ISYE-6414 Final Report - Air Quality Index

Ishika Arora, Ian Jiang, Tess Leggio

November 30, 2018

0 - Problem Setup

1 - Problem Statement

Air quality is a factor that is cited as both contributing to and being affected by climate change (1), and has been shown to have a direct impact on human health. As a factor with such important consequences, our team decided to study factors that impact air quality. For our project, we focus on answering the following question: *How is air quality across the United States related to demographic, emissions, and weather factors?* Specifically, we chose to investigate the impact of factors related to economic output and climate, including temperature, population, per capita income, greenhouse gas emissions, and industrial power consumption and generation. Our response variable is the Air Quality Index (AQI), a measure of local air quality calculated by the Environmental Protection Agency (EPA) based on the concentrations of 6 major classes of ground-level pollutants: ozone, carbon monoxide, nitrogen dioxide, sulfur dioxide, and two measures of particulate matter (PM2.5 and PM10). ##individual AQIs

The scope of our analysis ultimately included county-level data for all US counties with available AQI data. To limit the size of our dataset to a size manageable within R, we used daily data from 2016, the most recent year with data available for each of the factors included in our analysis.

(1): https://www.epa.gov/air-research/air-quality-and-climate-change-research

2 - Data Collection

Most of the data we used in our analysis is available from government websites. Data is available for download as CSV files from the following sources: per capita income data from bea.gov, population data from census.gov, powerplant energy consumption and generation data from eia.gov, and AQI data from aqs.epa.gov.

Weather data were significantly more challenging to collect. Daily weather sensor data is available from aqs.epa.gov, however the county data is far less complete than AQI data, so to avoid omitting a large portion of data from analysis, which could introduce bias into the model we instead collected temperature data from api.mesowest.net, an API which allows straight-forward programmatic extraction of data. By setting parameters in the URL, daily temperature and other weather averages can be scraped for any given day, state and county. We exploit this feature by looping through all counties in the AQI data and all days within our date range, sending a "get" request for each iteration. Data is available in JSON format, and includes daily averages for all stations within a given county. Therefore, a global measure of temperature and other weather factors is easily calculated by converting the JSON format into a Python dictionary and averaging across all stations. The speed of the script is highly dependent on local factors and internet speed, but on average it takes about 10 minutes per county for each of the 1053 counties included in our analysis.

3 - Data Cleaning

Upon collecting data from each of the 6 identified data sources, we were faced with the task of merging the datasets together. First and foremost, we had to ensure that the datasets merged together well. We had to merge data based on state name, county name, year, month, and date, so it was essential that these

fields were in the exact same format across data sources to avoid loss of data. In particular, we spent much time ensuring that county names were in the proper format. Since county name is a string field, it was not uncommon to see several different formats (for example, Saint John vs. St. John vs. StJohn). To avoid having several sparsely populated observations for each county, we wrote python scripts to format each dataset to use the same format and merge the datasets together.

After merging the datasets, we noted that there were several columns that were only sparsely populated, especially within weather data such as wind speed, relative humidity, and precipitation. As these data were sparse, we decided not to use them in our analysis because it may have introduced bias into the model. As such, we dropped the columns from our dataset. The remaining dataset was large and did not have a large number of missing values, so we decided to omit NAs from analysis rather than imputing missing data.

Next, we checked our merged data to ensure data quality. One of our more effective methods for investigating data quality was to plot individual predictors versus time to easily spot any clear outliers. Using this method, we were able to quickly spot clear outliers, such as the county Iberville, Louisiana, which reported temperatures of around -38C in the middle of the summer, which is not a reasonable value and did not align with checks on other websites. These values and few others were thus identified as clear data quality issues and were omitted from analysis.

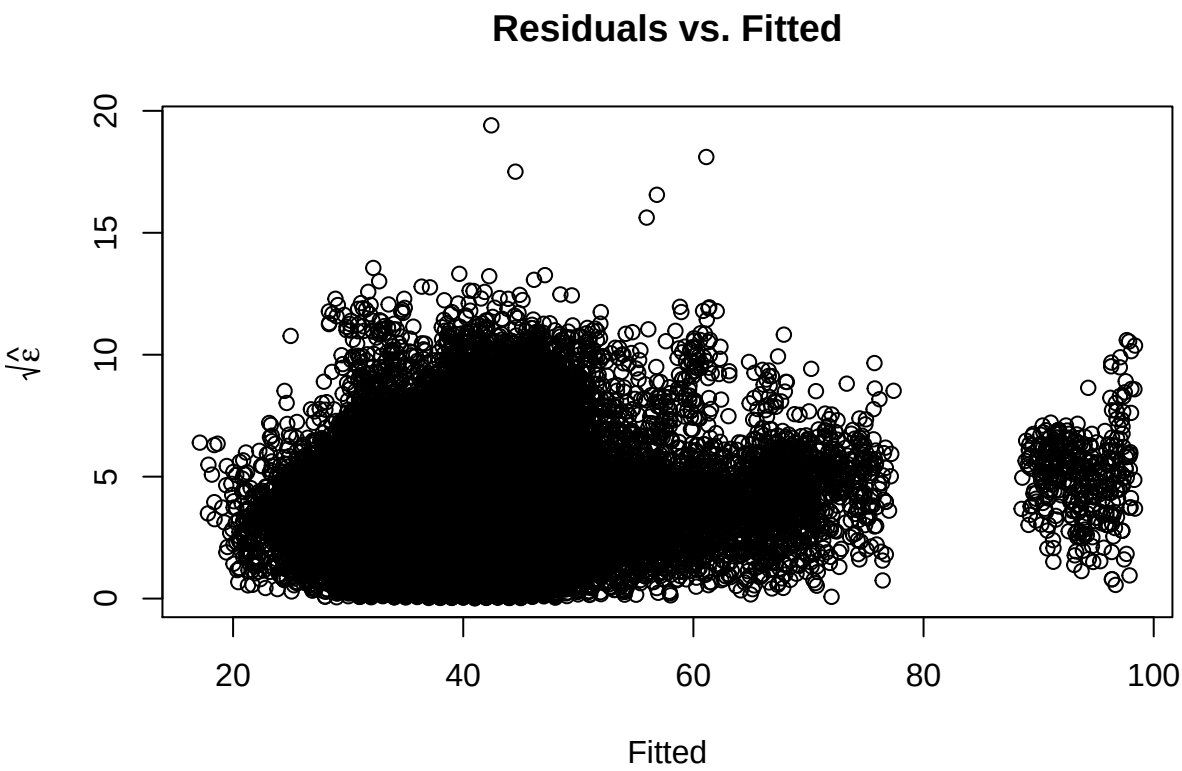
Finally, our initial dataset across multiple years included more than a million observations and was too large a file to easily work with within R, so our team made the decision to trim our dataset to a more reasonable size of approximately 330,000 observations by including only data from 2016.

