The Effects of Air Quality on U.S. Life Expectancy

Michael T. Moen

2024-12-09

Contents

1 Abstract

This report investigates the relationship between air quality and mortality rates across U.S. counties, with a focus on chronic respiratory and cardiovascular diseases. Using data from the Institute for Health Metrics and Evaluation (IHME) and the Environmental Protection Agency (EPA), multiple linear regression models were constructed to examine the effects of ground-level ozone (O_3) and fine particulate matter ($PM_{2.5}$) on mortality outcomes. Initial models show significant but weak associations between air quality indicators and mortality rates, suggesting potential confounding factors. Incorporating demographic variables such as smoking rates, poverty levels, and median household income significantly improved model performance, as evidenced by increases in adjusted R^2. Results indicate that demographic factors play a larger role in explaining mortality rate variations than air quality data alone, though O_3 remains a significant predictor for cardiovascular disease mortality.

2 Background

Ground-level ozone and fine particulate matter are two of the most significant air pollutants that impact public health. This analysis explores this issue by looking at the relationship of these two pollutants with the life expectancy of U.S. counties due to various mortality factors.

2.1 Data Sources

The data used in this analysis was collected from two main sources. The mortality data comes from the Institute for Health Metrics and Evaluation (IHME) and contains data for various mortality factors, with the most recent data being from 2014. Since this data is derived from death registration data, it contains complete data for every county in the United States (IHME, 2016).

The air quality data in this analysis was published by the Environment Protection Agency (EPA) and covers the presence of various pollutants recorded in the United States by county in 2023. Since this methodology only allows the EPA to monitor the pollutants in counties with the appropriate facilities to record such data, it is much more sparse. Only about one-third of counties have any recorded data, and the counties that do have recorded data typically only monitor select pollutants. However, the two pollutants that have the most significant impact on public health, ground-level ozone and fine particulate matter, are the most widely recorded pollutants. Because of this, these pollutants are the focus of this analysis (EPA, 2024).

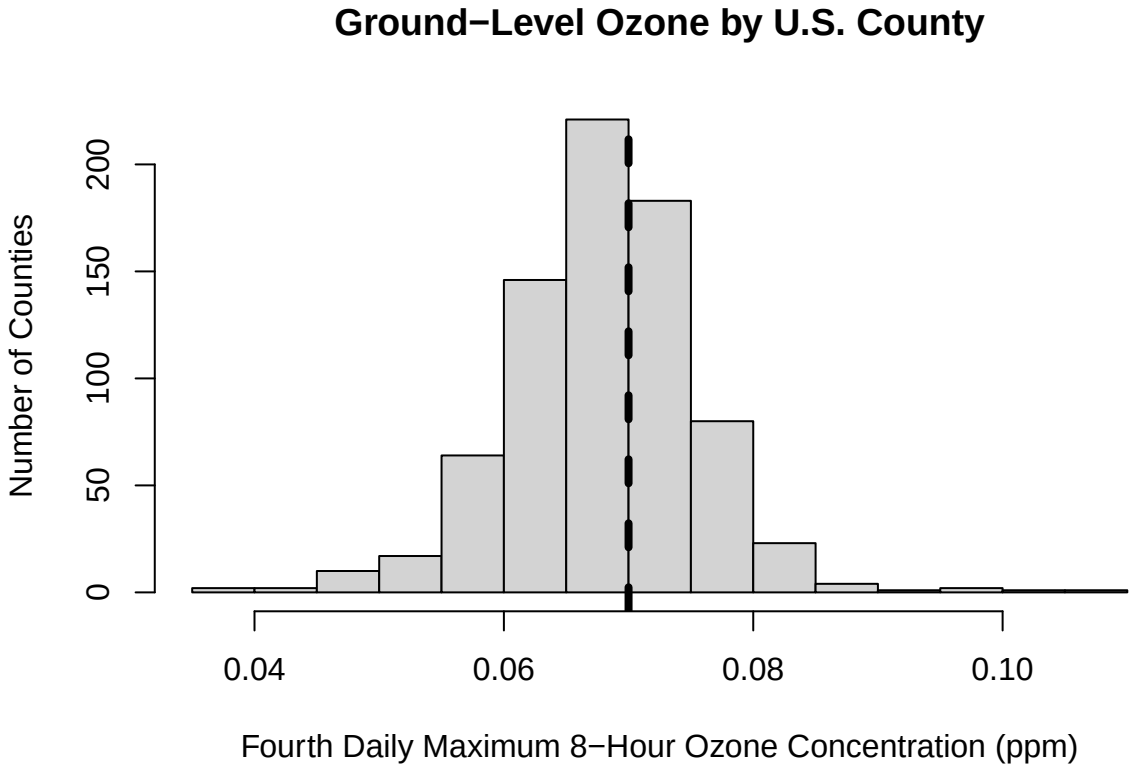
A variety of other demographic and health-related factors are analyzed in later sections of this report. This data comes from an online Kaggle data set, which contains county-level data as of 2019. This data comes from a variety of public sources, including the U.S. Census Bureau, Bureau of Labor Statistics, and the Center for Disease Control. This data contains no missing values to consider (He, 2023).

2.2 Ground-Level Ozone

Ground-level ozone (O_3) is a pollutant that has been shown to have negative effects on the health of individuals, particularly those with respiratory issues such as asthma.

The histogram shows the severity of the O_3 recorded by the EPA in 2023 for each county. Note that the EPA only recorded this data in 761 of the 3143 counties and county-equivalents (parishes in Louisiana, boroughs in Alaska, etc.) in the United States.

