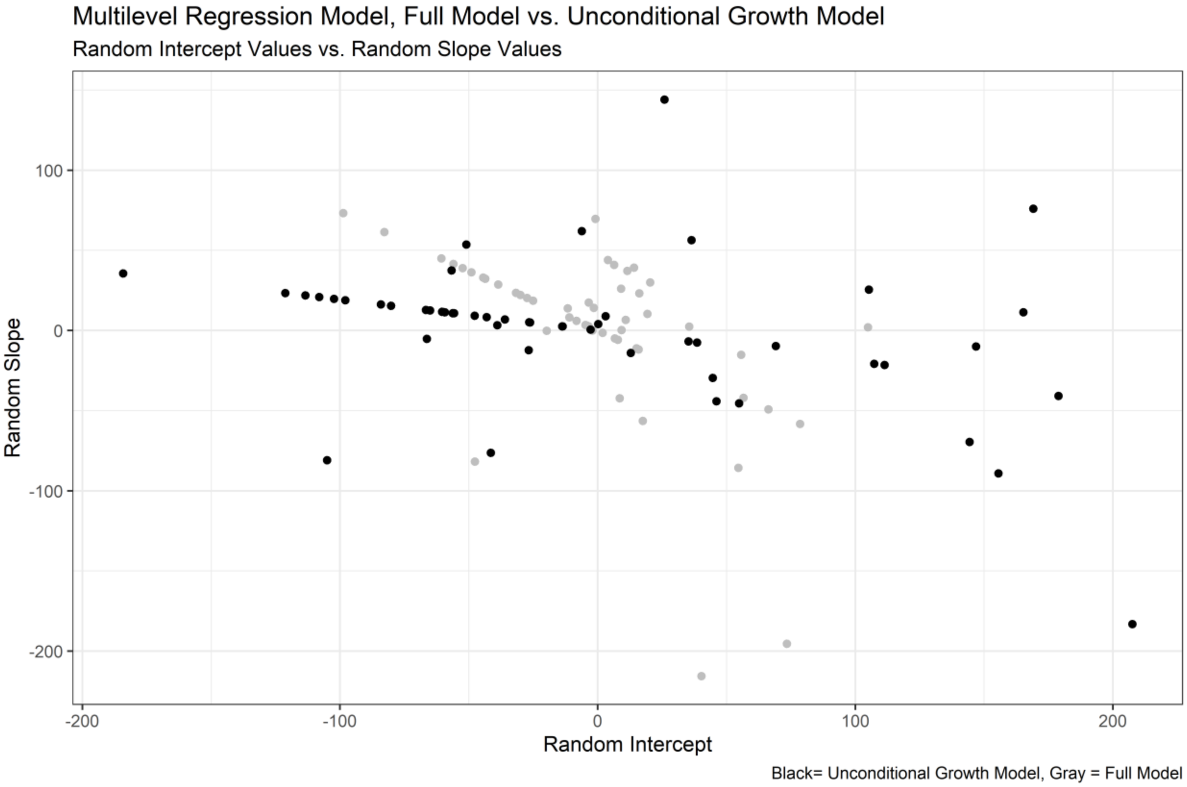


Figure 16. Regression Model, Correlation Random Slope and Intercept.

Table 12. Regression Model Output. All Models

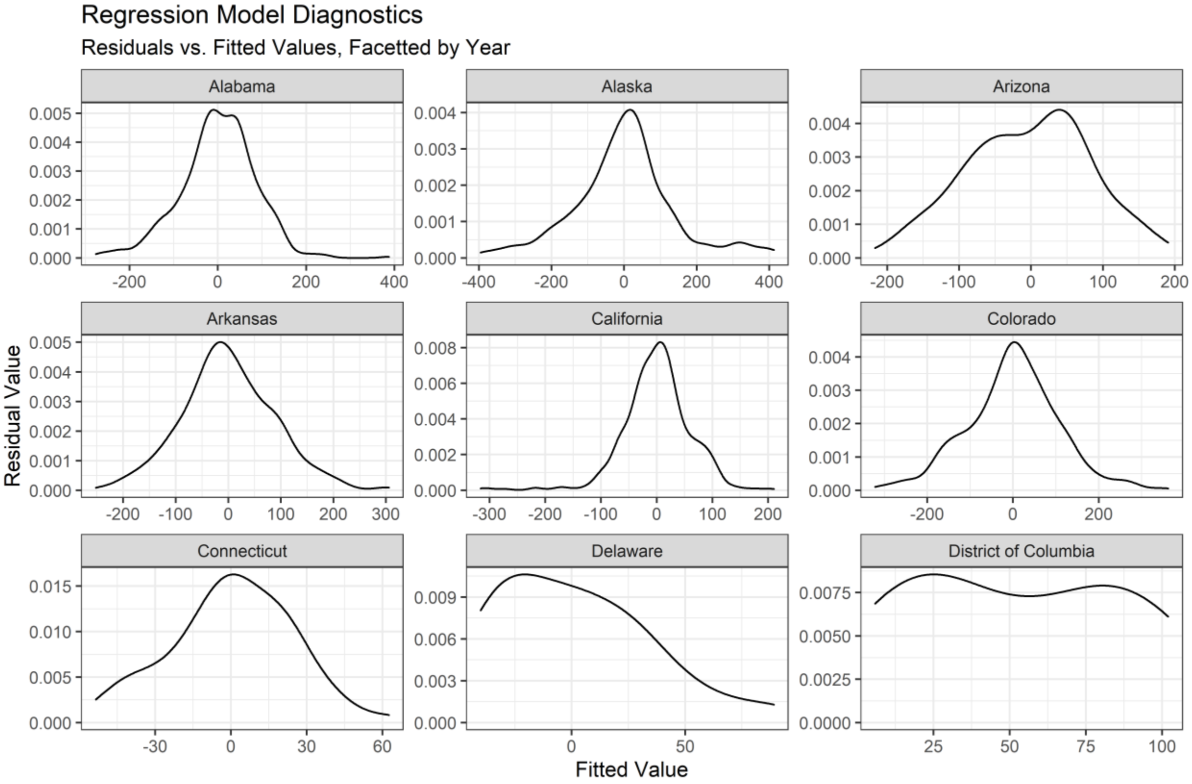
Parameter	Unconditional Mean Model		Unconditional Growth Model (Without Random Slope)		Unconditional Growth Model (With Random Slope)		Full Model	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept (α_i)	790.784*	13.155	787.219*	12.933	787.168*	12.855	791.311*	7.052
Coal Mining			65.569*	3.596	62.328*	14.747	-21.638	16.549
Above Median Mining							13.666	9.342
Appalachia							1.499	2.708
HS Grad Rate							11.584*	1.870
BA Grad Rate							-70.236*	2.304
Male Population							-18.952*	1.438
Hispanic Population							-60.315*	2.183
Coal Mining x Appalachia							47.504*	13.683
Above Median Mining x Appalachia							33.283*	11.827
Male Population x Hispanic Population							10.639*	1.946
Coal Mining x HS Grad Rate							-25.600*	7.568
Coal Mining x BA Grad Rate							-30.682*	8.009
Poverty Rate							38.469*	3.012
Median Age							-32.333*	1.799
Black Population							-12.639*	2.202
Southern State							60.708*	13.301
Rural County							-1.861	1.648
Unemployment Rate							24.220*	2.107
American Indian Population							27.526*	1.776
Median Income							-48.758*	3.074
Physician Access							-9.957*	1.352