4 - Model Diagnostics (Ian)

In the following analysis, we give results for diagnostics preformed on one model, the model for AQI_NO2. The data for this model does not include records with missing or spurious values of the response or predictors.

Checking Error Assumptions

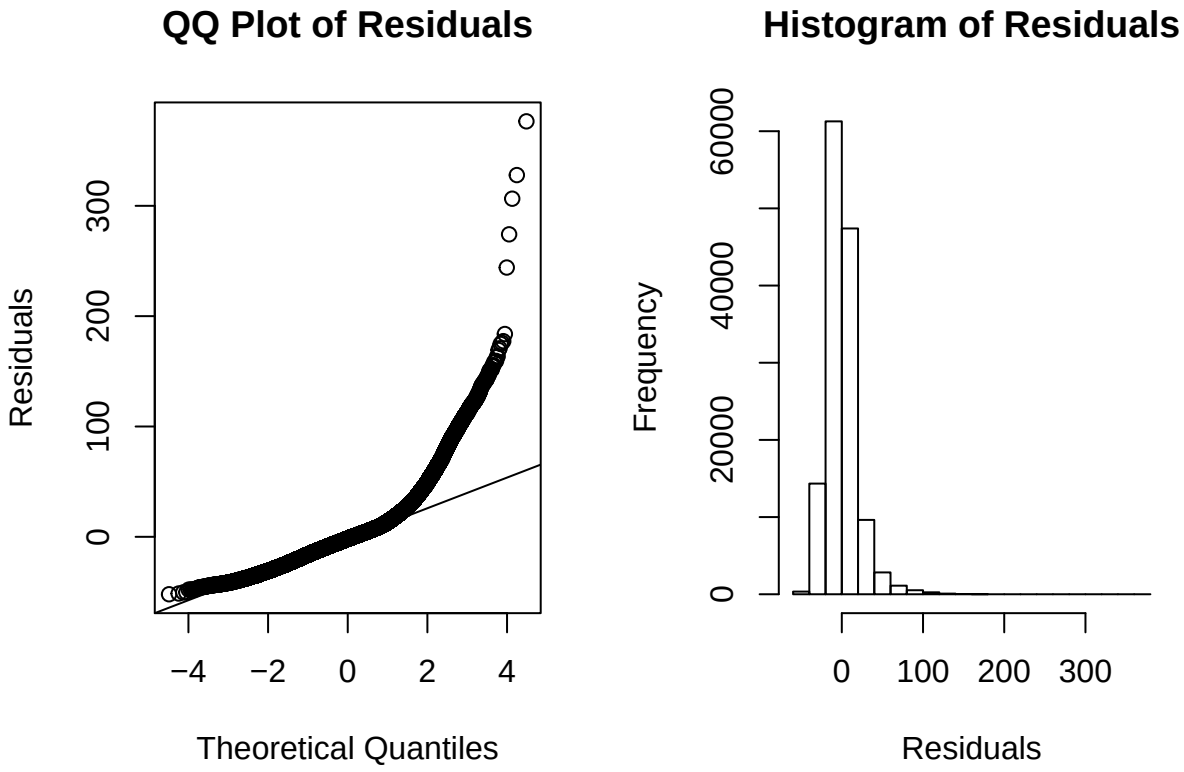
We check for the constancy of variance by examining the relationship of the residuals to the fitted values of the full model. If residuals follow a normal distribution, then their magnitudes would be half-normally distributed. Therefore, in order to mitigate the effects of the skewedness of this distribution, we also take the square root of the magnitudes, and plot it against the corresponding predictions:



When the square root residuals are fitted on the estimated response, we find a significant p-value, indicating that there is a dependence of the residuals. Therefore, transformation of either the response or factors are needed. For the moment, we leave the response alone, in order to proceed with the diagnostics.

