Metropolitan Mortality Predictors

Group 2

Anthony Zalev, Gabriel Wies, Will Fabian, Zhenguang Huang

11/13/2021

Introduction

Main goals of project

The existence of large population centers paired with the ability to study environmental and chemical variables opens the door to an interesting discussion. How do the chemical and environmental variables in metropolitan areas affect the mortality rate for those living there? The dataset from the U.S. Department of Labor Statistics tracks 60 Standard Metropolitan Statistical Areas around the United States. Using this data we have created a regression model that displays which factors affect mortality in urban areas.

Descriptive Variables

In the dataset we looked at there are 15 numerical variables that we investigated to determine what factors affect mortality rates. We categorized these variables into 3 groups: meteorological parameters, pollutants, and socioeconomic factors.

Age Adjusted Mortality

Age adjusted mortality controls for the effects of population age. This is typically done through weighted averages and gives insight into the percentage of premature deaths. Mortality in this model is the deaths per 100,000 people. This is going to be our main variable that we will use as the Y value in our regression model. This data was collected by the US Census Bureau.

Meteorological Parameters

The dataset was provided without descriptions of where the data was collected from but it can be assumed that meteorological and pollutant data was collected by local weather stations.

Temperature (JanTemp, JulyTemp) The dataset provides average temperatures for the months of January and July. January gives us a good representation of winter temperatures while July . Certain chemicals such as NOx are more detrimental to health in hotter urban areas due to increased reactivity. Colder temperatures in the winter lead to a larger amount of pollutants through an increase in demand for electricity. Cold weather has also been linked to an increase in natural as well as cardiovascular and respiratory deaths. Temperature is given in Fahrenheit. https://pubmed.ncbi.nlm.nih.gov/18952849/

Annual Rainfall (Rain) Rain typically reduces the amount of pollutants in the air. A high annual rainfall could also indicate more severe weather. Annual rainfall is given in inches.

Relative Humidity (RelHum) High levels of humidity lead to an increase in reactions of pollutants in the air. Lower humidity levels are known to cause issues in people with respiratory conditions.

Pollutants

Nitrous Oxides (NOx, NOxPot) NOx reacts with sunlight to produce ground level ozone. According to the EPA, ground level ozone can cause a multitude of respiratory issues. This is measured in parts per million (PPM). https://www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution

Sulfur Dioxide (SO2Pot) High concentration of SO2 in the air leads to the formation of particulate matter in the air. When inhaled it can lead to decreased lung function and premature death of people with lung or heart disease. This is measured in PPM. https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm

Hydrocarbon Potential (HCPot) Hydrocarbons are emitted from vehicles and react with the air to form NOx and other compounds that form ground level ozone. This is measured in PPM. https://www.epa.ohio.gov/dapc/echeck/whyecheck/healthef

Socioeconomic Variables

This data was provided by the US Department of Labor Statistics #### White Collar Workers Percentage (%WC) How labor intensive jobs are, and the environment in which people work for the majority of their lives might have an effect on mortality.

Race (%NonWhite) Many minority populations have been historically disadvantaged compared to white populations. According to the National Equity Atlas, In 2019, 16 percent of people of color lived in high-poverty neighborhoods compared to 4 percent of the white population. In addition to this, Scientific American says that people in lower income neighborhoods are more likely to experience higher levels of pollution.

Education People at different education levels might have different tendencies and behaviors. Education is represented by the average years of education.

Median Income (income) People's behavior and livelihood can change drastically as median income increases. Higher income individuals have access to better healthcare and can typically live healthier lifestyles. Income is given in dollar units.

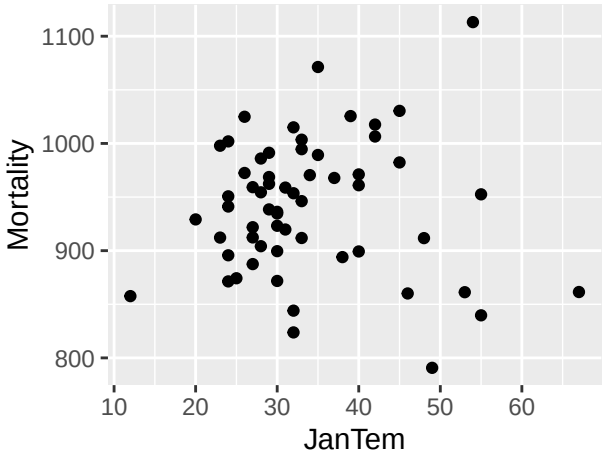
Population (pop) Cities that have smaller populations may have different characteristics than larger cities.

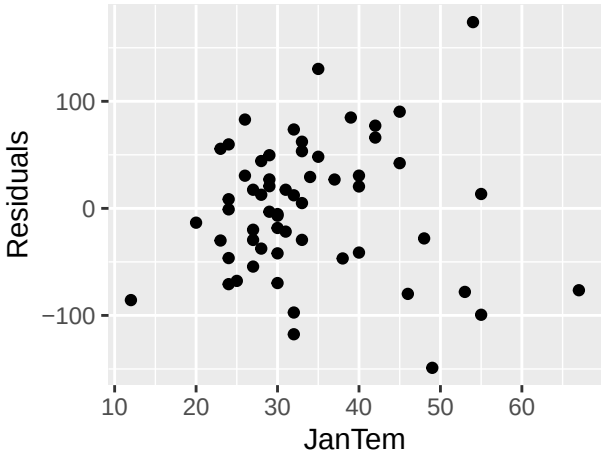
Population Density (PopDensity) As population density increases, the pollutants in a given area affect more people. Population density is given in people per square mile.

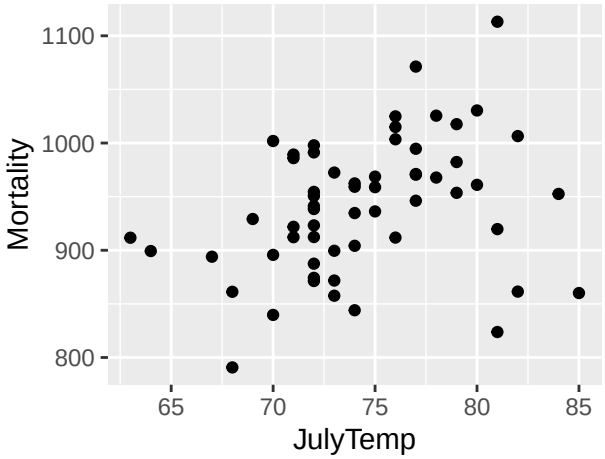
Residents per Household (pop/house) Similar to population density, having more people in a given arena, may affect people more. Also there is the potential to be more efficient with resources with higher population density per household.