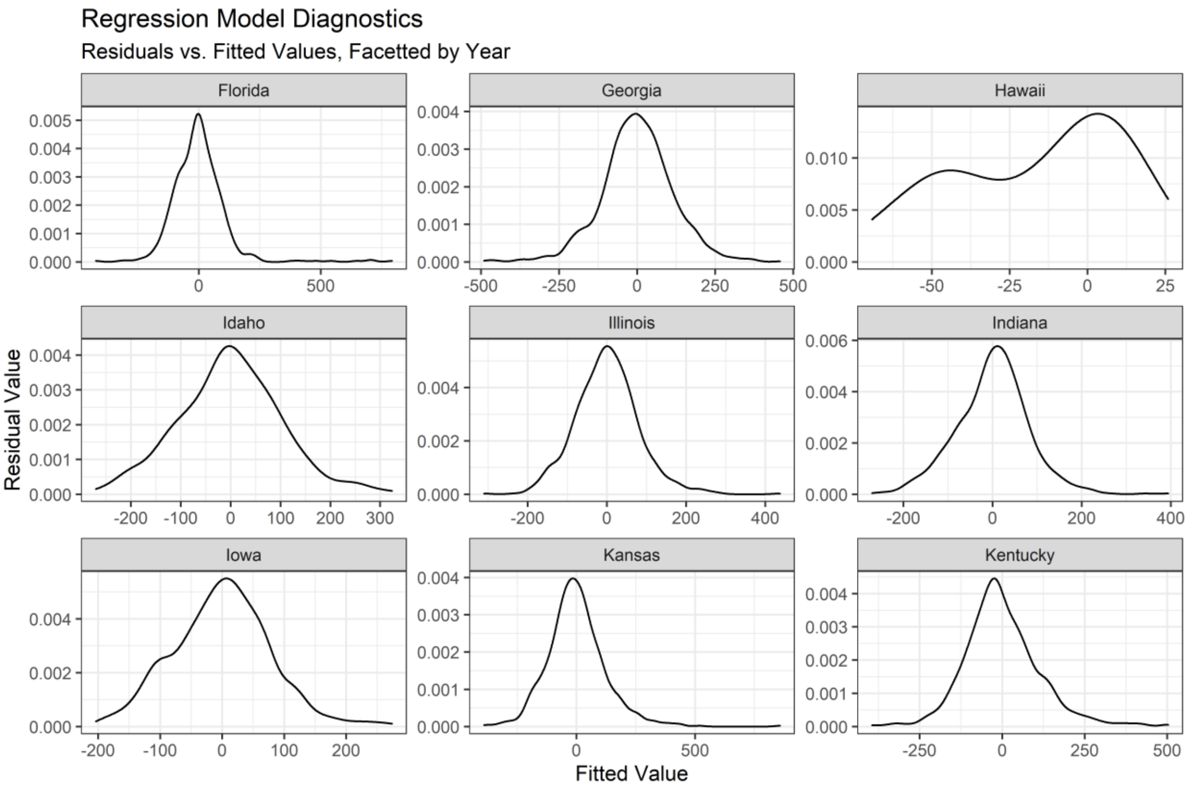
Table 13. Continuation – Regression Model Output. All Models

