

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

Michael T. Moen

2024-12-04

## Contents

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Data Sources . . . . .	3
2.2	Ground-Level Ozone . . . . .	3
2.3	Fine Particulate Matter . . . . .	4
2.4	Data Preparation . . . . .	5
<b>3</b>	<b>Predicting Mortality Outcomes from Air Quality Data</b>	<b>5</b>
3.1	Chronic Respiratory Disease Factors . . . . .	5
3.2	Cardiovascular Disease Factors . . . . .	6
<b>4</b>	<b>Predicting Mortality Outcomes with Air Quality and Demographic Data</b>	<b>6</b>
4.1	Chronic Respiratory Disease Factors . . . . .	6
4.2	Cardiovascular Disease Factors . . . . .	7
<b>5</b>	<b>Conclusion</b>	<b>7</b>
<b>6</b>	<b>Appendices</b>	<b>8</b>
6.1	Appendix A: Merging Datasets with Python . . . . .	8
6.2	Appendix B: Chronic Respiratory Disease Mortality Model Summary . . . . .	9
6.3	Appendix C: CVD Mortality Model Summary . . . . .	10
6.4	Appendix D: Chronic Respiratory Disease Demographic Model Summary . . . . .	11
6.5	Appendix E: CVD Demographic Model Summary . . . . .	12

# 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 and fine particulate matter (PM<sub>2.5</sub>) 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 ground-level ozone 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.

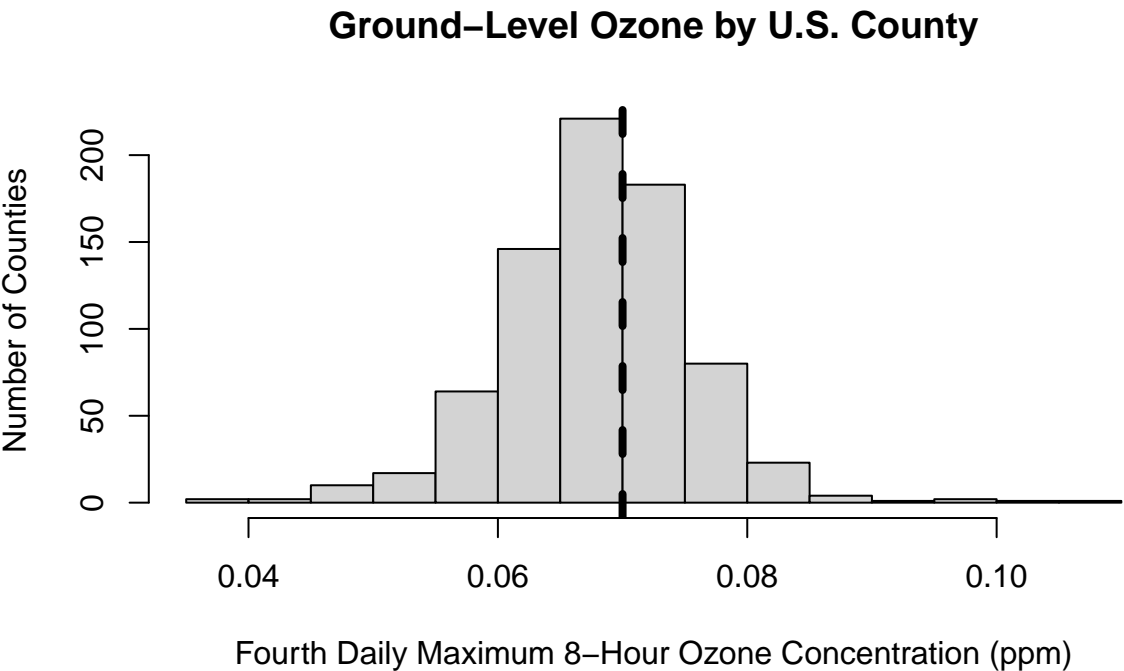
### 2.1 Data Sources

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 is not missing data for any counties.

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.