Checking Normality Assumption

Secondly, we investigate whether or not the residuals of the model are normal distributed. For this, we visually examine the histogram of residuals, as well as the qq-plot of residual quartiles against the corresponding Gaussian quartiles:



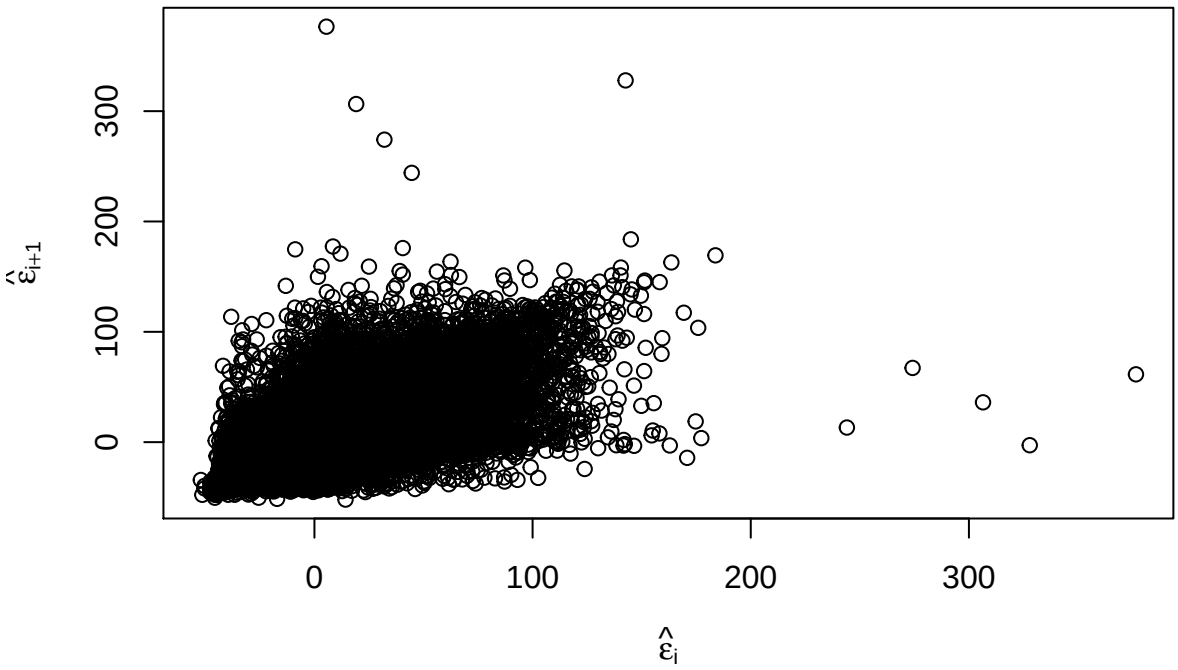
From the qq-plot, the quartiles appear to be skewed, with mass tending to the tails. Additionally, the histogram shows that the residuals appear to be right-skewed. We also use the shapiro-wilks test:

```
##
##  Shapiro-Wilk normality test
##
## data:  lmod$residuals[1:4999]
## W = 0.86576, p-value < 2.2e-16
```

The results reinforce the conclusion that residuals are non-normally distributed.

Checking for Serial Correlation

Next, we check the serial correlation in the error, to determine if there is any serial dependence of the residuals. This is achieved by simply regressing each residual on the previous, to determine if there is any trend:




```
##
## Call:
## lm(formula = tail(lmod$residuals, n - 1) ~ head(lmod$residuals,
##     n - 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -221.56   -7.41   -0.90    5.54  372.89
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -9.425e-05  3.845e-02  -0.002    0.998
## head(lmod$residuals, n - 1)  6.673e-01  2.006e-03 332.619   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 137831 degrees of freedom
## Multiple R-squared:  0.4453, Adjusted R-squared:  0.4453
## F-statistic: 1.106e+05 on 1 and 137831 DF,  p-value: < 2.2e-16
```

The R squared and the p-values show that this result is significant, and we can conclude that there is marked autocorrelation among the residuals which may be caused by a missing covariate.

5 - Model Selection

A) Stepwise selection

We tried running stepwise regression on the model before removing the collinear variables from the model, but the results were not very satisfactory. It removed only two factors out of the 20 predictors. However, when we looked at the pairwise correlation factors (and plots) and the VIFs, several factors seemed to be strongly correlated. This suggested that we had to remove more than just two factors from the model to

address multicollinearity.

B) Backward selection

Because of the complex interaction of the factors, we decided to use backward elimination for removing non-significant factors from the model. We started with 13 predictors in the model and then removed the predictor with highest p-value greater than 0.05. We removed two more factors from the model and finally used the following factors to train our model - - *other greenhouse gases, hydrofluorocarbons, biogenic CO2, population, perfluorinated chemicals, hexafluorethane, stationary combustion, industrial power consumption, Temperature, sulfur hexafluoride, and per capita income.*

6 - Transformations (Tess)

Based on the model diagnostics, specifically the violations of constant error variance and normality, it was clear we needed to modify the initial model. Several transformations were performed and tested to see if any transformations resulted in a better-fitting model. Some of these included square root transformations, adding polynomial terms for the predictors, investigating interactions between predictors, using log transformations, and others. Adding polynomial terms for Temperature improves the model fit. Most of the other transformations tested only improved model fit marginally while making it more difficult to interpret the model. Ultimately, we used Box Cox to identify the best transformation on the response.