One common way to measure the air quality in an area is the fourth daily maximum 8-hour concentration. This figure is calculated by first considering hourly measurements of the pollutant at the measurement sites. Then, the 8-hour period with the highest average concentration of ozone is calculated to find the daily maximum 8-hour ozone concentration. At the end of the year, every daily maximum 8-hour ozone concentration is ranked, and the fourth-highest value is considered for the yearly rating in order to limit the effect of outliers that are not indicative of long-term environmental effects. The EPA's National Ambient Air Quality Standards (NAAQS) uses this metric and sets the threshold for the ground-level ozone at 0.07 parts per million (ppm), which is represented by the vertical line in the histogram. The high number of counties above this threshold show that this pollutant has an unhealthy presence in many places in the United States.



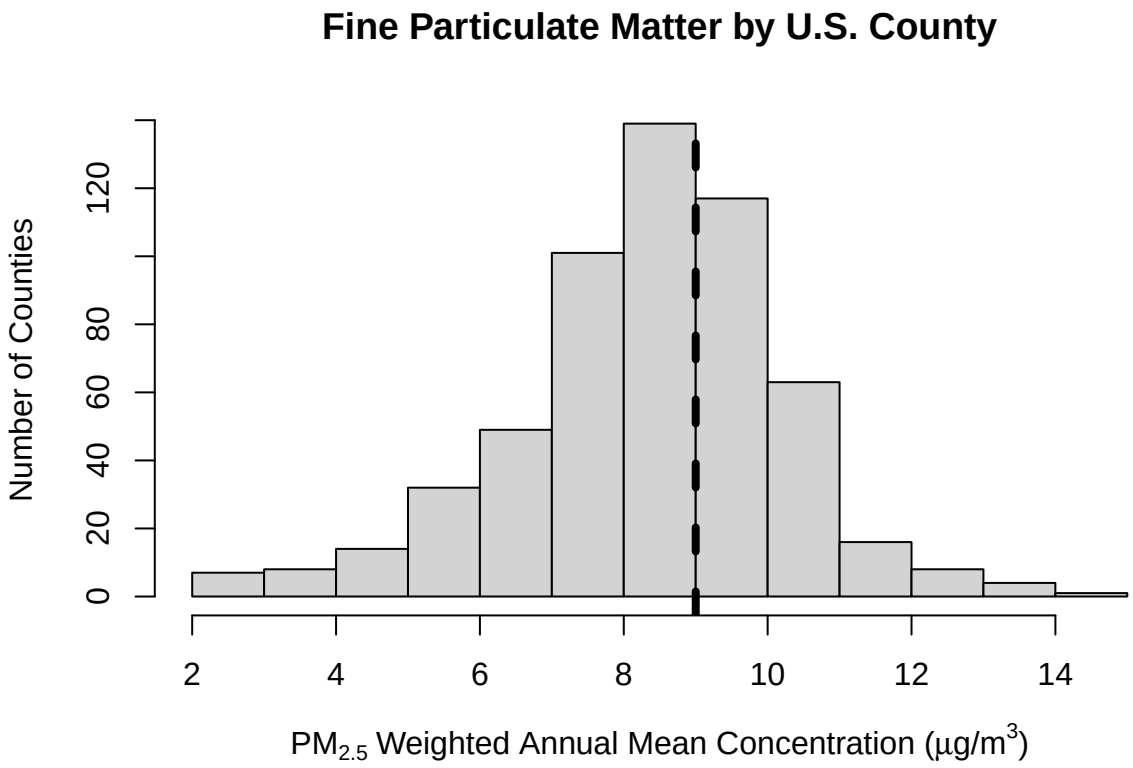
2.3 Fine Particulate Matter

Fine particulate matter ($PM_{2.5}$) refers to particles with a diameter of 2.5 micrometers or smaller. These fine particles can be made up of anything, with some of the notable sources of $PM_{2.5}$ including

vehicle emissions, industrial activity, and wildfires. $PM_{2.5}$ is small enough to penetrate deep into the lungs and into the bloodstream, posing a serious health concern, particularly among those with pre-existing respiratory issues.

The histogram shows the severity of the $PM_{2.5}$ recorded by the EPA in 2023 for each county. Note that the EPA only recorded this data in 559 of the 3143 counties and county-equivalents.

The $PM_{2.5}$ weighted annual mean concentration considers the population density near various reporting sites to more accurately reflect the conditions experienced by the population. The NAAQS sets the threshold for $PM_{2.5}$ weighted annual mean concentration at 9.0 $\mu g/m^3$ (micrometers per cubic meter of air), which is represented by the vertical line in the histogram. The high number of counties above this threshold show that this pollutant has an unhealthy presence in many places in the United States.



2.4 Data Preparation

The three data sets used in this analysis were found online as CSV and XLSX files. These file were downloaded and joined using Python's pandas library using the the county FIPS codes (see Appendix A for details). This merged data set was then imported into R for analysis.

Since the air quality data does not cover every U.S. county, only the cases that contain data for both O_3 and $PM_{2.5}$ are used in this analysis. A total of 444 U.S. counties contain data for both pollutants, making this the total number of observations considered in this analysis.

3 Predicting Mortality Outcomes from Air Quality Data

Air quality is a critical determinant of public health, with pollutants such as O_3 and $PM_{2.5}$ linked to various health outcomes. This section leverages county-level air quality and mortality data to investigate the relationship between these environmental factors and mortality outcomes across different causes of death. Using multiple linear regression, we explore the predictive power of the air quality indicators for mortality rates, focusing on chronic respiratory diseases and cardiovascular diseases.

