Large Differences in County-Level Mortality Rates Related to Race and Economic Advantage

Abstract

Mortality, and causes of death, can often be ascribed to causes such as lack of access to health care or environmental factors. Much research has described links between health outcomes, socioeconomic status and race. Furthermore, race and socioeconomic status are closely related. Using data from 1994-2003, we studied differences in death rates among counties in the United States. We determined four clusters of counties with similar mortality profiles. Employing principal component analysis and random forest classification, we determined which sets of county-level variables best classifies counties into the four mortality clusters. Economic and racial data are important in our classification model, and are also significantly related to rates of plausibly preventable deaths. In particular, economically disadvantaged counties with larger black population proportions have higher death rates. Our analysis supports the theory that economic and racial factors relate to health outcomes.

Introduction and Background

Mortality concerns all of us. Death is often considered as purely the result of easily identifiable causes such as cancer, heart disease, and car crashes. But there are other causes that are not immediately identifiable. Researchers have long been aware that mortality rates are related to socioeconomic status.[1,2] Access to health care can be the difference between life and death. Understanding what factors are correlated with high mortality rates can ultimately help explain why people die; it makes it possible to estimate where the rates of preventable deaths are highest, and thus know where to focus on reducing the number of such tragedies.

Scholars have studied the link between class and race, and the connection to health; for example, blacks are disproportionately poor and have higher mortality rates.[3] The counties in the US differ in racial composition, general economic variables, and in mortality rates for different causes of death. We expect death rates to relate to economic variables, race and the level of health care access. Research has demonstrated that mortality rates are higher for populations with fewer economic resources.[1] It is also clear that mortality rates are associated with race even after accounting for socioeconomic factors.[2,4,5]

We analyzed differences in mortality rates by studying health, economic, and demographic data for each county in the USA. By performing a clustering analysis, we identify counties that have similar death rates for some common causes of death: lung cancer, breast cancer, colon cancer, cardiovascular disease, injury, and motor vehicle accidents. With principal component analysis and random forest classification, we explore what types of data best separate the counties in our clusters. Using multiple linear regression, we study mortality rates for the two most common causes[7] of death in the US: cancer and cardiovascular disease.

Methods

We used R Studio 3.1.2 with the packages *stats*, *randomForest*, and *maps* for analysis.

Data Source and Dimension Reduction

Our data came from the Community Health Status Indicators to Combat Obesity, Heart Disease and Cancer, published by the Centers for Disease Control and Prevention on data.gov.[6] The dataset includes a large number of variables for each county in the United States, collected by the federal government between 1994 and 2003. These data include many missing values, which we sought to manage without introducing bias. We applied a randomized multiple imputation method based on multiple linear regression described by Gelman et al[8] to impute values for variables that had missing data for less than 10% of the counties.

To reduce the dimension of the data and make interpretations easier, we performed Principal Components Analysis (PCA). Because the set of variables is large and many variables are uncorrelated, we performed PCA on selected correlated subsets of variables. We chose our Principal Component indices (PCs) so that they explained at least 50% of the variance in their subsets. We created new variables for deaths due to cancer (breast, colon and lung cancer) and for deaths due to cardiovascular disease (coronary heart disease and stroke).

Clustering and Classification

To see if groupings of counties based on average life expectancy and mortality rates existed, we applied K-means clustering on these variables. After identifying clusters, we mapped them to better visualize the results. We also compared mortality rates between the clusters (see appendix).

We also wanted to see if we could classify the counties into the clusters of similar mortality rates using the indices (PCs) *not* based on mortality rates. It is important to note that the indices we used for classification do *not* contain any of the data that we used for clustering. We employed a Random Forest model for classification because, compared to other methods we tried, it had the lowest predicted true error rate. It also allowed us to easily compare the classification importance of the indices used in the model.

Multiple Linear Regression

To further identify which indices are most important for explaining differences in death rates, and the directions of their relationships to death rates, we performed Multiple Linear Regression (MLR). We used all of our indices to predict the cancer death rate, the cardiovascular disease death rate, and the accidental deaths index. We excluded 55 counties that had index values greater than 5 standard deviations from the mean for any index. We excluded the Medicare index from the accidental deaths model to avoid multi-collinearity issues (VIF values greater than 5). Our models met the conditions for inference (see appendix).