Box Cox identified a lambda value of 0.55, so for model interpretability we used a square root transformation on the response. This helped resolve some, but not all of the error assumption violations while improving the R squared value.

With a large dataset, even mild deviations from non-normality may be detected, but there would be little reason to abandon least squares because the effects of non-normality are mitigated by large sample sizes.

When the errors are not normal, least squares estimates may not be optimal. They will still be best linear unbiased estimates, but other robust estimators may be more effective. Also tests and confidence intervals are not exact. However, we can appeal to the central limit theorem which will ensure that the tests and confidence intervals constructed will be increasingly accurate approximations for larger sample sizes. Hence, we can afford to ignore the issue, provided the sample is sufficiently large or the violation not particularly severe.

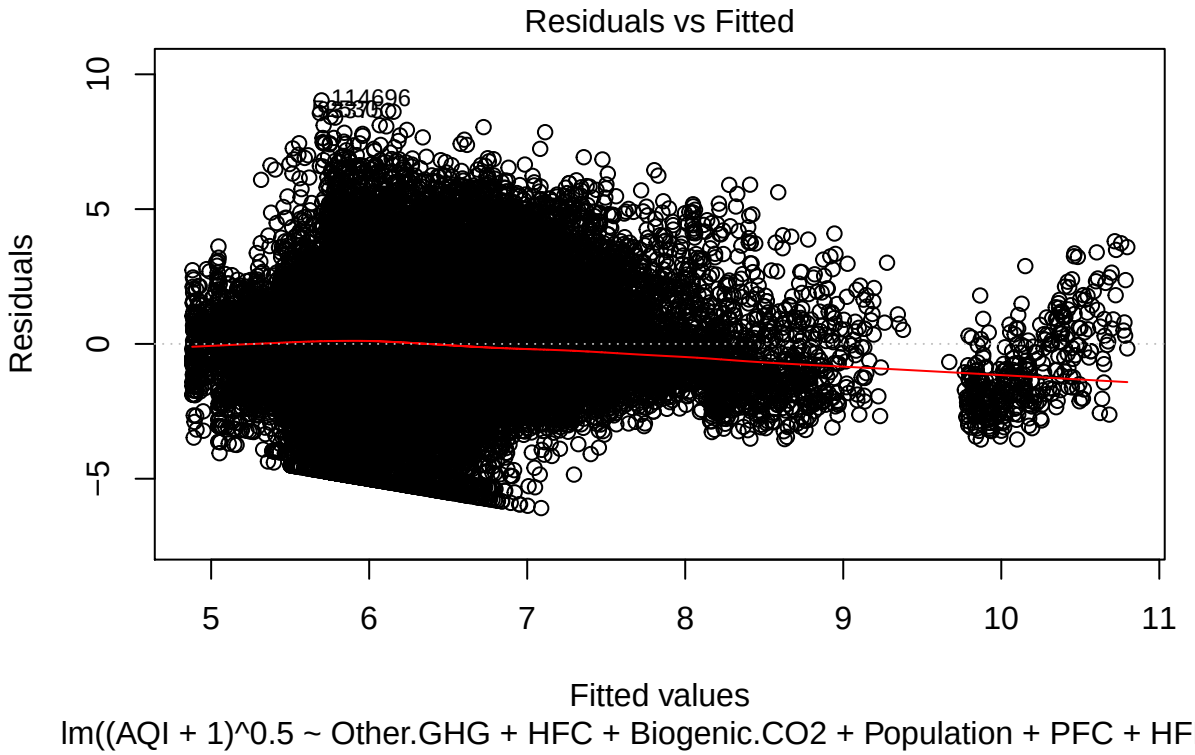
robust regression gives a similar residual standard error

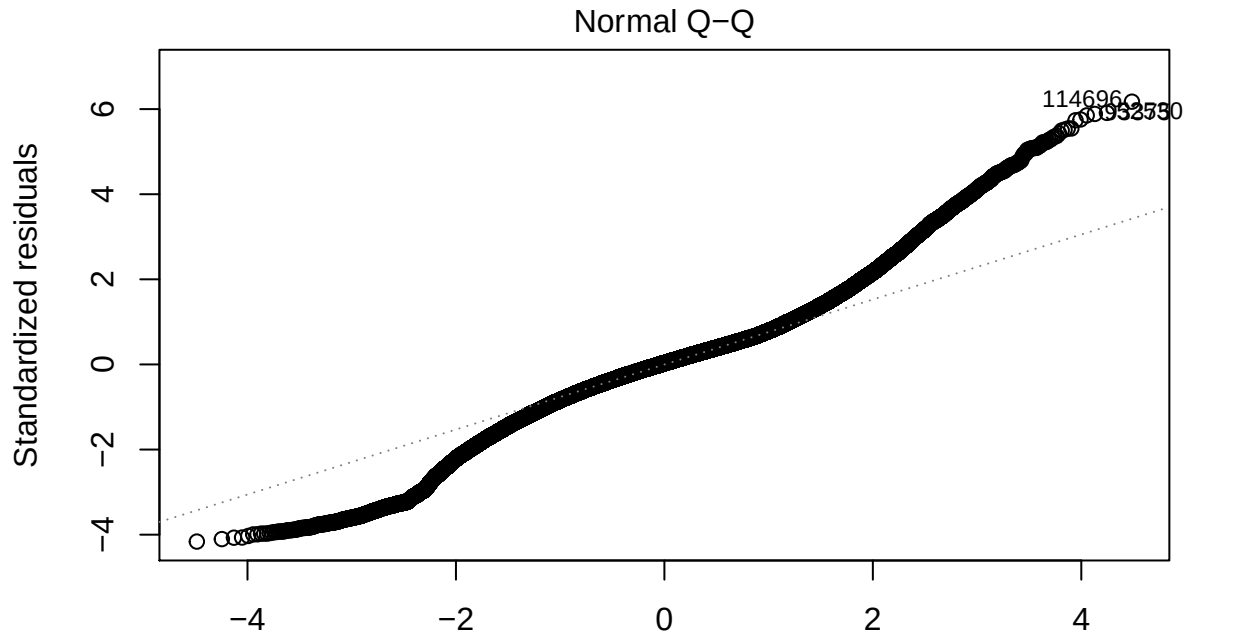
```
lmod.select <- lm(formula = (AQI+1)^.5 ~ Other.GHG + HFC +  Biogenic.CO2 + Population + PFC + HFE + Sta



summary(lmod.select)
```

