Predicting US Gun Deaths by County

Casey DeLano and Jimmy DeLano

December 4, 2019

Introduction

The United States has a gun violence crisis. Americans own nearly half the total number of civilian-owned guns worldwide[1]. We are first on the list of countries and territories with the highest number of civilian firearms per population with more than twice the number of guns as the second highest country[2]. With more guns than people, the US sees more than one mass shooting per day on average[3]. Compared to other high-income countries, our gun homicide rates are over 25 times higher[4]. As we approach a presidential election year, candidates are strategically highlighting – or shying away from – their gun legislation plans. No matter your political beliefs, the topic of gun violence is unavoidable. But do strict gun laws even help? Do states that allow the open carrying of firearms actually have lower gun death rates than those that require strict background checks[5]?

Our project aims to determine which factors explain the differences in gun deaths in the US at the county level. We use a variety of predictor variables, one of which is the severity of state gun laws, in our analysis. As we weigh interpretability with predictability, our goal is to create two models, one that is simple and easy to interpret and another that is more complex with the best predictability.

We hypothesize that states with tighter gun laws will have lower gun deaths per population. Mainstream media enforces the idea that "most gun violence victims come from communities that really need help[6];" therefore we also predict that counties with higher rates of poverty and income inequality will have higher rates of gun violence[7]. Potentially linked with the poverty predictor, and therefore with higher gun deaths, are lower levels of educational attainment and higher unemployment rates[8].

Data

Our data comes from three sources. First, our response variable comes from the Center for Disease Control and Prevention (CDC)[9]. Their Multiple Cause of Death database contains mortality and population counts by US county for counties with 20 or more gun deaths. 319 counties fit this criteria. We queried this database for all deaths from 2010 that were attributed to accidents, assaults, intentional self-harm, and undetermined events by any type of firearm. We use *Crude.Rate*, the number of gun deaths per 100,000 people as our response variable.

[1] https://www.cnn.com/2017/10/03/americas/us-gun-statistics/index.html
[2] https://www.vox.com/policy-and-politics/2017/10/2/16399418/us-gun-violence-statistics-maps-charts
[3] https://nationalinterest.org/blog/buzz/us-has-been-averaging-more-1-mass-shooting-day-71551
[4] https://www.cnn.com/2017/10/03/americas/us-gun-statistics/index.html
[5] https://lawcenter.giffords.org/wp-content/uploads/2010/07/Gun_Laws_Matter_Brochure.pdf
[6] https://www.aau.edu/research-scholarship/featured-research-topics/how-can-cities-reduce-gun-violence-invest-low-income
[7] https://luskin.ucla.edu/connection-poverty-inequality-firearm-violence
[8] https://www.cnbc.com/2018/04/20/unemployment-and-financial-distress-may-trigger-school-shootings.html
[9] https://wonder.cdc.gov/mcd-icd10.html

Second, all but one of our predictor variables come from the Census[10]. From a list of 6,600 potential county-level predictor variables, we selected 33 that we thought would best explain the variability in gun deaths based on our intuition and judgment. These variables range from race demographics to crime totals and are as close to 2010 as possible (dating back no earlier than 2007).

Our final predictor variable, gun law severity, comes from Gifford's Law Center[11]. Giffords Law Center is "the nation's leading policy organization dedicated to researching, writing, enacting, and defending proven laws and programs, [and] is on a mission to save lives from gun violence by shifting culture, changing policies, and challenging injustice." The center provides a yearly ranking of gun laws by state (see Figure 1). We used their 2010 rankings[12] as our final predictor variable, *Gun.Law.Ranking.2010*.

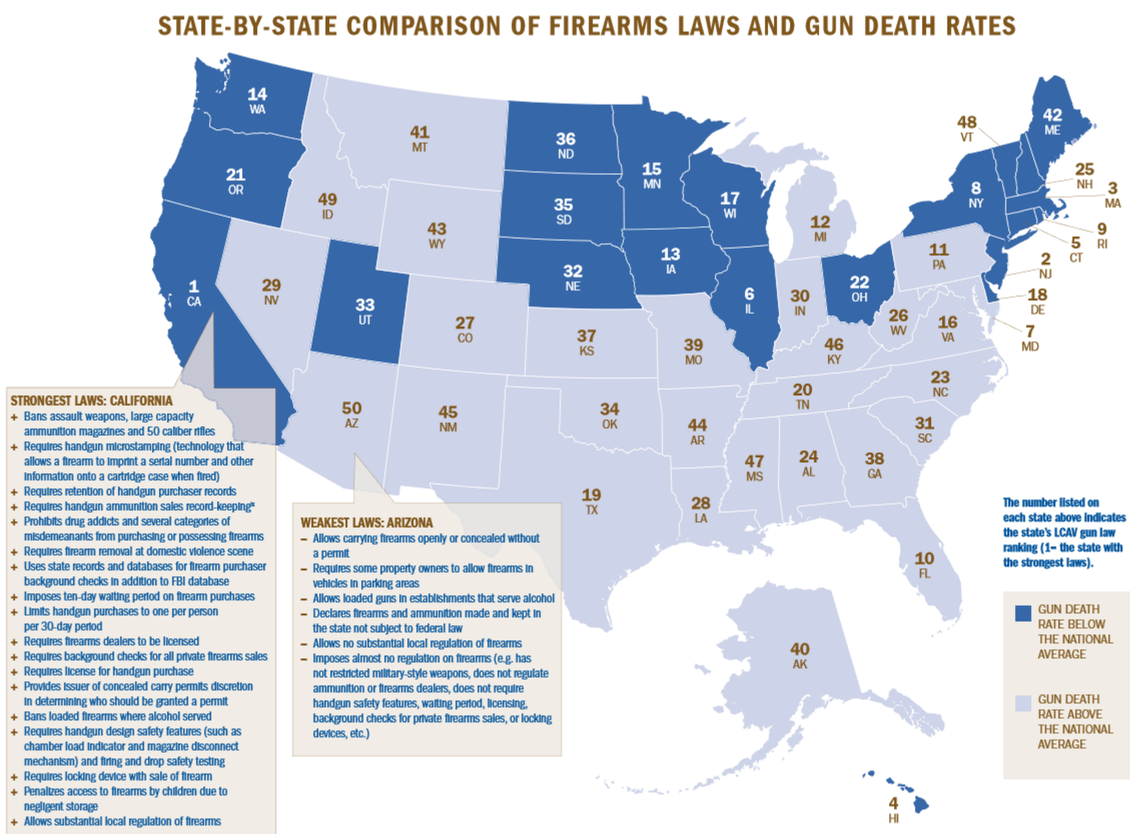


Figure 1: Gifford's Law Center 2010 State Rankings

We merged these three datasets together using their county identifiers. Our resulting dataset had 319 counties and 38 variables.

[10] https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html
[11] https://lawcenter.giffords.org/
[12] https://lawcenter.giffords.org/wp-content/uploads/2010/07/Gun_Laws_Matter_Brochure.pdf

Data Preparation

Feature Creation

See Appendix A for detailed data preparation information.

The first step we took was taking the absolute number of the different crime statistics and turning them into crime per 1000 population variables. We also scaled the Federal Government Expenditure by the total population in each county to create *Federal.Govt.Expenditure.Per.Person*.

We had nine crime-related predictor variables, all scaled to per 1000 people. After some exploration, we discovered that the violent crimes per 1000 people variable was a sum of four of the other crime variables: murders and manslaughters, rapes, robberies, and aggravated assaults. For this reason, we decided to omit those four more granular variables and include only the aggregate violent crimes variable.