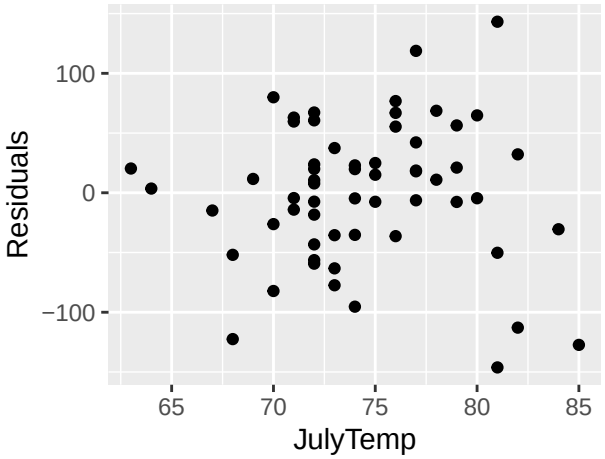
Data Exploration

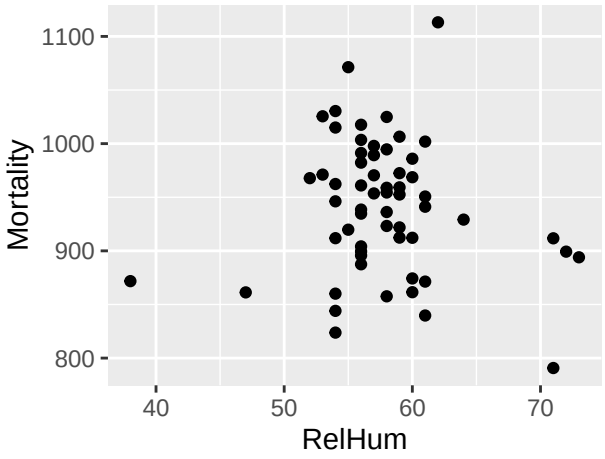
To explore the data we plotted each predictor variable against mortality followed by each predictor variable against the residuals with mortality. By looking at this we will be able to determine if the data has excessive outliers, needs to be transformed, or has other obvious issues.

