Abstract

The Robert Wood Johnson Foundation's "US County Health Rankings" dataset (2014) was used to investigate the interactions between economic and health factors, as well as the interaction of regional location on those measures. Two sets of hypotheses were proposed: (1) there is a correlational relationship between specific paired economic and health factors and (2) there is a correlational relationship between economic and health factors and location (using US Census divisions and regions). Linear regression modeling showed that there was a significant positive correlation between several rates within the data set. As well, there is a significant correlation between location and rates. These results are consistent with current research findings that violence and poverty negatively impact health factors. As well, that physical activity has an inverse relationship with obesity.

Keywords: data analysis, economic factors, public health, childhood poverty, premature death, hospitalization, adult obesity, unemployment, uninsured, violent crime rate

The Interaction of Economic Factors and Public Health

Many research studies have explored the relationship between violence and health factors. For example, Beck et al. (2016) performed a retrospective cohort study analyzing the correlation between pediatric asthma-related emergency hospital room visits and hospitalizations and factors including crime rates, violent crime rates, and poverty rates. They found that "the violent crime rate explained 35% of the population-level asthma utilization variance." Doleac (2018) cited evidence that increased rates of insurance related to Medicaid expansion correlated with reduced rates for violent crime and property crime.

There is also a growing awareness of the impact of violence on mental and physical health. In particular, within health care, there is a growing awareness of the injurious impact of trauma and chronic stress on health and longevity. For example, McFarlane (2010) wrote about the long-term impact of traumatic stress on both mental and physical health, citing chronic pain, high blood pressure, obesity, and cardiovascular disease. Similarly, Tucker (2018) wrote, "The evidence linking poverty with ill-health is unequivocal. Birthweights in the most deprived areas are on average 200g lower than in the richest, and children in disadvantaged families are more likely to die suddenly in infancy, to suffer acute infections, and to experience mental ill-health." Moreover, she cited a survey of pediatricians who overwhelmingly agreed that "poverty and low income contribute 'very much' to ill-health among their patients," noting that "inadequate housing, homelessness, food insecurity, and the stress and stigma of poverty are affecting children's physical and mental health in a myriad of ways." Current research also indicates a strong correlation between obesity and physical activity. For example, Sarma, Zaric, Campbell, and Gilliland (2014) found that both leisure time and work time physical activity exerts a

negative effect on body mass index. This study will analyze publicly available data to further analyze these relationships.

Study Methodology

Analysis of Dataset

The Robert Wood Johnson Foundation's "US County Health Rankings" dataset (2014) (a comma-separated values (CSV) file) was analyzed using the R Foundation's R Version 3.5.1 software and Tableau's Public software. The "US County Health Rankings" dataset provides ranked numbers for US counties for a variety of health and community rates including preventable hospital stays, premature death, obesity, physical inactivity, uninsured, violent crime, unemployment, and children in poverty. The data represents calendar years 2003 through 2012.

As a preliminary step, R and Tableau were used to analyze the data to better understand the dataset. The CSV file contains 14 columns representing identity and numeric data. The state code, county code, measure, and year span together can serve as a compound primary key for each row. In the rare instances where a specific location-year span-measure combination does not have data points, a row is still included in the CSV file, but numeric fields in the row are left blank. "Year span" data may be a single year, or it may be a span of years. Within the dataset, several variables were provided with year spans of single years, including the rates of preventable hospital stays (for three years), childhood poverty, obesity, unemployment, physical inactivity, and uninsured. Some variables were provided with rolling two- and three-year averages, including rates for preventable hospital stays (for most years in the data set), premature death, and violent crime. There is variation in the measures that are reported, with gaps for some

measures on some years (see Appendix A for details about the measures that are represented for each year span).

For the purpose of this study, a subset of the dataset was used. Only county-level data were included in this study to avoid duplication of data points as composites at the state and national levels. Data for a few measures (diabetic monitoring, mammography screening, daily fine particulate matter, and sexually transmitted infections) were omitted in order to focus the study on the remaining variables. Rates for Puerto Rico were very limited in the dataset, as data represented only a single year for a single measure. Because of its limited representation, data for Puerto Rico was removed prior to analysis. Similarly, since data prior to 2002 represented a single measure, and so was not useful for correlation analysis, data prior to 2002 were also excluded from analysis.

County Health Ranking's "2017 Trends Data Documentation" was reviewed to better understand the data. A few warnings, in particular, were noted. Breaks in data collection strategies were noted, resulting in a caution for comparisons made over time (for adult obesity in 2010 when cell phones were first used, and for uninsured rates in 2008 and children in poverty after 2002-2004, when the sources of the data changed). Since this study does not include a series analysis, these breaks did not result in excluded data. There were recommendations for caution when comparing adult obesity and physical inactivity across states and a stern warning that violent crime cannot be compared across states.

Importing Data

Five fields were imported from the County Health Rankings CSV file for in this study:

1. State: the two character state abbreviation

2. County code: a number; each state reports statewide rates using county code “0” and starts counting specific counties with “1”
3. Year span: a single year or a range of years
4. Measure name: text
5. Raw value: a standard percentile rate for most measures; For preventable hospital stays, it is the number per 1000 fee-for-service Medicare enrollees; for violent crime, it is the number of reported violent crime offenses per 100,000 population; for premature death rates, it is the years of potential life lost before age 75 per 100,000 population

The imported fields were used to compute additional fields for in this study:

1. Location-state: a concatenation of state, county code, and year span values for use as a primary key
2. Normed rate: In preparation for regression analysis, each rate was plotted on a histogram and Q-Q plot to assess for normal distribution. Where the distribution followed a normal distribution, the rate was duplicated in a new field containing normed rates. Where the rate was skewed, the data were transformed (using `sqrt()` for mildly right-skewed distribution and `log()` for stronger right-skewed distribution) before saving it in the new normed rate field (see Appendix B for the graphs and formulations that were used to analyze for normal distribution for each measure)
3. Scaled rate: the scaled rate was computed for use in a line chart showing multiple measures, since rates varied widely

4. Region: assigned based on state abbreviation using the United States Census Bureau's classification
5. Division: assigned based on state abbreviation using the United States Census Bureau's classification

After rows were imported, they were assembled into a cross table for analysis using `merge()` with the attribute `all=TRUE` so that all desired rates in the CSV file would be represented in the table even when some measures did not have corresponding rates (resulting in "NA" values). When creating the cross table, the computed key (state abbreviation-county code-year span, such as "WY-43-2005") was used as a primary key.

Exploratory Analysis: Visualizing the Data

In order to visualize relationships between measures, three scatterplot matrices were created by plotting normed data in paired plots using the R function `pairs()`. Measures were grouped together in a way that allowed their year spans to overlap (see Appendix C for the scatterplot matrices). A visual review of the scatterplot matrices suggested that it was likely that there were correlations between several measures, including: physical inactivity and adult obesity, physical inactivity and childhood poverty, adult obesity and unemployment, adult obesity and childhood poverty, unemployment and childhood poverty, uninsured and childhood poverty, preventable hospital stays and childhood poverty, uninsured and preventable hospital stays, violent crime and premature death, and preventable hospital stays and premature death.

In addition, using the R function `boxplot()`, two notched boxplots were created for each measure: one showing the measure factored by region, and one showing the measure factored by division. Boxplots were not created for violent crime rate since the data was not

meant to be used to make comparisons between states. Doyle (n.d) explained, “although not a formal test, if two boxes' notches do not overlap there is ‘strong evidence’ (95% confidence) their medians differ.” These boxplots suggest that there were likely real differences in some of the measures based on location.

