ISYE 6414

REGRESSION ANALYSIS

TEAM PROJECT

UNDERSTANDING MORTALITY AMONGST RACES IN THE UNITED STATES

TEAM 6

Kunaal Ahuja

William Bass

Parvathy Sarat

Table of Contents

Abstract
Recently, research has shown that the all-cause mortality rate in the United States has been increasing for middle-aged white non-Hispanic men and women. This report expands on this finding by exploring the demographic factors that may impact age at the time of death for the broader population of the United States, with the goal of identifying populations at-risk of earlier death. Following regression analysis using Whites as the baseline, with all variables present in the model, it was independently confirmed that Black and American Indian populations still die at a younger age. From these findings, we recommend more public health initiatives targeting the relationship between systemic racism and race-related mortality rate disparities, with the goal of improving the life expectancy of minority populations in the United States.

Introduction
The mortality rate in the United States has been in steady decline since the 1970s. This success has largely been attributed to improved prevention and treatment of diseases as well as increased public knowledge of health risks (1). The overall decrease in mortality has not been limited to any demographic subset. In fact, life expectancy has been on the rise globally and has been expected to continue improving.

Recently, however, an increase in the all-cause mortality of middle-aged white non-Hispanic men and women in the United States has been reported. This increase has been found to be unique to this subpopulation of people with no parallel in similarly wealthy nations (2). Further investigation has attributed this phenomenon to a rise in suicides and drug poisonings. This public health reversal has been reported on extensively in popular media and is often paired with reports on the “opioid crisis” currently affecting the United States (3).

Despite this media attention, minority and indigenous populations in the US have long been reported to die at an increased rate compared with Whites (4). In this report, we examine the mortality data set produced yearly by the Centers for Disease Control and Prevention (CDC) under the National Vital Statistics Systems. The mortality data set contains extensive information on the demographic details and causes of deaths for every death recorded in the US. Using a combination of regression techniques, this report seeks to independently confirm that minority populations are still at risk of premature death, despite the recently documented increase in white mortality, by exploring the demographic factors that may impact age at time of death.

Methods

Figure 1 below displays an overview of the sequence of regression techniques applied to the mortality data set.

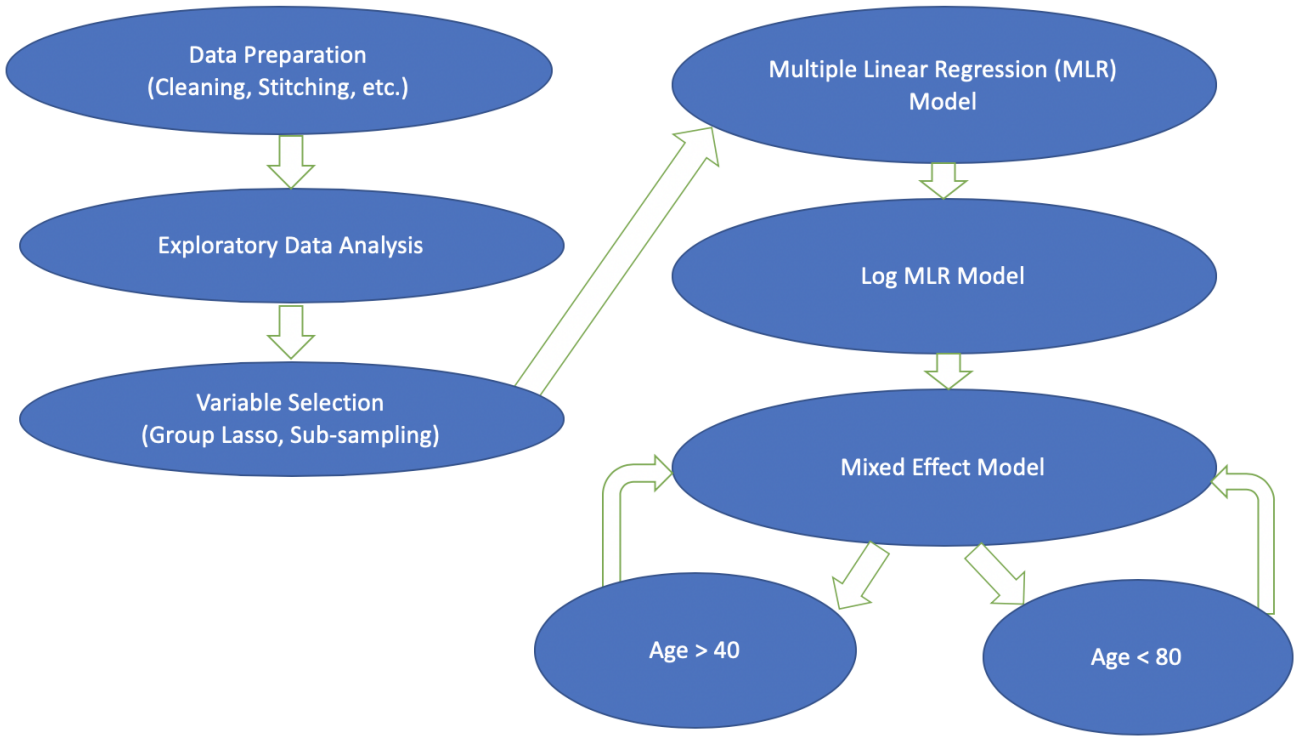

Figure 1: Progression of Regression Analysis

Mortality Data Set

All data used are publicly available from the CDC's National Vital Statistics Systems. The data were pulled from kaggle.com/cdc/mortality. The mortality data set contains extensive information on the demographic details and causes of deaths for every death recorded in the United States. The predicting variables used to predict age (at time of death) were population characteristics (race, hispanic, education), and year (at time of death). Cause_of_death, marital_status, and resident_status (reported as 39 grouped ICD10 codes) were included as controlling variables. See the supplementary figures for further information.

Data Wrangling

Predicting variables were recoded from the json file in the dataset to make them into categorical variables. Data from 2011-2015 was included for this analysis and sub-sampled across the years to make a 400K data samples data set. Few predictors were categorized into bins to make the inference better. For example, education column was grouped into 4 bins - 'Up to High School','College < 4 years', 'College > 4 years' and 'unknown'. See the supplementary figures and tables for more details on columns and their labels.

Exploratory Data Analysis

Exploratory data analysis was conducted by plotting a histogram of the response variable (Figure 2) and by plotting each covariate against the response. Figure 3 and Figure 4 show the boxplots for the marital status and cause of death, respectively. Boxplots for the remaining variables are included in the supplementary figures.