Similarly, we noticed that the other four crime-related predictors were highly correlated amongst each other (see Figure 2).

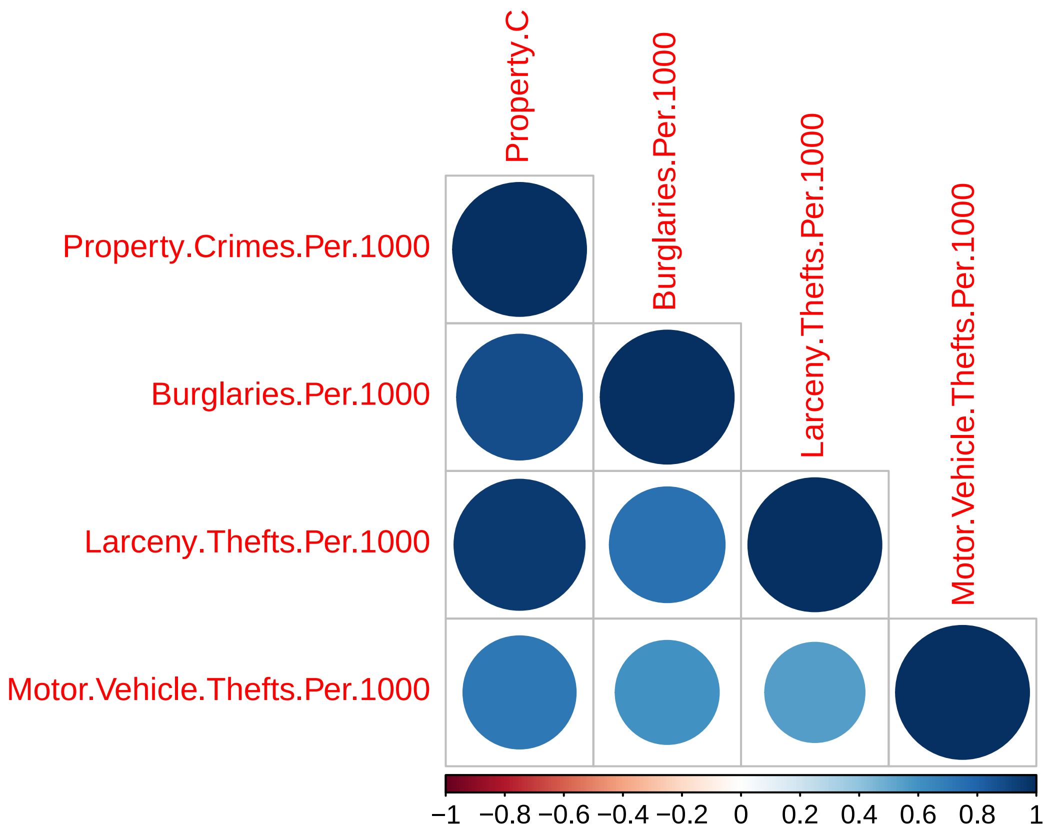


Figure 2: Correlation plot of crime-related predictors not accounted for in violent crimes

We summed the other four crime-related variables to create a non-violent crimes per 1000 people variable. In all, we boiled down these nine crime variables into two larger categories: *Violent.Crimes.Per.1000* people and *Non.Violent.Crimes.Per.1000* people.

None of our predictor variables were categorical, so we grouped the percent of Republican voters in the 2008 presidential election into three buckets to create our own categorical variable. We considered counties that voted less than 40% Republican to be *Democrat*, those that voted 40-60% Republican to be *Swing*, and those above 60% to be *Republican*.

Data Quality Control

Four percent of the entries for one of our predictors, *Average.Land.Value*, were missing. We decided to omit this variable entirely.

In addition, we omitted five counties from our analysis due to the following reasons.

Bronx County, NY, New York County, NY, and Queens County, NY had missing values for all columns of crime data and federal government expenditure.

The predictor variable rapes per 1000 people has a median value of 0.297 and a third quartile value of 0.407. Navajo County, AZ had 3.83 rapes per 1000 people. Based on the extremity of this outlier, and on external sources that refuted its accuracy[13], we omitted the county entirely.

District of Columbia, DC had no gun law rank in 2010. Additionally, its scaled federal government expenditure was orders of magnitude larger than other counties.

Data Exploration

Our response variable, *Crude.Rate* looks relatively normal except for four counties which have very high gun deaths. This causes the distribution of the variable to be skewed right (see Figure 3).