Hypotheses

Correlation between measures.

The set of null hypotheses for the pairs of rates identified in the exploratory phase is that, for each paired set of measures listed above, the correlation between the paired measures is zero. The alternative hypothesis is that there is a correlation between the paired measure rates. All paired measures that were analyzed were found to be significantly correlated (see Table 1).

Table 1
Regression Statistics for Paired Measures

Measure 1	Measure 2	p-value *	r-squared
Premature Death	Preventable Hospitalization	3.48E-226	0.304
Uninsured	Preventable Hospitalization	6.15E-150	0.0552
Adult Obesity	Unemployment	0	0.149
Adult Obesity	Physical Inactivity	0	0.458
Adult Obesity	Childhood Poverty	0	0.216
Physical Inactivity	Childhood Poverty	0	0.304
Uninsured	Childhood Poverty	0	0.246
Unemployment	Childhood Poverty	0	0.239
Preventable Hospitalization	Childhood Poverty	0	0.159
Premature Death	Violent Crime	0	0.127

* p-value less than 0.01 indicates a significant correlation

For each of the pairs tested in Table 1, using an alpha value 0.01, the correlation was found to be significant, providing strong evidence against the null hypothesis. The coefficient of determination (r-squared values in table 1), describes “the proportion of the variation in the

response that can be attributed to the predictor" (Davies, 2016, p. 460). While all paired measures inspected showed significant correlations, they varied in the correlation of determination. Paired measures with the highest correlations of determination are most closely predictive of each other. For example, 46% of the variation in adult obesity can be accounted for by physical inactivity. Similarly, 30% of the variation in physical inactivity can be anticipated given childhood poverty. It is important to note that correlation does not suggest causation.

To visually confirm the apparent linear relationship between measures in the small paired plots, measures were plotted as pairs using the `smoothScatter()` function in the `ggplot2` library. This function was selected as it is helpful for cases with a large number of observations (INWT-Blog-RBloggers, 2018) (see figure 1 for two examples with strong correlations of determination and Appendix D for additional smooth scatter plots). Plotting with `smoothScatter()` confirmed that the measures did not appear to have more complicated non-linear relationships.

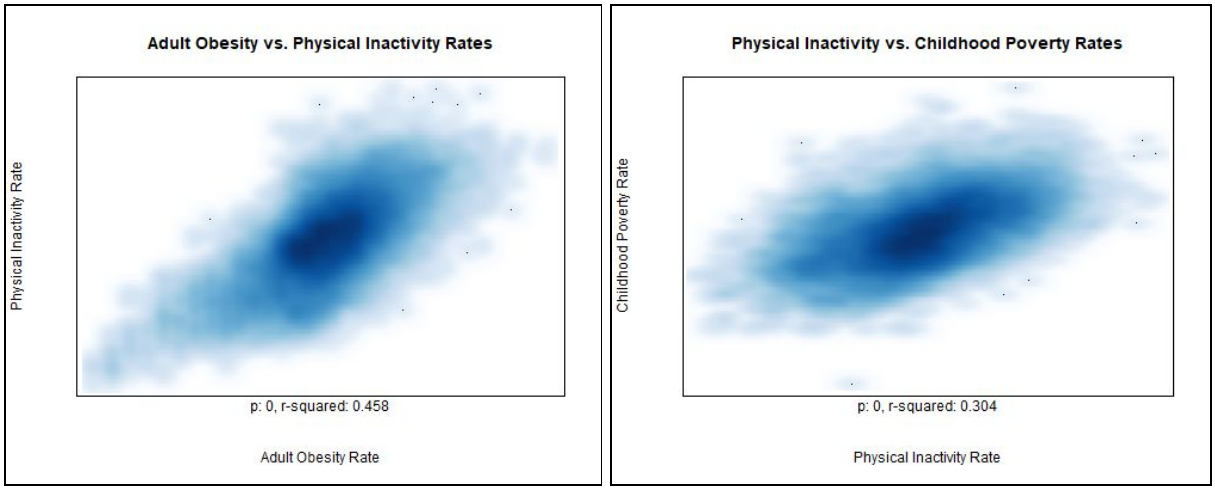

Figure 1: Smooth scatter plots showing adult obesity rates against physical inactivity rates (left) and childhood poverty rates against physical inactivity rates (right). Both plots show a fairly linear relationship with noticeable noise.

Differences by location.

The null hypotheses concerning location are that all population means are equal when comparing each measures' normed rate, categorized by region and by division. The alternative hypothesis is that at least one population mean for each measure's normed rate is different from the rest when measures are grouped by region and by division. All measures except violent crime were analyzed using `aov()`, using the region to factor rates and again using division to factor rates. Each of those cases was found to be statistically significant, with p <2e-16 (see table 2).

Table 2
Sum of Squares for each Measure analyzed by Region and Division

Measure 1	Region	Division	p-value *
Adult Obesity	7.6	9.8	p <2e-16
Childhood Poverty	59.2	66.5	p <2e-16
Physical Inactivity	16.2	21.4	p <2e-16
Premature Death	421.5	462.9	p <2e-16
Preventable Hospitalization	459.1	650.6	p <2e-16
Unemployment	71.0	883.0	p <2e-16
Uninsured	32.2	41.7	p <2e-16

* p-value less than 0.01 indicates a significant correlation

Finally, scaled rates for each measure (except violent crime) were grouped by division and plotted (see figure 2). Divisions were arranged manually to highlight regional affiliations. Rates were scaled so that measures could be compared visually. A horizontal line was drawn at scaled rate = 0 to emphasize whether data points fell above or below division averages for each

of the measures. A visual comparison showed clear differences across regions. For example, divisions in the Northeast region fell below average for the health and economic measures that were visualized. In contrast, divisions in the South region were above average for most measures.