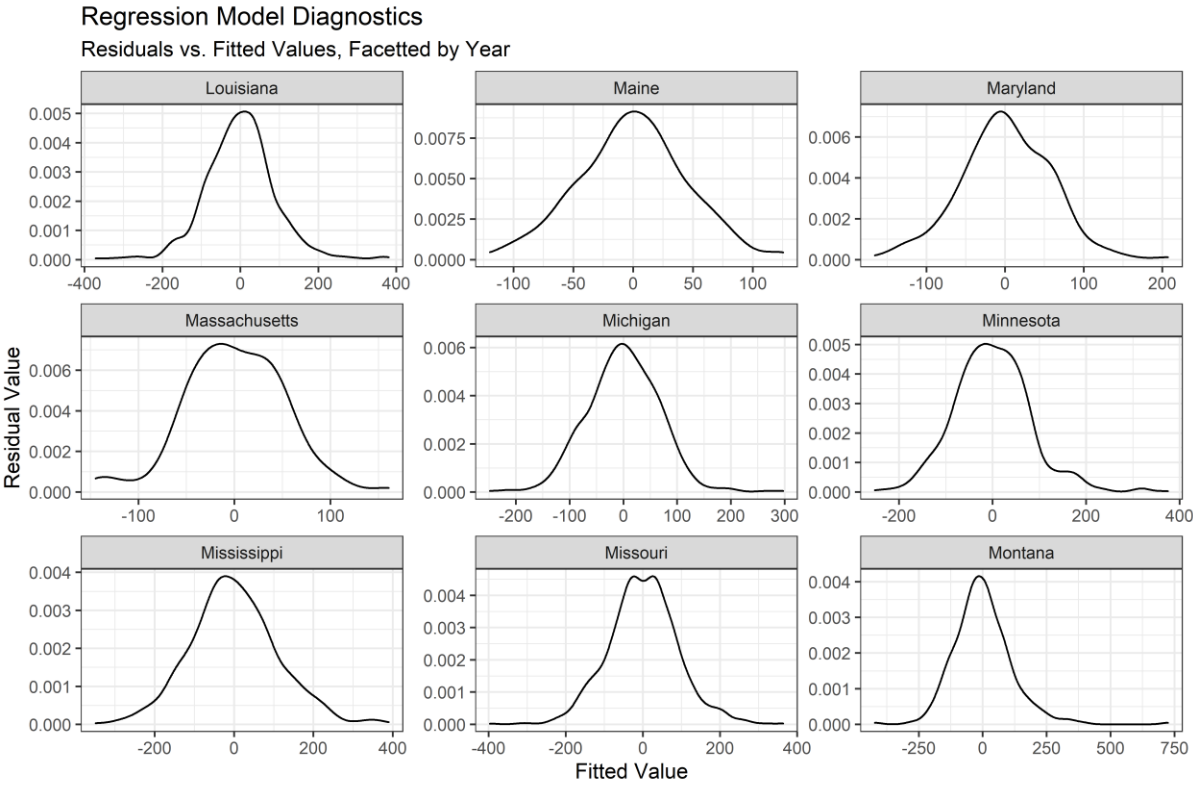
Parameter	Unconditional Mean Model		Unconditional Growth Model (Without Random Slope)		Unconditional Growth Model (With Random Slope)		Full Model	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Uninsured Population							-34.272*	2.205
Alcoholism Rate							-12.864*	1.947
Obesity Rate							12.011*	2.061
Smoking Rate							35.257*	1.792
Time							-6.471*	1.081
Time squared							1.635*	0.144
County Size							-3.936*	1.833
σ_{ς_1}	93.371		91.774		91.216		41.676	
σ_{ς_2}					69.958		75.008	
$\rho_{\varsigma_1 \varsigma_2}$					-0.252		-0.412	
σ_{ε}	122.358		121.523		120.837		95.925	
ICC	0.368		0.363		0.475		0.445	
AICc	298509.199		298176.690		297965.464		286797.204	
BIC	298533.450		298209.024		298013.963		287063.875	
N	23952		23952		23952		23952	
Groups	51		51		51		51	

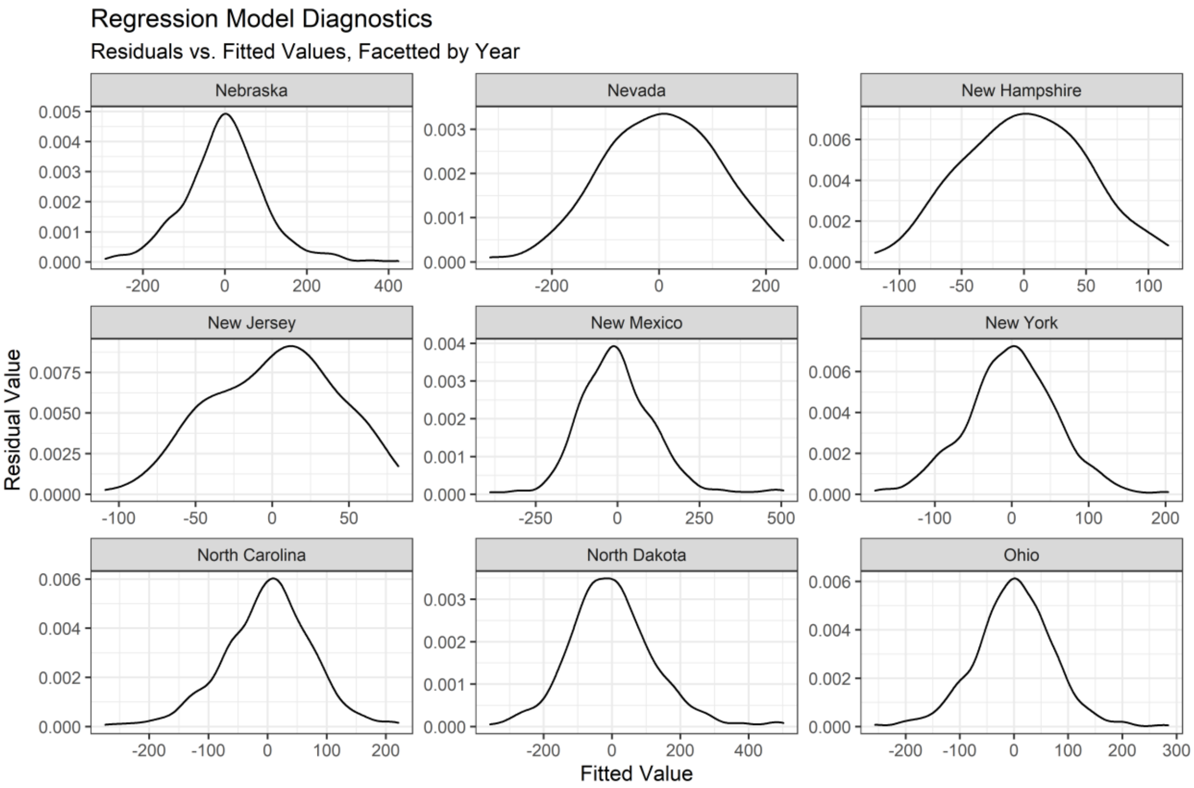
Appendix E – Residual Diagnostics

The following graphics show the residual distribution broken down by state.