Results

We created the following PCs with accompanying interpretations: PopPC, an index of population size and density; YoungPopPC, an index of the youth of the population; WhitePC and BlackPC, indices of the proportions of white residents and black residents, respectively; InfantHealthPC, an index of the health of infants; EconAdvantagePC, an index of the residents' economic advantage; HCaccessPC, an index of health care access; MentalHealthPC, an index of mental health; IllnessReportsPC, an index of reporting rates of common illnesses; MedicarePC, an index of Medicare patients; AccidentDeathsPC, an index for accidental deaths.

Clustering and Classification

We chose to split our counties into four clusters based on interpretive value and the shape of the within-group sum of squares versus number of clusters plot (see appendix). The clusters form a gradient of mortality rates: cluster 1 has the lowest rates and cluster 4 the highest rates. Differences between clusters are larger for more preventable causes of death such as breast cancer and heart disease (see appendix). As expected, average life expectancy differs substantially between the clusters, as shown in Figure 1. Figure 2 maps the counties. In addition to death rates, the clusters from 1 to 4 form gradients of decreasing economic advantage, decreasing health care access, and increasing Black population index (see appendix). Remember that these variables were not used for clustering.

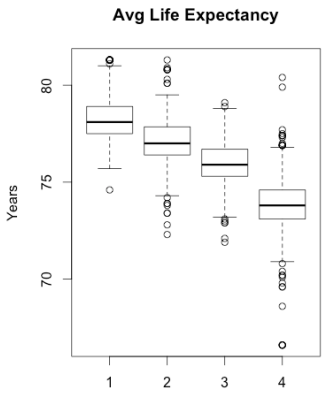

Fig. 1. Average life expectancy (in years) across clusters.

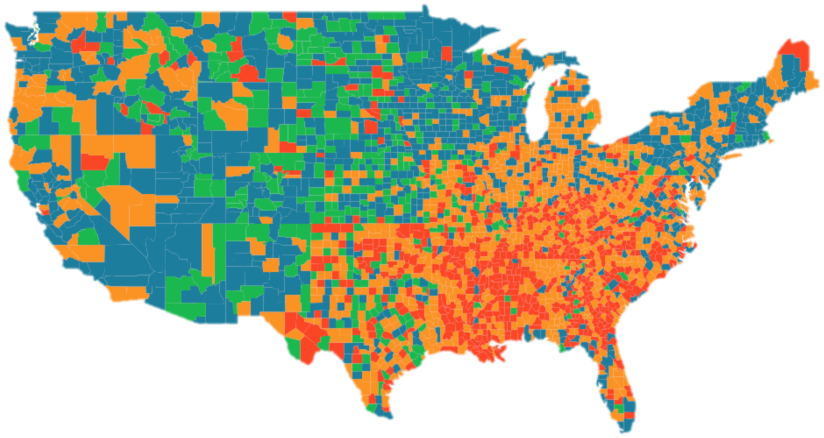

Fig. 2. Map of County Clusters. Colors correspond to increasing mortality rates: blue, green, yellow and red.

Using the indices not based on mortality rates, the Random Forest model classified counties into the clusters with an estimated true error rate of 30.4%, which is much lower than the root node error of 65.2%. Thus, with decent accuracy, the model predicts which mortality clusters the counties fall into despite using *none* of the information used to create these clusters.

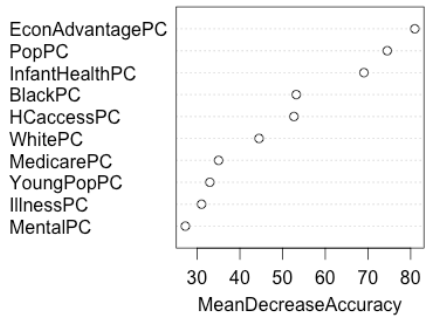

Fig. 3. (Left) The most important PCs for classifying counties (see appendix for more details).