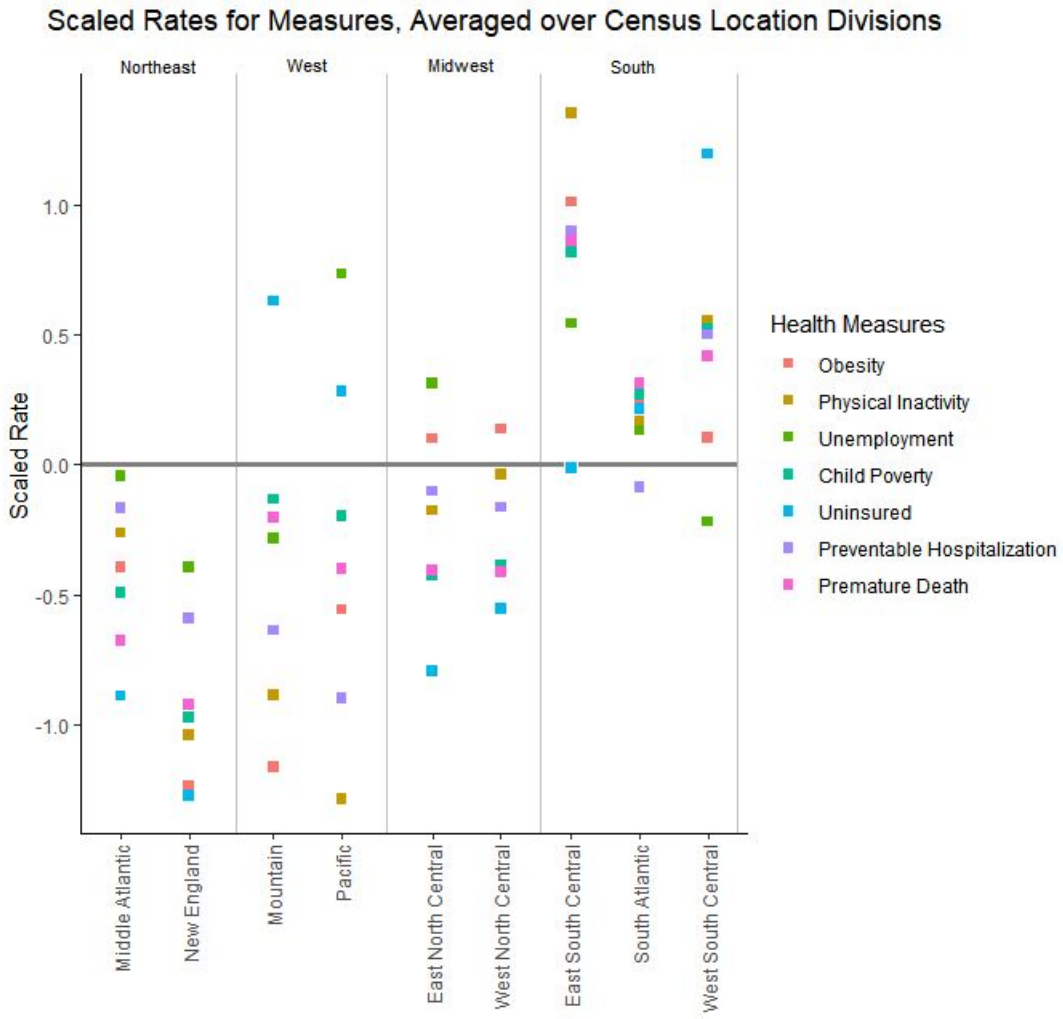

Figure 2: Scaled rates for measures, grouped by region and division.

Discussion

An analysis of data from The Robert Wood Johnson Foundation's "US County Health Rankings" showed patterns in the data. As expected, there are correlations between many of the

health and economic measures. As stated, all analyzed linear correlations were found to be significant. The coefficient of determination (see Table 1) provided further information about the variation, for example showing that 15.9% of the variation in Preventable Hospitalization is associated with Childhood Poverty, 30.4% of the variation in Physical Inactivity is associated with Childhood Poverty, and 45.8% of the variation in Adult Obesity is associated with Physical Inactivity. Moreover, there are differences in health and economic measures across regions and divisions.

Directions for Further Study

A further study could develop rolling years for all measures to allow for more comparisons, in particular examinations of the relationship between violent crime and premature death with the other economic and health factors. As well, single regression comparisons were made in this study due to the limitations in overlap of year spans for the various measures. By creating rolling year averages, multiple linear regression techniques could be used to reveal more fine-tuned relationships between economic and health rate factors using multiple linear regression. It could also prove fruitful to categorize counties, such as by rural, suburban, and urban, to look for further relationships.

While this study revealed several correlations within the data, it also suggested questions with regard to causation. For example, correlations were found between several interconnected factors. And yet, the nature of this observational study was unable to uncover causational relationships. It would be interesting to understand whether, for example, addressing childhood poverty could impact the 16% variation in preventable hospitalization.

References

2017 trends data documentation. County Health Rankings [PDF]. Retrieved from http://www.countyhealthrankings.org/sites/default/files/2017TrendsDocumentation.pdf

Beck, A. F., Huang, B., Ryan, P. H., Sandel, M. T., Chen, C., and Kahn, R. S. (2016, June). Areas with high rates of police-reported violent crime have higher rates of childhood asthma morbidity. *The Journal of Pediatrics, 173*, 175–182. https://dx.doi.org/10.1016%2Fj.jpeds.2016.02.018

Beckett. L. (2014, February 4). Living in a violent neighborhood is as likely to give you PTSD as going to war: Yet this at-home PTSD crisis remains largely ignored. *ProPublica*. Retrieved from https://www.motherjones.com/politics/2014/02/ptsd-among-wounded-americans-in-violent-neighborhoods/

Cady, F. (2017). *The data science handbook*. Hoboken, NJ: Wiley Publishing. ISBN: 9781119092940

Davies, T. (2016). *The book of R: A first course in programming*. San Francisco, CA: No Starch Press. ISBN: 9781593276515

Doleac, J. L. (2018, January 3). New evidence that access to health care reduces crime. *Brookings*. Retrieved from https://www.brookings.edu/blog/up-front/2018/01/03/new-evidence-that-access-to-health-care-reduces-crime/

Doyle, D. Notched Box Plots. [web log comment]. Retrieved from https://sites.google.com/site/davidsstatistics/home/notched-box-plots

"Geographic Terms and Concepts - Census Divisions and Census Regions." United States Census Bureau. Retrieved from https://www.census.gov/geo/reference/gtc/gtc_census_divreg.html

INWT-Blog-RBloggers. (2018, September). SmoothScatter with ggplot2. [web log comment]. Retrieved from https://www.r-bloggers.com/smoothscatter-with-ggplot2/

McFarlane, A. (2010, Feb 9). The long-term costs of traumatic stress: intertwined physical and psychological consequences. *World Psychiatry*, *9*(1): 3–10. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2816923/

Robert Wood Johnson Foundation. 2014. US County Health Rankings [data file]. Retrieved from https://public.tableau.com/s/sites/default/files/media/County_Health_Rankings.csv

Sarma S., Zaric G. S., Campbell M. K., and Gilliland J. (2014 July). The effect of physical activity on adult obesity: evidence from the Canadian NPHS panel. *Economics and Human Biology*, 1-21. doi: 10.1016/j.ehb.2014.03.002

Tableau. (2018). Public [Software]. Available from https://public.tableau.com/en-us/s/

The R Foundation. (2018). R (Version 3.5.1) [Software]. Available from https://www.r-project.org/

Tucker, J. (2018, 20 April). "The impact of poverty on child health." Royal College of Paediatrics and Child Health. Retrieved from https://www.rcpch.ac.uk/news-events/news/impact-poverty-child-health

Appendix A

Year Spans and Measures Represented in US County Health Rankings

Table 3
Year Spans and Measures Represented in US County Health Rankings

Year Span	Measures with Rates included in the Year Span							
	Adult obesity	Children in Poverty	Physical inactivity	Pre-mature Death	Prevent. Hospital Stays	Unemp-loyment	Un-insured	Violent Crime
1997-1999				X				
1998-2000				X				
1999-2001				X				
2000-2002				X				
2001-2003		X		X		X		
2002								
2002-2004				X				
2003		X				X		
2003-2005				X	X			X
2004	X	X	X			X		
2004-2006				X				X
2005	X	X	X			X		
2005-2007				X				X
2006	X	X	X			X	X	
2006-2007					X			
2006-2008				X				X
2007	X	X	X			X	X	
2007-2009				X				X
2008	X	X	X		X	X	X	
2008-2010				X				X
2009	X	X	X		X	X	X	
2009-2011								X
2010	X	X	X		X	X	X	
2011		X			X	X	X	
2012		X				X		

Appendix B

Histograms and Q-Q Plots

Each of the variables was plotted with a histogram and a QQPlot to visually assess the distribution of the data. Variables that were not normal were transformed to data that was roughly normally distributed for use in regression analyses.