[13] http://recordspedia.com/Arizona/Navajo-County/Crime-Statistics

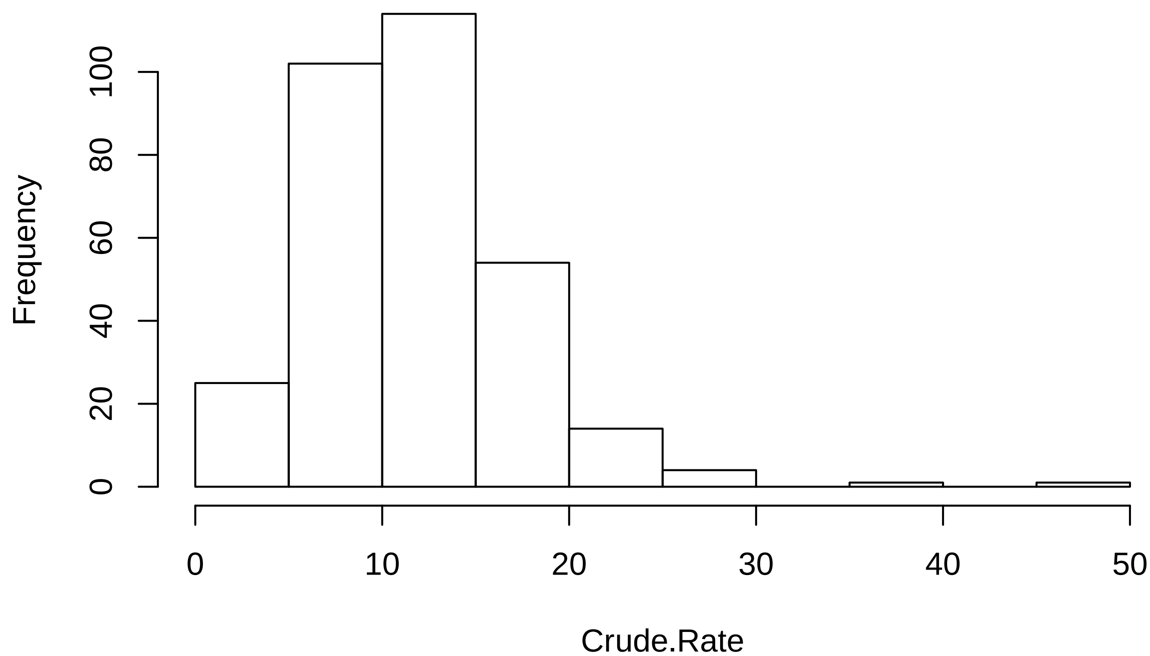
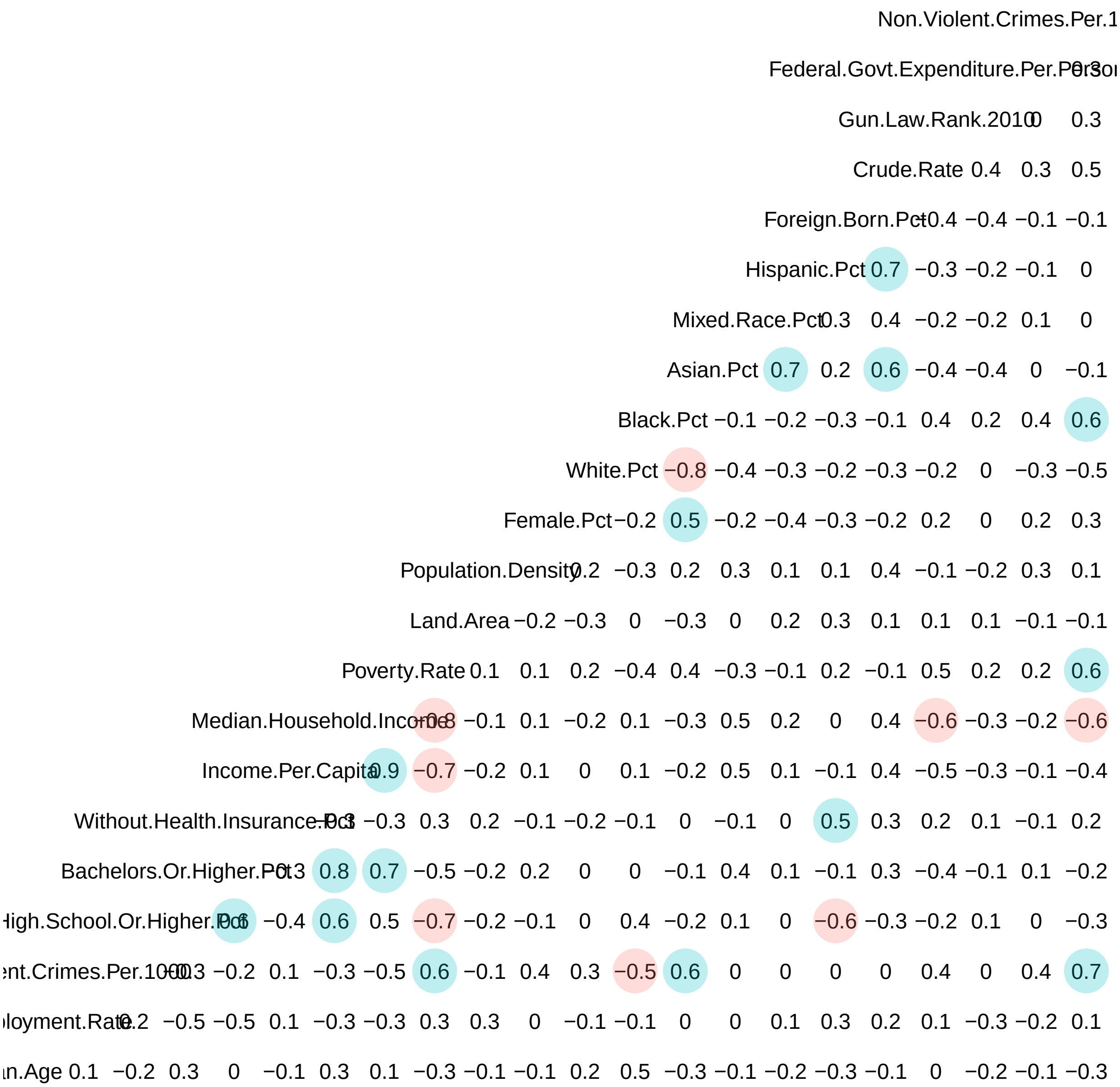


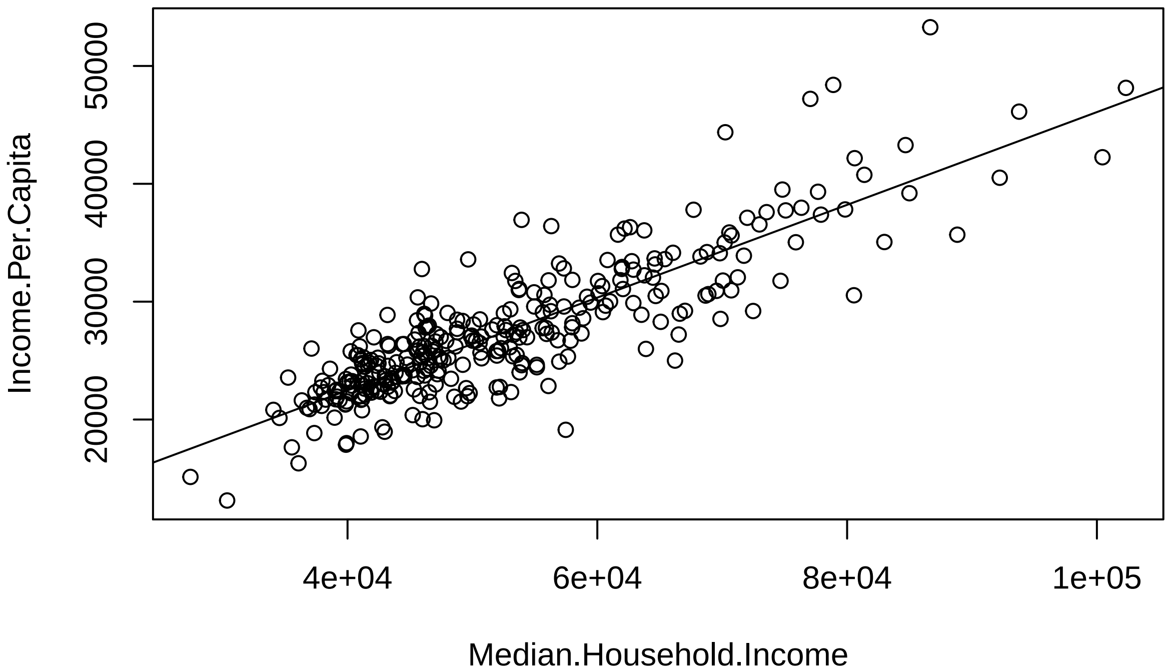
Figure 3: Histogram of response variable: Crude Rate

In our data exploration, we are most concerned with correlated predictors as this will make our model very unstable. We tackle this issue in two ways. First, we look at correlations between each of the predictors. Second, we look at variance inflation factor (VIF) values.

Given our high number of predictor variables, creating a scatterplot matrix is too difficult to interpret. Instead, we began by looking at a correlation matrix between our variables, highlighting those correlations above 0.5 or below -0.5 so that we could look into simplifying our predictors (**see Appendix B.1**).