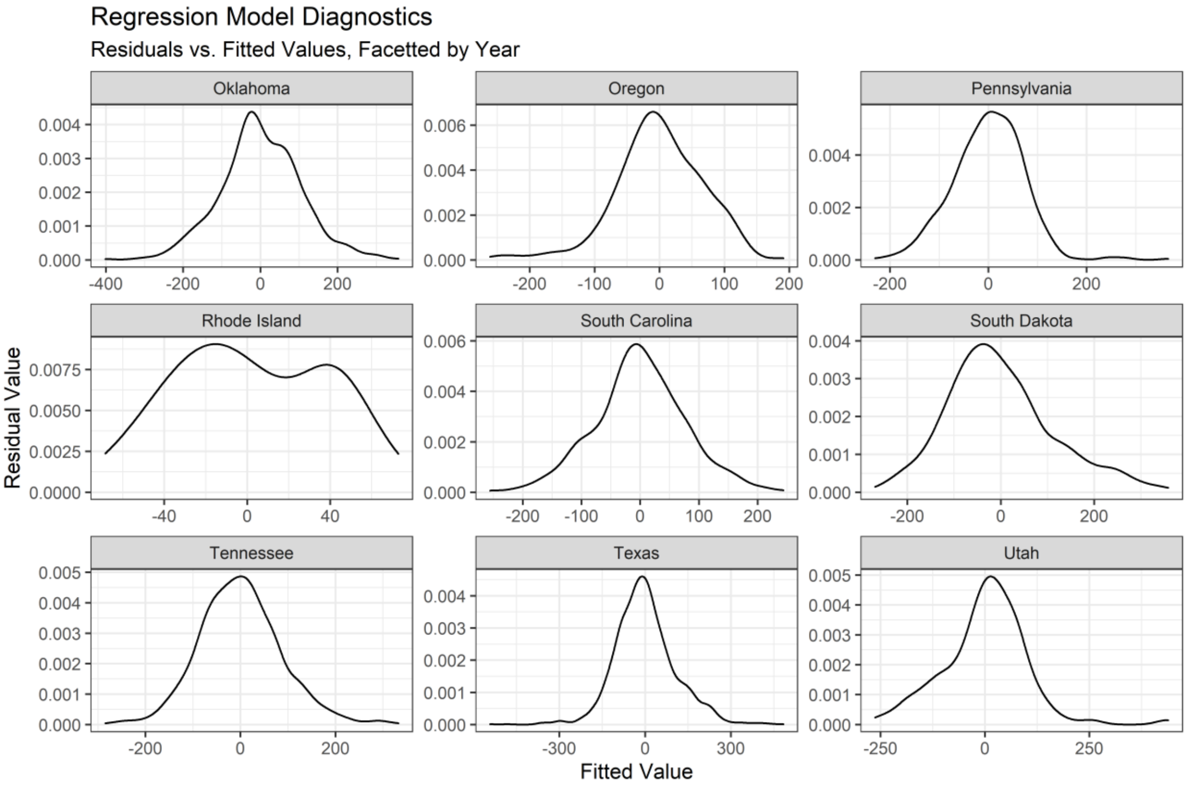



Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Residual Value
Fitted Value
0.000
0.001
0.002
0.003
0.004
0.005
0.010
0.006
0
500
-500
-250
250
-50
-25
25
-200
-100
100
200
300
400