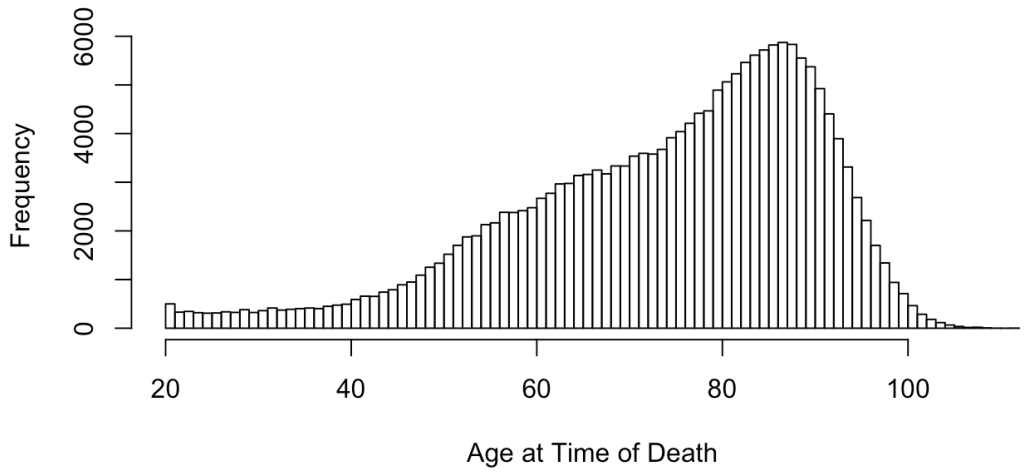

Figure 2: Histogram of the response variable - age

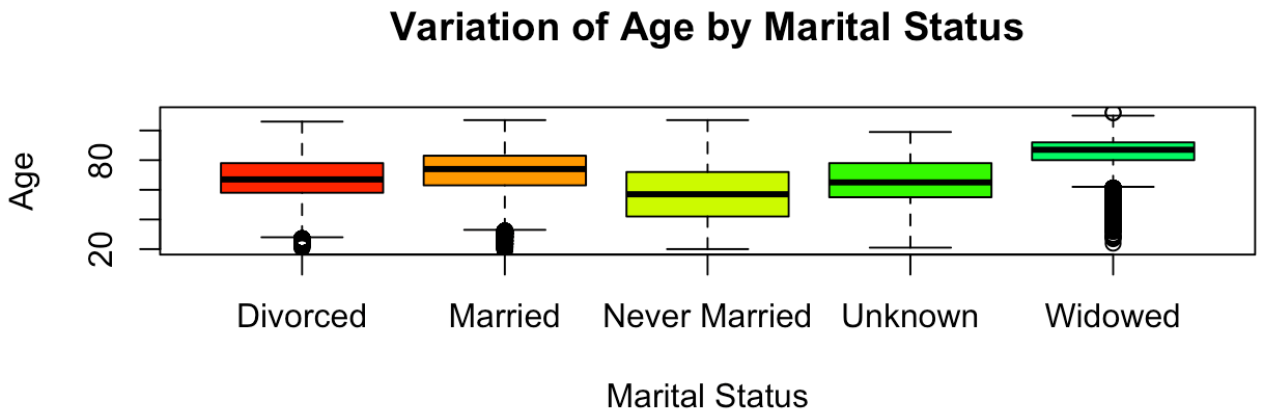

Figure 3: Box Plot of age v/s Marital Status

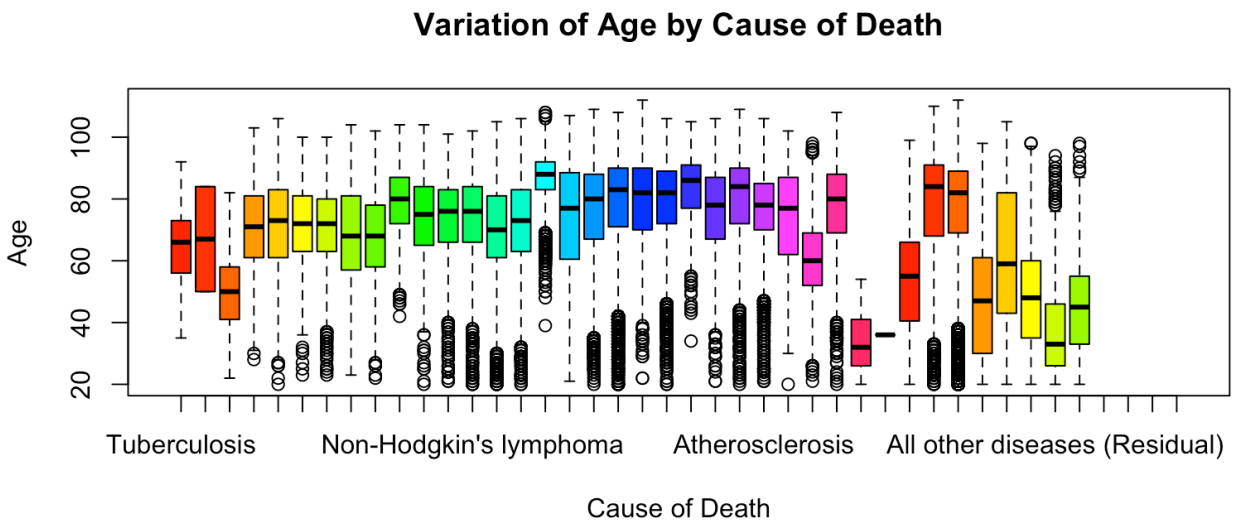

Figure 4: Box Plot of age v/s Cause of Death

Correlation

The 'hetcor' function was used from the R library 'polycor' to estimate the correlations between the categorical variables. As shown in Table 1, no high correlations were observed between the predicting variables.

Correlation	resident_ status	age	education	sex	marital_ status	race	hispanic	cause_of_ death	year
resident_status	1.00	-0.12	-0.01	0.06	-0.05	-0.01	0.06	0.07	0.01
age	-0.12	1.00	-0.03	-0.25	0.38	NA	0.16	NA	0.00
education	-0.01	-0.03	1.00	0.11	-0.12	-0.05	0.20	-0.03	0.02
sex	0.06	-0.25	0.11	1.00	-0.34	0.01	-0.05	0.01	0.01
marital_status	-0.05	0.38	-0.12	-0.34	1.00	-0.02	0.05	0.04	-0.02
race	-0.01	NA	-0.05	0.01	-0.02	1.00	0.38	-0.03	0.02
hispanic	0.06	0.16	0.20	-0.05	0.05	0.38	1.00	-0.02	-0.02
cause_of_death	0.07	NA	-0.03	0.01	0.04	-0.03	-0.02	1.00	0.01
year	0.01	0.00	0.02	0.01	-0.02	0.02	-0.02	0.01	1.00

Table 1: Correlation between the variables

Multicollinearity

Additionally, a VIF test revealed no multicollinearity among the variables (Supplementary Figures).

MLR

The first model fit to the data set was a multiple linear regression model using the ordinary least squares method.

Variable Selection

Group LASSO

Simple LASSO regression does not recognize dummy variables as being a part of a single feature. To ensure all dummy variables were included for significant categorical covariates, variable selection was conducted using the group lasso method ('grplasso' package in R).