Non.Violent.Crimes.Per.1
Federal.Govt.Expenditure.Per.Perso 0.3
Gun.Law.Rank.2010 0 0.3
Crude.Rate 0.4 0.3 0.5
Foreign.Born.Pct −0.4 −0.4 −0.1 −0.1
Hispanic.Pct 0.7 −0.3 −0.2 −0.1 0
Mixed.Race.Pct 0.3 0.4 −0.2 −0.2 0.1 0
Asian.Pct 0.7 0.2 0.6 −0.4 −0.4 0 −0.1
Black.Pct −0.1 −0.2 −0.3 −0.1 0.4 0.2 0.4 0.6
White.Pct −0.8 −0.4 −0.3 −0.2 −0.3 −0.2 0 −0.3 −0.5
Female.Pct −0.2 0.5 −0.2 −0.4 −0.3 −0.2 0.2 0 0.2 0.3
Population.Density 0.2 −0.3 0.2 0.3 0.1 0.1 0.4 −0.1 −0.2 0.3 0.1
Land.Area −0.2 −0.3 0 −0.3 0 0.2 0.3 0.1 0.1 0.1 −0.1 −0.1
Poverty.Rate 0.1 0.1 0.2 −0.4 0.4 −0.3 −0.1 0.2 −0.1 0.5 0.2 0.2 0.6
Median.Household.Income −0.8 −0.1 0.1 −0.2 0.1 −0.3 0.5 0.2 0 0.4 −0.6 −0.3 −0.2 −0.6
Income.Per.Capita 0.9 −0.7 −0.2 0.1 0 0.1 −0.2 0.5 0.1 −0.1 0.4 −0.5 −0.3 −0.1 −0.4
Without.Health.Insurance.Pct −0.3 −0.3 0.3 0.2 −0.1 −0.2 −0.1 0 −0.1 0 0.5 0.3 0.2 0.1 −0.1 0.2
Bachelors.Or.Higher.Pct 0.3 0.8 0.7 −0.5 −0.2 0.2 0 0 −0.1 0.4 0.1 −0.1 0.3 −0.4 −0.1 0.1 −0.2
igh.School.Or.Higher.Pct 0.6 −0.4 0.6 0.5 −0.7 −0.2 −0.1 0 0.4 −0.2 0.1 0 −0.6 −0.3 −0.2 0.1 0 −0.3
ent.Crimes.Per.1000 −0.3 −0.2 0.1 −0.3 −0.5 0.6 −0.1 0.4 0.3 −0.5 0.6 0 0 0 0 0.4 0 0.4 0.7
loyment.Rate 0.2 −0.5 −0.5 0.1 −0.3 −0.3 0.3 0.3 0 −0.1 −0.1 0 0 0.1 0.3 0.2 0.1 −0.3 −0.2 0.1
n.Age 0.1 −0.2 0.3 0 −0.1 0.3 0.1 −0.3 −0.1 −0.1 0.2 0.5 −0.3 −0.1 −0.2 −0.3 −0.1 0 −0.2 −0.1 −0.3

We used this in addition to variance inflation factor analysis to determine two sets of variables with high colinearity. First, *Income.Per.Capita* and *Median.Household.Income* have a correlation of 0.86. Due to the fact that its VIF is higher and it is less predictive of *Crude.Rate*, we omitted *Income.Per.Capita*.



Similarly, *Black.Pct* and *White.Pct* have a correlation of 0.77. *Black.Pct* is more predictive of *Crude Rate* and its VIF is lower than *White.Pct*, so we kept *Black.Pct* and left out *White.Pct* **(see Appendix B.1)**.

Analysis

Best Interpretable Model

For our model creation, we began by creating a simple model that includes all of our predictor variables and no interaction terms. The boxcox command in *R* recommended to raise our response variable, *Crude.Rate*, to the power $\frac{1}{4}$. This transformation made the distribution much more normal (see Figure 4). We made this adjustment to our response variable, and then ran a stepwise function (using minimum BIC) to get a simple and interpretable model. Our resulting model has an R^2 of 0.7117 and includes ten predictor variables. We consider this our first model **(see Appendix B.2)**. The summary of this model is as follows:

##

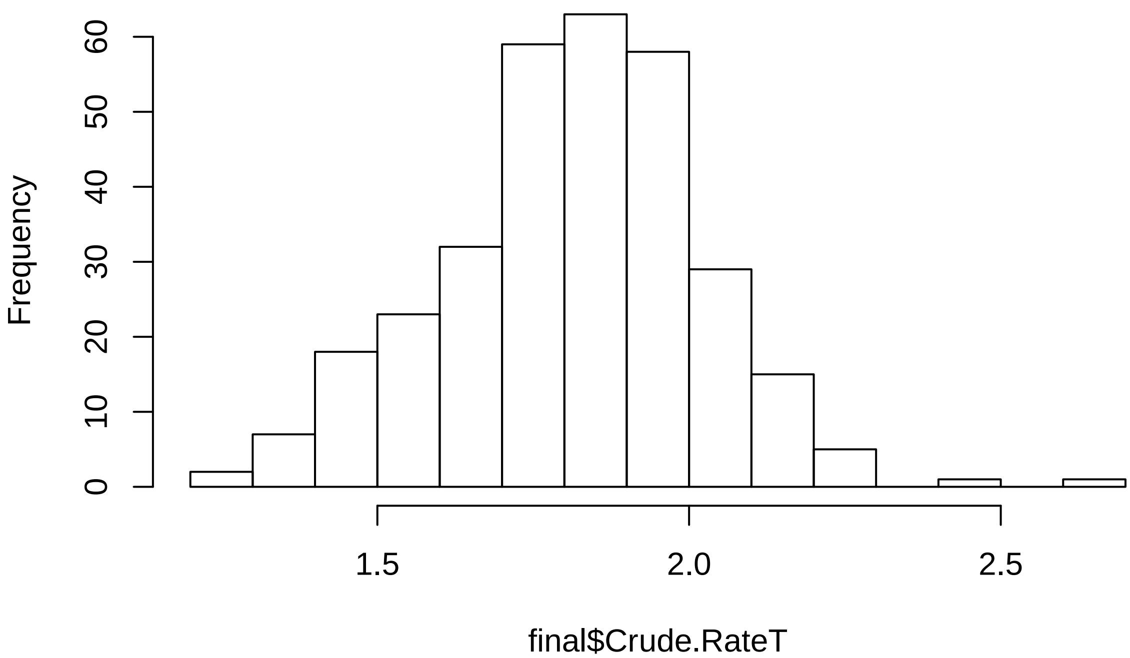


Figure 4: Histogram of transformed response variable: Crude Rate to the 1/4

```
## Call:
## lm(formula = Crude.RateT ~ Median.Age + Unemployment.Rate + Violent.Crimes.Per.1000 +
##     High.School.Or.Higher.Pct + Without.Health.Insurance.Pct +
##     Female.Pct + Black.Pct + Foreign.Born.Pct + Gun.Law.Rank.2010 +
##     Federal.Govt.Expenditure.Per.Person, data = final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27446 -0.07309  0.00091  0.06700  0.39695
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         3.6264396  0.5247698   6.911 2.87e-11
## Median.Age                          0.0105042  0.0019726   5.325 1.97e-07
## Unemployment.Rate                   0.0125719  0.0035357   3.556 0.000437
## Violent.Crimes.Per.1000             0.0156875  0.0029845   5.256 2.78e-07
## High.School.Or.Higher.Pct          -0.0061079  0.0017563  -3.478 0.000580
## Without.Health.Insurance.Pct        0.0106760  0.0015056   7.091 9.49e-12
## Female.Pct                         -0.0411993  0.0098548  -4.181 3.81e-05
## Black.Pct                           0.0036128  0.0006931   5.212 3.47e-07
## Foreign.Born.Pct                   -0.0131767  0.0010226 -12.886  < 2e-16
## Gun.Law.Rank.2010                   0.0039595  0.0006131   6.458 4.24e-10
## Federal.Govt.Expenditure.Per.Person  0.0052091  0.0010919   4.770 2.87e-06
##
## (Intercept)                        ***
## Median.Age                         ***
## Unemployment.Rate                  ***
## Violent.Crimes.Per.1000            ***
## High.School.Or.Higher.Pct          ***
## Without.Health.Insurance.Pct       ***
## Female.Pct                         ***
## Black.Pct                          ***
## Foreign.Born.Pct                   ***
## Gun.Law.Rank.2010                  ***
## Federal.Govt.Expenditure.Per.Person ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1132 on 302 degrees of freedom
## Multiple R-squared:  0.7117, Adjusted R-squared:  0.7022
## F-statistic: 74.56 on 10 and 302 DF,  p-value: < 2.2e-16
```

Best Predictive Model

Next, we tackled creating a more complex model that predicts *Crude.Rate* as best we can. We ran a model with all interactions and squared terms in JMP using forward stepwise minimum BIC. We put the resulting model back into R and ran boxcox again to see if we should transform *Crude.Rate* with this more complicated model. The range of recommended transformations included raising crude rate to the power $\frac{1}{4}$. We re-ran the forward minimum BIC stepwise model in JMP using this transformed predictor variable to get our second, more complicated model. This model has 33 predictors **(see Appendix B.3)**. Our JMP output is below.