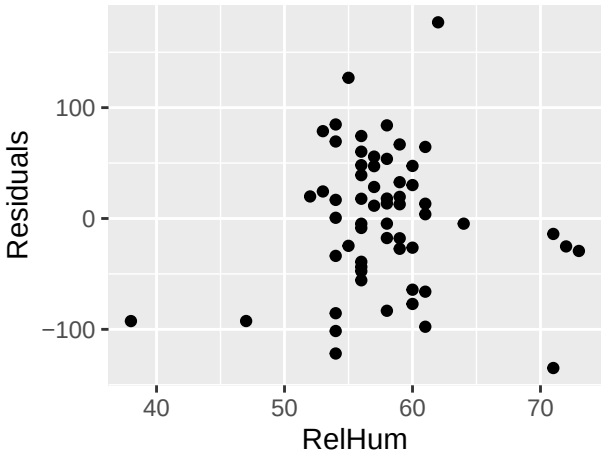


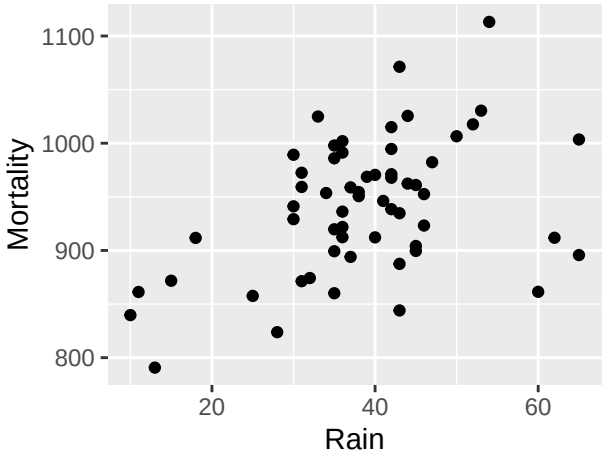


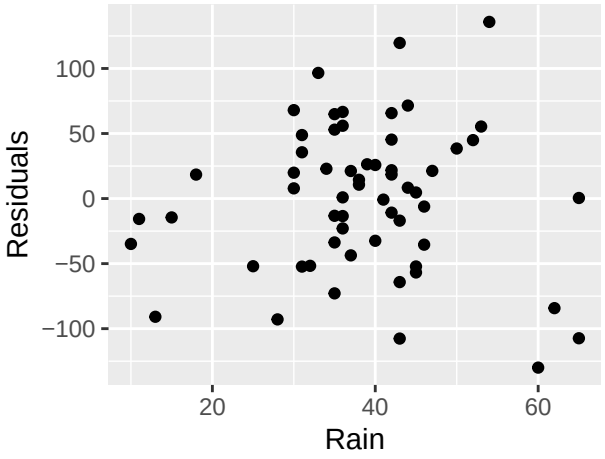


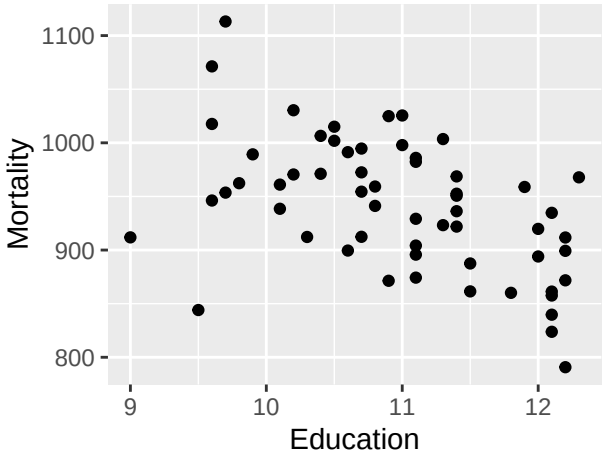


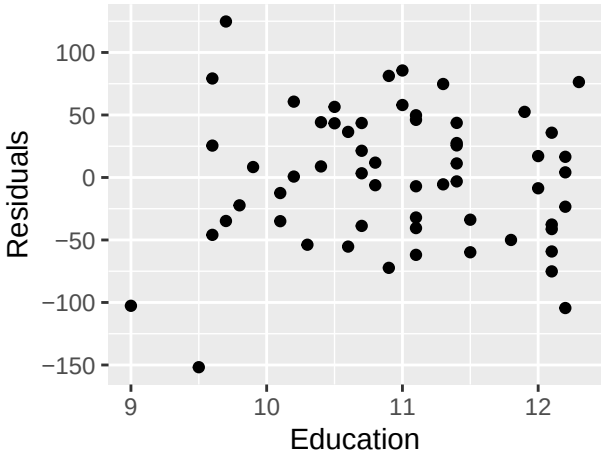


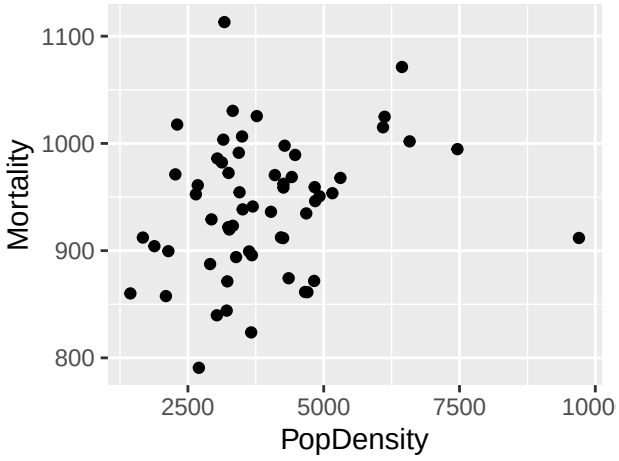


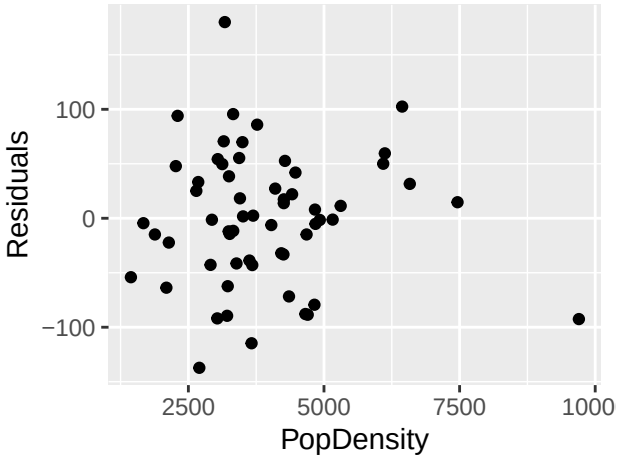

Mortality
1100
1000
900
800
20
40
60
Rain

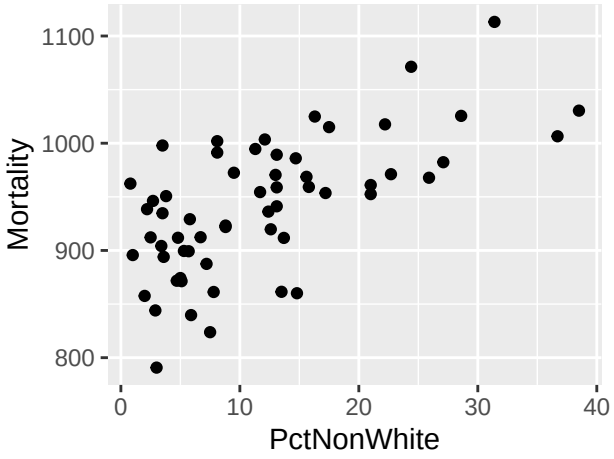

Residuals
100
50
0
−50
−100
20
40
60
Rain

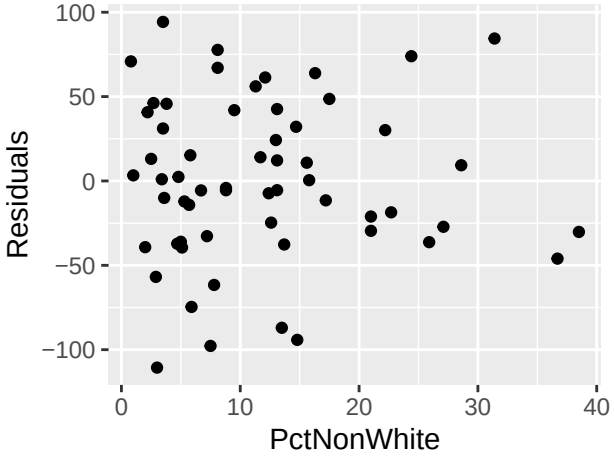

Mortality
1100
1000
900
800
9
10
11
12
Education

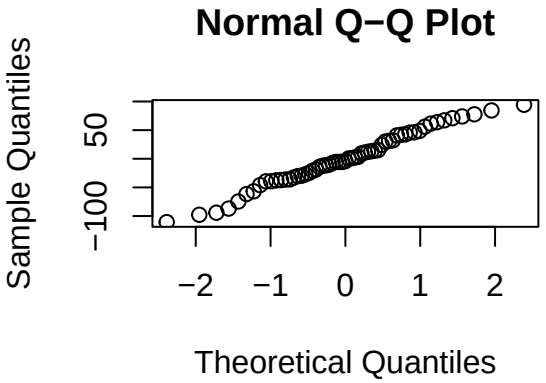

Residuals
100
50
0
−50
−100
−150
9
10
11
12
Education

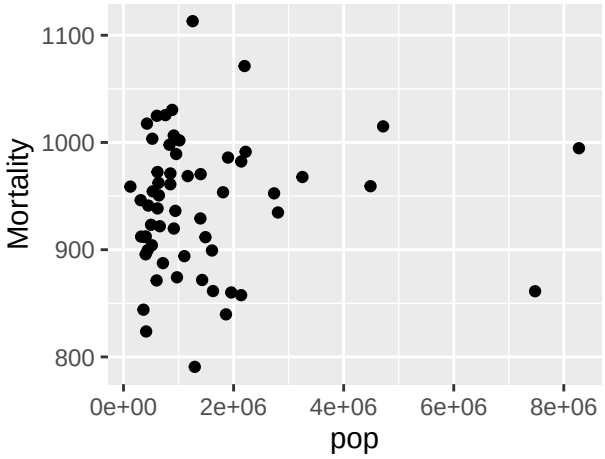

Mortality
1100
1000
900
800
2500
5000
7500
1000
PopDensity

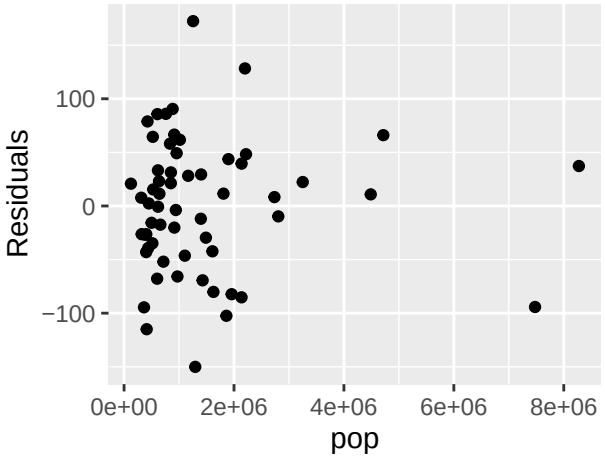

Residuals
100
0
−100
2500
5000
7500
1000
PopDensity

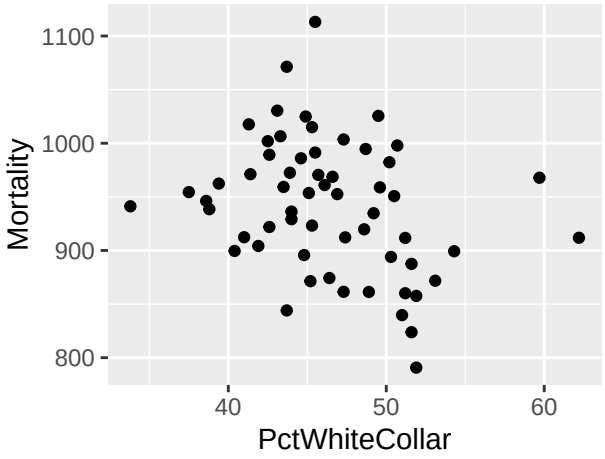

Mortality
1100
1000
900
800
0
10
20
30
40
PctNonWhite

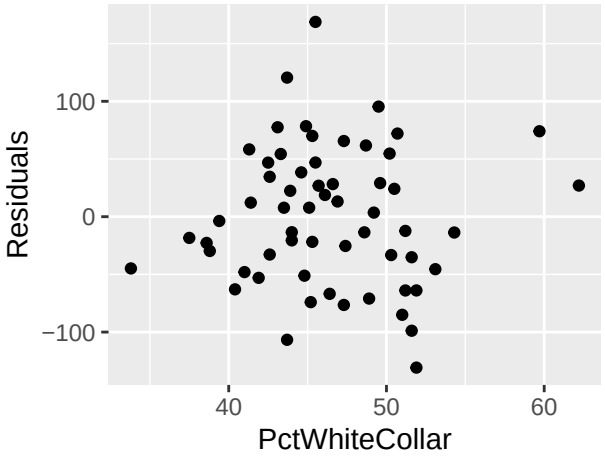

Residuals
100
50
0
−50
−100
0
10
20
30
40
PctNonWhite

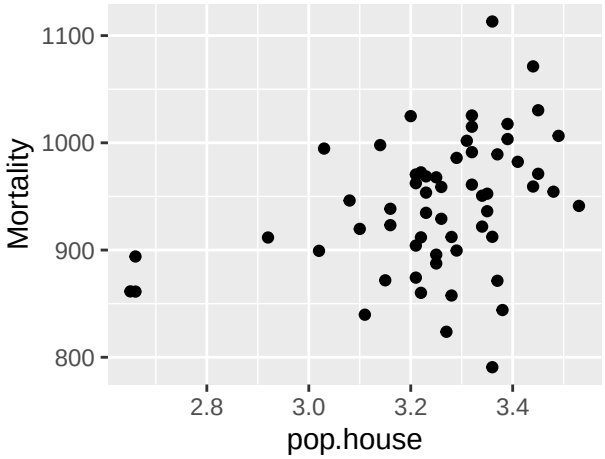

Normal Q−Q Plot
Sample Quantiles
50
−100
−2
−1
0
1
2
Theoretical Quantiles

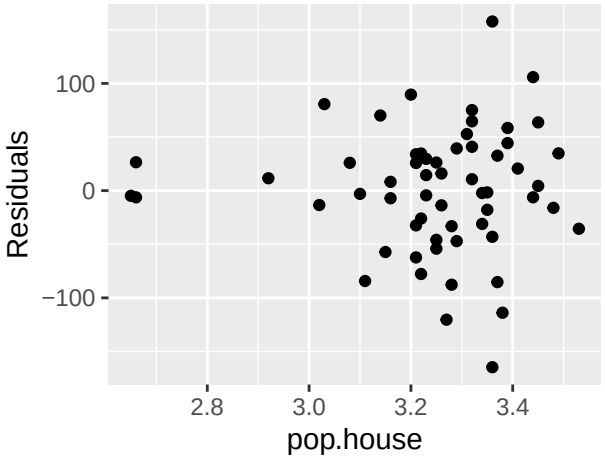

Mortality
1100
1000
900
800
0e+00
2e+06
4e+06
6e+06
8e+06
pop

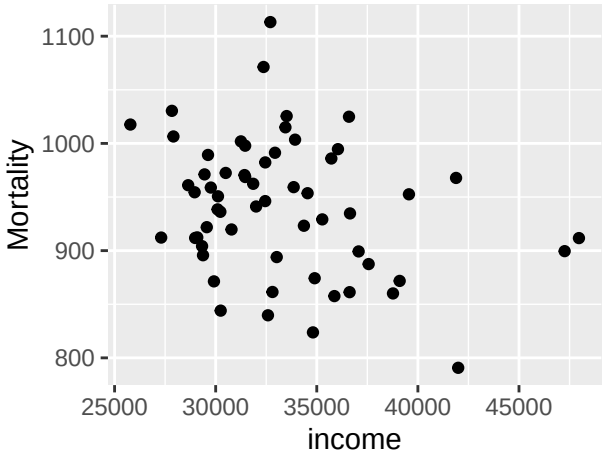

Residuals
100
0
−100
0e+00
2e+06
4e+06
6e+06
8e+06
pop

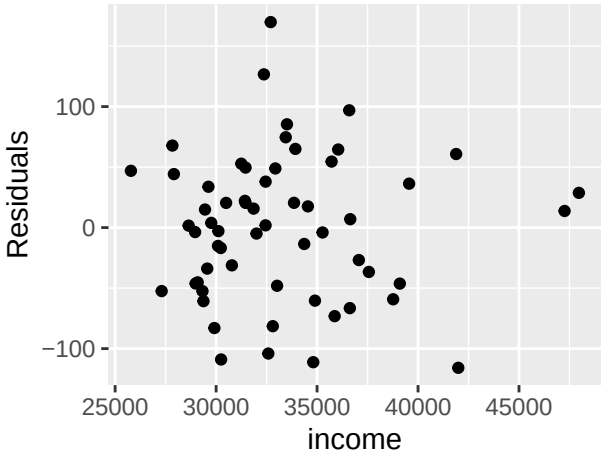

Mortality
1100
1000
900
800
40
50
60
PctWhiteCollar

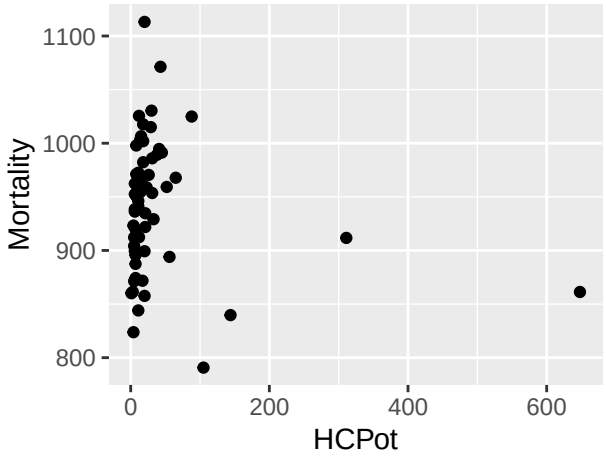

Residuals
100
0
−100
40
50
60
PctWhiteCollar

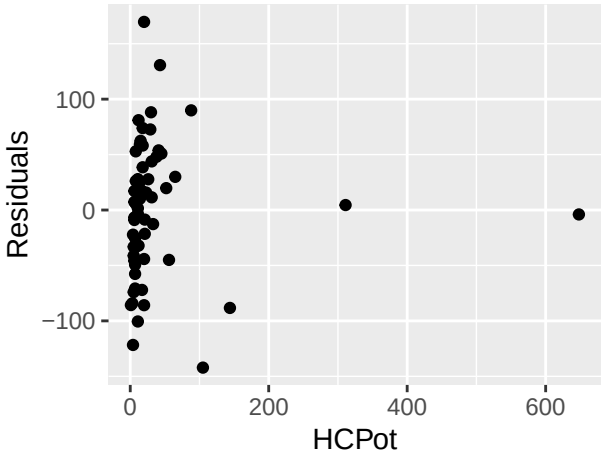

Mortality
1100
1000
900
800
2.8
3.0
3.2
3.4
pop.house

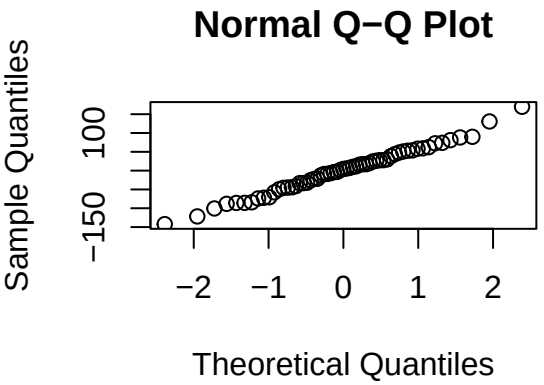

Residuals
100
0
−100
2.8
3.0
3.2
3.4
pop.house