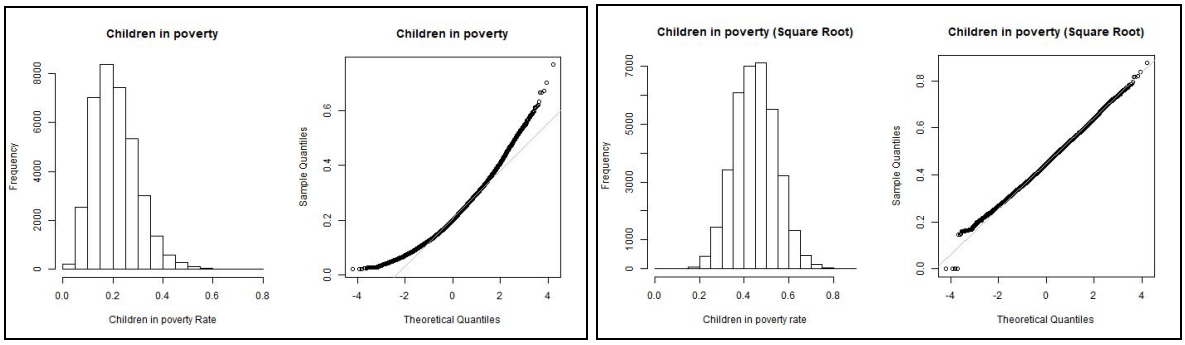

Figure 4: Children in Poverty Rate. The histogram and Q-Q plot of the rate of children in poverty has a right-skewed distribution (left image). The rate's square root is roughly normally distributed (right image).

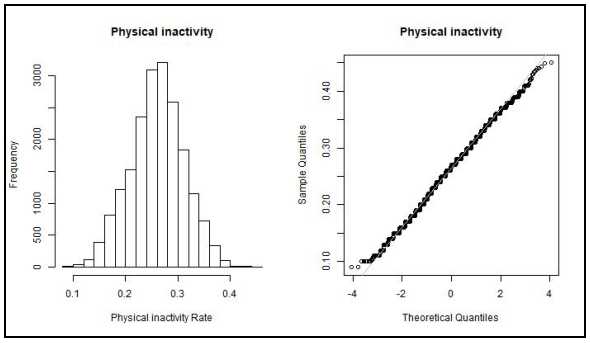

Figure 5: Physical Inactivity Rate. The rate of physical inactivity is roughly normally distributed.

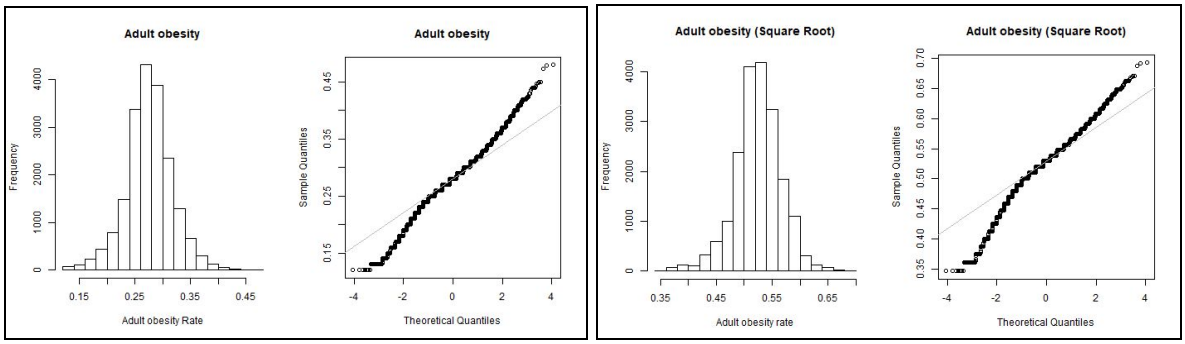

Figure 6: Obesity Rate: The rate of adult obesity is roughly normally distributed (left image). Its square root was visually inspected (right image). Since it appeared to be less normal than the untransformed rate, the untransformed rate was used in regression analyses.

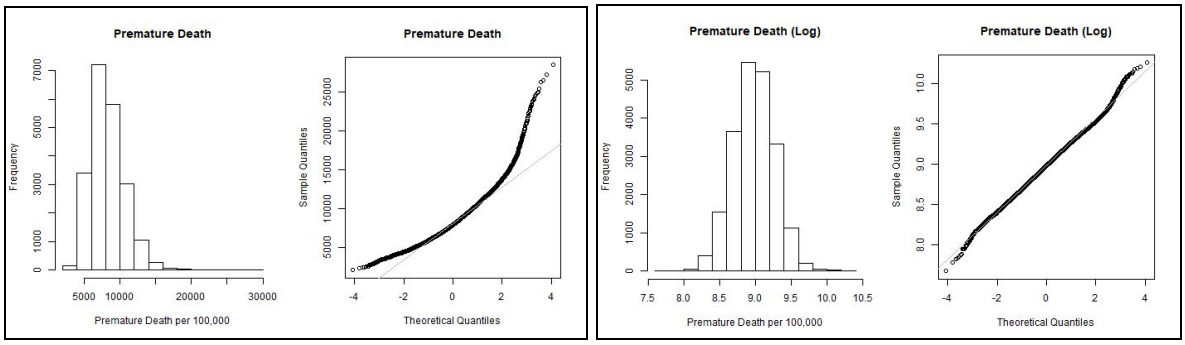

Figure 7: Premature Death Rate. The histogram and Q-Q plot of the rate of premature death has a right-skewed distribution (left image). The rate's log is roughly normally distributed (right image).

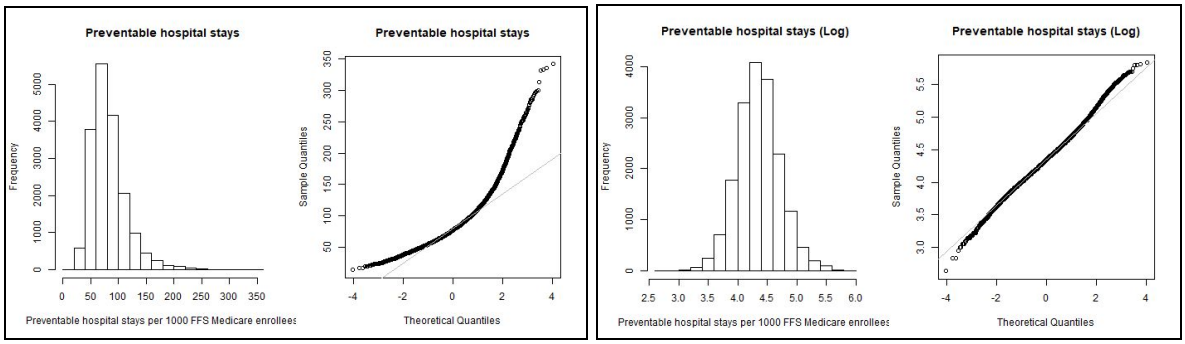

Figure 8: Preventable Hospital Stays rate. The histogram and Q-Q plot of the rate of preventable hospital stays has a right-skewed distribution (left image). The rate's log is roughly normally distributed (right image).