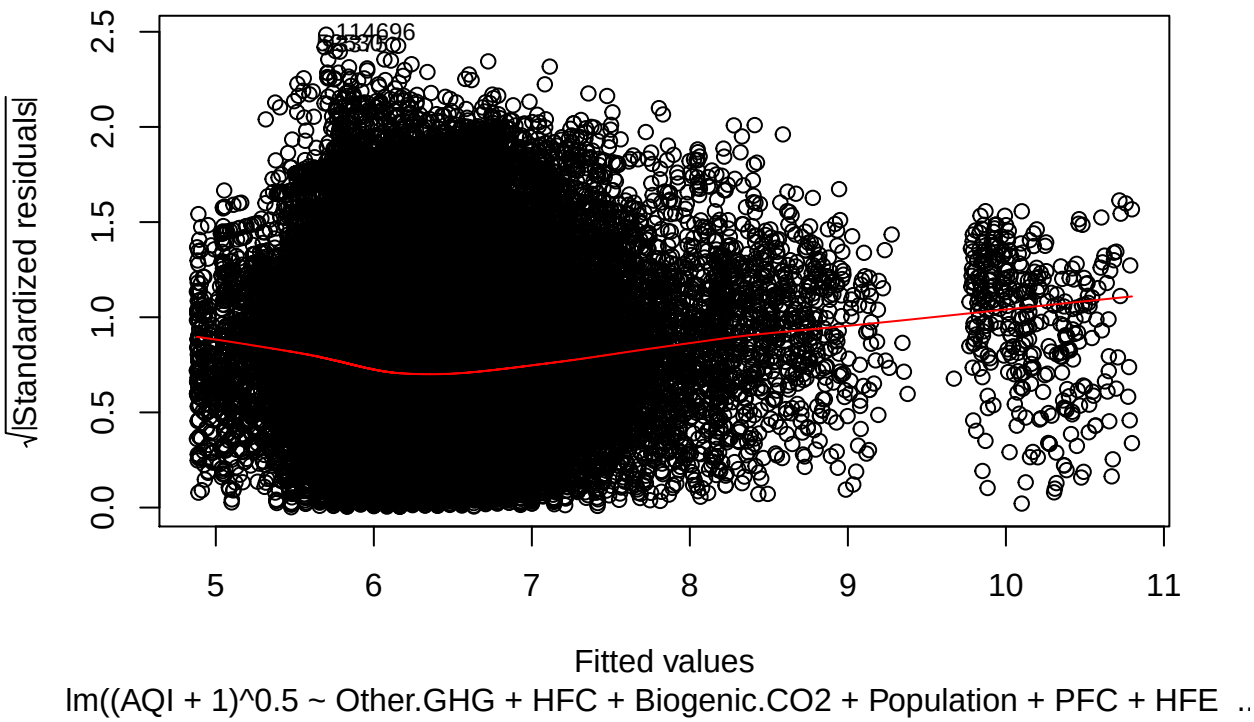
```
##
## Call:
## lm(formula = (AQI + 1)^0.5 ~ Other.GHG + HFC + Biogenic.CO2 +
##     Population + PFC + HFE + Stationary.Combustion + pp_consumed_MMBtu +
##     poly(Temperature, 2) + SF6 + Income, data = data_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0884 -0.7554  0.0447  0.7518  9.0322
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

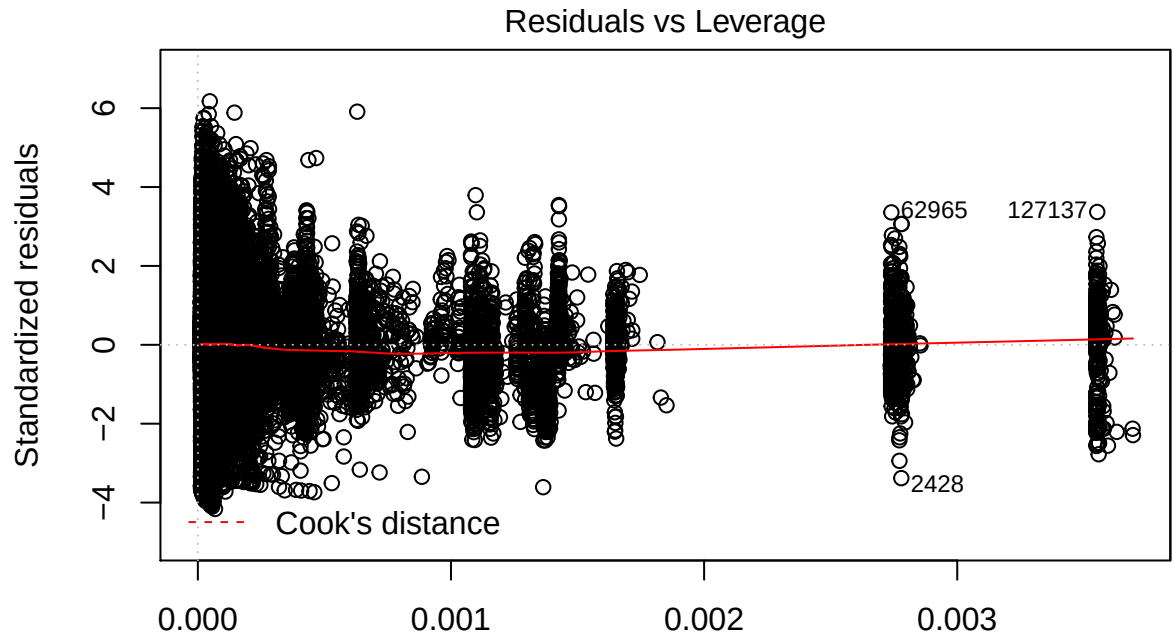
```
## (Intercept)            6.001e+00  1.771e-02 338.923  < 2e-16 ***
## Other.GHG              1.565e-04  2.805e-05   5.580 2.41e-08 ***
## HFC                    3.030e-07  2.749e-08  11.020  < 2e-16 ***
## Biogenic.CO2          -2.464e-07  1.062e-08 -23.198  < 2e-16 ***
## Population             4.280e-07  6.172e-09  69.346  < 2e-16 ***
## PFC                   -1.059e-06  6.778e-08 -15.629  < 2e-16 ***
## HFE                   -4.894e-06  9.196e-07  -5.322 1.03e-07 ***
## Stationary.Combustion  8.244e-09  1.942e-09   4.244 2.20e-05 ***
## pp_consumed_MMBtu     -8.627e-09  5.264e-10 -16.390  < 2e-16 ***
## poly(Temperature, 2)1  1.159e+02  1.490e+00  77.751  < 2e-16 ***
## poly(Temperature, 2)2  7.088e+01  1.471e+00  48.203  < 2e-16 ***
## SF6                    3.991e-06  2.509e-07  15.907  < 2e-16 ***
## Income                 3.087e-06  3.843e-07   8.032 9.63e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.463 on 137816 degrees of freedom
## Multiple R-squared:  0.1098, Adjusted R-squared:  0.1097
## F-statistic:  1416 on 12 and 137816 DF,  p-value: < 2.2e-16
```