The models presented in this section evaluate both the statistical significance and practical implications of the relationships between air quality indicators and mortality outcomes. By analyzing the coefficients, R^2_{adj} values, and p-values, we assess the strength and direction of these associations. Furthermore, the findings highlight potential confounding variables and areas for further investigation.

Subsections below discuss findings for chronic respiratory disease mortality and cardiovascular disease mortality in detail, referencing full model summaries provided in the appendices.

3.1 Chronic Respiratory Disease Factors

The first question that we seek to answer in this analysis is how well county-level air quality indicators can predict the county-level mortality rates due to chronic respiratory disease. To answer this question, we create a multiple linear regression model with the mortality rate due to cardiovascular factors as the target and O_3 and $PM_{2.5}$ as the predictors. The full summary of this model can be found in Appendix B.

The low p-value on this model's F-statistic shows that there is a statistically significant relationship between the air quality indicators and the mortality rate due to chronic respiratory diseases. However, the low R^2_{adj} value of 0.01358 indicates that the predictors are a poor fit, explaining very little of the variation in the mortality rates.

The most suspicious aspect of this model's summary is the negative coefficient associated with the O_3. This relationship is counter-intuitive, since an increase in the pollutant is expected to worsen health outcomes. However, this model predicts that the presence of the pollutant lowers the mortality rate in a county. This indicates that there may be issues with confounding variables.

3.2 Cardiovascular Disease Factors

Air quality has been shown to have an impact on cardiovascular health. This section examines this relationship by creating a multiple linear regression model that uses O_3 and $PM_{2.5}$ to predict mortality due to cardiovascular disease (CVD). The full summary of this model is available in Appendix C.

This model's very low p-value for the F-statistic indicates that there is a statistically significant relationship between the air quality and mortality due to CVD. The R^2_{adj} value of 0.05078 indicates that the O_3 and $PM_{2.5}$ predict the mortality rate due to CVD more accurately than the mortality rate to chronic respiratory disease.

4 Predicting Mortality Outcomes with Air Quality and Demographic Data

Some suspicious outcomes in the previous section suggest that there are confounding factors influencing the mortality rates investigated. This section aims to examine this issue by using demographic data, including poverty, unemployment, median household income, and percentage of adult smokers, alongside the air quality data to further examine the relationship. Including these predictors provides a more complete understanding of the factors influencing mortality rates. As a result, we can improve the predictive power of our model and understand the relative importance of air quality as a predictor for the observed mortality categories.

The following sections once again examine the mortality rate due to chronic respiratory disease and CVD using multiple linear regression. Each of the models below is initially created using O_3, $PM_{2.5}$, percentage of adults who smoke, percent of the area in the county that is developed, median household income, poverty rate, unemployment rate, and uninsured rate as predictors. Then, variable selection is performed using backward elimination with BIC is performed to address issues with multicollinearity, since much of the demographic data is highly correlated. These resulting models are what are analyzed in the subsections below. For a further exploration of the initial predictors included in these models, see Appendix D.

4.1 Chronic Respiratory Disease Factors

The variable selection process for predicting mortality due to chronic respiratory disease reduces the model from the eight initial predictors to the following three: percentage of adults who smoke, percentage of area that is developed, and median household income. Notably, both of the air quality indicators included in the initial model are dropped in the variable selection. This suggests that both the O_3 and $PM_{2.5}$ observations at the county level have little effect on mortality due to chronic respiratory disease. See Appendix E for the full model summary.

The regression model using these three predictors has an R^2_{adj} value of 0.5036, indicating that the model accounts for 50.36% of the variation in the results. This is significantly higher than the R^2_{adj} of the model that only considers the air quality metrics, which has an R^2_{adj} value of 0.01358, further indicating that the mortality rate is better explained by the demographic factors than the air quality data.

The three predictors in this model have low p-values, indicating that they are each statistically significant in predicting the mortality rate due to chronic respiratory disease. The coefficients of the predictors indicate that higher rates of smoking increases this mortality rate, while higher rates of development and higher household incomes decrease the mortality rate.

4.2 Cardiovascular Disease Factors

The variable selection process for predicting mortality due to CVD reduces the model from the eight initial predictors to the following three: percentage of adults who smoke, poverty rate, and ground-level ozone. Notably, the O_3 predictor is retained in this variable selection. This suggests that both the ground-level ozone observations at the county level have have a noticeable effect on mortality due to CVD. See Appendix F for the full model summary.

The regression model using these three predictors has an R^2_{adj} value of 0.5616, indicating that the model accounts for 56.16% of the variation in the results. This is significantly higher than the R^2_{adj} of the model that only considers the air quality metrics, which has an R^2_{adj} value of 0.05078, further indicating that the mortality rate is better explained by the demographic factors than the air quality data.

The three predictors in this model have extremely low p-values, indicating that they are each statistically significant in predicting the mortality rate due to chronic respiratory disease. The coefficients of the predictors indicate that higher rates of smoking, poverty, and O_3 all increase this mortality rate.

5 Conclusion

This analysis shows that while O_3 and $PM_{2.5}$ exhibit statistically significant relationships with mortality outcomes, their contribution is overshadowed by demographic factors such as smoking prevalence, poverty, and median household income. Models incorporating these demographic predictors showed substantial improvements in explanatory power, with adjusted R^2 values exceeding 0.5 for both chronic respiratory and cardiovascular disease mortality rates. Therefore, we conclude that while O_3 and $PM_{2.5}$ have a statistically significant effect on mortality outcomes, they are not very useful in constructing a model to predict such outcomes.