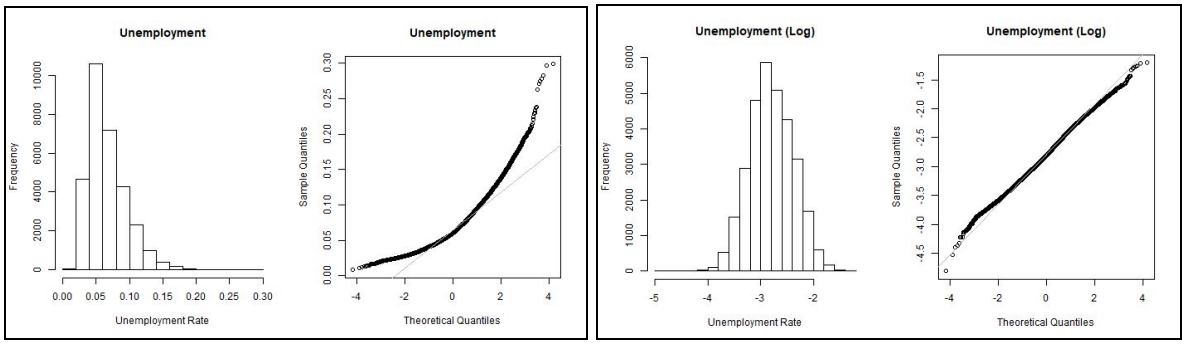

Figure 9: Unemployment Rate. The histogram and Q-Q plot of the rate of unemployment has a right-skewed distribution (left image). The rate's square root is roughly normally distributed (right image).

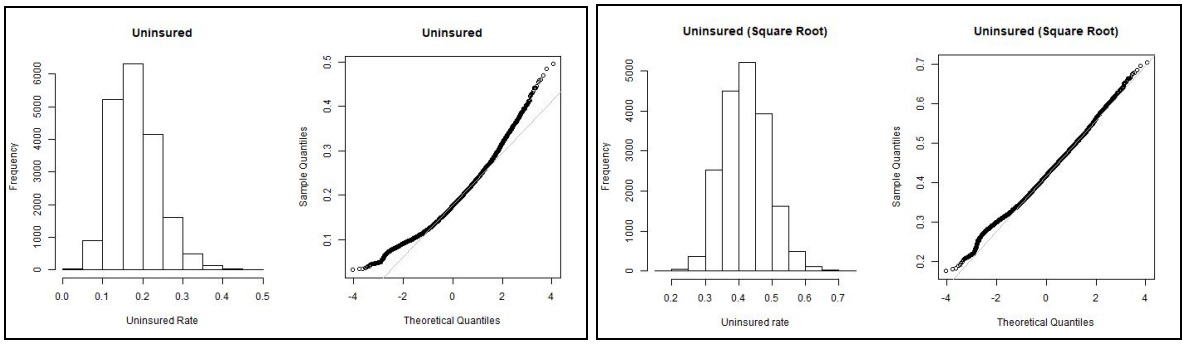

Figure 10: Uninsured Rate. The histogram and Q-Q plot of the rate of the uninsured has a right-skewed distribution (left image). The rate's square root is roughly normally distributed (right image).

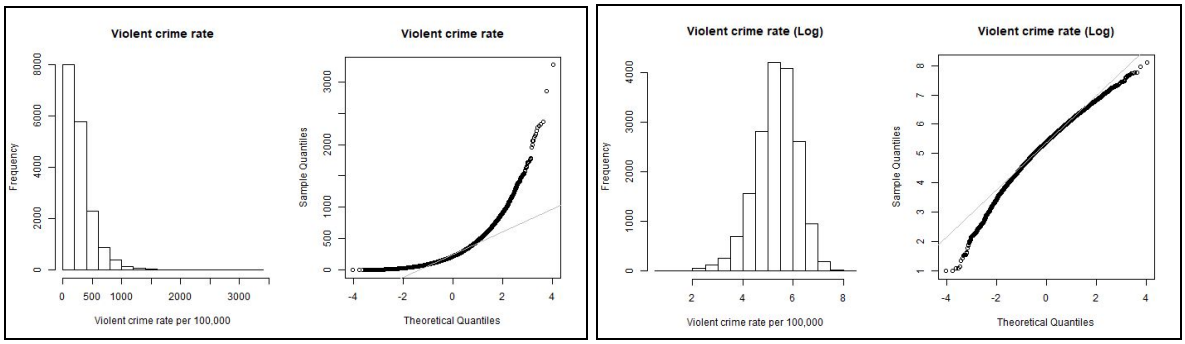

Figure 11: Violent Crime Rate. The histogram and Q-Q plot of the rate of violent crime has a right-skewed distribution (left image). The rate's log is roughly normally distributed (right image).

Appendix C

Scatterplot Matrices

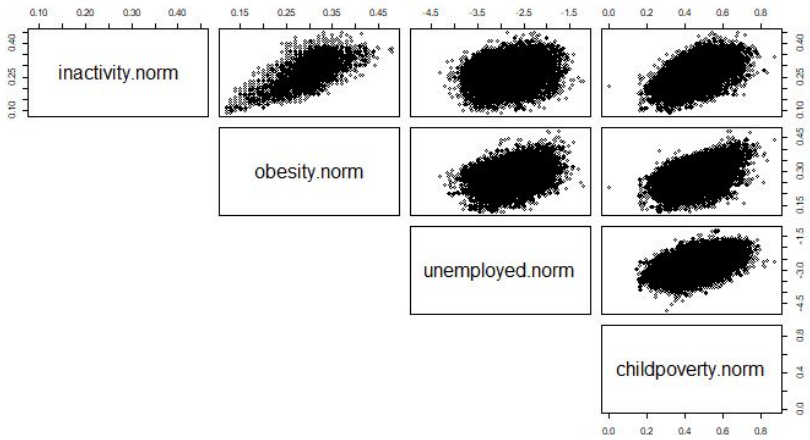

Figure 12: Scatterplot matrix showing relationships between normalized rates for Physical Inactivity, Adult Obesity, Unemployment, and Childhood Poverty.

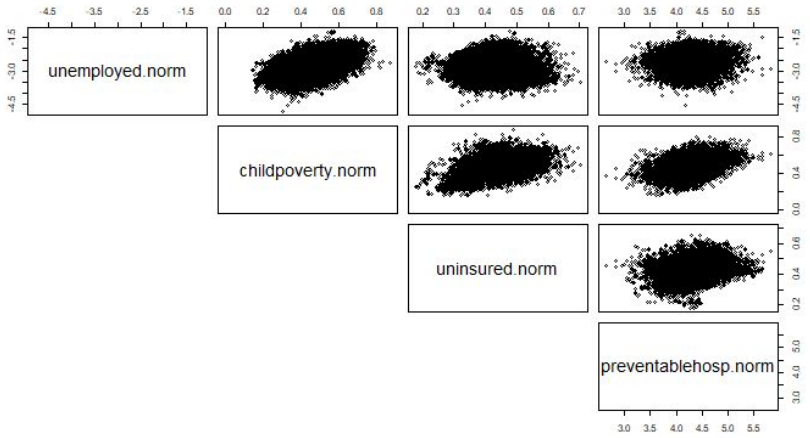

Figure 13: Scatterplot matrix showing relationships between normalized rates for Unemployment, Childhood Poverty, Uninsured, and Preventable Hospitalizations.