Sub-sampling

Because the number of observations used for the model was so large, a t-test would be insufficient to determine the significance of the variables. Therefore, repeated subsampling was done on the dataset to confirm the significance of the covariates by examining the histograms of p-values.

Mixed-Effects Model

The modeled data set was constructed using death statistics from multiple years. This created a need to account for the possible random effects introduced by different sampling years. Using the 'lme4' package in R, a mixed-effects model was fit to the data . The same variables were used as previous models except year was treated as a possible source of random effects.

Results

Variable Selection

All variables were found to be significant by group lasso as well as by subsampling on random subsets of the data.

MLR

A multiple linear regression model was fit to estimate age with all predicting variables. The resulting R-squared value was 0.448. The constant variance and mean zero assumptions were found to be violated with a clearly decreasing trend (Figure 5). The QQ plot indicated heavy tails (Supplementary Figures).

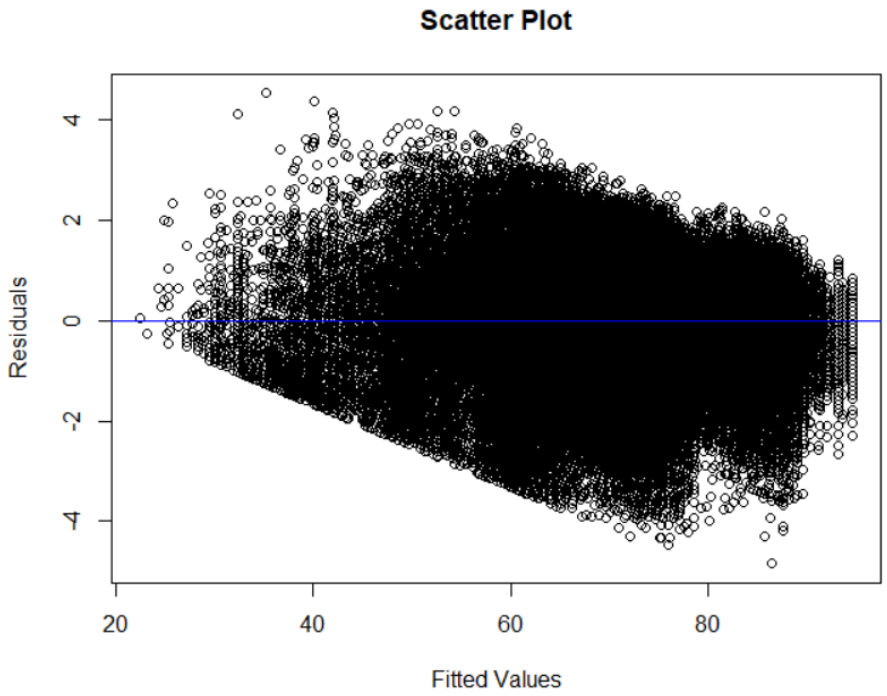

Figure 5: Model Assumption Check - Fitted Values vs. Residuals - MLR Model

The log-transformed response was fit to the variables to see if there would be any improvements in fit, but the non-constant variance persisted (Supplementary Figures).

Because the data used was drawn from multiple years, a mixed-effects model was built that treated year as a random effect. The truncated coefficient estimates are given below in Table 2.

Fixed effects:

	Estimate
(Intercept)	62.29598
resident_statusIntrastate NR	-2.27684
resident_statusInterstate NR	-2.89283
resident_statusForeign Residents	-7.43598
sexM	-1.05347
marital_statusMarried	3.46072
marital_statusNever Married	-7.93473
marital_statusUnknown	-1.12344
marital_statusWidowed	14.52354
raceBlack	-4.30227
raceAmerican Indian	-6.28138
raceAsian/Pacific	-1.72871
hispanicnon-hispanic	3.74668
educationCollege < 4 years	-1.76871
educationCollege > 4 years	1.49314

Table 2: Coefficient estimates for Mixed-Effects model with year as random effect

The QQ plot indicated a marginally better fit, but the non-constant variance persisted (Figure 17 in Supplementary Figures).

Distinct Age Subpopulations

It was hypothesized that the previous poor fits were due to separable, distinct age subpopulations for which a single model would be insufficient. To test this, two separate mixed-effects models were built for age greater than or equal to 40 and age less than or equal to 80. Visual residual analysis revealed a better overall fit for the under 40 model than seen previously (Figure 6). Partial F-tests conducted indicated significance for each of the variables. Additionally, the QQ plot indicated a normal distribution (Figure 7).