Figure 3 shows the indices that are most useful for classifying counties, with more important indices being listed in higher positions. It is not surprising that the economic, population, infant health and health care access indices are useful. We expect health outcomes to be correlated with these sets of data. It is more surprising that the race indices are more important than the illness, Medicare and mental health indices.

Multiple Linear Regression (MLR)

We explored the data further by performing MLR to predict mortality rates using our indices. First, we predicted the cancer mortality rate using the other non-death indices as predictors (R^2 value of 0.36; all coefficients' p-values < 0.02 except MedicarePC).

$$\begin{aligned}\text{Cancer} = {} & 106.1 + 6.1*\text{BlackPC} - 7.1*\text{EconAdvantagePC} + 3.0*\text{PopPC} \\ & - 3.9*\text{InfantHealthPC} - 1.2*\text{IllnessReportsPC} + 0.8*\text{MentalHealthPC} \\ & + 2.7*\text{YoungPopPC} + 1.9*\text{WhitePC} + 1.2*\text{HCaccessPC} + 0.7*\text{MedicarePC}\end{aligned}$$

We standardized each index, so that we can compare the effect of a one-unit increase in each index (equivalent to a one standard deviation difference). The economic advantage, black, and infant health indices have the largest effects on the predicted cancer death rates.

We created similar models to predict the cardiovascular disease mortality rate and the accidental death rate index (see appendix). Both models were significant (R^2 values of 0.34 and 0.38, respectively). In the accidental deaths model, increases in economic advantage, population and access to health care—the indices with the largest coefficients—decrease the predicted death index. In the cardiovascular disease model, the economic advantage, infant health and race indices have the largest coefficients—the first two in the negative direction while the white and black indices have large positive coefficients.

Discussion

Research has demonstrated relationships between class and race, as well as between class, race and mortality. We find that we can identify four groups of counties with similar mortality rates. The most important data for separating the clusters are those on economic advantage, population, infant health, race and health care access.

Our MLR models further demonstrate that these variables explain differences in mortality rates. We can rationalize the accident death model as those living in wealthier counties with better access to health care can likely get emergency care more quickly after an injury or motor vehicle accident. The negative coefficients on the economic advantage index in the cancer deaths model is expected. But the large positive coefficient on the black index is harder to explain. It reflects that even after controlling for economic conditions, counties with larger black population proportions have substantially higher rates of death from cancer. This is not a simple case of multi-collinearity as the correlations between the black index and the other indices are low (see appendix for full index correlation table). The high coefficient on BlackPC really does appear to be an issue of race. Better economic conditions lead to a dramatic reduction in cardiovascular disease, likely reflecting better preventative care and better long-term care.

Our work demonstrates the importance of racial and economic data for explaining differences in county-level mortality rates. Poorer and denser counties with a greater black proportion of the population have increased mortality rates, especially for more preventable causes of death, and likely suffer a large number of deaths that could be prevented with access to quality health care.

References

1. Aaron Antonovsky. 1967. “Social Class, Life Expectancy and Overall Mortality.” *The Milbank Memorial Fund Quarterly* 45 (2): 31–73. doi:10.2307/3348839.

2. McLaughlin, Diane K., and C. Shannon Stokes. 2002. “Income Inequality and Mortality in US Counties: Does Minority Racial Concentration Matter?” *American Journal of Public Health* 92 (1): 99–104. doi:10.2105/AJPH.92.1.99.

3. Murray, Christopher J. L, Sandeep C Kulkarni, Catherine Michaud, Niels Tomijima, Maria T Bulzacchelli, Terrell J Iandiorio, and Majid Ezzati. 2006. “Eight Americas: Investigating Mortality Disparities across Races, Counties, and Race-Counties in the United States.” *PLoS Med* 3 (9): e260. doi:10.1371/journal.pmed.0030260.

4. DeSantis, Carol, Rebecca Siegel, Priti Bandi, and Ahmedin Jemal. 2011. “Breast Cancer Statistics, 2011.” *CA: A Cancer Journal for Clinicians* 61 (6): 408–18. doi:10.3322/caac.20134.