Interestingly, O_3 emerged as a significant predictor in the CVD model, even after accounting for demographic variables, reinforcing its role as a critical air pollutant affecting public health. The absence of $PM_{2.5}$ as a retained predictor in any of the final models suggests that its impact may be confounded by other variables at the county level or less direct.

Major limitations in the approach used in this analysis arise from the limitations in the data. While the historical air quality data would provide deeper insights into the long-term affects of air quality on mortality, such data was not readily available. In addition to improving our understanding of the link between air quality and mortality, such data would allow for the construction of models to identify areas that will be impacted by such effects in the future. That being said, the results of the latter models indicate that other demographic and health-related factors do a much better job of predicting these outcomes and are easier to collect for all counties than air quality metrics, which require specific monitoring stations.

6 References

1. EPA. (2024). *Air Quality Statistics by County, 2023*. Retrieved from https://www.epa.gov/air-trends/air-quality-cities-and-counties on 2024-12-09.
2. He, Lawrence. (2023). *United States County Level Health Data*. Retrieved from https://www.kaggle.com/datasets/lawrencehe/county-level-health-data?resource=download on 2024-12-09.
3. IHME. (2016). *US county-level mortality*. Kaggle. Retrieved from https://www.kaggle.com/datasets/IHME/us-countylevel-mortality/data on 2024-12-09.

7 Appendices

7.1 Appendix A: Merging Data Sets with Python

The three data sets used in this analysis were joined using Python's pandas library. Note that some columns were dropped and renamed for ease of use, but this was omitted from this code block for brevity.

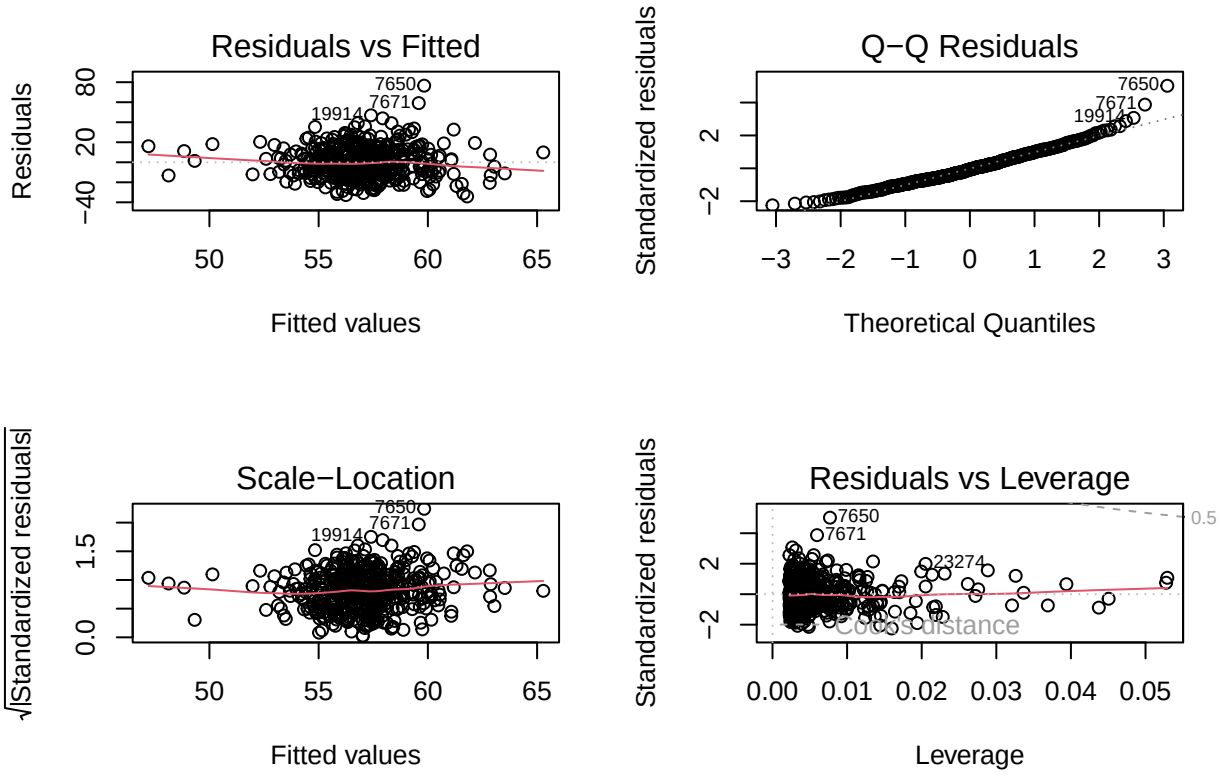
```
import pandas as pd

# Read and clean mortality data
mortality_df = pd.read_csv('mort.csv')
mortality_df = mortality_df.dropna(subset=['FIPS'])
mortality_df['FIPS'] = mortality_df['FIPS'].astype(int)

# Read and clean air quality data
aqi_df = pd.read_excel('ctyfactbook2023.xlsx',
                       sheet_name='County Factbook 2023',
                       skiprows=2)
aqi_df = aqi_df.dropna(subset=['County FIPS Code'])
aqi_df['County FIPS Code'] = aqi_df['County FIPS Code'].astype(int)

# Read demographic data
demographic_df = pd.read_csv('raw_data.csv')

# Join data sets on FIPS codes
merged_df = pd.merge(mortality_df, aqi_df,
                     left_on='FIPS', right_on='County FIPS Code', how='inner')
merged_df2 = pd.merge(trimmed_df, demographic_df,
                      left_on='FIPS', right_on='FIPS', how='inner')

# Export merged dataframe to CSV
merged_df2.to_csv('air_demographic.csv', index=False)
```