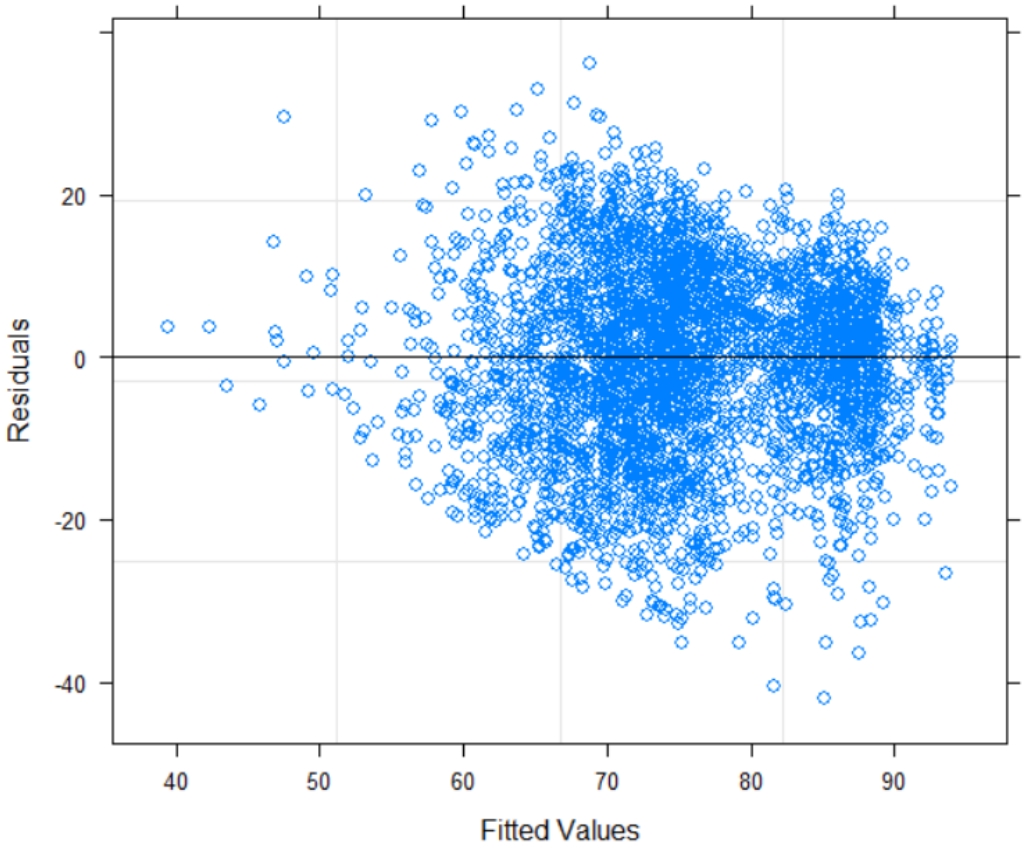

Figure 6: Mixed-Effects >=40 years - Fitted Values vs. Residuals

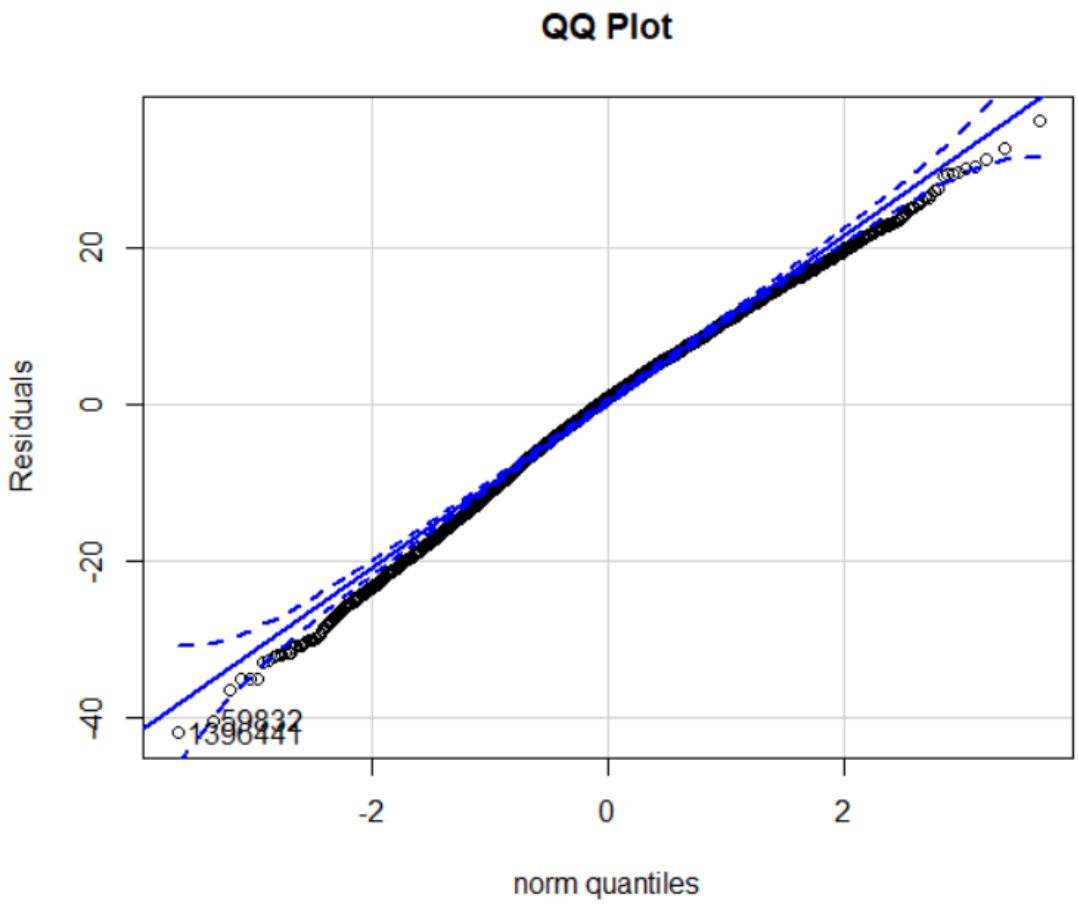

Figure 7: Mixed-Effects >=40 years - QQ plot

The model fit, however, for the less than 80 cohort was worse than that of the previous subpopulation tested. (Figures 19, 20 in Supplementary Figures)

Discussion

Many regression models were created and evaluated on the CDC mortality data set. Despite the mixed-effects model fit on observations with age greater than or equal to 40 displaying the best overall fit, the mixed-effects model for the full data set was ultimately decided upon. As the goal was to build an explanatory model linking population demographics with age at time of death, a model utilizing only a subset of the overall population would be inadequate to capture the more general trend.

Regression analysis revealed that, with all other variables in the model held constant, Blacks and American Indians died an estimated 4.3 and 6.3 years earlier than their white counterparts. This finding was very consistent with earlier research indicating that Blacks and American Indians die on average 5 years earlier (4, 5, 6). Cause-specific mortality gaps between Whites and minority populations have previously been noted. For instance, it has been shown that Blacks suffer a death rate from HIV-related disease almost 10 times greater than Whites (4). However, this report clearly demonstrates that the race related mortality gap persists even when cause of death is included in the model as a controlling variable. Also, of

particular note, the mortality gap has not been effaced by the recent rise in premature middle-aged death among Whites.

As shown, the age at time of death disparity among Whites and minority races has been relatively consistent despite the recent downward trend among Whites. Much research has been focused on how to close this gap. Recently, a nationwide study of rural counties with majority Black or American Indian populations concluded that these counties "hav[e] younger populations, lower median incomes, fewer females, poorer access to healthy food, and higher unemployment rates. Each of those is an important social determinant of health, and all tend to be associated with poorer health" (7). Further, systemic racism has been proposed as the fundamental cause of race-related health inequalities and any public health initiative that only addresses the "proximate risk factors for disease and death" will fail to result in any long-lasting reduction in this inequity (8). With this research in mind, the authors of this paper propose further public health initiatives targeting systemic racism and its impact on health outcome disparities with the goal of closing the mortality gap between Whites and Black/American Indian populations observed in this report.

Conclusion

While the recent increase in all-cause mortality rate for middle-aged white non-Hispanic men and women has been well documented in both research journals and popular media, much less coverage has been given to the health inequities that still persist between Whites and minority populations in the United States. This report demonstrates that despite this reported trend, with all other variables in the model held constant, Blacks and American Indians die an estimated 4.3 and 6.3 years earlier than their white counterparts, even with cause of death used as a controlling variable. While attention has justifiably been focused on this recent phenomenon, public health efforts must also address the remaining fundamental causes of health inequities based upon racial identity.

Sources

1. Cutler D, Deaton A, Lleras-Muney A. “The Determinants of Mortality”. *Journal Economic Perspectives* 20(3):97-120. 2006.

2. Case A, Deaton A. “Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century”*. Proc Natl Acad Sci* 112(49):15078–83. 2015.

3. https://www.nytimes.com/2015/11/03/health/death-rates-rising-for-middle-aged-white-americans-study-finds.html

4. Spalter-Roth R, Lowenthal TA, Rubio M. “Race, ethnicity, and the health of Americans”. *Am. Sociol. Assoc.*, 2005. www.asanet.org/images/research/docs/pdf/race_ethnicity_health.pdf

5. Shiels, M.S., Chernyavskiy, P., Anderson, W.F., Best, A.F., Haozous, E.A., Hartge, P., et al.,2017. “Trends in premature mortality in the USA by sex, race, and ethnicity from 1999 to 2014: an analysis of death certificate data”. *Lancet* 389:1043–1054. https://doi.org/10.1016/s0140-6736(17)30187-3.