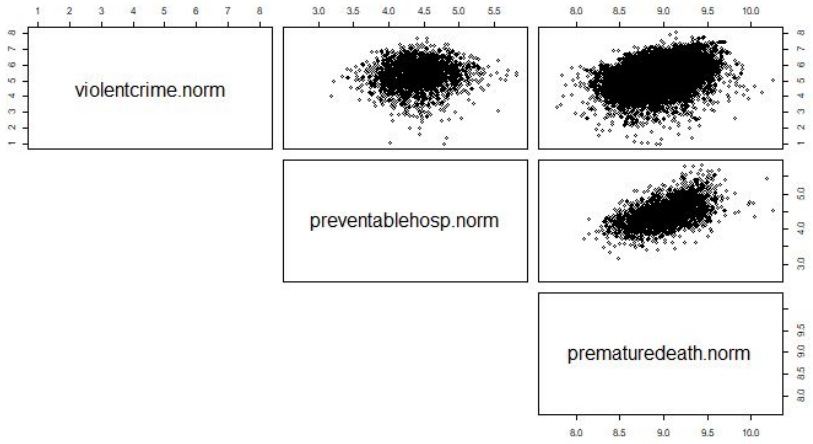

Figure 14: Scatterplot matrix showing relationships between normalized rates for Violent Crime, Preventable Hospitalizations, and Premature Death.

Appendix D

Boxplots of Measures by Location

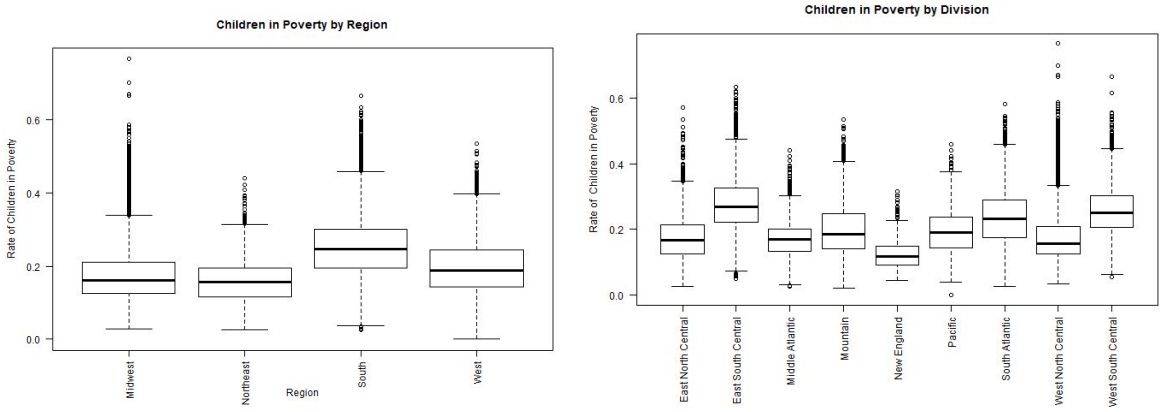

Figure 15: Boxplots of Children in Poverty by Location. By region (left) and division (right).

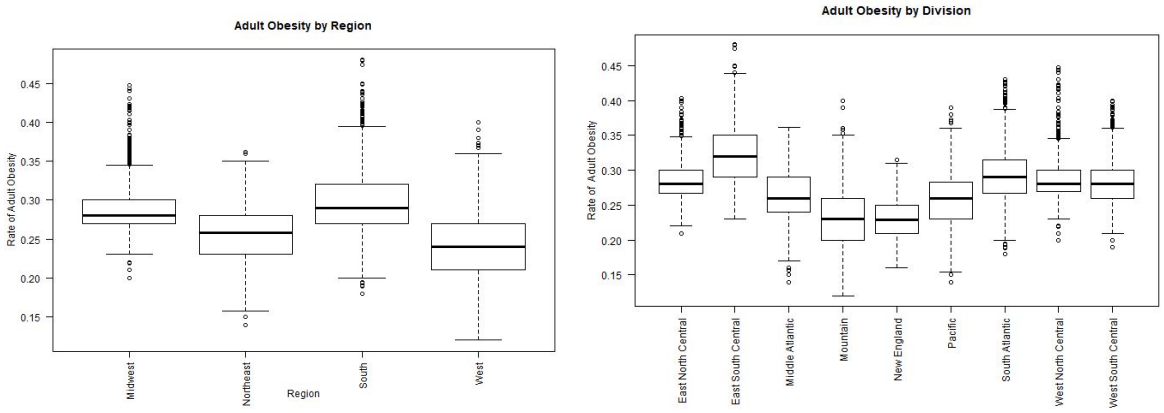

Figure 16: Boxplots of Adult Obesity by Location. By region (left) and division (right).

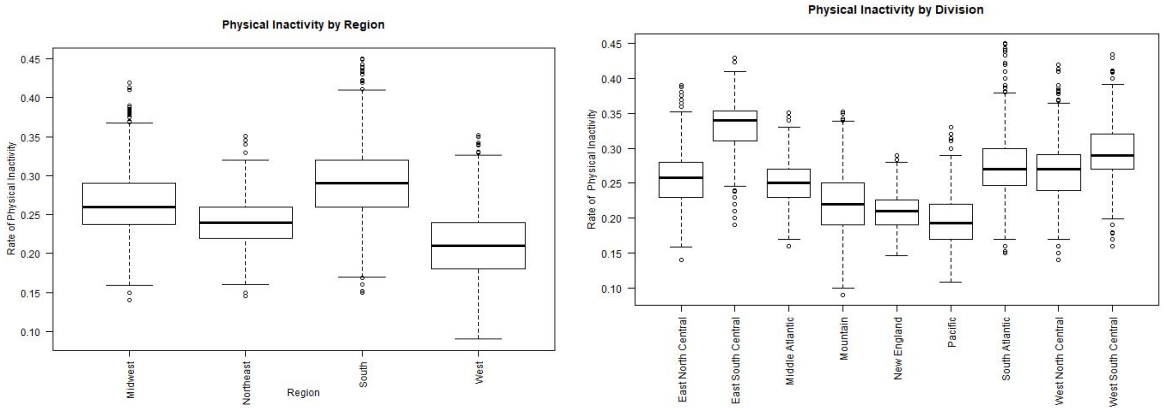

Figure 17: Boxplots of Physical Inactivity by Location. By region (left) and division (right).

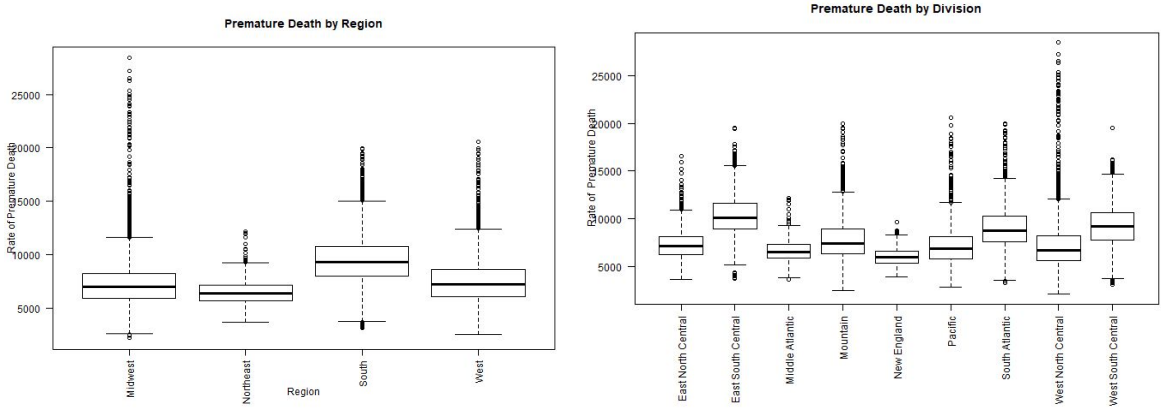

Figure 18: Boxplots of Premature Death by Location. By region (left) and division (right).