```
plot(lmod.select)
```

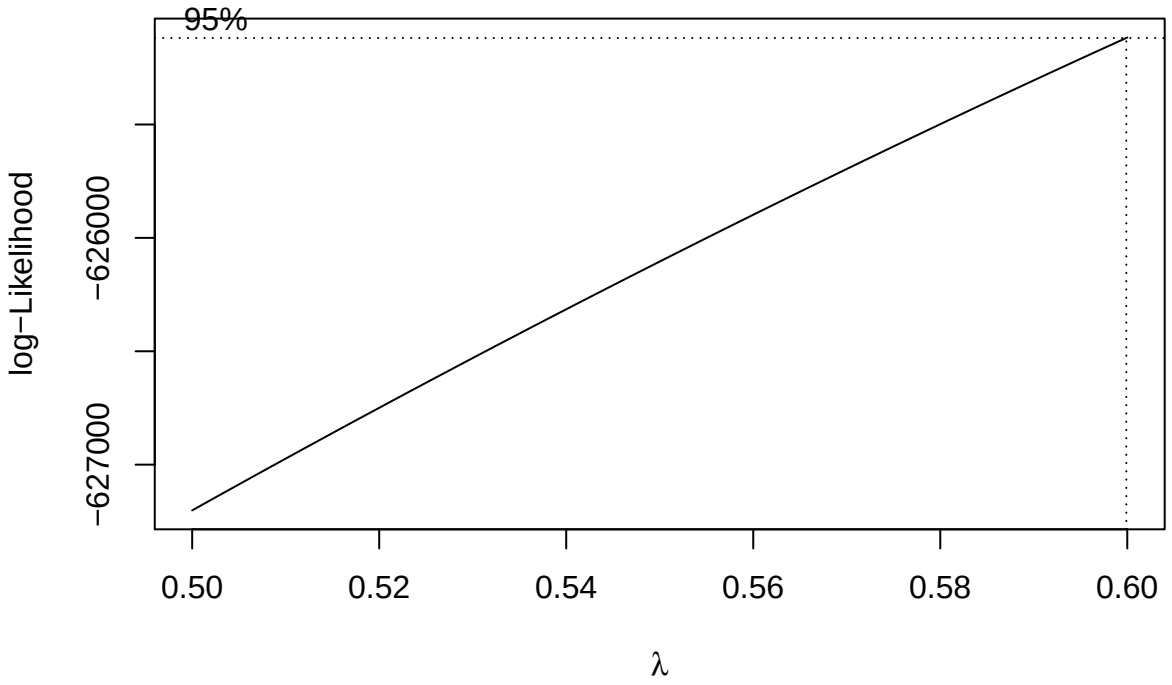



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm((AQI + 1)^0.5 ~ Other.GHG + HFC + Biogenic.CO2 + Population + PFC + HFE ..


Scale−Location
√|Standardized residuals|
Fitted values
lm((AQI + 1)^0.5 ~ Other.GHG + HFC + Biogenic.CO2 + Population + PFC + HFE ..