5. Ward, Elizabeth, Ahmedin Jemal, Vilma Cokkinides, Gopal K. Singh, Cheryll Cardinez, Asma Ghafoor, and Michael Thun. 2004. “Cancer Disparities by Race/Ethnicity and Socioeconomic Status.” *CA: A Cancer Journal for Clinicians* 54 (2): 78–93. doi:10.3322/canjclin.54.2.78.

6. Centers for Disease Control and Prevention. Community Health Status Indicators to Combat Obesity, Heart Disease and Cancer. http://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer. October 21, 2014.

7. Centers for Disease Control and Prevention – FastStats - Leading Causes of Death. http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. December 9, 2014.

8. Gelman, A., and Hill, J., *Data Analysis Using Regression and Multilevel/Hierarchical Models,* 2012, Cambridge University Press, Cambridge, UK. 648 pp.

Appendix

1. K-means Clustering

Variables used for clustering:

- Average Life Expectancy
- Total Mortality Rate
- Infant Mortality Rate
- Breast Cancer Mortality Rate
- Colon Cancer Mortality Rate
- Coronary Heart Disease Mortality Rate
- Lung Cancer Mortality Rate
- Motor Vehicle Accident Mortality Rate
- Stroke Mortality Rate
- Injury Mortality Rate

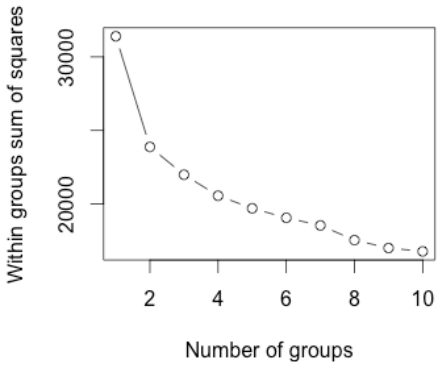

Figure A1 (left): Within Groups Sum of Squares vs. number of clusters for k-means clustering. The plot shows that two clusters is probably the best clustering solution, but we chose four in order to get a better gradient of death rates. Choosing four is also not completely unreasonable, as the slope of the line from 3 to 4 is slightly steeper than to the right of 4, and the slope stays fairly constant to the right of 4. For the solution with 4 clusters, between_SS / total_SS = 34.5%.

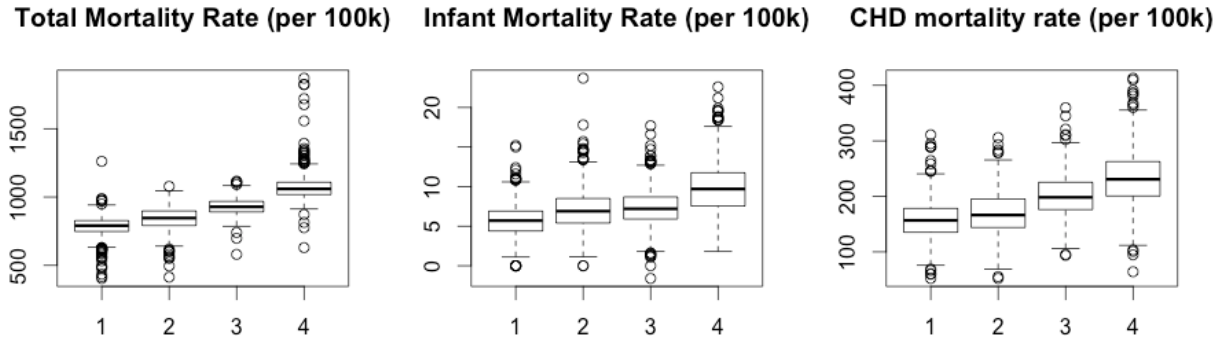

Figure A2: Boxplots of mortality rates by cluster. CHD stands for coronary heart disease. The plots show a clear gradient of mortality rates from lowest in cluster 1 to highest in cluster 4. Mortality rates are given for each county as the number of deaths per 100,000 residents.