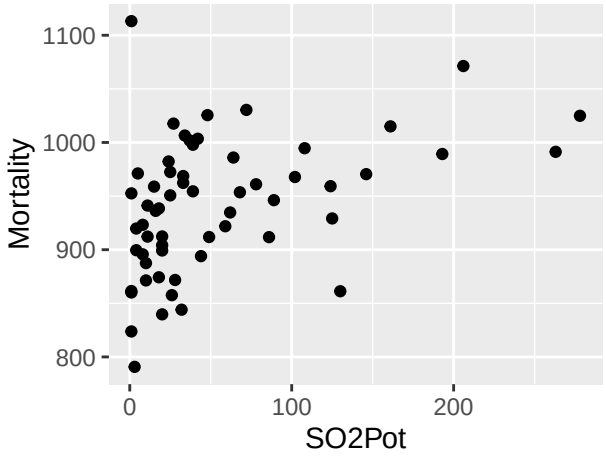

Mortality
1100
1000
900
800
25000
30000
35000
40000
45000
income

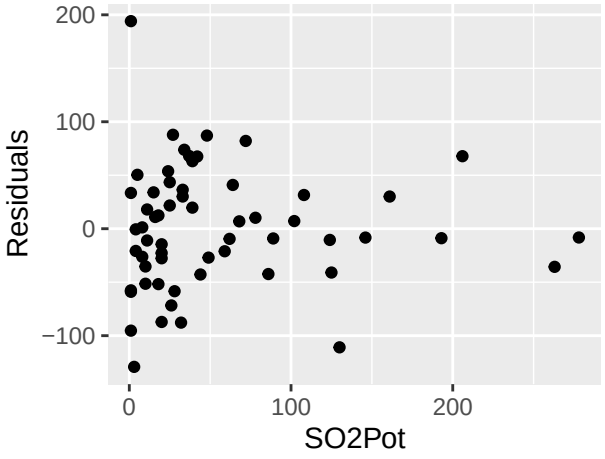

Residuals
100
0
−100
25000
30000
35000
40000
45000
income

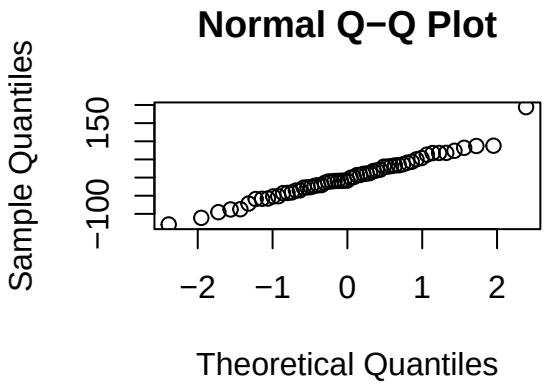

Mortality
1100
1000
900
800
0
200
400
600
HCPot


Residuals
100
0
−100
0
200
400
600
HCPot


Normal Q−Q Plot
Sample Quantiles
100
−150
−2
−1
0
1
2
Theoretical Quantiles


Mortality
1100
1000
900
800
0
100
200
SO2Pot


Residuals
200
100
0
−100
0
100
200
SO2Pot

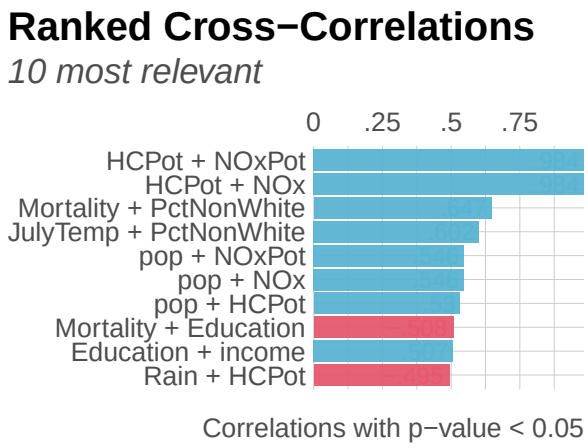


Based on these plots, we can see that some of the plots seemingly have outliers which could affect how we interpret the skew. We will determine if these values are actually outliers after we have fit the model and can use the cooks distance. Beyond this, the residuals for these plots seem to be normally distributed with constant variance which implies that we do not need to transform them. The QQ-Norm plots show that the pollutant plots do have randomly distributed residuals despite what the graphs appear like due to possibly outlying points.

Check Correlations

We find a high degree of correlation between the following predictors: NoxPot and HCPot have a correlation of .98 NOxPot and Nox have a correlation of 1.

So we can use Nox to represent Nox, HCPot, NoxPot.



Model Building and Diagnostics

Due to the small data set we have decided not to split the data. Ideally you need 6 to 10 times the number of samples per final predictor in your model and with a N = 60, we simply don't have enough data points.

Stepwise Regression

To find the best model we do regsubsets on the training data. We then can compare some common metrics across the best representative models for which they in this case have 1 to 8 predictors.