```
bx <- boxcox(lmod.select, plotit=T, lambda=seq(0.5, 0.6, by=0.01))
```




```
#Getting best boxcox parameter -- lambda ~ 0.5
lambda <- bx$x[which.max(bx$y)]


rlmod <- rlm(AQI^0.5 ~ Other.GHG + HFC +  Biogenic.CO2 + Population + PFC + HFE + Stationary.Combustion
summary(rlmod)
```

```
##
## Call: rlm(formula = AQI^0.5 ~ Other.GHG + HFC + Biogenic.CO2 + Population +
##     PFC + HFE + Stationary.Combustion + pp_consumed_MMBtu + Temperature +
##     SF6 + Income, data = data)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5104 -0.7584  0.0393  0.7279 14.2029
##
## Coefficients:
##                        Value    Std. Error t value
## (Intercept)              5.6635   0.0167   338.7490
## Other.GHG                0.0001   0.0000     6.0972
## HFC                      0.0000   0.0000    10.9057
## Biogenic.CO2             0.0000   0.0000   -29.8741
## Population               0.0000   0.0000    71.3555
## PFC                      0.0000   0.0000   -18.6792
## HFE                      0.0000   0.0000    -2.8018
## Stationary.Combustion    0.0000   0.0000     2.1069
## pp_consumed_MMBtu        0.0000   0.0000   -12.0061
## Temperature              0.0234   0.0003    72.0555
## SF6                      0.0000   0.0000    17.5063
## Income                   0.0000   0.0000     1.2440
##
## Residual standard error: 1.099 on 137822 degrees of freedom
```

```
## Generalized least squares fit by REML
##   Model: AQI^0.5 ~ Other.GHG + HFC + Biogenic.CO2 + Population + PFC +      HFE + Stationary.Combusti
##   Data: na.omit(data)
##        AIC      BIC    logLik
##   22444.98 22538.98 -11208.49
##
## Variance function:
##  Structure: Power of variance covariate
##  Formula: ~fitted(.)
##  Parameter estimates:
##   power
## 1.52393
##
## Coefficients:
##                            Value  Std.Error   t-value p-value
## (Intercept)             7.987760 0.13294190  60.08459  0.0000
## Other.GHG              -0.023723 0.01028409  -2.30677  0.0211
## HFC                     0.000063 0.00001350   4.64986  0.0000
## Biogenic.CO2           -0.000001 0.00000006  -8.88471  0.0000
## Population              0.000000 0.00000001  20.85923  0.0000
## PFC                    -0.000010 0.00000234  -4.45382  0.0000
## HFE                    -0.000762 0.00026262  -2.90037  0.0037
## Stationary.Combustion   0.000000 0.00000001  -0.84817  0.3964
## pp_consumed_MMBtu       0.000000 0.00000000  -9.30771  0.0000
## Temperature             0.025449 0.00179190  14.20242  0.0000
## SF6                     0.000005 0.00000181   2.97430  0.0029
## Income                 -0.000030 0.00000278 -10.85314  0.0000
##
##  Correlation:
```

```
##                       (Intr) Ot.GHG HFC    Bg.CO2 Popltn PFC    HFE
## Other.GHG             -0.183
## HFC                   -0.231  0.802
## Biogenic.CO2          -0.085  0.123  0.157
## Population             0.248 -0.105 -0.220 -0.292
## PFC                    0.233 -0.892 -0.932 -0.128  0.162
## HFE                   -0.053  0.552  0.395 -0.003  0.094 -0.552
## Stationary.Combustion  0.058  0.021  0.126  0.131 -0.495 -0.033 -0.066
## pp_consumed_MMBtu     -0.155 -0.027 -0.133 -0.197 -0.160  0.058  0.035
## Temperature           -0.276 -0.097 -0.141 -0.023 -0.079  0.117  0.044
## SF6                    0.146 -0.113 -0.302  0.136 -0.012  0.135 -0.122
## Income                -0.951  0.230  0.320  0.058 -0.310 -0.309  0.040
##                       Sttn.C p__MMB Tmprtr SF6
## Other.GHG
## HFC
## Biogenic.CO2
## Population
## PFC
## HFE
## Stationary.Combustion
## pp_consumed_MMBtu     -0.228
## Temperature            0.057 -0.167
## SF6                   -0.099  0.070  0.118
## Income                -0.047  0.145  0.048 -0.247
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.3243242 -0.6807286 -0.1994687  0.5107010  7.0381029
##
## Residual standard error: 0.07119146
## Degrees of freedom: 6103 total; 6091 residual
```

7 - Final Model (Ishika)

Based on the diagnostics and transformation results (as shown above), we built the final models for each defining parameter by transforming the response as square root. We used the following set of factors to train the model, which were selected based on the results of multicollinearity tests in the diagnostics section - *nitrogen trifluoride, other greenhouse gases, hydrofluorocarbons, biogenic CO2, population, perfluorinated chemicals, hexafluorethane, stationary combustion, industrial power consumption, Temperature, methane, sulfur hexafluoride, and per capita income.*

The R-squared value of the full final model comes out to be 9.5%. All the predictors are statistically significant in this model as shown in the model summary below.