Table 1: Some Summary Statistics By Cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Counties in Cluster	954	407	1089	691
Median Average Life Expectancy	78.1 years	77.0 years	75.9 years	73.8 years
Median Total Mortality Rate	791 / 100k	846 / 100k	928 / 100k	1060 / 100k
Median CHD Mortality Rate	157 / 100k	166 / 100k	198 / 100k	231 / 100k

2. Principle Component Analysis – Indices

Table 2A: PC Indices Used for Classification and as Predictors for MLR

Name of PC	Underlying Variables (loading signs in parenthesis; in order of loadings)
Population	Population Size(+), Population Density(+)
YoungPop	Pop% Age 65-84(-), Pop% Age 19-64(+), Pop% Age <19 (+)
White	Pop% White(+), Pop% Black(-), Pop% Asian(-), Pop% Hispanic(+)
Black	Pop% Asian(-), Pop% Hispanic(-), Pop% Black (+)
InfantHealth	Premature Birth Rate(-), Low Birth Weight Rate(-), Birth Rate to Unmarried Mothers(-), Birth Rate to Mothers Under 18(-), Very Lower Birth Weight Rate(-), Birth Rate to Mothers over 40(+)
EconAdvantage	Pop% without HS Diploma(-), Poverty Rate(-), Pop% on Medicare Disability(-), Pop% Severely Work Disabled(-), Pop% Unemployed(-)
HCaccess	Pop% Uninsured(-), Dentist Rate(+), Physician Rate(+), Late Prenatal Care Rate (-)
MentalHealth	Major Depression Rate(-), Rate of Recent Drug Use(-)
IllnessReports	Hepatitis A Rate(+), Syphilis Rate(+), Hepatitis B Rate(+), Flu Rate(+), Pertussis Rate(+), Measles Rate(+)
Medicare	Pop% with Elderly Medicare(+), Pop% on Medicare Disability(+)

Table 2B: PC Indices Used as Response Variables for MLR

Name of PC	Underlying Variables (loading signs in parenthesis; in order of loadings)
AccidentDeaths	Motor Vehicle Accident Mortality Rate(+), Injury Mortality Rate(+)

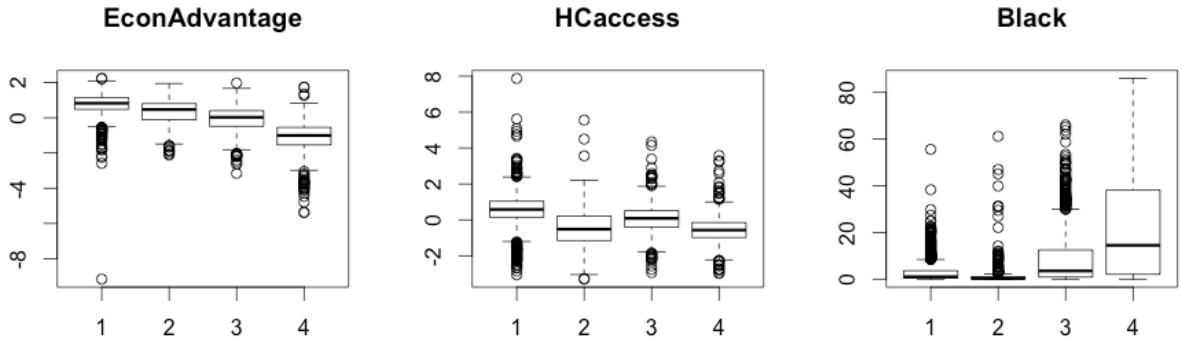

Figure A3: Boxplots of three PC indices rates by cluster. The EconAdvantage and Black indices are clearly correlated with the mortality rates shown in Figure A2. The HCaccess and Black boxplots highlight the large number of outliers – counties that have unusually large black populations for instance.