The first metric we have is R^2. The best model of 8 predictors has the highest value at .744. However, all the models with 6 or more predictors have an R^2 greater than .724.

The second metric is adjusted R^2. Often it is considered a more accurate metric than just R^2 since it describes the percent of variation explained by only the independent variables that effect the dependent variable. This makes up for some of the weaker predictors in a given model. Here, the best model of 7 predictors is considered the best with an adjusted R^2 value of .701.

The third indicator is residual sum of squares (SSR). Its no suprise that the best model of 8 predictors has the lowest SSR. Adding predictors to the model always has a less or equal to SSR as the model with $p-1$ predictors.

To address this we look at Mallows's C_p which addresses the overfitting that relying on SSR creates. Model 6 - 1 is best for Cp. It is simpler than 8 -1 and within very close margins on R^2. After that, within a close margin of error the best model of 7 predictors is the next best model.

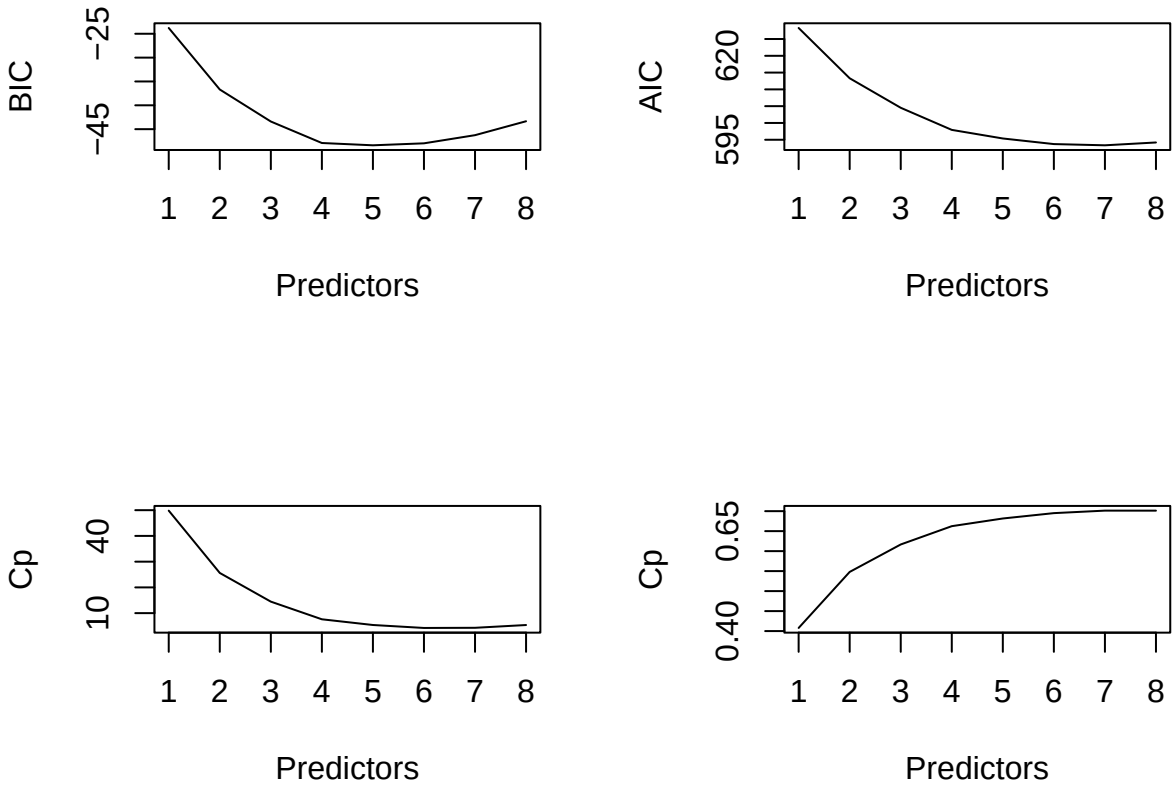
Next we look at Bayesian Information Criterion(BIC). We are looking for models with a BIC closest to 0. BIC also tries to solve the overfitting problem of relying on SSR. BIC introduces a larger penalty term for more predictors than AIC. The second best model of 1 predictor suprisingly has the best BIC, however due to its poor other metrics we look on towards the others. In general BIC seems to weight heavily in favor of the lower predictor count samples that have shown to have low adjusted R^2 and $high C_p$ values.