```
lambda <- 0.5
lmod.T <- lm(formula = AQI ^ lambda ~ Other.GHG + HFC +  Biogenic.CO2 + Population + PFC + HFE + Statio

summary(lmod.T)
```

```
##
## Call:
## lm(formula = AQI^lambda ~ Other.GHG + HFC + Biogenic.CO2 + Population +
##     PFC + HFE + Stationary.Combustion + pp_consumed_MMBtu + Temperature +
##     SF6 + Income, data = data_mod)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5817 -0.7588  0.0541  0.7556  9.1872
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.444e+00  2.008e-02 271.067  < 2e-16 ***
## Other.GHG              1.532e-04  2.917e-05   5.252 1.50e-07 ***
## HFC                    3.056e-07  2.859e-08  10.691  < 2e-16 ***
## Biogenic.CO2          -2.791e-07  1.103e-08 -25.313  < 2e-16 ***
## Population             4.373e-07  6.418e-09  68.124  < 2e-16 ***
## PFC                   -1.068e-06  7.049e-08 -15.153  < 2e-16 ***
## HFE                   -5.329e-06  9.563e-07  -5.572 2.52e-08 ***
## Stationary.Combustion  7.361e-09  2.020e-09   3.645 0.000268 ***
## pp_consumed_MMBtu     -7.550e-09  5.465e-10 -13.814  < 2e-16 ***
## Temperature            2.976e-02  3.903e-04  76.244  < 2e-16 ***
## SF6                    4.261e-06  2.609e-07  16.331  < 2e-16 ***
## Income                 2.168e-06  3.988e-07   5.436 5.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.521 on 137817 degrees of freedom
## Multiple R-squared:  0.09343,    Adjusted R-squared:  0.09336
## F-statistic:  1291 on 11 and 137817 DF,  p-value: < 2.2e-16
```

Similarly, we built the final component models for each defining parameter with response transformation and selected factors. We are only showing the results here for the NO2 model, for which the final R-squared value came out to be 20.5%.

```
lambda <- 0.5
lmod.T.NO2 <- lm(formula = AQI ^ lambda ~ Other.GHG + HFC +  Biogenic.CO2 + Population + PFC + HFE + Sta

summary(lmod.T.NO2)
```

```
##
## Call:
## lm(formula = AQI^lambda ~ Other.GHG + HFC + Biogenic.CO2 + Population +
##     PFC + HFE + Stationary.Combustion + pp_consumed_MMBtu + Temperature +
##     SF6 + Income, data = data_mod_NO2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6282 -0.9608 -0.0290  0.9684  7.0773
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.849e+00  3.090e-02  92.203  < 2e-16 ***
## Other.GHG             -9.195e-03  3.151e-03  -2.918  0.00352 **
## HFC                    4.377e-07  2.545e-08  17.197  < 2e-16 ***
## Biogenic.CO2          -2.424e-07  2.369e-08 -10.233  < 2e-16 ***
## Population             4.924e-07  6.596e-09  74.649  < 2e-16 ***
## PFC                    8.655e-07  1.248e-07   6.934 4.14e-12 ***
## HFE                   -1.649e-05  6.091e-05  -0.271  0.78655
## Stationary.Combustion -2.187e-08  1.923e-09 -11.373  < 2e-16 ***
```

```
## pp_consumed_MMBtu      5.198e-09  6.137e-10   8.470  < 2e-16 ***
## Temperature           -1.779e-02  5.898e-04 -30.159  < 2e-16 ***
## SF6                    6.415e-08  7.932e-07   0.081  0.93555
## Income                 2.032e-05  6.001e-07  33.857  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.35 on 48444 degrees of freedom
##   (89373 observations deleted due to missingness)
## Multiple R-squared:  0.2053, Adjusted R-squared:  0.2051
## F-statistic:  1138 on 11 and 48444 DF,  p-value: < 2.2e-16
```

8 - Future Research (Ishika)

9

A) Additional predictors

Future research work can include additional predictors in the model, since only about 20% of the variation in the model is explained by the current set of predictors. Some additional factors that can be included are - *more weather factors (wind speed, humidity and precipitation), geographical data like elevation or use city-level data rather than county-level data.*

B) Additional scope

We can further investigate air quality with more data collected across the globe so that we have more informational dataset. We can also run the analysis over a larger time interval (current analysis is for an year's data). We couldn't process more than an year's data in R, so we decided to build the model for lesser data.

C) Nonlinear models

We tried running the general additive model. Future work could involve expanding on this, since we think it would produce promising results. >>>>>>> 10b54100cd32fddef01cd331553330169e1d8300:Final_Report/Report_Tess_Secti
4.Rmd