6. Sequist TD, Cullen T, Acton KJ. “Indian Health Service Innovations Have Helped Reduce Health Disparities Affecting American Indian And Alaska Native People.” *Health Affairs*. 30(10). 2011.

7. Henning-Smith CE, Hernandez AM. “Rural Counties With Majority Black Or Indigenous Populations Suffer The Highest Rates Of Premature Death In The US”. *Health Affairs* 38(12). 2019.

8. Phelan JC, Link G. "Is Racism a Fundamental Cause of Inequalities in Health?" *Annual Review of Sociology* 41.1: 311-30. 2015.

Supplementary Figures and Tables

Variables in the model

Variable	Type		Labels
Age	Response Variable	Continuous	Range: (20 - 120)
Resident Status	Predicting Variable	Categorical	Resident
			Intrastate NR
			Interstate NR
			Foreign Residents
Sex	Predicting Variable	Categorical	F
			M
Education	Predicting Variable	Categorical	Up to High School
			College < 4 years
			College > 4 years
Marital Status	Predicting Variable	Categorical	Divorced
			Married
			Never Married
			Unknown
			Widowed
Race	Predicting Variable	Categorical	Puerto Rican
			White
			Black
			American Indian
			Asian/Pacific
Year	Predicting Variable	Categorical	2011
			2012
			2013
			2014
			2015
Hispanic	Predicting Variable	Categorical	Hispanic
			Non Hispanic

Cause of Death	Predicting Variable	Categorical	Tuberculosis
			Syphilis
			HIV
			Malignant neoplasms
			Malignant neoplasm Stomach
			Malignant neoplasms of Colon
			Malignant neoplasm of Pancreas
			Malignant neoplasms of Lung
			Malignant neoplasm of breast
			Malignant neoplasms of Ovary
			Malignant neoplasm of prostate
			Malignant neoplasms of urinary tract
			Non-Hodgkin's lymphoma
			Leukemia
			Other malignant neoplasms
			Diabetes
			Alzheimer's
			Major cardiovascular diseases
			Diseases of heart
			Hypertensive heart disease
			Ischemic heart diseases
			Other diseases of heart
			hypertension and hypertensive renal disease
			Cerebrovascular diseases
			Atherosclerosis

			Other diseases of circulatory system
			Influenza and pneumonia
			Chronic lower respiratory diseases
			Peptic ulcer
			Chronic liver disease and cirrhosis
			Nephritis
			Pregnancy
			perinatal period
			Congenital malformations
			Sudden infant death syndrome
			Symptoms
			All other diseases (Residual)
			Motor vehicle accidents
			All other and unspecified accidents
			Suicide
			Assault
			external causes