Report: Fit Model - JMP Pro

Fit Group

Response Crude.RateT

Analysis of Variance

Source	DF	Squares	Prob > F

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0247629	0.491426	6.16	<.0001*
Median.Age	0.0091251	0.002066	4.42	<.0001*
Unemployment.Rate	0.0045143	0.003545	1.27	0.2039
Violent.Crimes.Per.1000	0.0114594	0.003471	3.30	0.0011*
High.School.Or.Higher.Pct	-0.001691	0.002376	-0.71	0.4774
Without.Health.Insurance.Pct	0.0102551	0.001702	6.03	<.0001*
Median.Household.Income	-4.129e-6	1.508e-6	-2.74	0.0066*
Poverty.Rate	-0.005459	0.003305	-1.65	0.0997
Population.Density	1.2044e-5	6.608e-6	1.82	0.0694
Female.Pct	-0.029117	0.008707	-3.34	0.0009*
Black.Pct	0.0041667	0.000755	5.52	<.0001*
Mixed.Race.Pct	0.0143871	0.005841	2.46	0.0144*
Hispanic.Pct	0.0005141	0.00118	0.44	0.6634
Foreign.Born.Pct	-0.014276	0.002158	-6.62	<.0001*
Gun.Law.Rank.2010	0.0027114	0.000557	4.87	<.0001*
Federal.Govt.Expenditure.Per.Person	0.004435	0.000959	4.62	<.0001*
Non.Violent.Crimes.Per.1000	0.000159	0.000381	0.42	0.6769
Voter.Group.2008{Democrat&Swing-Republican}	-0.005858	0.009856	-0.59	0.5527
(Unemployment.Rate-9.74249)*(High.School.Or.Higher.Pct-85.9383)	0.0025716	0.000546	4.71	<.0001*
(High.School.Or.Higher.Pct-85.9383)*(High.School.Or.Higher.Pct-85.9383)	0.0006543	0.000249	2.63	0.0090*
(High.School.Or.Higher.Pct-85.9383)*(Without.Health.Insurance.Pct-17.1304)	-0.000121	0.000369	-0.33	0.7429
(High.School.Or.Higher.Pct-85.9383)*(Median.Household.Income-52334.8)	4.8892e-7	1.497e-7	3.27	0.0012*
(Without.Health.Insurance.Pct-17.1304)*(Hispanic.Pct-14.307)	-0.000524	0.000121	-4.32	<.0001*
(Mixed.Race.Pct-2.95048)*(Hispanic.Pct-14.307)	0.0011917	0.000475	2.51	0.0126*
(Violent.Crimes.Per.1000-4.97379)*(Foreign.Born.Pct-10.4837)	0.0017771	0.000319	5.56	<.0001*
(Population.Density-1213.25)*(Foreign.Born.Pct-10.4837)	-1.382e-6	2.727e-7	-5.07	<.0001*
(Female.Pct-51.0709)*(Foreign.Born.Pct-10.4837)	-0.003154	0.001002	-3.15	0.0018*
(Foreign.Born.Pct-10.4837)*(Foreign.Born.Pct-10.4837)	0.0003533	0.000072	4.91	<.0001*
(Black.Pct-14.962)*(Gun.Law.Rank.2010-19.0415)	-0.000165	3.382e-5	-4.88	<.0001*
(Mixed.Race.Pct-2.95048)*(Federal.Govt.Expenditure.Per.Person-10.5352)	-0.001547	0.000648	-2.39	0.0177*
(Gun.Law.Rank.2010-19.0415)*(Federal.Govt.Expenditure.Per.Person-10.5352)	0.0003304	0.000089	3.71	0.0002*
(Without.Health.Insurance.Pct-17.1304)*(Non.Violent.Crimes.Per.1000-70.7369)	-0.000192	5.361e-5	-3.58	0.0004*
(Median.Household.Income-52334.8)*(Voter.Group.2008{Democrat&Swing-Republican}-0.65495)	4.2725e-6	1.415e-6	3.02	0.0028*
(Poverty.Rate-14.1003)*(Voter.Group.2008{Democrat&Swing-Republican}-0.65495)	0.0134856	0.00375	3.60	0.0004*