Table 3: Correlations Between Indices (PCs)

	Illness									
AccidentDeaths	-0.19	Acc..								
White	-0.18	-0.10	White							
Black	-0.33	0.12	0.00	Black						
EconAdvantage	0.04	-0.39	0.34	-0.09	Econ					
Hcaccess	0.17	-0.51	0.13	0.04	0.46	HC				
Medicare	-0.17	0.26	0.02	0.30	-0.56	-0.08	MC			
MentalHealth	0.01	0.06	-0.15	-0.09	0.01	-0.15	-0.03	Mental		
InfantHealth	-0.01	-0.25	0.70	-0.08	0.66	0.39	-0.20	-0.15	Infant	
YoungPop	0.17	-0.15	-0.29	-0.26	0.08	0.06	-0.65	-0.06	-0.15	Young
Pop	0.83	-0.23	-0.17	-0.35	0.07	0.23	-0.16	0.01	0.03	0.17

3. Random Forest Classification

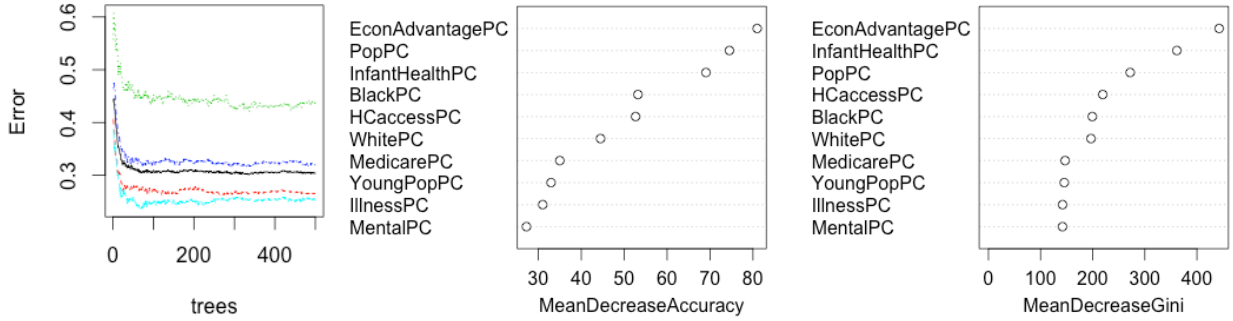

Figure A4: The leftmost graph shows the out-of-bag error rates as the number of trees increases. The rates are fairly constant approaching 500 trees, which means that we probably do not need to fit more trees. The charts on the middle and right show rankings of the most important variables for classification, with the most important variables at the top. *MeanDecreasyAccuracy* and *MeanDecreaseGini* are measures of how well the variables separate counties into different clusters. Higher values mean that the variables are more useful for classification.