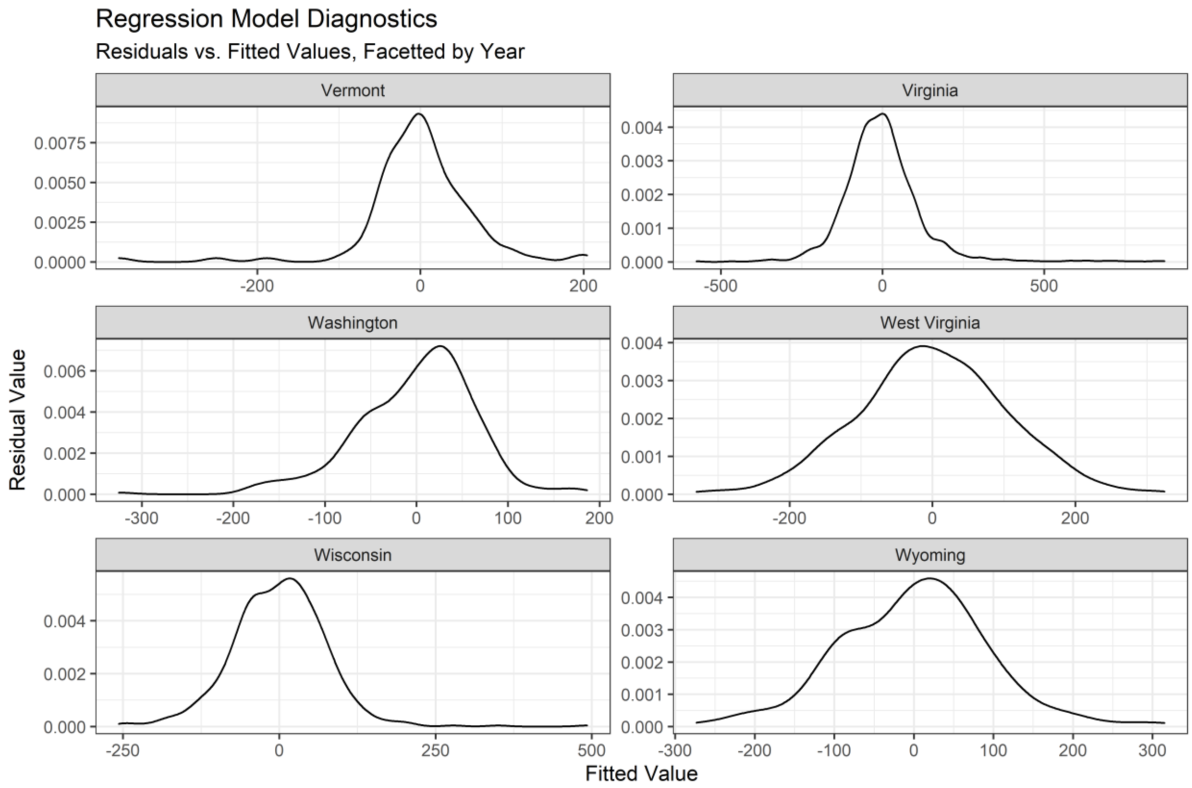

Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Residual Value
Fitted Value
0.0000
0.0025
0.0050
0.0075
0.000
0.001
0.002
0.003
0.004
0.005
0.006
-400
-200
0
200
400
-100
-50
50
100
300
-250
250
500
750