Effect Tests

Effect Details

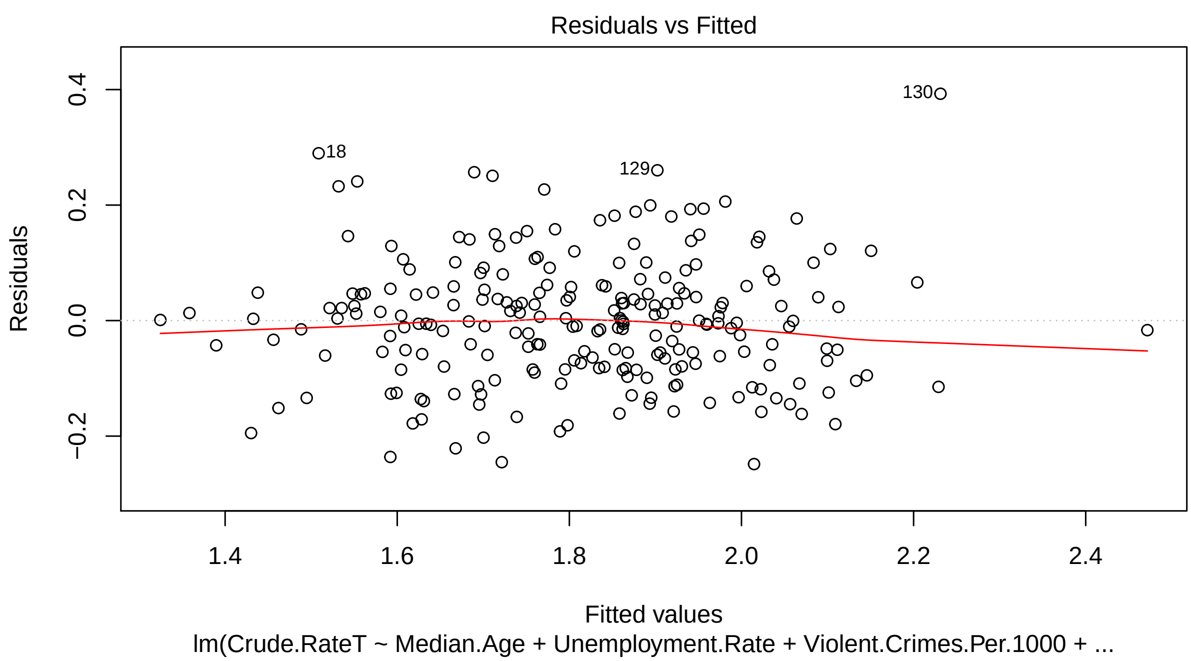
Cross-Validation

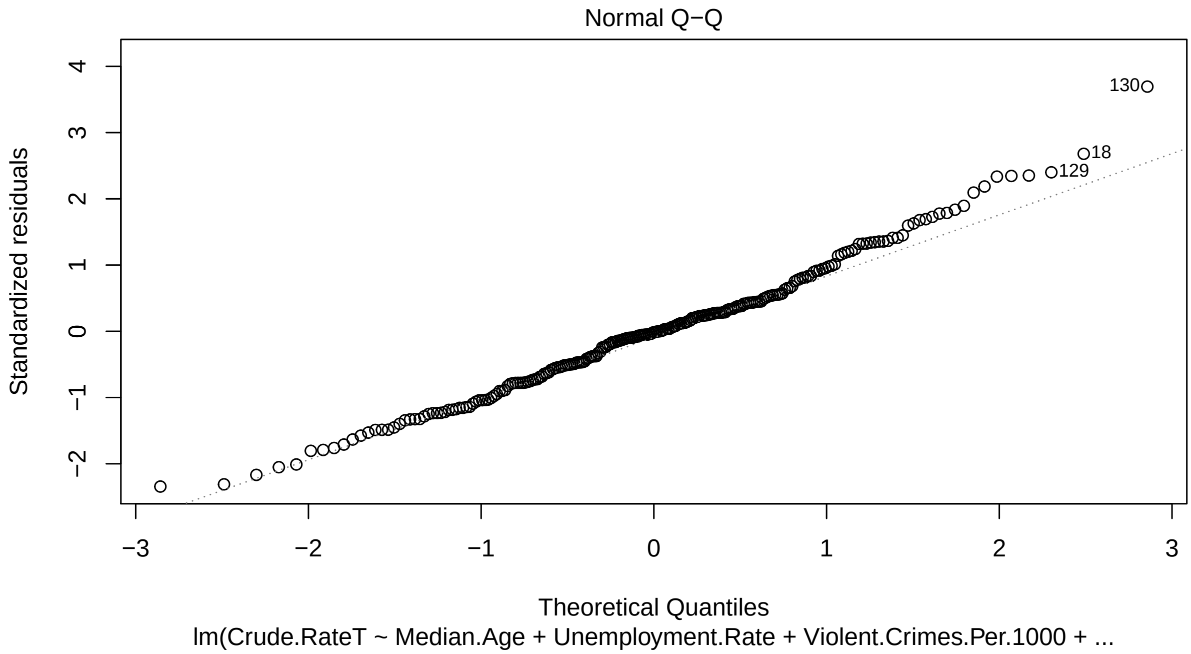
Now we'll run a cross-validation for both of our models. We created a test set (25% of total data) and a training set (75% of total data). We trained our model on the training set and predicted on the test set, calculating SSE, R^2, and two MSE in the process. We repeated this process ten times and took the average of the four summary values. The following table provides those averages. The "means_simple" column shows the results for our best interpretive model and the "means_complex" column shows the results for our best predictive model **(see Appendix B.4)**.

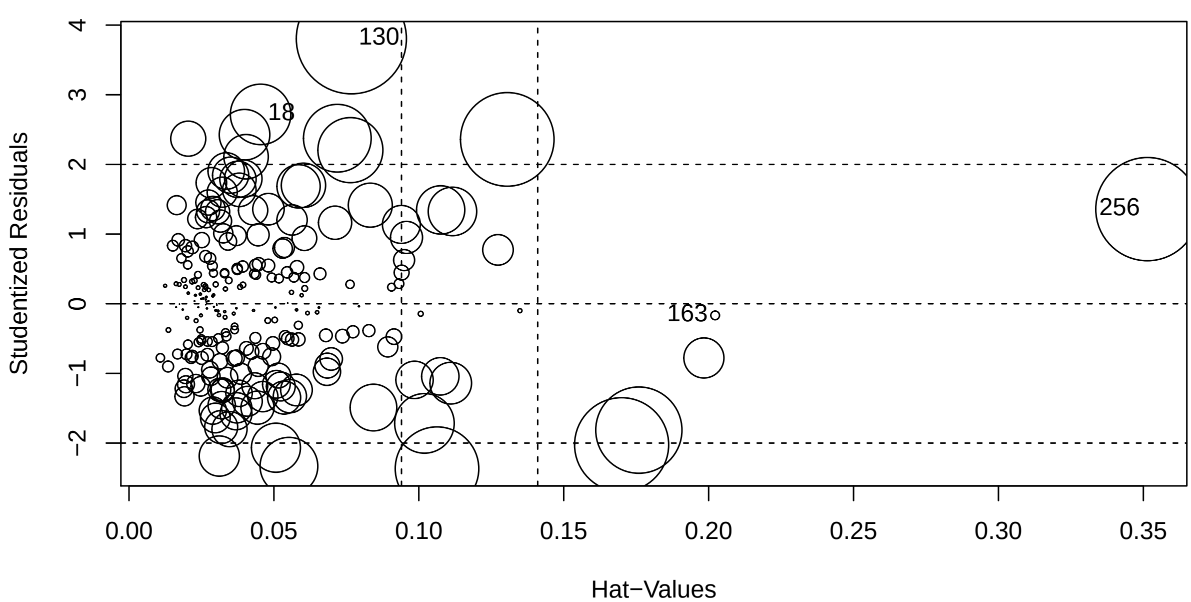
	means_simple	means_complex
sse	1.1172060	0.8391551
r2	0.7028361	0.7775684
train.mse	0.0120176	0.0075293
test.mse	0.0141418	0.0106222

Diagnostics

We ran diagnostics on both of our models, beginning with the simpler one **(see Appendix B.5)**.