Finally we look at the Akaike Information Criterion(AIC) which is similiar to BIC besides putting a smaller penalty on the number of predictors. With the lower penalty for predictors we see that the best model of 6 predictors looks like the best model. The best model of 8 predictors comes in close second.

Taking all of this into account we will take a closer look at the best model of 6 (lowest AIC and C_p), 7 (highest Adj R^2), and 8 predictors (lowest SSR)

Model	R^2	Adjusted R^2	Residual Sum of Squares	Cp	BIC	AIC	PRESS
1	0.4180348	0.4078249	131519.79	49.821673	-23.784255	628.2879	-11.10089
1	0.2581521	0.2451372	167652.08	78.619218	-9.462976	642.6092	-39.94498
2	0.5635397	0.5479518	98636.78	25.613810	-36.681795	613.3128	-56.59202
2	0.5420791	0.5257248	103486.70	29.479210	-33.849861	616.1447	-46.14257
3	0.6365134	0.6166869	82145.25	14.470011	-43.398614	604.5184	-58.53693
3	0.6249647	0.6045083	84755.16	16.550123	-41.553230	606.3638	-38.38113
4	0.6857109	0.6624302	71026.98	7.608721	-47.901395	597.9381	-82.76509
4	0.6804678	0.6567988	72211.88	8.553086	-46.925258	598.9143	-45.89257
5	0.7090854	0.6816407	65744.52	5.398579	-48.383589	595.3784	-77.34775
5	0.7058326	0.6780810	66479.63	5.984463	-47.727554	596.0344	-97.65348
6	0.7266009	0.6950548	61786.16	4.243756	-47.969768	593.7147	-96.93193
6	0.7243113	0.6925011	62303.57	4.656138	-47.477742	594.2067	-162.03373
7	0.7373596	0.7013109	59354.77	4.305935	-46.260893	593.3460	-182.98574
7	0.7313239	0.6944468	60718.79	5.393060	-44.920375	594.6865	-118.46455
8	0.7424629	0.7012569	58201.47	5.386749	-43.341050	594.1883	-392.87019

Model	R^2	Adjusted R^2	Residual Sum of Squares	Cp	BIC	AIC	PRESS
8	0.7392640	0.6975463	58924.38	5.962907	-42.612739	594.9166	-251.18884



Best models predictor coeffieicent table

Here we see the chosen predictors of the 3 best models.

Predictors that didn't make it into the largest model include: Relative Humidity, Population, Pop House, Income, and NOx.

In model 8.1 SO2 Potential and Population density have relatively low coefficients. In model 7.1 PctWhiteCollar and Education have low coefficients. In model 6.1 Pct NonWhite and Rain Have low relative coefficients.

	Model 8.1	Model 7.1	Model 6.1
(Intercept)	1179.285	1098.928	1214.028
Education	-7.742	NA	-14.734
JanTem	-1.429	-1.446	-1.559
JulyTemp	-2.304	-2.134	-2.479
PctNonWhite	4.945	4.939	4.942
PctWhiteCollar	-1.750	-2.399	NA

	Model 8.1	Model 7.1	Model 6.1
PopDensity	0.006	0.007	NA
Rain	1.412	1.616	1.387
SO2Pot	0.211	0.225	0.252

Check for Multicollinearity

Multicollinearity is low across all predictors.

We will move forward with model 6 since its the simplest and is the best for C_p and AIC criteria. In the initial anova we see most variables are signifigant except January temperature. There must be some hidden relation that the correlation and multicollinearity plot missed.

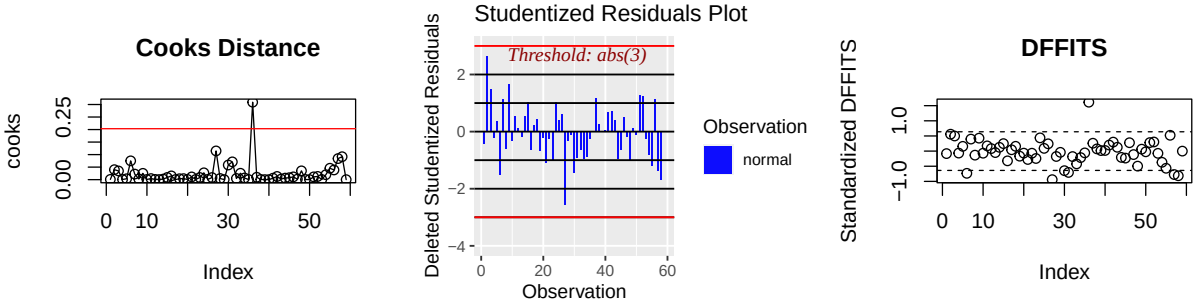
	VIF 8.1	VIF 7.1	VIF 6.1
Education	2.181	NA	1.498
JanTem	1.440	1.439	1.386
JulyTemp	1.927	1.896	1.907
PctNonWhite	2.048	2.048	1.995
PctWhiteCollar	1.775	1.231	NA
PopDensity	1.606	1.442	NA
Rain	1.696	1.416	1.655
SO2Pot	1.513	1.472	1.250

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
JanTem	1	57.50507	57.50507	0.048397	0.8267379
JulyTemp	1	26957.86478	26957.86478	22.688074	0.0000157
Rain	1	21257.96382	21257.96382	17.890967	0.0000951
PctNonWhite	1	90549.71645	90549.71645	76.207767	0.0000000
SO2Pot	1	19301.77852	19301.77852	16.244617	0.0001824
Education	1	6081.53709	6081.53709	5.118297	0.0278803
Residuals	52	61786.15953	1188.19538	NA	NA