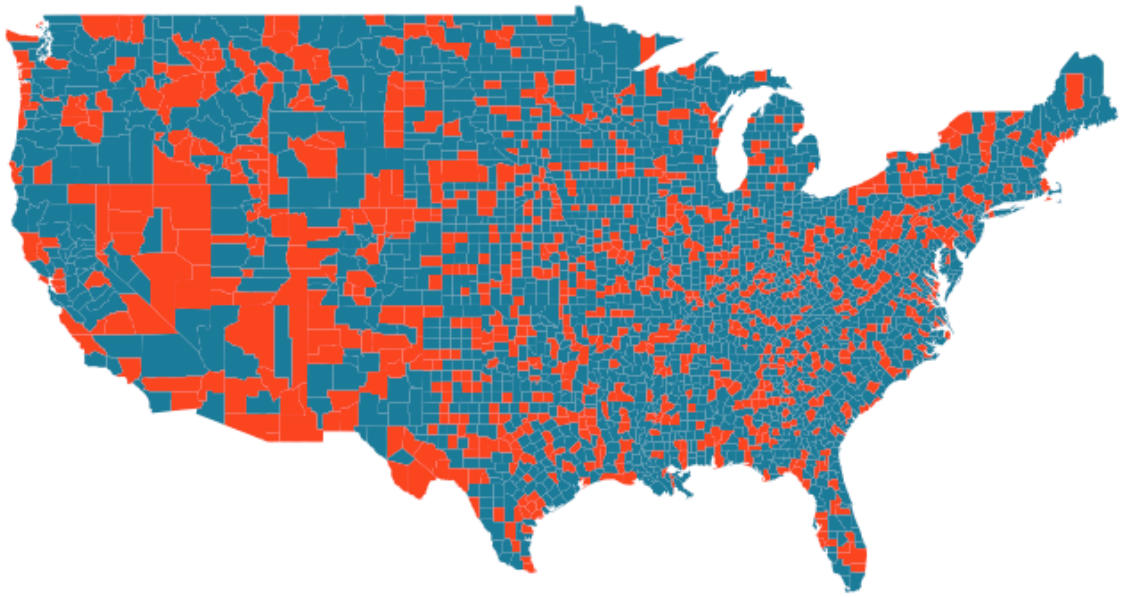

Figure A5: Map showing which counties the Random Forest model predicted right (blue) and which it predicted wrong (red). This is based on 'out-of-bag' predictions.

4. Multiple Linear Regression Models

We removed 55 obvious outliers before fitting models. This left us with 3093 observations (counties). We rescaled the indices after removing the outliers so that they still have mean 0 and standard deviation 1. For each model, the residuals have mean of zero by using least squares and counties are independent. We assume linear relationships.

Table 4: MLR Output – Predicting Cancer Mortality Rates

Cancer ~ PopPC + YoungPopPC + WhitePC + BlackPC + InfantHealthPC + EconAdvantagePC + HCaccessPC + MentalPC + IllnessPC + MedicarePC				
	F(10,3099)	Prob > P	R-Squared	Adj-R^2
	177.228	0	0.364	0.362

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	106.079	0.258	411.847	<0.001
PopPC	3.03	0.512	5.914	<0.001
YoungPopPC	2.693	0.422	6.383	<0.001
WhitePC	1.914	0.434	4.414	<0.001
BlackPC	6.135	0.335	18.331	<0.001
InfantHealthPC	-3.858	0.499	-7.726	<0.001
EconAdvantagePC	-7.125	0.517	-13.776	<0.001
HCaccessPC	1.198	0.352	3.408	0.001
MentalPC	0.783	0.273	2.867	0.004
IllnessPC	-1.157	0.459	-2.522	0.012

(Omitted non-significant coefficients)

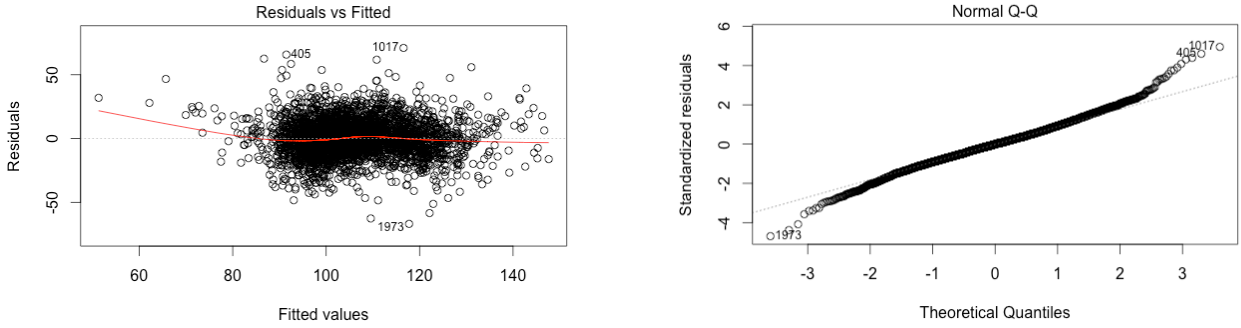

Figure A6: Residuals vs. fitted plot and residual QQ plot for the Cancer Deaths model. Inference conditions seem to be met.