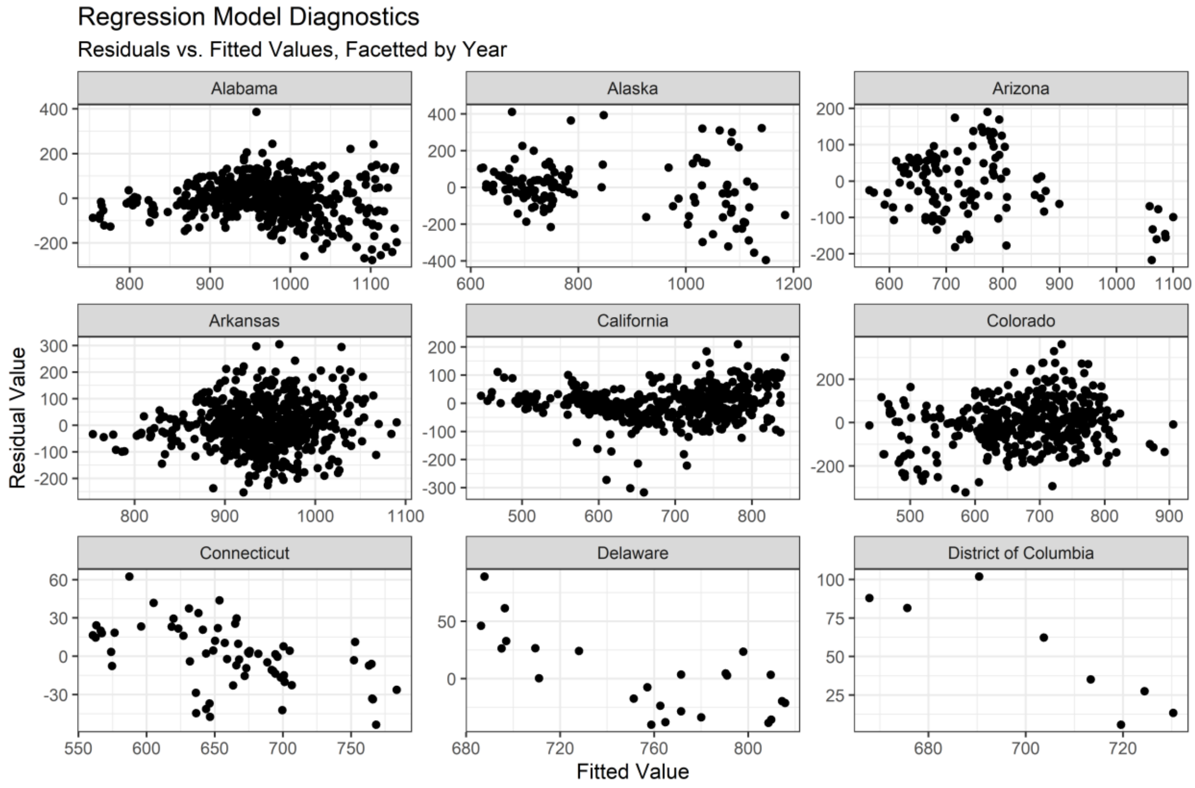

Regression Model Diagnostics

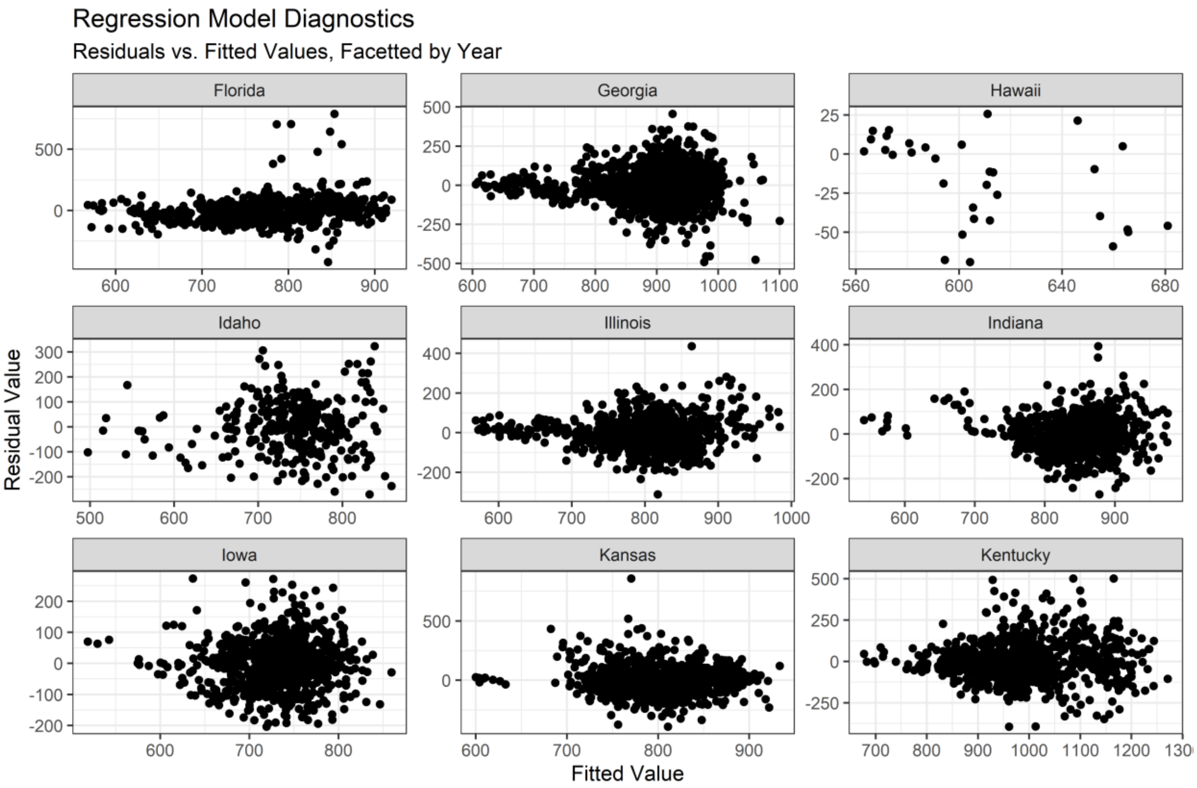
Residuals vs. Fitted Values, Facetted by Year