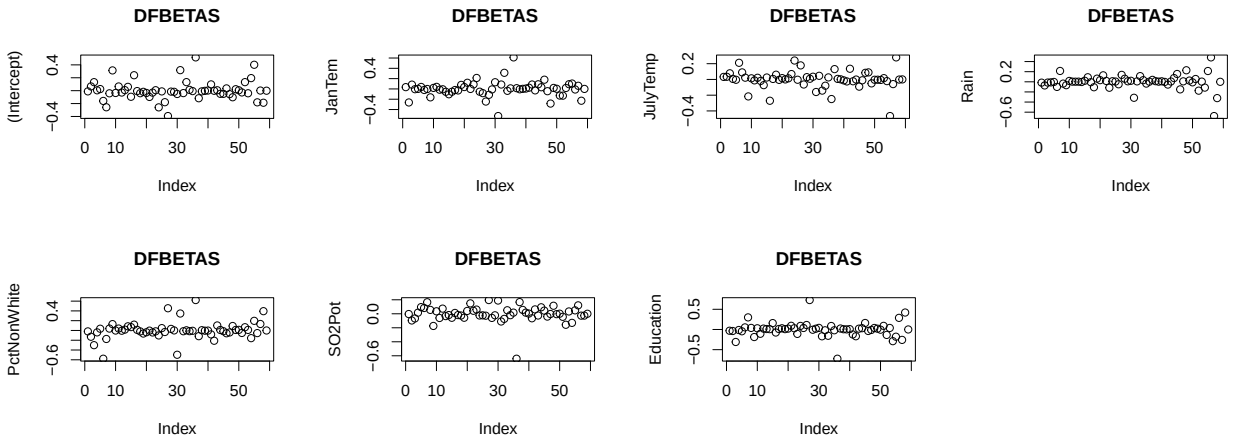
Check Model for Outliers, Leveredge, and Influential Points

In the cooks model we see one outlier. Which means this datapoint has high leverage. No outlier in the studentized residual plot. Quite a few outliers in the DFFITS plot. Which means there are few influential points. The point with a high positive DFFITS value is equal to our cooks outlier. In this case we also see a few outliers in the negatives.

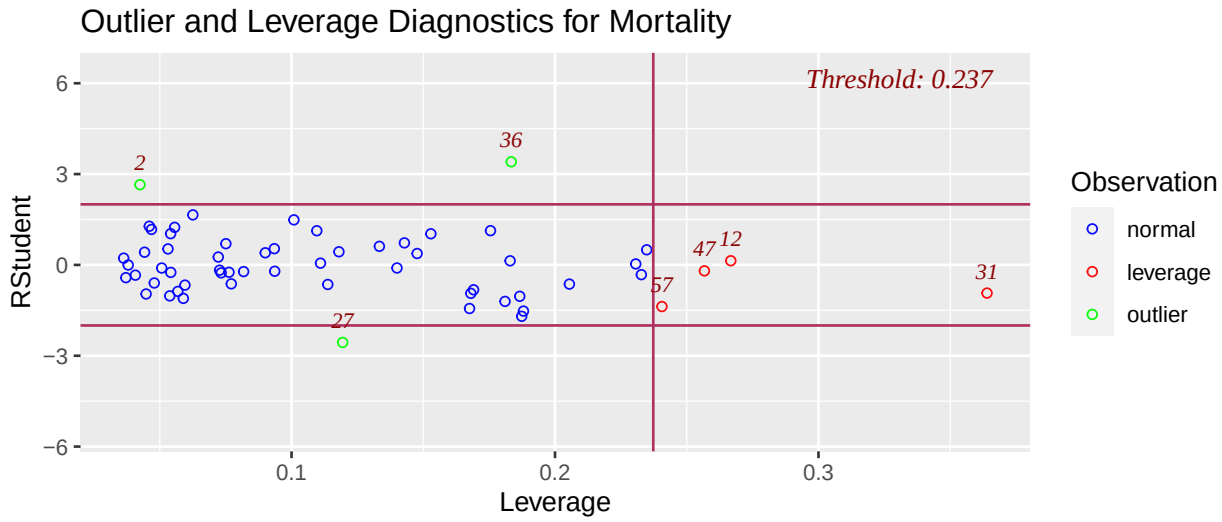
Threshold is $\frac{2P}{N} = \frac{2*6}{59} = .203$



No Influential points in DFBetas



We have outliers, and we have points with leverage, but no points with both, so we will leave all points in and simply do a robust regression as a point of reference as a remedial measure.

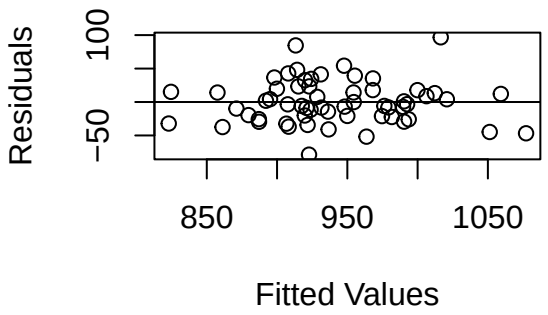


While its not appropriate to calculate R^2 for Robust linear regression, we see a reduction is residual error when we do the regression, indicating the robust regression account for the outliers and influential points better than the simple regression. The RSE of the original regression being 34.47 and the robust regression RSE being 30.8.

	Robust Regression 6.1	6.1 Old
(Intercept)	1189.981	1214.028
JanTem	-1.559	-1.559
JulyTemp	-2.241	-2.479
Rain	1.421	1.387
PctNonWhite	4.583	4.942
SO2Pot	0.279	0.252
Education	-14.146	-14.734

Weighted Least Squares Regression

To see if we need to do this test we check for heteroscedasticity. We do a Breush Pagan Test and see signifigant results for non constant variance. So we will move forward to the weighted regression