7.2 Appendix B: Chronic Respiratory Disease Mortality Model Summary

The first model examined in this analysis uses the ground-level ozone and $PM_{2.5}$ measurements to predict the mortality rates due to chronic respiratory disease. This simple multiple linear regression model is described by the summary below:

```
##
## Call:
## lm(formula = Mortality.Rate..2014. ~ O3.8.hr..ppm. + PM2.5.Wtd.AM..mu.g.m3.,
##     data = filtered_resp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.138 -10.455  -1.027   9.983  76.530
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               74.6883     6.3478  11.766  < 2e-16 ***
## O3.8.hr..ppm.           -273.5233   103.3434  -2.647  0.00842 **
## PM2.5.Wtd.AM..mu.g.m3.    0.1454     0.4375   0.332  0.73971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 441 degrees of freedom
## Multiple R-squared:  0.01803,    Adjusted R-squared:  0.01358
## F-statistic:  4.05 on 2 and 441 DF,  p-value: 0.01808
```

The diagnostic plots suggest that the linearity assumptions hold, and that there are no significant outliers or heavy tails affecting the model. Due to this, there are no issues with goodness-of-fit in the model.

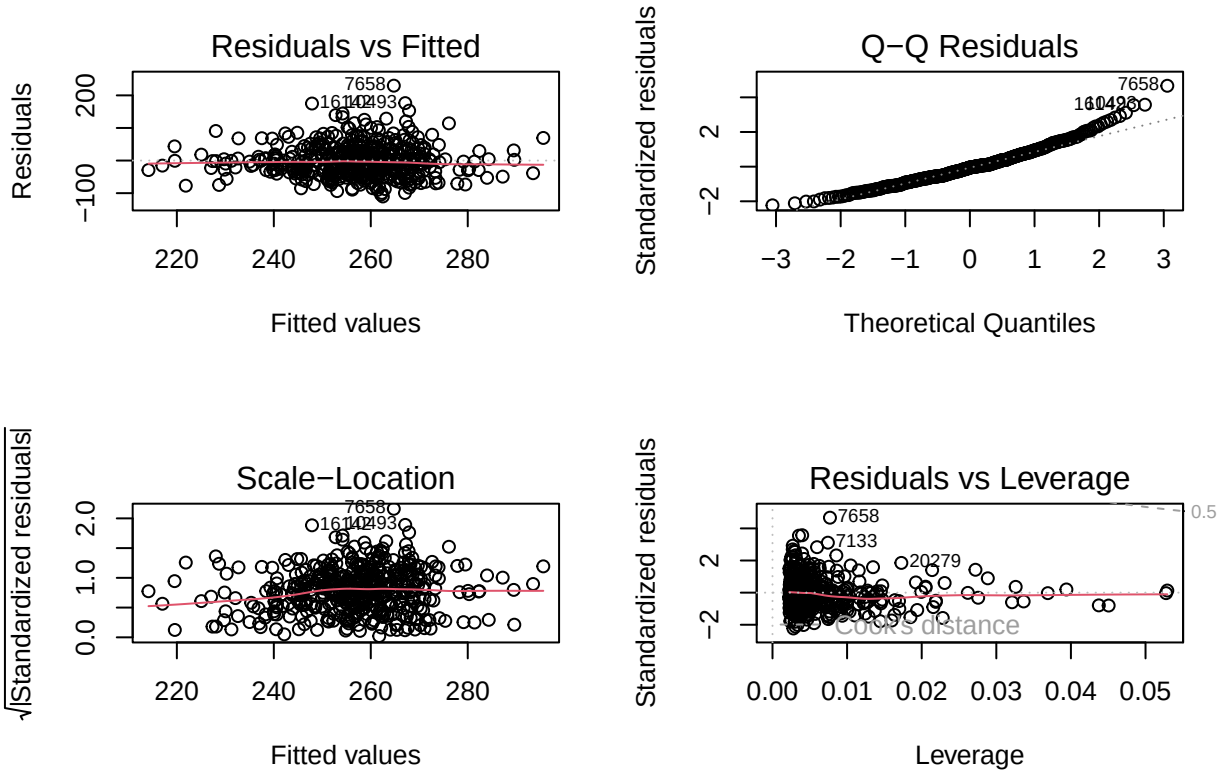

Residuals vs Fitted
Residuals
Fitted values
7650
7671
19914
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
23274
Leverage

7.3 Appendix C: CVD Mortality Model Summary

The second model examined in this analysis uses the O_3 and $PM_{2.5}$ measurements to predict the mortality rates due to CVD. This simple multiple linear regression model is described by the summary below:

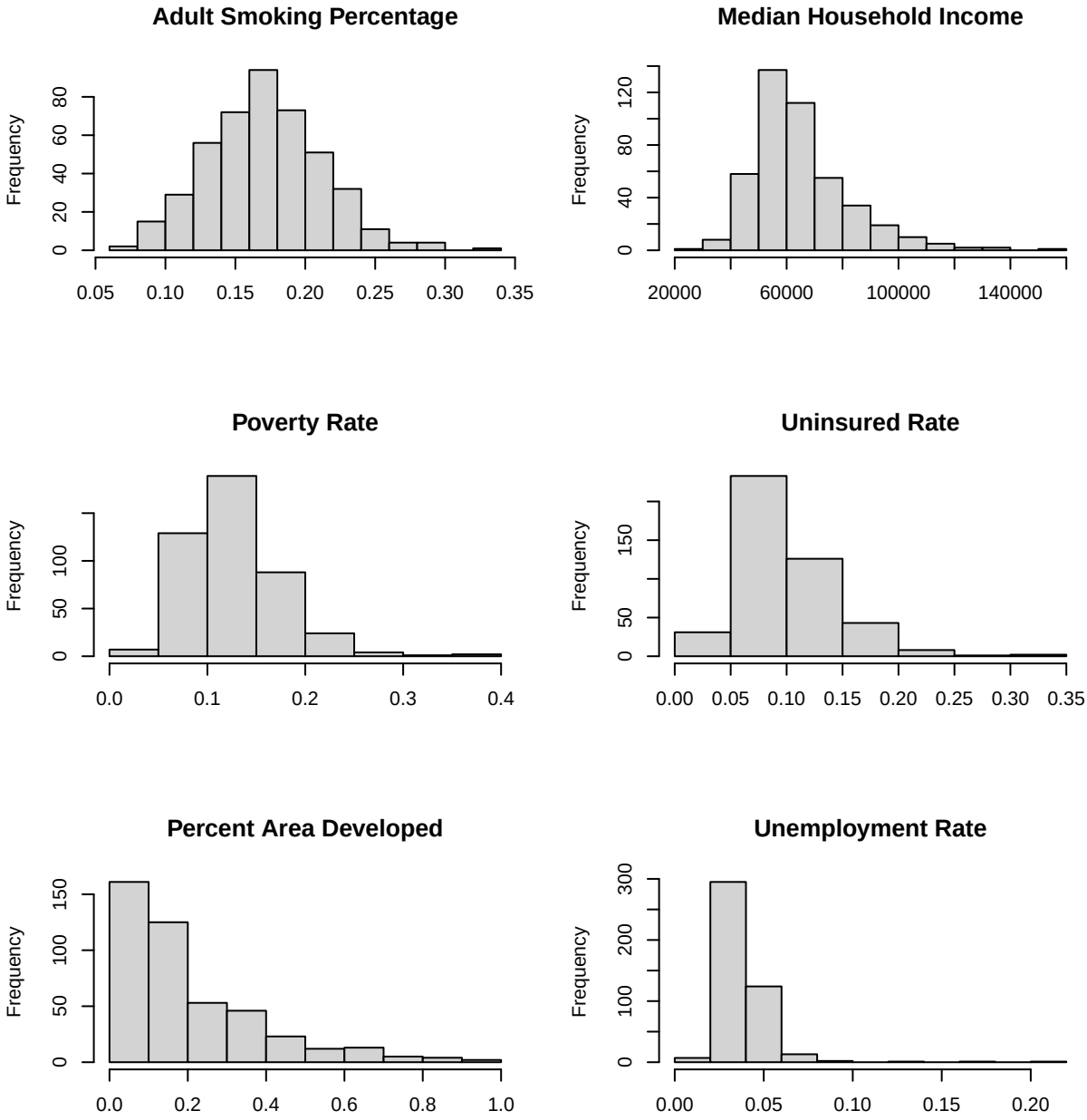
```
##
## Call:
## lm(formula = Mortality.Rate..2014. ~ O3.8.hr..ppm. + PM2.5.Wtd.AM..mu.g.m3.,
##     data = filtered_cvd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.129  -33.264   -2.883   27.225  229.768
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               234.574     20.499  11.443  < 2e-16 ***
## O3.8.hr..ppm.            -541.703    333.722  -1.623    0.105
## PM2.5.Wtd.AM..mu.g.m3.    7.061      1.413   4.998 8.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.42 on 441 degrees of freedom
## Multiple R-squared:  0.05506,    Adjusted R-squared:  0.05078
## F-statistic: 12.85 on 2 and 441 DF,  p-value: 3.77e-06
```

The diagnostic plots suggest that the linearity assumptions hold, and that there are no significant outliers or heavy tails affecting the model. Due to this, there are no issues with goodness-of-fit in the model.