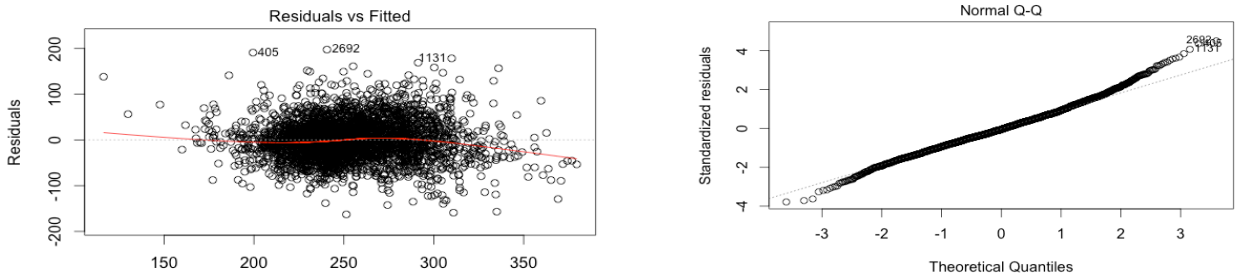

Figure A7: Residuals vs. fitted and residual QQ plot for the Cardiovascular disease model. Inference conditions seem to be met.

Table 5: MLR Output – Predicting Cardiovascular Disease Mortality Rates

Cardio ~ PopPC + YoungPopPC + WhitePC + BlackPC + InfantHealthPC + MentalPC + EconAdvantagePC + HCaccessPC + IllnessPC + MedicarePC				
	F(10,3105)	Prob > P	R-Squared	Adj-R^2
	159.379	0	0.339	0.337

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	254.109	0.787	322.744	<0.001
YoungPopPC	5.005	1.29	3.879	<0.001
WhitePC	9.251	1.314	7.04	<0.001
BlackPC	9.543	1.006	9.49	<0.001
InfantHealthPC	-12.462	1.521	-8.194	<0.001
EconAdvantagePC	-21.737	1.579	-13.764	<0.001
MentalPC	8.48	0.834	10.17	<0.001

(Omitted non-significant coefficients)

Table 6: MLR Output – Predicting Accidental Mortality Rates

AccidentPC ~ PopPC + YoungPopPC + WhitePC + BlackPC + InfantHealthPC + EconAdvantagePC + HCaccessPC + MentalPC + IllnessPC				
	F(9,3083)	Prob > P	R-Squared	Adj-R2
	209.445	0	0.379	0.378

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.032	0.013	-2.503	0.012
HCaccessPC	-0.387	0.017	-22.307	<0.001
PopPC	-0.219	0.047	-4.629	<0.001
EconAdvantagePC	-0.217	0.02	-10.848	<0.001
BlackPC	0.108	0.02	5.487	<0.001
YoungPopPC	-0.095	0.015	-6.457	<0.001
InfantHealthPC	0.08	0.025	3.21	0.001
WhitePC	-0.053	0.022	-2.46	0.014

(Omitted non-significant coefficients)

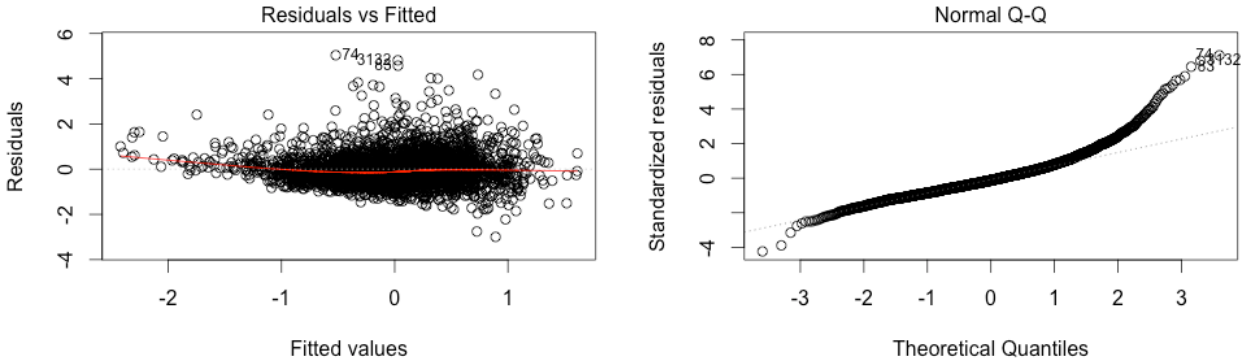

Figure A8: Residuals vs. fitted plot and residual Normal QQ plot for the Accidental Deaths model. Inference conditions seem to be met, although the upper tail of the QQ plot is deviating somewhat.