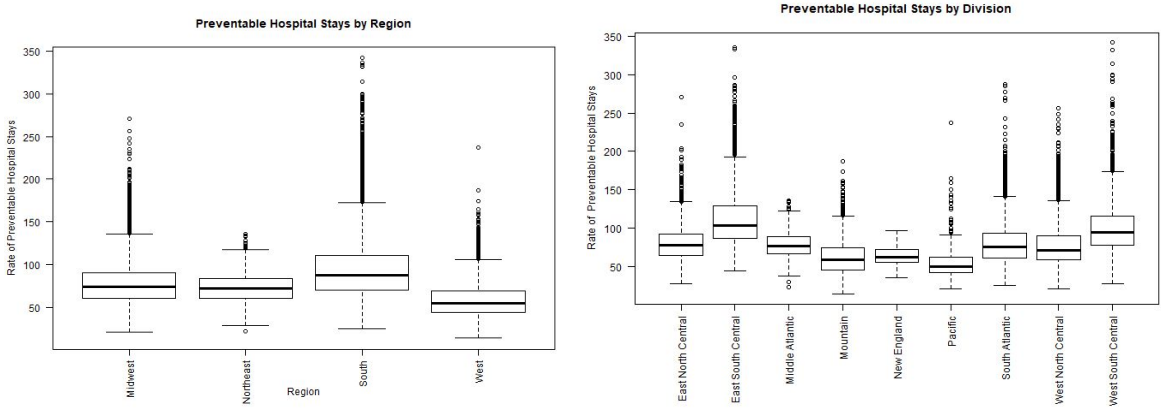

Figure 19: Boxplots of Preventable Hospital Stays by Location. By region (left) and division (right).

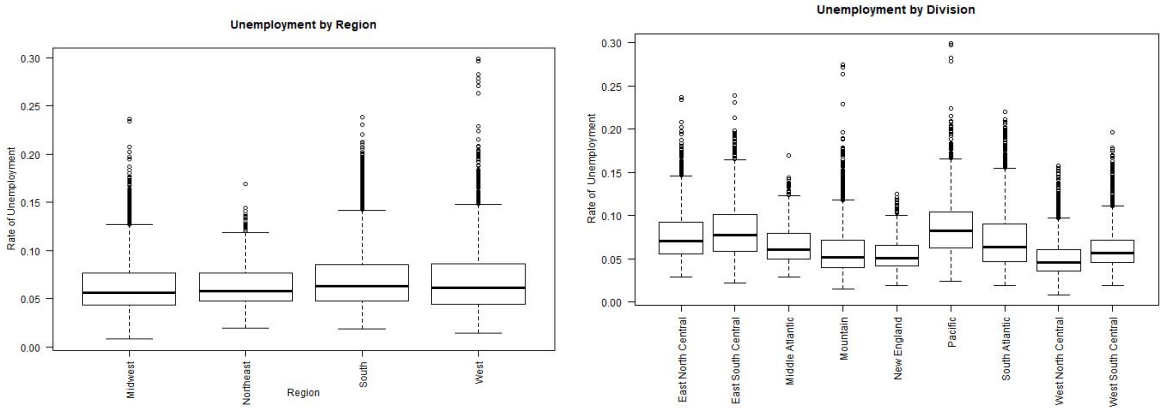

Figure 20: Boxplots of Unemployment by Location. By region (left) and division (right).

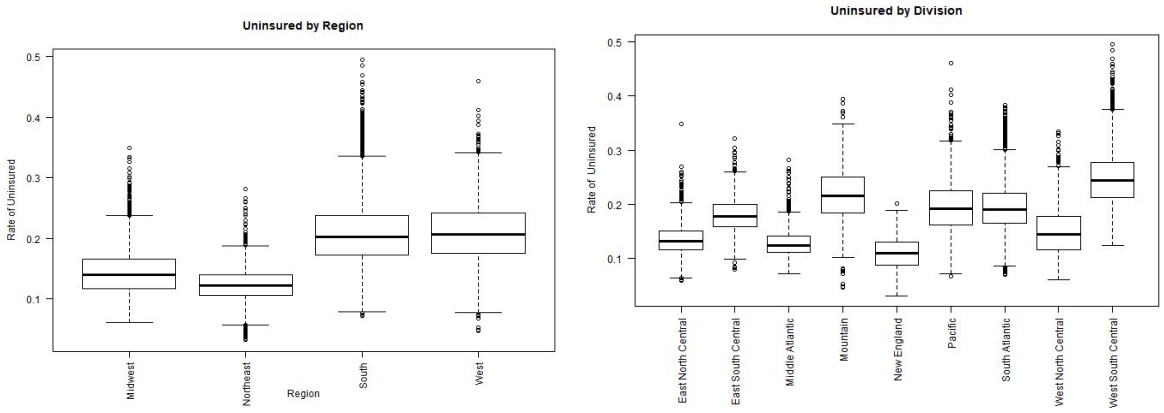

Figure 21: Boxplots of Uninsured Rates by Location. By region (left) and division (right).

Appendix D

Smooth Scatter Plots for Paired Measures

Scatter plots of normed rates were drawn to assess whether relationships between measures appeared to be linear.

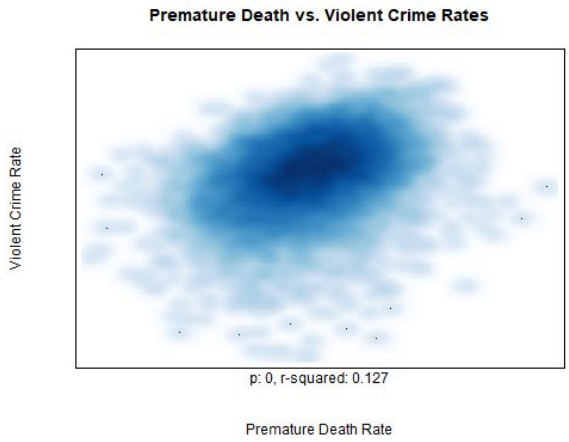

Figure 22: Scatter plot of violent crime versus premature death rates

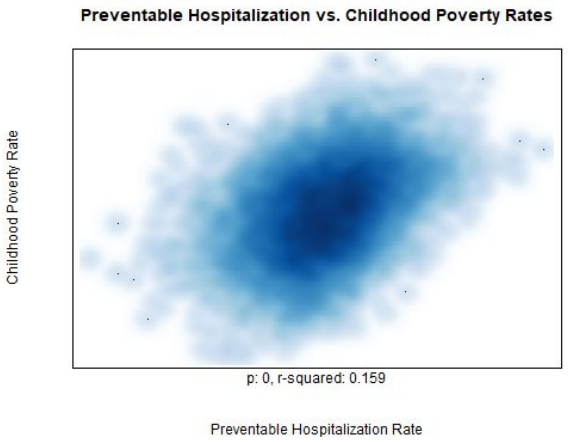

Figure 23: Scatter plot of childhood poverty versus preventable hospitalization rates