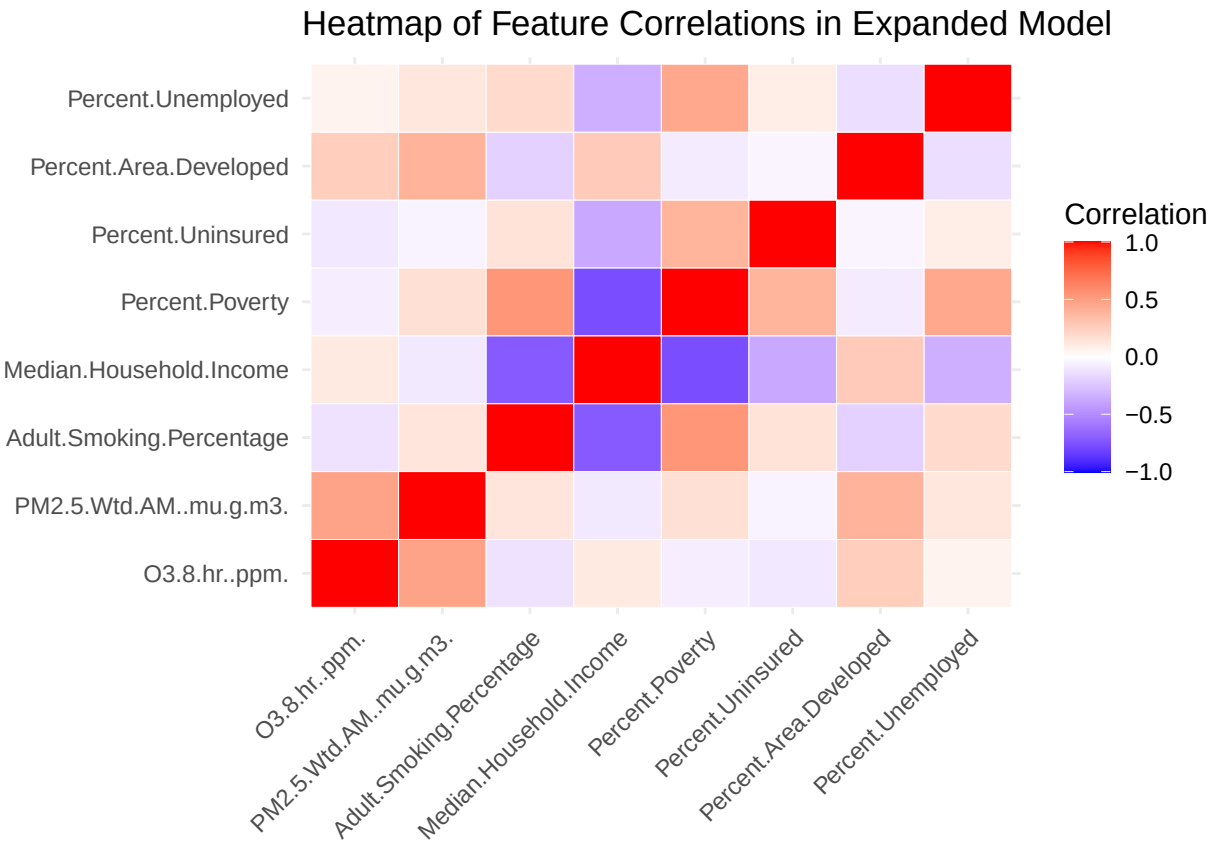

Residuals vs Fitted
Residuals
Fitted values
7658
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
7658
7133
20279

7.4 Appendix D: Demographic and Health Features

The following histograms show the distributions of the predictors considered in the expanded models of this analysis. Most of these distributions are right-skewed.



The following heat map shows the correlation between the predictors included in the initial expanded models.



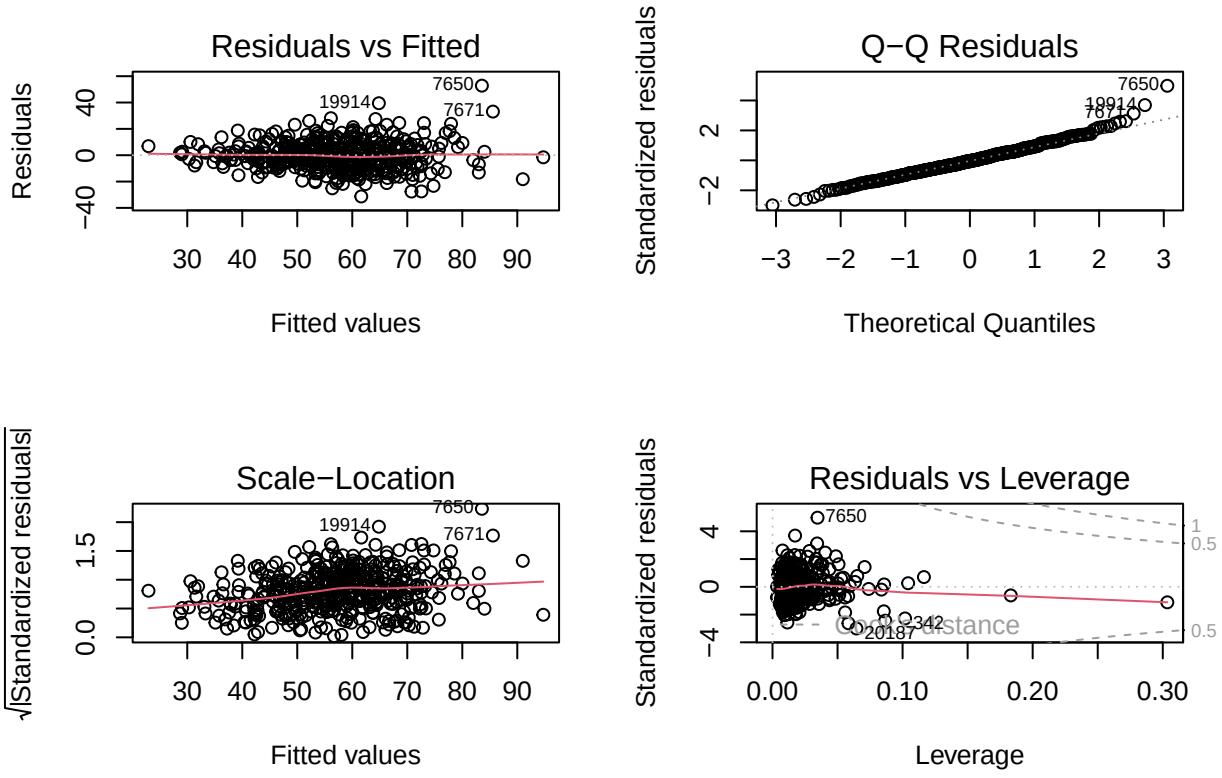
The correlation matrix shows that many of the demographic and health-related features, particularly `Median.Household.Income`, `Percent.Poverty`, `Adult.Smoking.Percentage` are strongly correlated. Additionally, the two air quality indicators, O_3 and $PM_{2.5}$ are strongly correlated with each other, but weakly correlated with the other features.