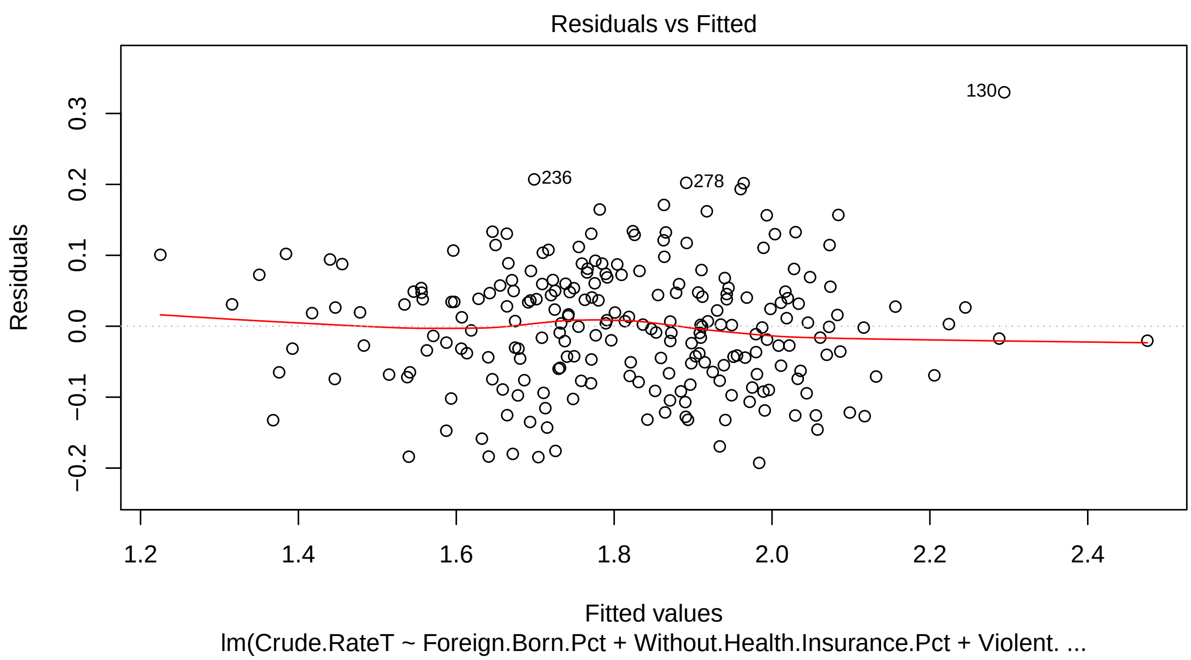


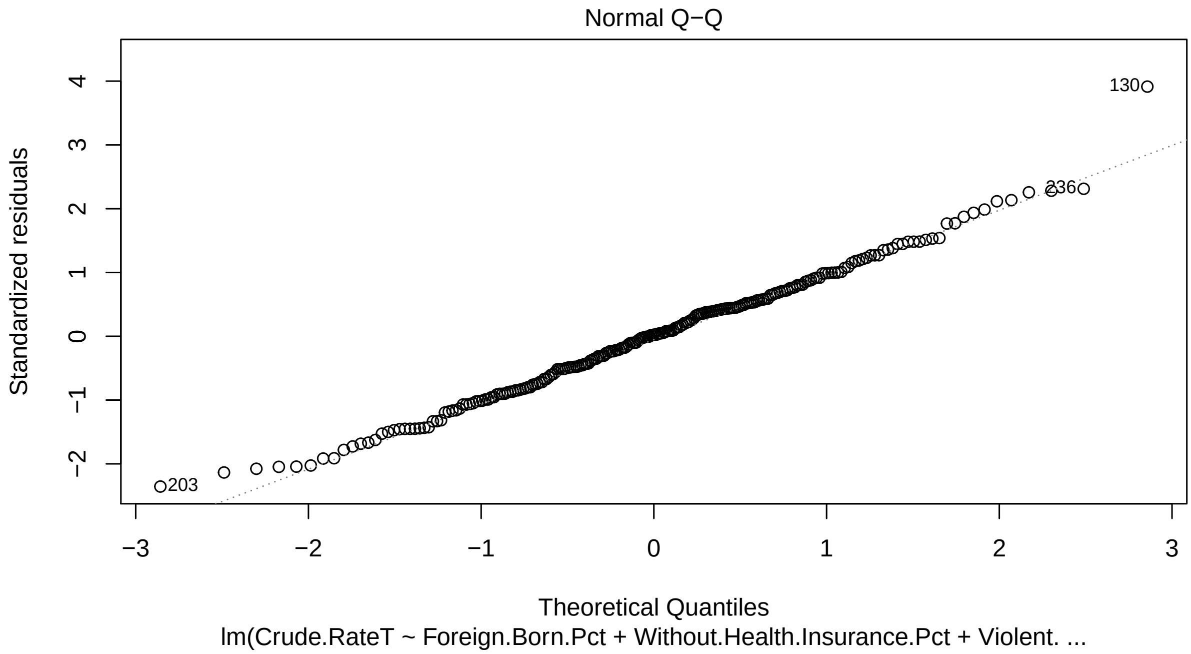

```
##          StudRes        Hat         CookD
## 130  3.8041332 0.07672751 0.1031017788
## 256  1.3583222 0.35137986 0.0905225425
## 163 -0.1670623 0.20225033 0.0006460771
```

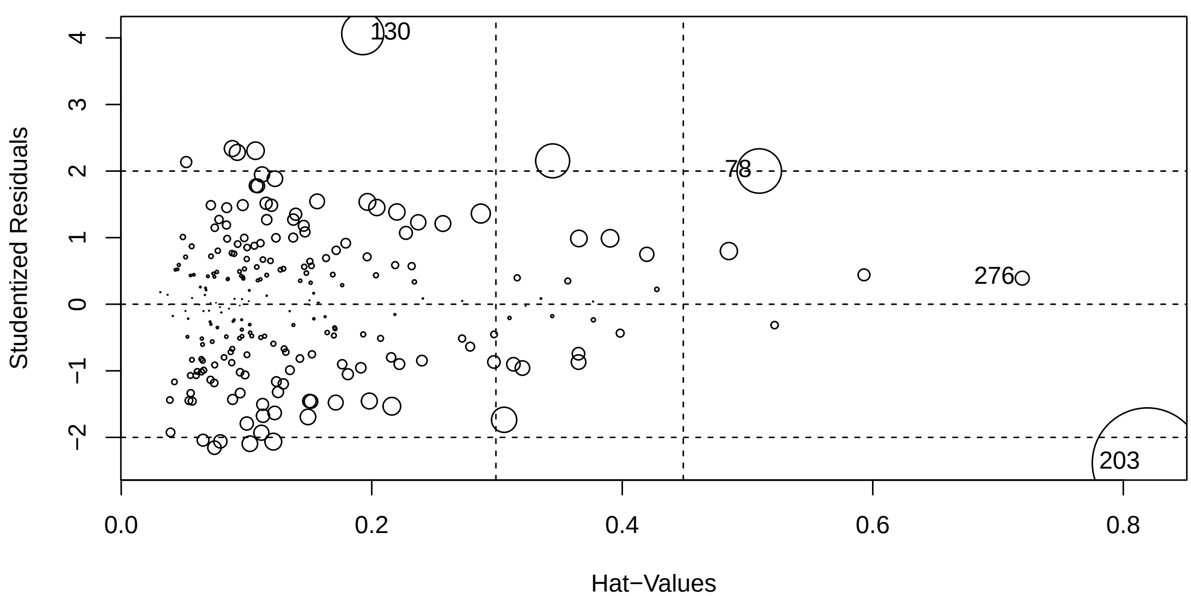
```
## 18   2.7179304 0.04539853 0.0310484163
```

For our more interpretable model, the residual plot looks fairly random and centered at zero, as we want it to be. While there is some curvature to the smoother here, it is of no major concern. The normal plot is straight with the exception of the tails. Especially at the high end, there is some deviation from the expected quantiles which indicates some heteroscedasticity. There are four points identified in the influence plot. Implications of these influential points are discussed later in this section.

We ran these same diagnostics for our more predictive model.








```
##          StudRes       Hat      CookD
## 130  4.0635763 0.1928288 0.10455752
## 78   2.0010003 0.5093743 0.11700545
## 276  0.3916652 0.7193314 0.01128103
```

```
## 203 -2.3834178 0.8192001 0.71850087
```

Notably, the normal plot looks much better at the tail ends then in the first model; with the exception of the point 130, the points form a straight line. The influence plot identifies four points here as well.

In the case of our more interpretable model, removing the identified four influential points has nearly no effect on the coefficients of the model. For the more predictive model, removing the identified points changes some of the coefficients slightly, but the signs and significance values of all coefficients remain the same **(see Appendix B.5)**.

While we do not think it is necessary to look into all of the identified influential points, note that row 130 is the only one that is identified in both models as a high studentized residual in both diagnostic plots. Recall in the histogram of *Crude.Rate* that a few points make the data right skewed (see Figure 3). Row 130 is Orleans Parish, LA and its *Crude.Rate* is 47.41, far above the third quartile value of 14.56. Both models underpredict the *Crude.Rate* for this county. Indeed, a few Google searches confirm that Louisiana leads the nation in gun deaths[14], far above the national average in both homicides and suicides[15]. 2010, the year of the data we use, is the first year that Louisiana recorded more gun deaths than motor vehicle deaths[16]. While this is quite unusual, after verifying its accuracy, we see no reason to omit it from our models.

Model Recaps

Let's start with interpreting the coefficients of our simple model. An increase in the following predictors is associated with an **increase** in the rate of gun deaths per 100,000 when we control for everything else: median age, unemployment rate, violent crimes per 1000, percent of people without health insurance, percent of population that is black, gun law rank, and federal government expenditure per person. Notice this means that if a state increases its gun law rank (laws become less restrictive), we expect an increase in gun deaths. An increase in the following predictors is associated with a **decrease** in the rate of gun deaths per 100,000, all else equal: percent of people who have completed high school, percent of population that is female, and the percent of population that is foreign born.