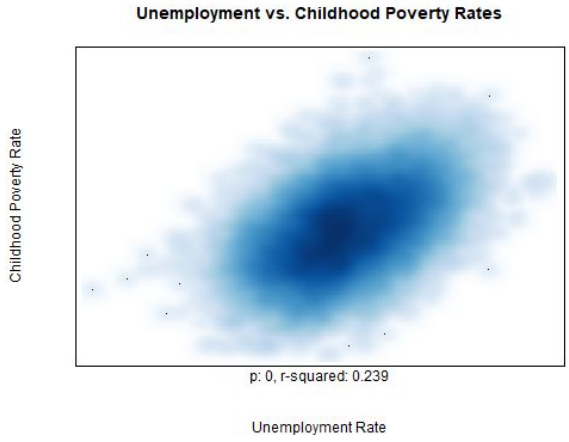

Figure 24: Scatter plot of childhood poverty versus unemployment rates

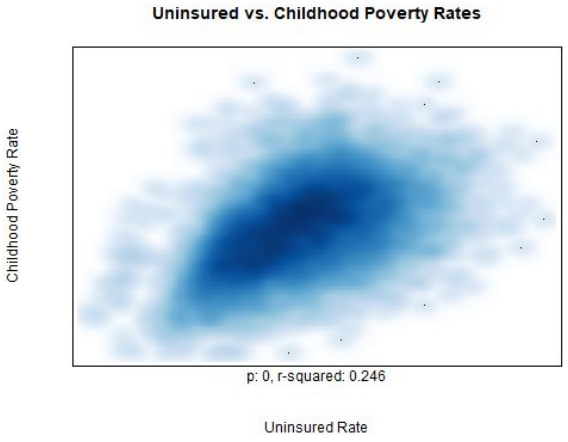

Figure 25: Scatter plot of childhood poverty versus uninsured rates

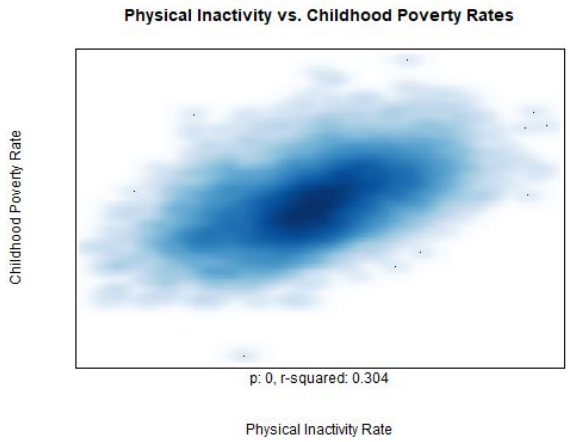

Figure 26: Scatter plot of childhood poverty versus physical inactivity rates

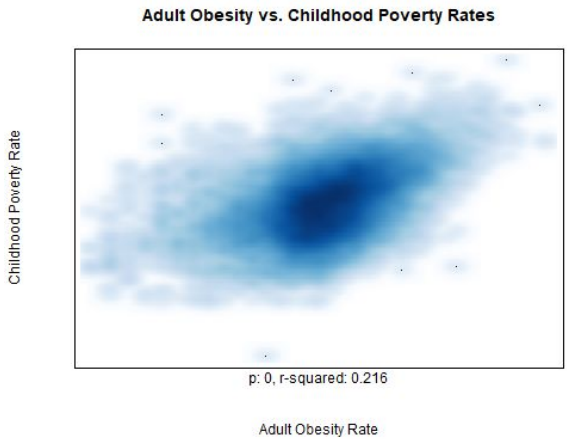

Figure 27: Scatter plot of childhood poverty versus adult obesity rates

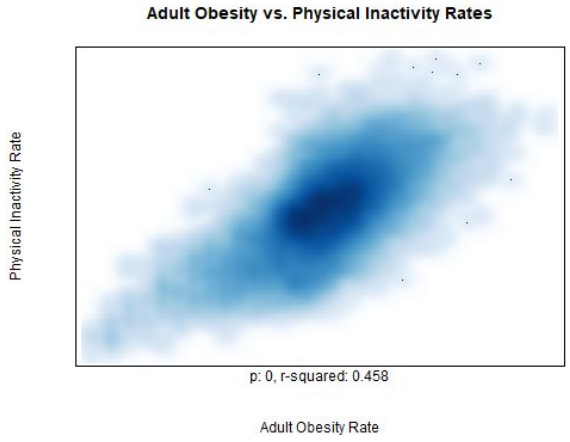

Figure 28: Scatter plot of physical inactivity versus adult obesity rates

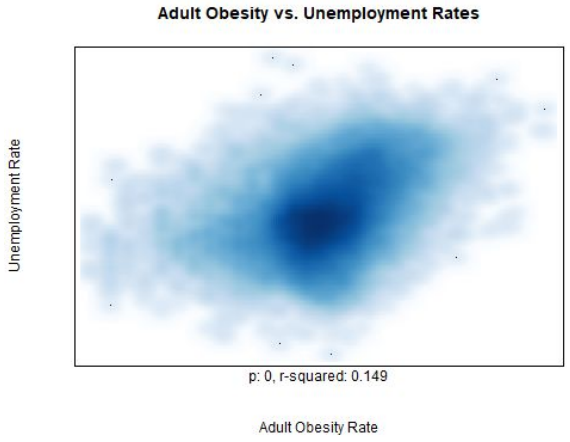

Figure 29: Scatter plot of unemployment versus adult obesity rates

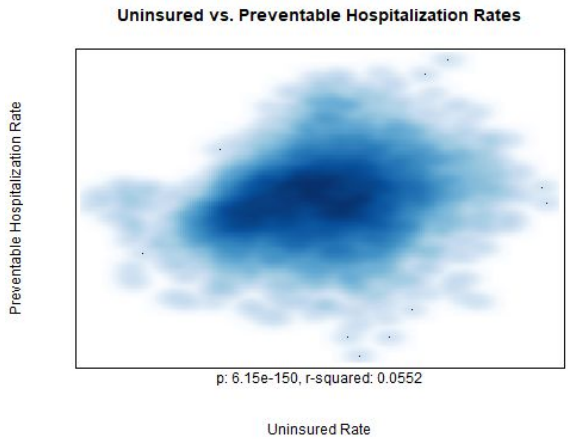

Figure 30: Scatter plot of preventable hospital versus uninsured rates

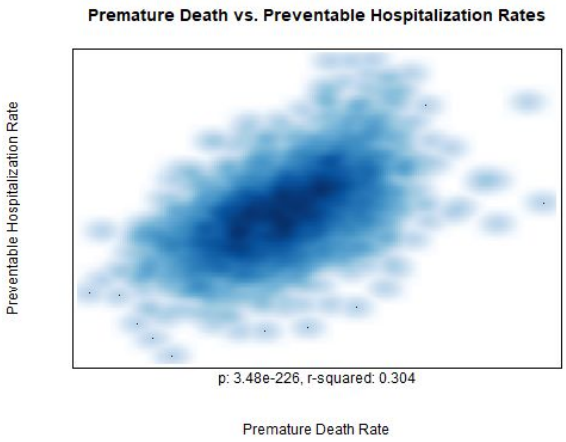

Figure 31: Scatter plot of preventable hospital versus premature death rates