7.5 Appendix E: Chronic Respiratory Disease Demographic Model Summary

The following model predicts the 2014 mortality rate due to chronic respiratory diseases. First, a model containing several demographic, health, and air quality figures is constructed. Then, backward elimination using BIC is performed to remove predictors that are weak or whose effect on the mortality rate is also explained by other predictors. The resulting model's summary is given below:

```
##
## Call:
## lm(formula = Mortality.Rate..2014. ~ Adult.Smoking.Percentage +
##     Percent.Area.Developed + Median.Household.Income, data = filtered_resp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.186  -7.174  -0.770   6.178  50.732
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.022e+01  5.510e+00   5.485 6.99e-08 ***
## Adult.Smoking.Percentage   2.098e+02  1.791e+01  11.713  < 2e-16 ***
## Percent.Area.Developed    -1.423e+01  2.910e+00  -4.891 1.41e-06 ***
## Median.Household.Income   -1.121e-04  4.347e-05  -2.579   0.0102 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.86 on 440 degrees of freedom
## Multiple R-squared:  0.507,  Adjusted R-squared:  0.5036
## F-statistic: 150.8 on 3 and 440 DF,  p-value: < 2.2e-16
```

The diagnostic plots suggest that the linearity assumptions hold, and that there are no significant outliers or heavy tails affecting the model. Due to this, there are no issues with goodness-of-fit in the model.

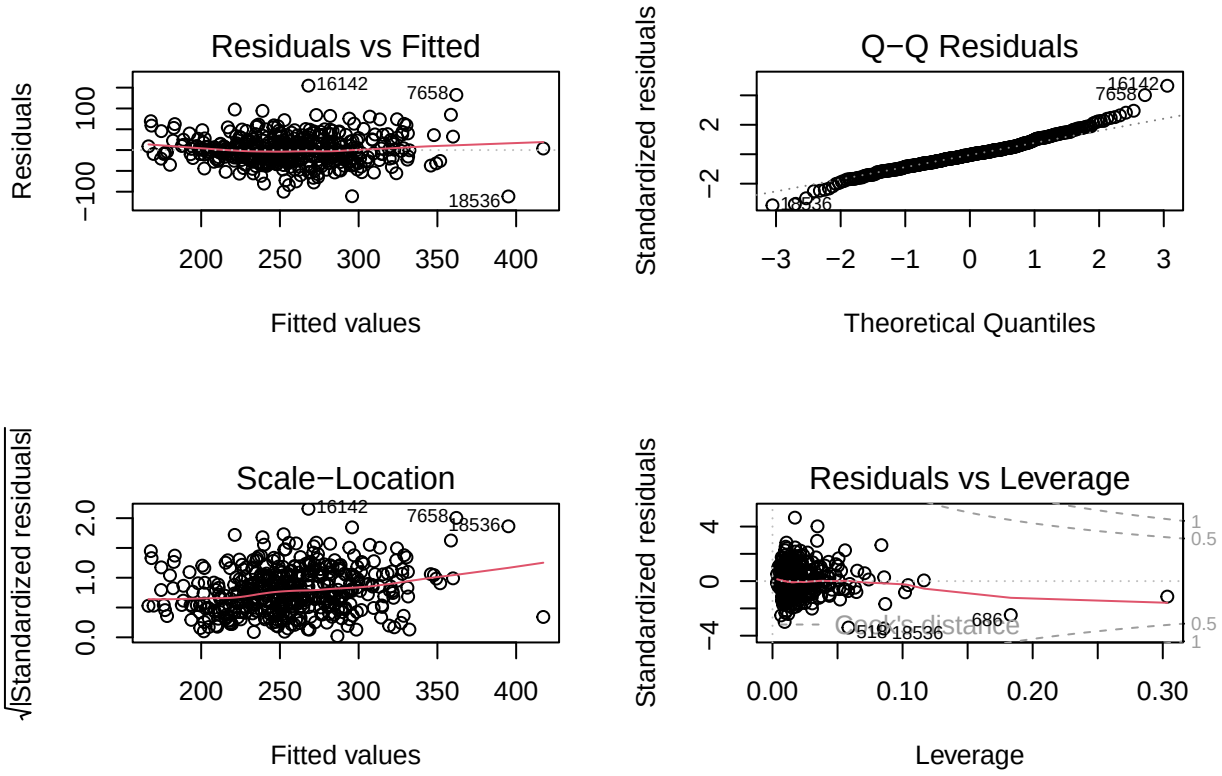

Residuals vs Fitted
Residuals
Fitted values
7650
19914
7671
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
20187
342

7.6 Appendix F: CVD Demographic Model Summary

The following model predicts the 2014 mortality rate due to CVD. First, a model containing several demographic, health, and air quality figures is constructed. Then, backward elimination using BIC is performed to remove predictors that are weak or whose effect on the mortality rate is also explained by other predictors. The resulting model's summary is given below:

```
##
## Call:
## lm(formula = Mortality.Rate..2014. ~ O3.8.hr..ppm. + Adult.Smoking.Percentage +
##     Percent.Poverty, data = filtered_cvd_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117.25  -21.14   -1.96   17.53  153.81
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  40.08      16.36   2.450   0.0147 *
## O3.8.hr..ppm.               810.77     201.23   4.029 6.60e-05 ***
## Adult.Smoking.Percentage    690.45      46.27  14.922  < 2e-16 ***
## Percent.Poverty             308.05      39.69   7.762 5.89e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.59 on 440 degrees of freedom
## Multiple R-squared:  0.5646, Adjusted R-squared:  0.5616
## F-statistic: 190.2 on 3 and 440 DF,  p-value: < 2.2e-16
```

The diagnostic plots suggest that the linearity assumptions hold, and that there are no significant outliers or heavy tails affecting the model. Due to this, there are no issues with goodness-of-fit in the model.


Residuals vs Fitted
Residuals
Fitted values
16142
7658
18536
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
686
518