Table 3: Variables used in the model

Exploratory Data Analysis

Box Plots

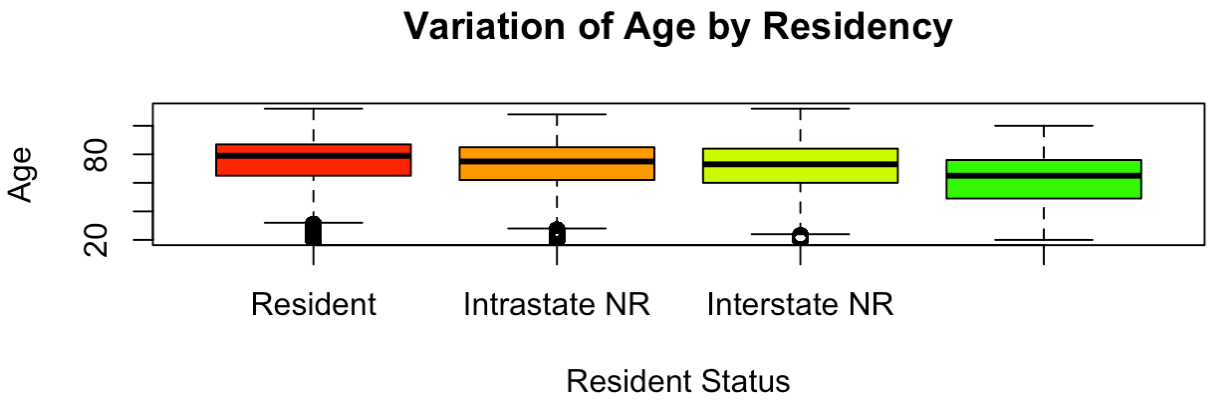

Figure 8: EDA - Box Plot of age v/s Residency

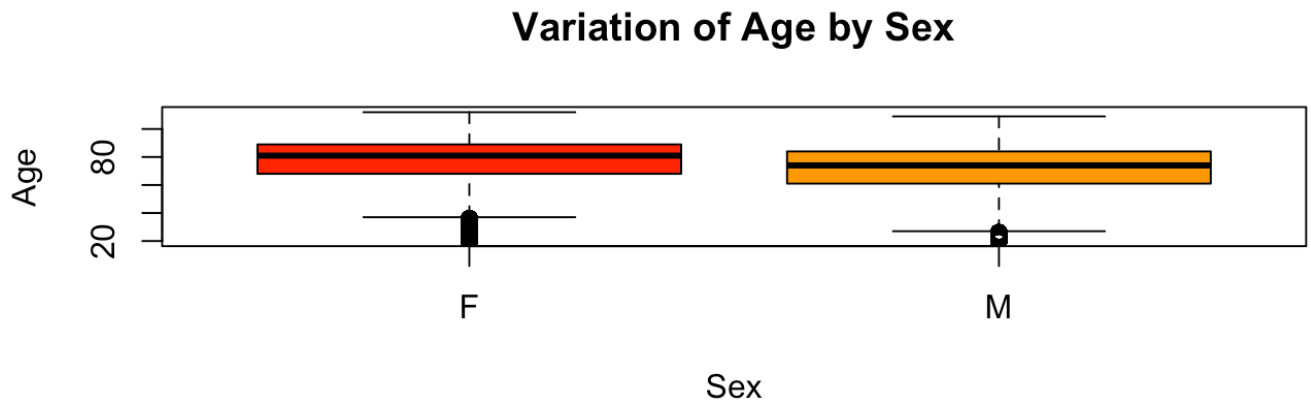

Figure 9: EDA - Box Plot of age v/s Sex

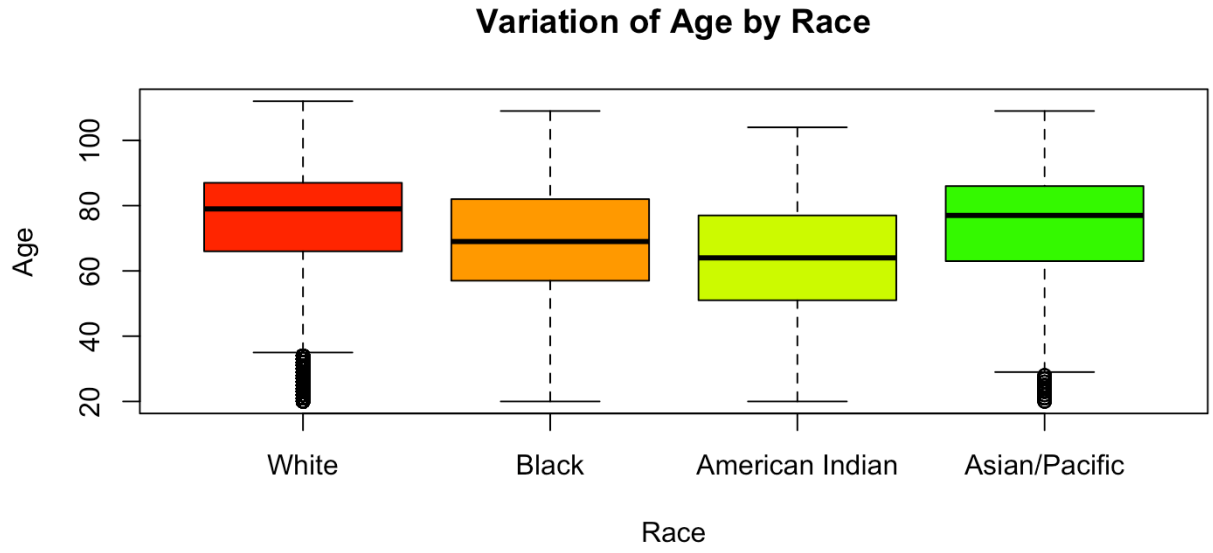

Figure 10: EDA - Box Plot of age v/s Race

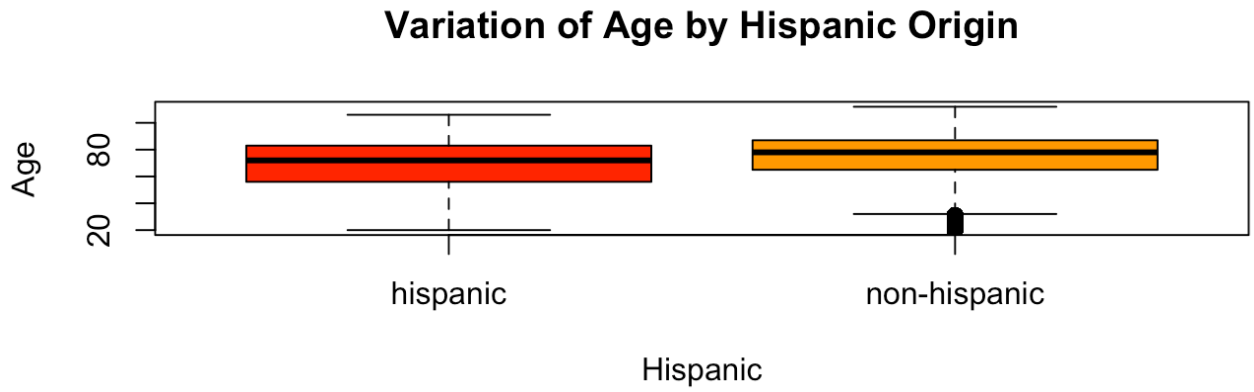

Figure 11: EDA - Box Plot of age v/s Hispanic Origin

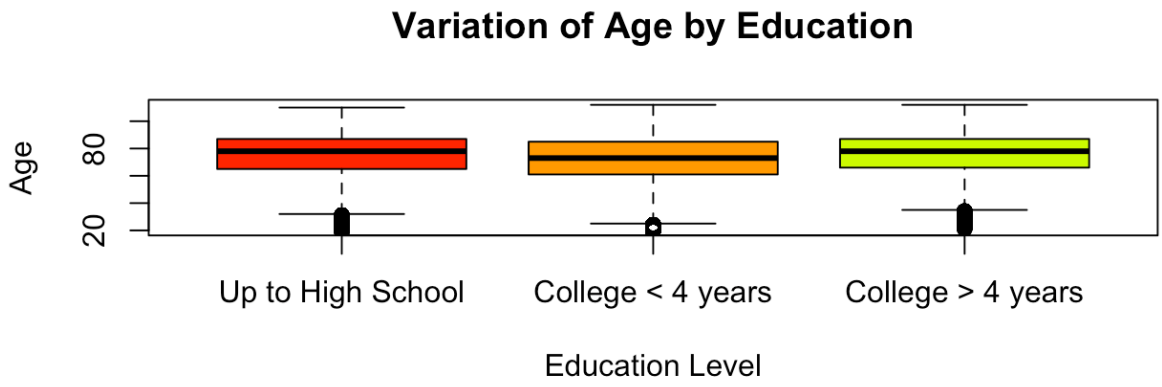

Figure 12: EDA - Box Plot of age v/s Education

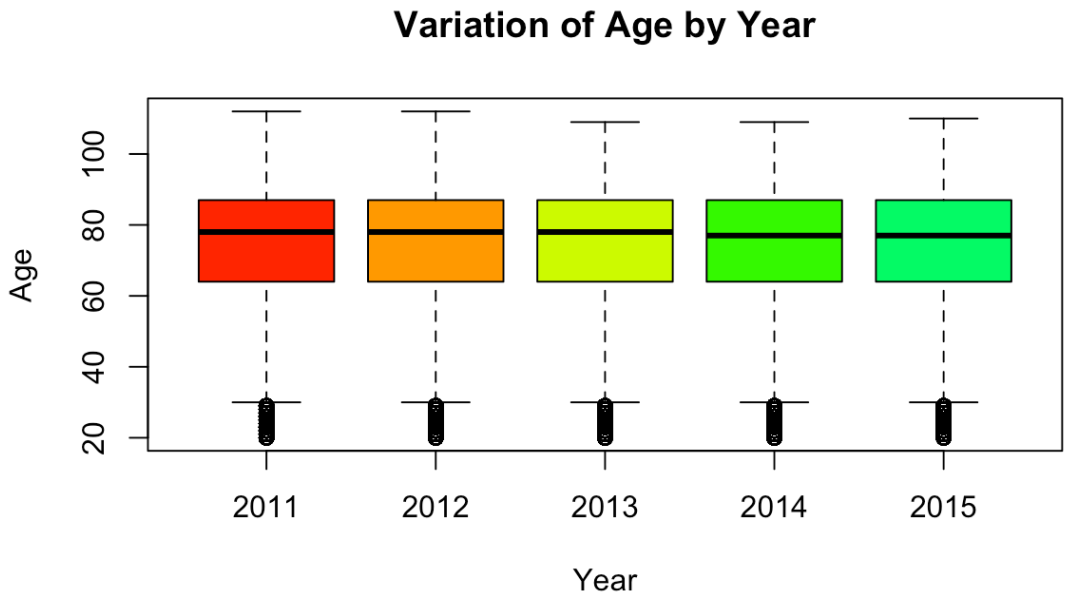

Figure 13: EDA - Box Plot of age v/s Year

Multicollinearity Table

Multicollinearity	GVIF	Df	GVIF^(1/(2*Df))
resident_status	1.024027	3	1.003965
sex	1.218689	1	1.103943
marital_status	1.318562	4	1.035172
race	1.073685	3	1.01192
hispanic	1.032794	1	1.016265
education	1.044449	2	1.010932
cause_of_death	1.254249	37	1.003066
year	1.00508	4	1.000634