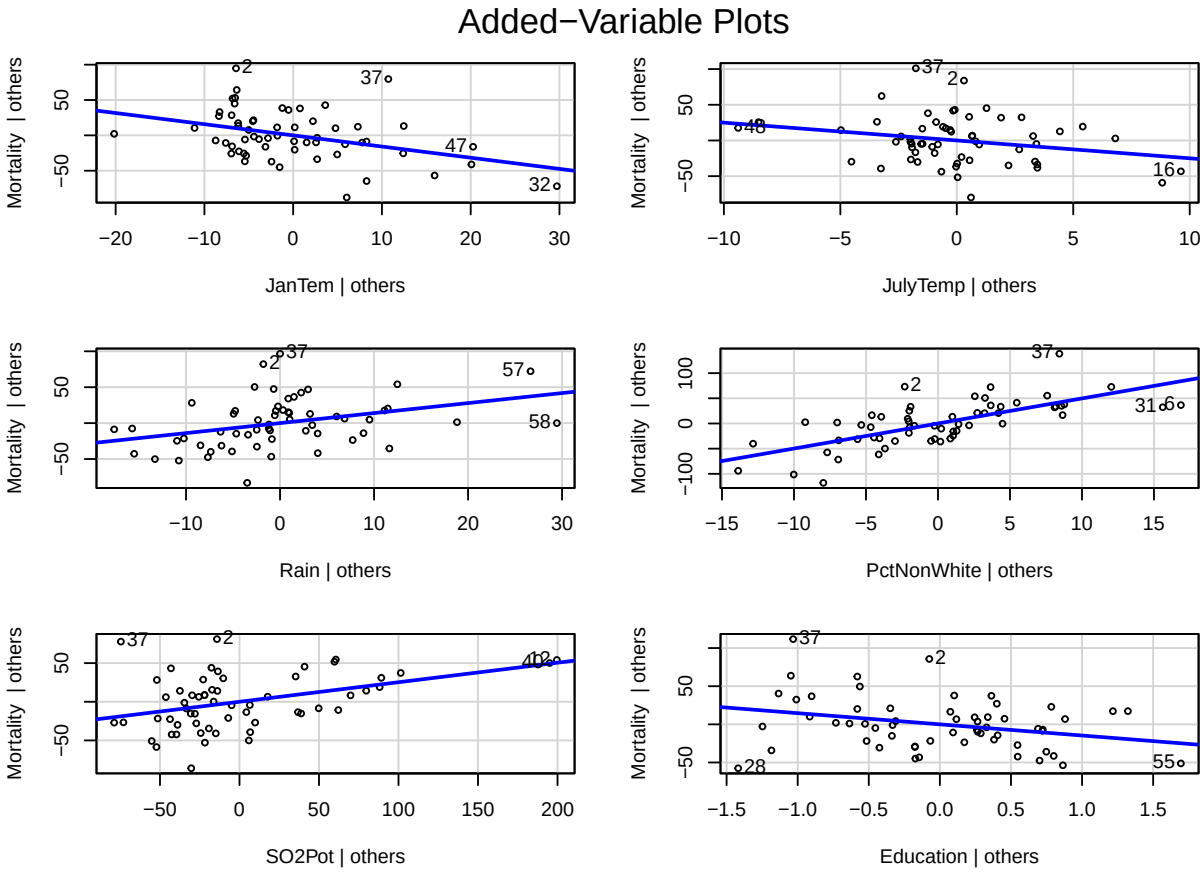

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                            Data
##  -------------------------------------------------------------
##  Response : Mortality
##  Variables: JanTem JulyTemp Rain PctNonWhite SO2Pot Education
##
##         Test Summary
##  ----------------------------
##  DF            =    6
##  Chi2          =    13.99547
##  Prob > Chi2   =    0.02968679
```

Here we perform our weighted least square regression. We will choose to move forward with this model. As shown in the anova, the signifigance the predictors went up, even though January tempurature is still non signifigant.

	Weighted Regression 6.1 Coef	6.1 Old Coef
(Intercept)	1214.576	1214.028
JanTem	-1.580	-1.559
JulyTemp	-2.499	-2.479
Rain	1.394	1.387
PctNonWhite	4.977	4.942
SO2Pot	0.253	0.252
Education	-14.651	-14.734

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
JanTem	1	0.4331719	0.4331719	0.2331189	0.6312473
JulyTemp	1	41.3256941	41.3256941	22.2401329	0.0000184
Rain	1	35.0974711	35.0974711	18.8883076	0.0000646
PctNonWhite	1	142.0712063	142.0712063	76.4580626	0.0000000
SO2Pot	1	29.7971209	29.7971209	16.0358330	0.0001984
Education	1	9.4175023	9.4175023	5.0681908	0.0286213
Residuals	52	96.6242471	1.8581586	NA	NA

In the final added variable plot all the predictors look to have a strong relationship with mortality so it indicates a good model.



With our added variable plot of this model indicating a good, the anova indicating significance of available, and any relevant remdial measures explored. Our final model is as follows:

$$Mortality = 1214.57 - 1.58 JanTemp - 2.49 JulyTemp + 1.39 Rain + 4.97 PctNonWhite - 14.65 Education + .252 SO2Pot$$

Here is our final summary. We see an $AdjR^2$ of .6962 which for a model with human elements in pretty good. Nearly all of the predictors are significant to at least $\alpha = .05$ with the exception of July Temperature.

```
mod <- summary(wls_model_6.1.lm)
mod
```

```
##
## Call:
## lm(formula = Mortality ~ JanTem + JulyTemp + Rain + PctNonWhite +
##     SO2Pot + Education, data = data, weights = wt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1203 -0.8603 -0.1338  0.6885  3.7432
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1214.57616  126.42322   9.607 4.07e-13 ***
## JanTem        -1.58047    0.51731  -3.055  0.00354 **
## JulyTemp      -2.49866    1.34264  -1.861  0.06840 .
## Rain           1.39408    0.49855   2.796  0.00723 **
## PctNonWhite    4.97657    0.71450   6.965 5.63e-09 ***
## SO2Pot         0.25260    0.08042   3.141  0.00278 **
## Education    -14.65095    6.50788  -2.251  0.02862 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.363 on 52 degrees of freedom
## Multiple R-squared:  0.7276, Adjusted R-squared:  0.6962
## F-statistic: 23.15 on 6 and 52 DF,  p-value: 4.237e-13
```

Model Interpretation and Implication

Conclusion

In conclusion, we carefully studied and sifted through the data, using the stepwise method and criterion method to select the best model. We then used DFFITS and Cook's statistics to test the influential observations in our linear regression model. Finally, we use weighted and robust regression to test if there are remedial measures for variances and outliers. For our model we chose these six variables:

Temperature

Climate change is the single biggest health threat facing humanity, and health professionals worldwide are already responding to the health harms caused by this unfolding crisis. The Intergovernmental Panel on Climate Change (IPCC) has concluded that in order to avert catastrophic health impacts and prevent millions of climate change-related deaths, the world must limit temperature rise to 1.5°C. Global heating of even 1.5°C is not considered safe. However, every additional tenth of a degree of warming will take a severe toll on people's lives and health.

Rainfall

Extreme precipitation events have become more common since the 1950s in many regions of the world including much of the United States. The Midwest and Northeast have seen the most substantial increases in heavy precipitation events. Scientists expect these trends to continue as the planet continues to warm. The extreme rainfall events cause flooding and play an important role in exacerbating public health problems, namely the spread of water-related infectious diseases.

Education

The popularization of education can make people realize the importance of environmental protection and learn to protect themselves simultaneously. Moreover, the spread of this knowledge across age groups could significantly reduce mortality.

Race

Data from the Federal Reserve Board's Survey of Consumer Finance show that white households have more wealth than other racial groups, with black and Hispanic families having the least wealth. The average white home, for example, is four to six times richer than the average black household. There is a direct correlation between wealth and access to health care.

SOx Pollution

Sulfur pollution can contribute to respiratory illness by making breathing more difficult. Longer exposures can aggravate existing heart and lung conditions, as well. Sulfur dioxide and other SOx are partly culpable in the formation of thick haze and smog, which can impair visibility and impact health.

Suggested Actions

Based on our model we give the following suggestions:

1. Strengthen education and publicity so that people build up the awareness of protecting the environment and learn to prevent natural disasters

2. Reduce carbon emissions, the earth's future temperature rise will largely depend on cumulative greenhouse gas emissions.

3. Give priority to low-sulfur fuels, such as low-sulfur coal and natural gas, and improve coal-burning technology, reduce the emissions of sulfur dioxide and nitrogen oxide in the process of coal burning.

Bibleography


```
## [1] _ How Smog Affects Health_. <URL:
## https://www.epa.ohio.gov/dapc/echeck/whyecheck/healthef>.
##
## [2] _Health and Environmental Effects of Particulate Matter (PM)_.
## 2021. <URL:
## https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm>.
##
## [3] _Health Effects of Ozone Pollution_. 2021. <URL:
## https://www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution>.
```