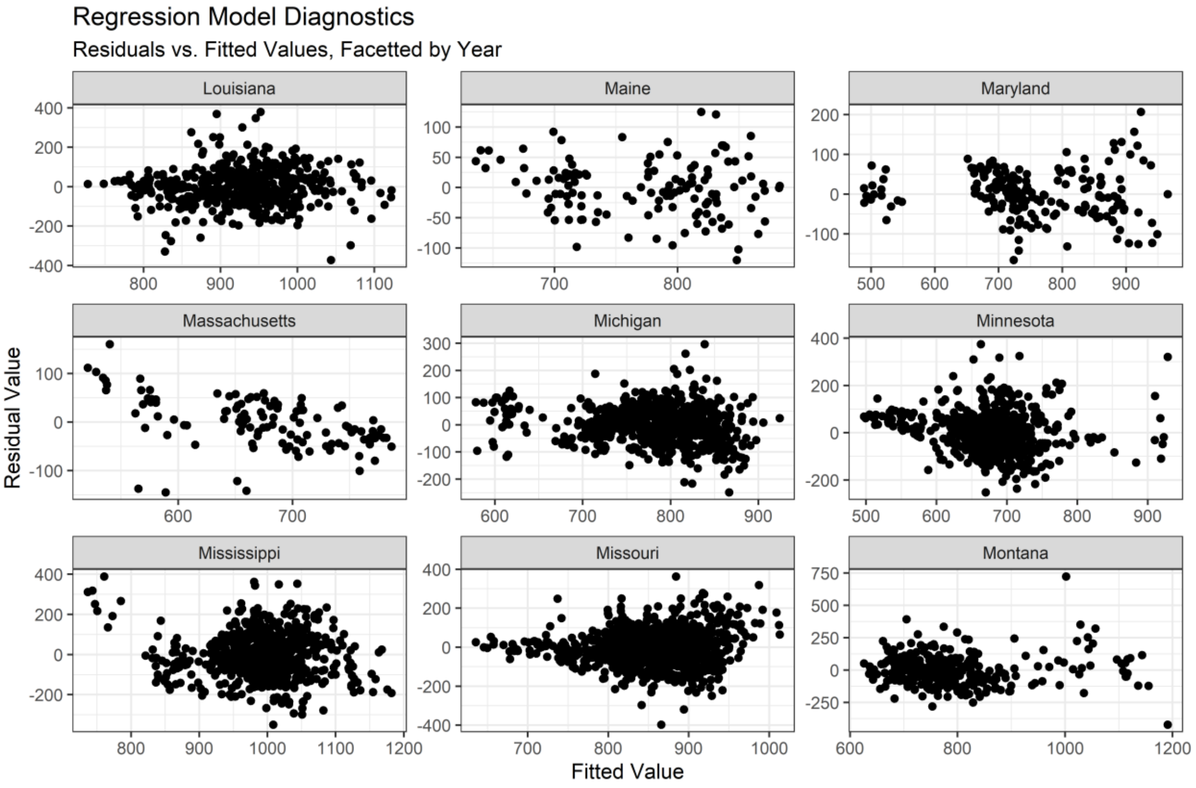

Regression Model Diagnostics

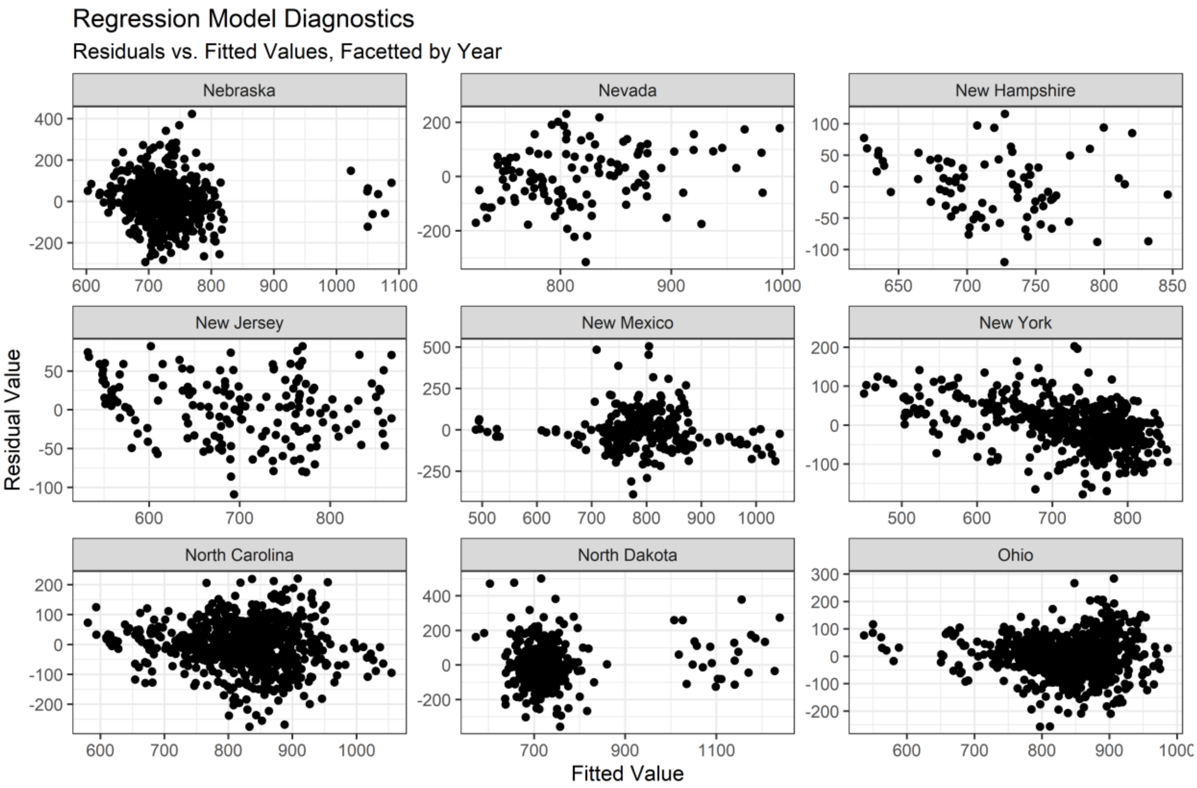
Residuals vs. Fitted Values, Facetted by Year

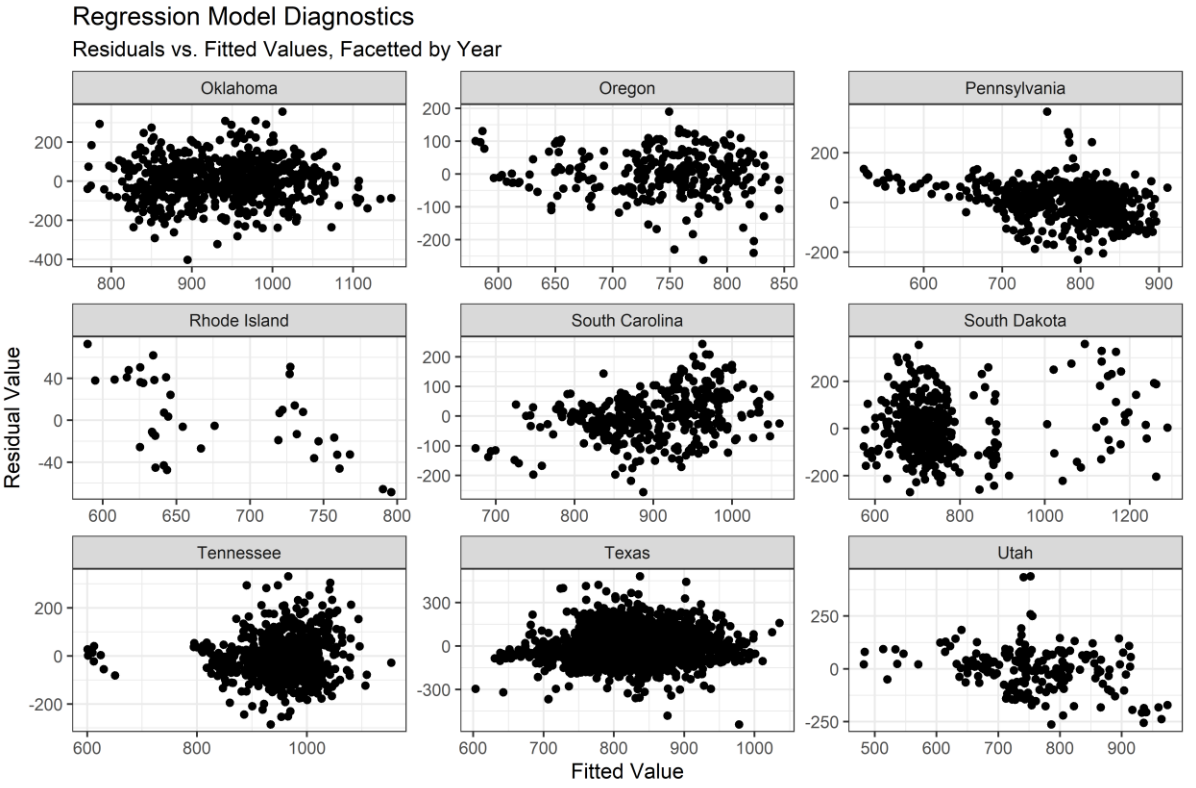

Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Residual Value
Fitted Value
0.0000
0.0025
0.0050
0.0075
0.000
0.001
0.002
0.003
0.004
0.006
-500
0
500
-300
-250
-200
-100
100
200
250
300
500