Table 4: Multicollinearity table of the predicting variables

Model Building and Checking Assumptions

Multiple Linear Regression model

Figure 14: QQ plot for the initial MLR model showing heavy tails

MLR model on log-transformed response

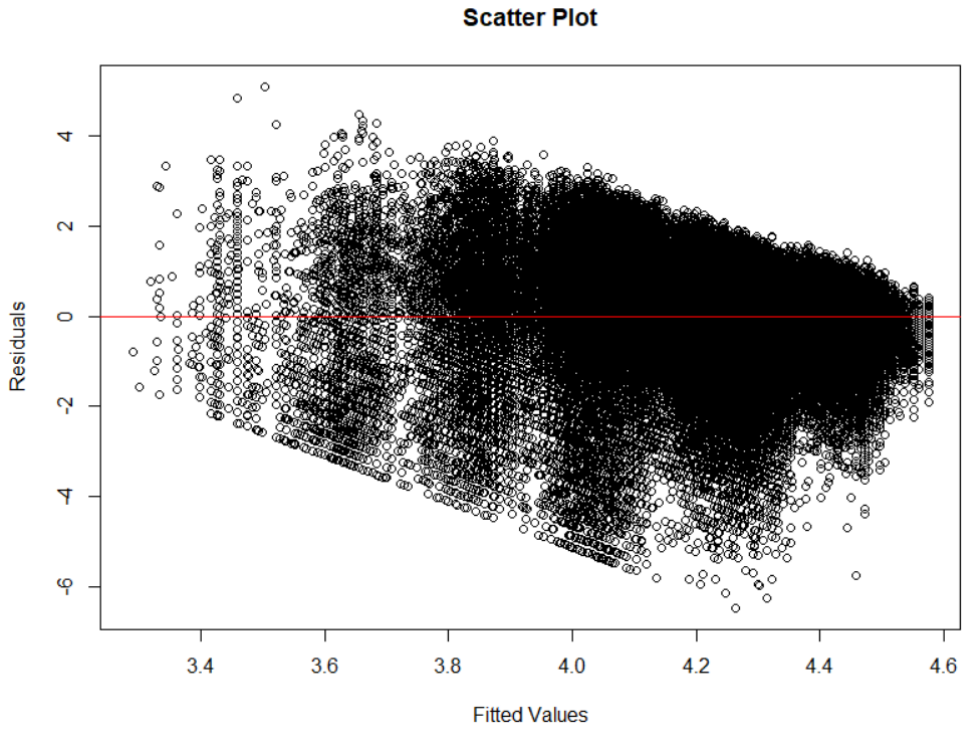

Figure 15: Fitted Values vs. Residuals for MLR model using response with transformation