Recall that our best predictive model has 33 predictors while our best interpretable model has 10. Our best predictive model has an R^2 value of 0.82 while our best interpretable model gives $R^2 = 0.71$ when we use the full dataset. After performing cross-validation ten times for each of the models, the best predictive model gets an average R^2 on the test set of 0.78 while the best interpretable model gives 0.70 **(see Appendix B.4)**.

The difference between the mean squared error (MSE) for the training and test set on our best interpretive model is smaller than that for the best predictive model (0.0021 versus 0.0031). This difference isn't concerningly large and since the average of the test set MSE is smaller for the complex, best predictive model, we know that it is actually a better predictor of *Crude.Rate* and is not overfitting.

[14] https://www.nola.com/news/politics/article_6dc1b261-5e67-51fc-8703-9f30d14e36df.html

[15] https://www.livestories.com/statistics/louisiana/gun-firearm-violence-deaths-mortality

[16] https://www.theadvocate.com/baton_rouge/news/article_c81ec05d-6dc8-5edf-9ff7-3f4539c43bfe.html

Summary

The goal of this analysis was to determine factors that are relevant in predicting the number of gun deaths per 100,000 people at the county level. We sought to create two models, one that is the most interpretable and another that is the most predictive. We used stepwise minimum BIC to find both of these models. Each of them used a transformed response variable ($Crude.Rate^{\frac{1}{4}}$) per the boxcox command's recommendation.

Interpretations

Of course we cannot draw any causal relationships between any of our predictors and gun deaths. But we can think through why some variables explain the variability in *Crude.Rate* while others do not. We can interpret the coefficients of our best interpretable model (hence the name) while we can only give more general conclusions from our best predictive model.

Before our analysis we hypothesized counties that are "worse off" socioeconomically would have higher gun death rates. We can see this is true in our best interpretable model: counties with higher unemployment rates, higher percentages of people not on health insurance, and more federal government expenditure per person have higher gun death rates. We assume that counties with more federal government expenditure are those that need the most government assistance (think welfare programs). Less educated counties can also expect higher gun death rates.

The positive coefficient on *Gun.Law.Rank.2010* confirms our hypothesis that stricter gun laws are associated with fewer gun related deaths. This predictor is particularly relevant as we look towards the upcoming 2020 presidential election and candidates' plans for gun law reform.

There is overwhelming historical evidence that males are responsible for the vast majority of gun homicides (80-90% of the total)[17]. Male-on-male homicides account for almost 75% of all US homicides[18]. Our negative coefficient on the percent of females in a county supports this evidence.

Finally, the only race-related variable included in our simpler model is *Black.Pct*. The model suggests that counties with higher percentages of black individuals will have higher gun death rates. The variable *Black.Pct* is correlated with other factors in our model that contribute to the gun death rate, such as government expenditure per person (correlation = 0.36) and violent crimes (correlation = 0.56). It is also correlated with the percent of people in poverty, a variable that is excluded from this simpler model (correlation = 0.37). Many sources provide evidence of a "cradle-to-prison" pipeline, "particularly for youths of color living in poverty and in disadvantaged urban areas, that triggers a cascade of events that increase the likelihood of gun violence[19]." This idea tells a more complete story for why *Black.Pct* is predictive of gun deaths.

Again, our best predictive model is much harder to interpret. But we do see much more nuance in the inclusion of race-related variables here than in the first model. Also of note here is the fact that our categorical voter group variable is included while it was not relevant in our first model. Democratic counties (<40% Republican voters) can expect fewer gun deaths than swing states (40-60% Republican voters) and even fewer than Republican counties(>60% Republican), all else

[17] https://www.kff.org/other/state-indicator/firearms-death-rate-by-gender/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

[18] https://www.liebertpub.com/doi/abs/10.1089/vio.2017.0016?journalCode=vio

[19] https://www.apa.org/pubs/info/reports/gun-violence-prevention

equal. The Republican coefficient is significant at the 0.001 level while the swing state coefficient only at the 0.05 level.

Limitations and Future Work

There are a few limitations to our analysis. First, while most variables were collected from the 2010 Census report, other variables were not available for that year. Some of our predictor variables are from 2007-2009. For future work, using more recent data for all variables would be relevant due to the sharp increase in US gun deaths after 2010[20]. After the Census' new report in 2020, rerunning the analysis and comparing it to the 2010 model could illuminate factors that have contributed to the historical shift in *Crude.Rate*.

Updating our data would also mean updating the gun law rank variable. A limitation of our analysis here is that this variable is cardinal when it should not be: for example, the difference in gun law severity between the two strictest states may not be the same as the difference for a pair of states with an average strictness. More recent publications from Gifford's Law Center of this variable solve this issue. Rather than ranking states from 1 to 50, Gifford's Law Center gives states a letter grade (A+ through F-). There are some intervals of this new system where no states fall while other grades are given to many states. For example, the 2018 scorecard has 4 states in the "B" range and 22 in the "F" range[21]. This new system more accurately represents differences in state gun laws.

Remember that during our data cleaning process, we had to omit three major New York counties due to missing data. These three counties are particularly important due to their large population size. Excluding these counties from our dataset takes away from the full picture of our analysis as we hope to capture gun death predictors for the entire US. In addition to getting data for these omitted counties, our analysis covered only 44 of 50 states because the CDC reported gun deaths only for counties with 20 or more. In an ideal world, we would have data for all 3,007 US counties.

Finally, our response variable, *Crude.Rate* does not distinguish between types of gun deaths, namely between suicides and homicides. Firearm suicide attempts are about 80-90% fatal while firearm homicide attempts are about 20% fatal[22]. This means that we miss the large number of firearm injuries due to attempted homicide. In future analysis, we would first separate *Crude.Rate* to distinguish between suicides and homicides; suicides accounted for about 60% of US gun deaths in 2017[23]. Secondly, and this is data permitting, we would look into including gun-related injuries. Including these could give a more accurate picture of the burden of gun violence than does just looking at gun-related deaths.

Conclusions

In light of the US gun violence crisis and the upcoming presidential election, the goal of this project was to create two models that predict gun deaths at the county level. We created an interpretable model with 10 variables. Of note, state gun laws, percent of the population that is

[20] https://www.pbs.org/newshour/health/gun-deaths-started-to-rise-after-more-than-a-decade-of-being-stable
[21] https://lawcenter.giffords.org/scorecard/
[22] https://www.cnn.com/2018/04/23/health/gun-deaths-in-men-by-state-study/index.html
[23] https://www.pewresearch.org/fact-tank/2019/08/16/what-the-data-says-about-gun-deaths-in-the-u-s/

female, and federal government expenditure (as a proxy for welfare need) are relevant in predicting *Crude.Rate*. Our more complex model has 33 predictors and explains more of the variation in the response variable; more Republican voters are associated with more gun deaths, all else equal, in this model. Cross-validation shows that the best predictive model is not overfitting and diagnostic tests on both the models confirm we have valid assumptions of linearity, homoscedastic residuals, and normal errors. The major limitations of our analysis include the cardinality of the *Gun.Law.Rank* variable and the inability to distinguish between types of gun deaths. Future work could involve curating an updated dataset that includes all 50 states and distinguishes between suicides and homicides. Although data would be nearly impossible to obtain, including information on gun-related injuries could more wholistically depict the effects of US gun violence.