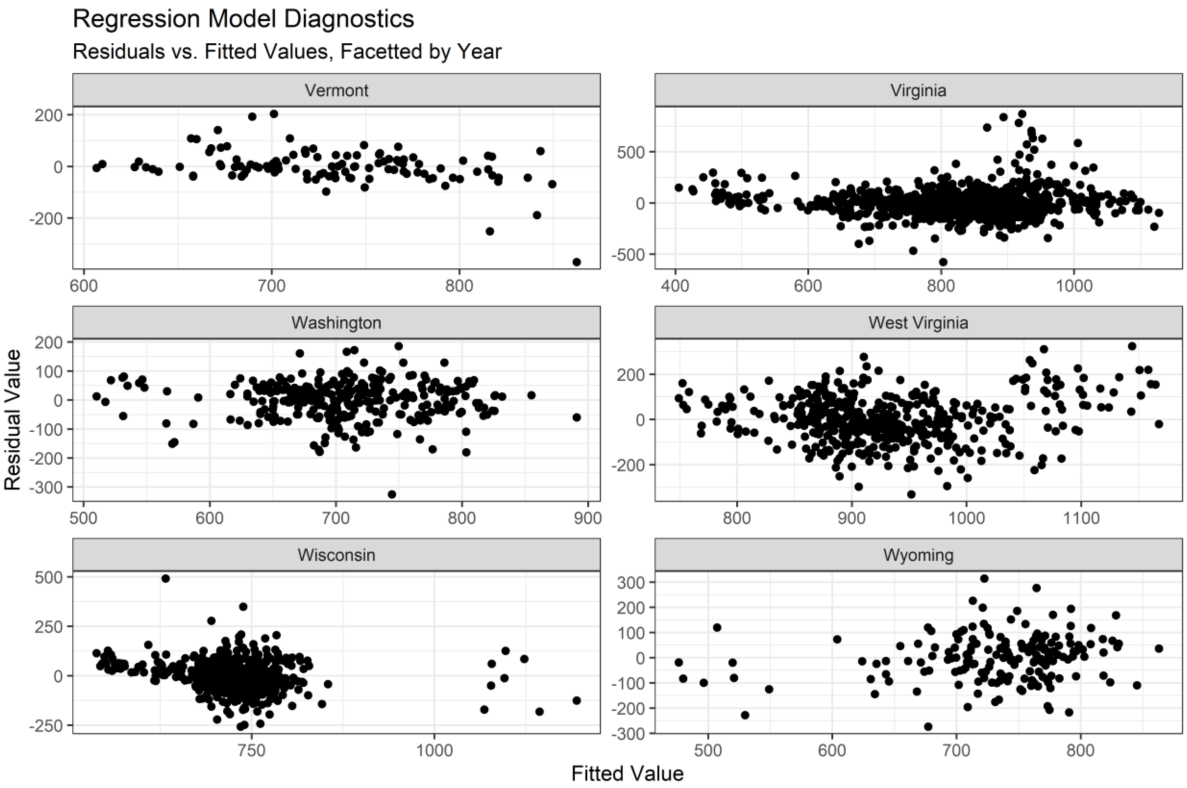

Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Residual Value
Fitted Value
-400
-300
-200
-100
0
100
200
300
400
-30
30
60
50
25
75
500
550
600
650
680
700
720
750
760
800
900
1000
1100
1200


Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Residual Value
Fitted Value
500
0
600
700
800
900
500
250
0
-250
-500
600
700
800
900
1000
1100
25
0
-25
-50
560
600
640
680
300
200
100
0
-100
-200
500
600
700
800
400
200
0
-200
600
700
800
900
1000
400
200
0
-200
600
700
800
900
200
100
0
-100
-200
600
700
800
500
0
600
700
800
900
500
250
0
-250
700
800
900
1000
1100
1200
130


Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Residual Value
Fitted Value
400
200
0
-200
-400
800
900
1000
1100
100
50
0
-50
-100
700
800
200
100
0
-100
500
600
700
800
900
100
0
-100
600
700
300
200
100
0
-100
-200
600
700
800
900
400
200
0
-200
500
600
700
800
900
400
200
0
-200
800
900
1000
1100
1200
400
200
0
-200
-400
700
800
900
1000
750
500
250
0
-250
600
800
1000
1200


Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Residual Value
Fitted Value
400
200
0
-200
600
700
800
900
1000
1100
200
0
-200
800
900
1000
100
50
0
-50
-100
650
700
750
800
850
50
0
-50
-100
600
700
800
500
250
0
-250
500
600
700
800
900
1000
200
100
0
-100
500
600
700
800
200
100
0
-100
-200
600
700
800
900
1000
400
200
0
-200
700
900
1100
300
200
100
0
-100
-200
600
700
800
900
1000


Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Residual Value
Fitted Value
200
0
-200
-400
800
900
1000
1100
200
100
0
-100
-200
600
650
700
750
800
850
200
0
-200
600
700
800
900
40
0
-40
600
650
700
750
800
200
100
0
-100
-200
700
800
900
1000
200
0
-200
600
800
1000
1200
200
0
-200
600
800
1000
300
0
-300
600
700
800
900
1000
250
0
-250
500
600
700
800
900


Regression Model Diagnostics
Residuals vs. Fitted Values, Facetted by Year
Vermont
200
0
-200
600
700
800
Virginia
500
0
-500
400
600
800
1000
Washington
200
100
0
-100
-200
-300
500
600
700
800
900
West Virginia
200
0
-200
800
900
1000
1100
Wisconsin
500
250
0
-250
750
1000
Wyoming
300
200
100
0
-100
-200
-300
500
600
700
800
Residual Value
Fitted Value

VITA

Henning Tovar is a graduate student in Applied Data Analytics (MS) and Political Science (MA) at Appalachian State University. After finishing his degree in Political Science and Philosophy at Friedrich-Alexander University in Nuremberg, Germany, Henning chose to start graduate school in the United States. Growing up close to the coal mining region of Germany, Henning has always shared an interest in the effect of coal mining on the environment and the people living in mining communities.