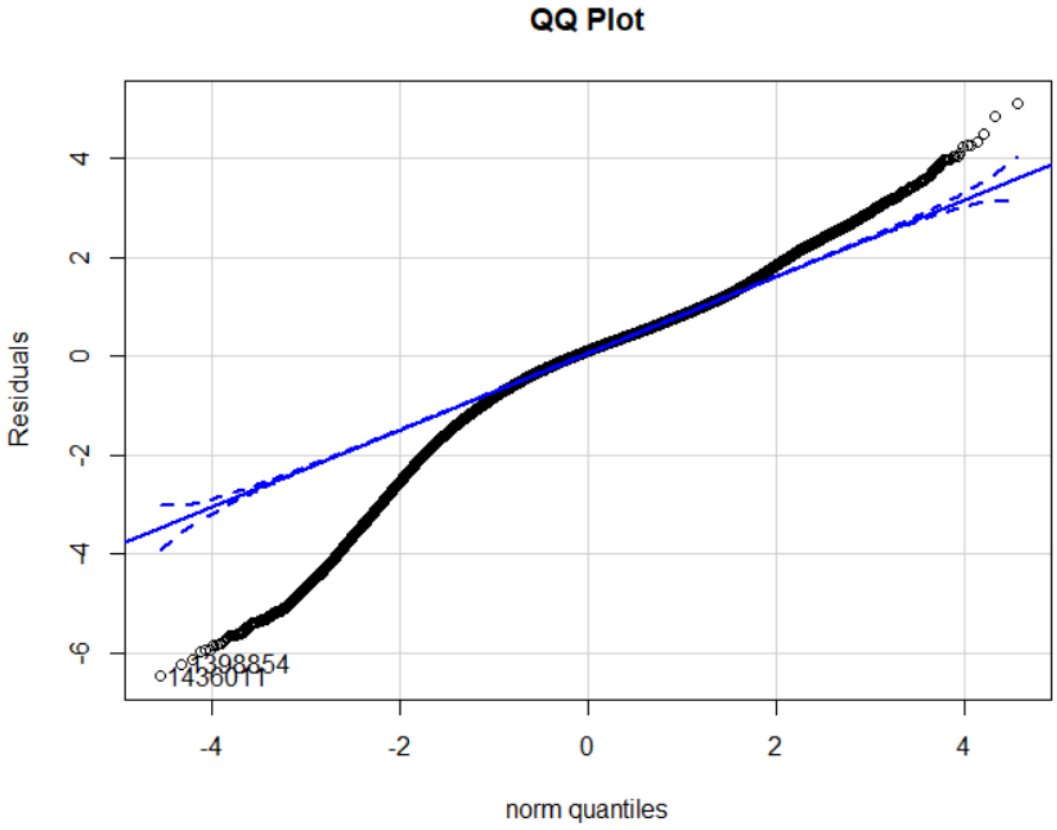

Figure 16: QQ plot for MLR model using response with transformation

Mixed-Effects model with year as random effect

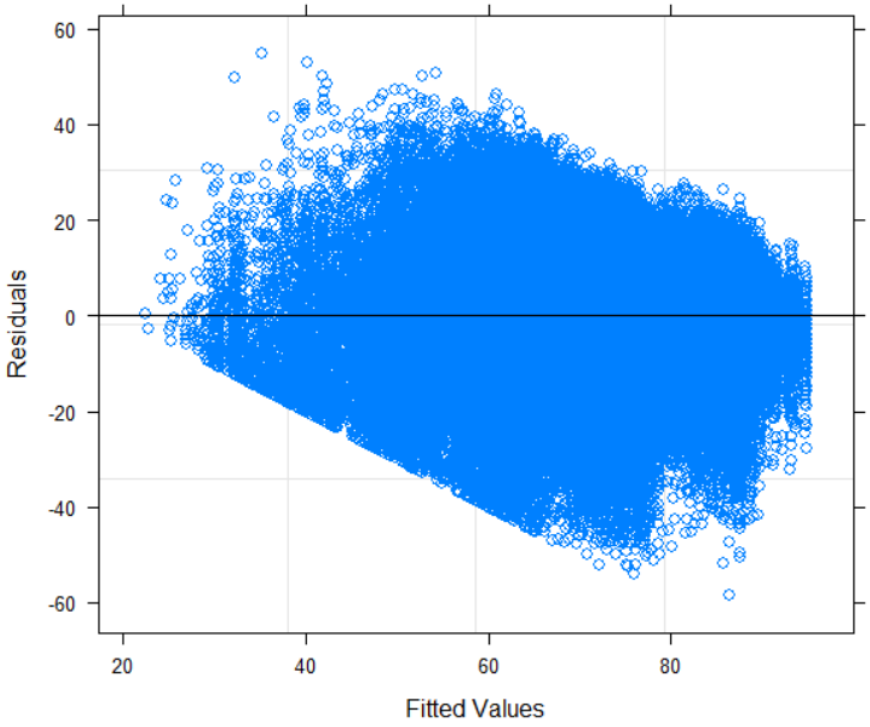

Figure 17: Fitted Values vs. Residuals for Mixed-Effects model with year as random effect

Figure 18: QQ plot for Mixed-Effects model with year as random effect

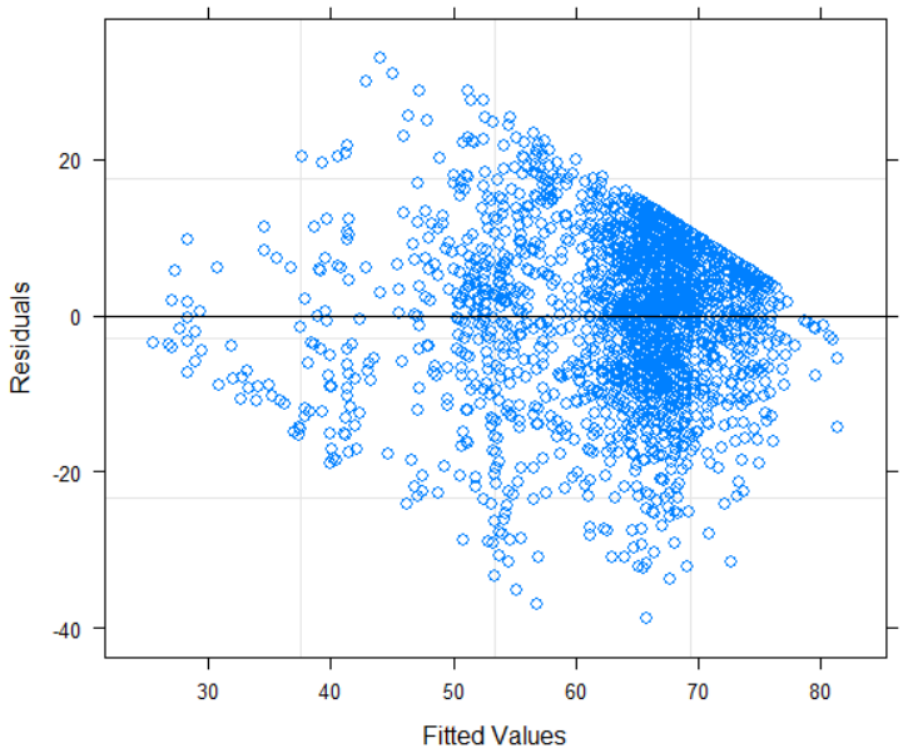

Figure 19: Fitted Values vs. Residuals for Mixed-Effects model on subpopulation with age <=80

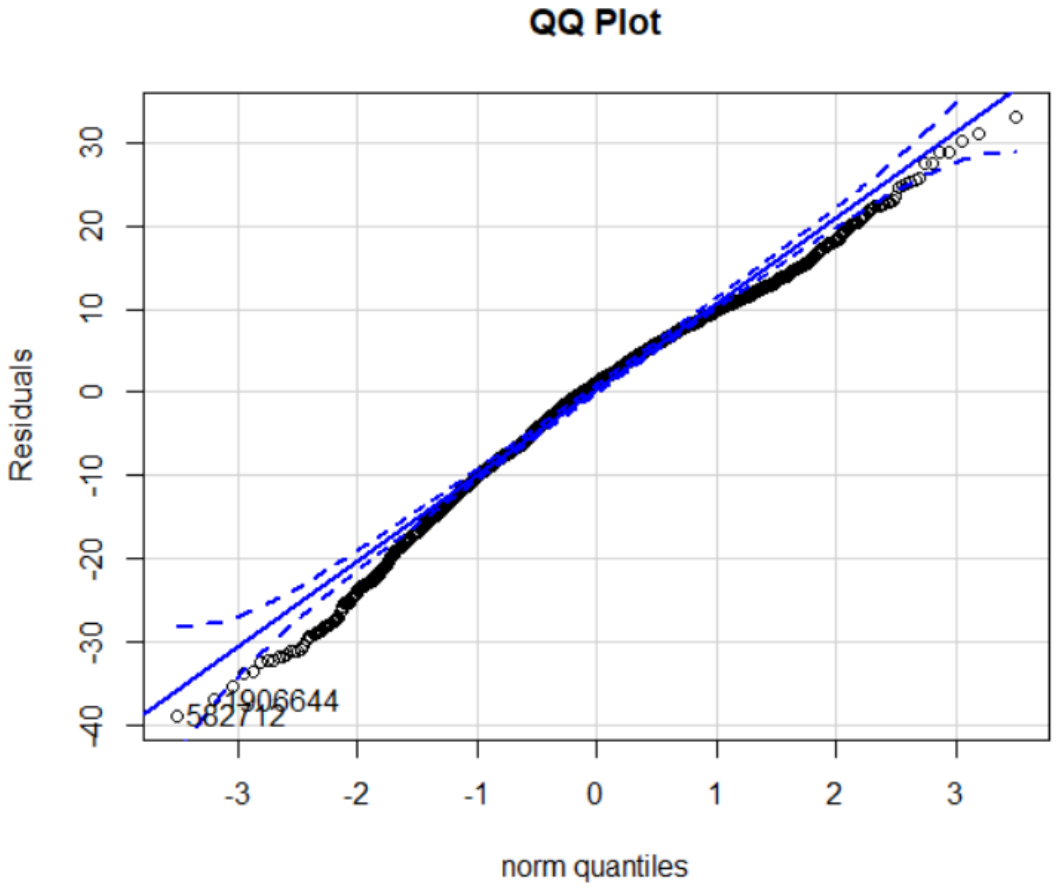

Figure 20: QQ plot for Mixed-Effects model model on subpopulation with age <=80