### 2.2 Ground-Level Ozone

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

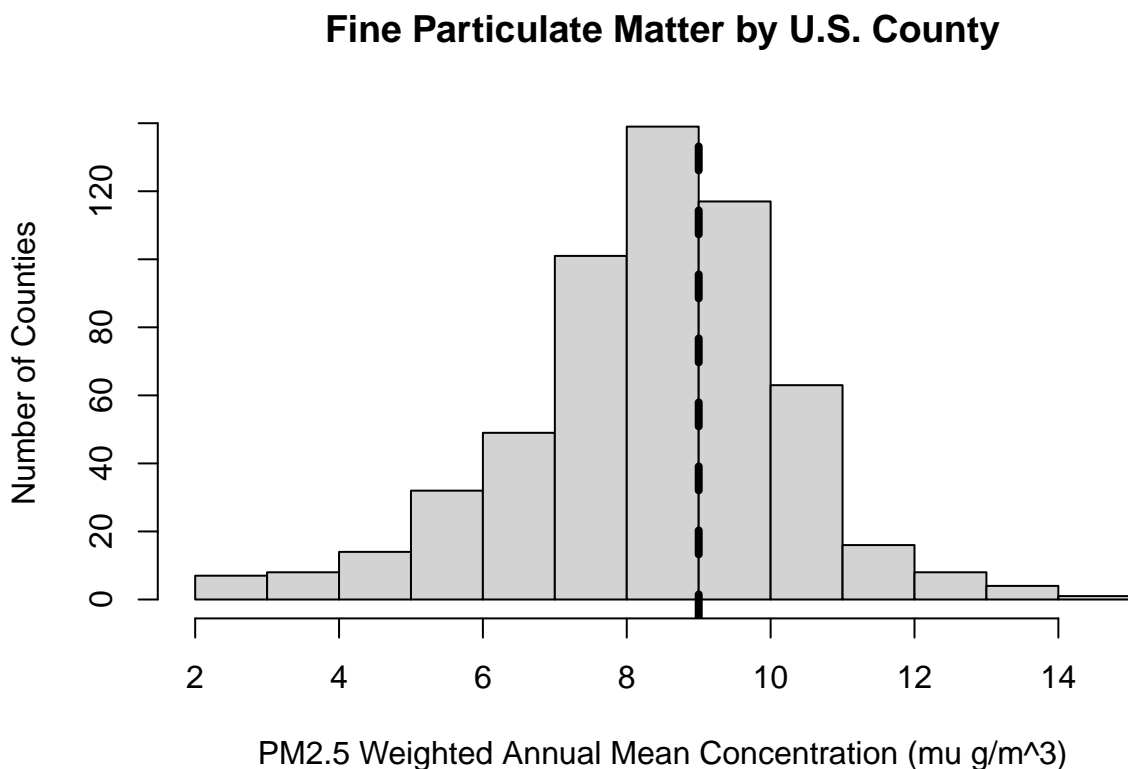


The histogram shows the severity of the ground-level ozone recorded by the EPA in 2023 for each county. Note that the EPA only recorded this data in 761 of the 3143 counties and county-equivalents (parishes in Louisiana, boroughs in Alaska, etc.) in the United States.

The EPA's National Ambient Air Quality Standards (NAAQS) 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 (PM2.5) 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 PM2.5 including vehicle emissions, industrial activity, and wildfires. Like ground-level ozone, fine particulate matter has been shown to have negative impact on health, particularly among those with pre-existing respiratory issues.



The histogram shows the severity of the PM2.5 recorded by the EPA in 2023 for each county. Note that the EPA only recorded this data in 562 of the 3143 counties and county-equivalents (parishes in Louisiana, boroughs in Alaska, etc.) in the United States.

The NAAQS sets the threshold for the fine particulate matter at 9.0  $\mu\text{g}/\text{m}^3$ , 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 two datasets used in this analysis were found online as CSV and XLSX files. These files were downloaded and joined using Python’s pandas library using the county FIPS codes (see Appendix A for details). This merged dataset 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 ground-level ozone and PM2.5 are used in this analysis.

## 3 Predicting Mortality Outcomes from Air Quality Data

Air quality is a critical determinant of public health, with pollutants such as ground-level ozone and PM2.5 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,  $R^2_{\text{adj}}$  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 ground-level ozone and PM2.5 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  $R^2_{\text{adj}}$  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 ground-level ozone. This relationship is counterintuitive, 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 ground-level ozone and PM2.5 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  $R^2_{\text{adj}}$  value of 0.05078 indicates that the ground-level ozone and PM2.5 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 ground-level ozone, PM2.5, 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.

### 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 ground-level ozone and PM2.5 observations at the county level have little effect on mortality due to chronic respiratory disease. See Appendix D for the full model summary.

The regression model using these three predictors has an  $R^2_{\text{adj}}$  value of 0.5036, indicating that the model accounts for 50.36% of the variation in the results. This is significantly higher than the  $R^2_{\text{adj}}$  of the model that only considers the air quality metrics, which has an  $R^2_{\text{adj}}$  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 ground-level ozone 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 E for the full model summary.

The regression model using these three predictors has an  $R^2_{\text{adj}}$  value of 0.5616, indicating that the model accounts for 56.16% of the variation in the results. This is significantly higher than the  $R^2_{\text{adj}}$  of the model that only considers the air quality metrics, which has an  $R^2_{\text{adj}}$  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 ground-level ozone all increase this mortality rate.

## 5 Conclusion

This analysis highlights the limited predictive power of air quality indicators alone in explaining variations in mortality rates at the county level. While ground-level ozone and PM2.5 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.

Interestingly, ground-level ozone emerged as a significant predictor in the cardiovascular disease model, even after accounting for demographic variables, reinforcing its role as a critical air pollutant affecting public health. The absence of PM2.5 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.

## 6 Appendices

### 6.1 Appendix A: Merging Datasets with Python

The two datasets 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)

# Join datasets on FIPS codes
merged_df = pd.merge(mortality_df, aqi_df,
                    left_on='FIPS', right_on='County FIPS Code', how='inner')

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



## 6.2 Appendix B: Chronic Respiratory Disease Mortality Model Summary

The first model examined in this analysis uses the ground-level ozone and PM2.5 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
```

### 6.3 Appendix C: CVD Mortality Model Summary

The second model examined in this analysis uses the ground-level ozone and PM2.5 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
```

## 6.4 Appendix D: Chronic Respiratory Disease Demographic Model Summary

```
##
## 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+00  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
```

## 6.5 Appendix E: CVD Demographic Model Summary

```
##
## Call:
## lm(formula = Mortality.Rate..2014. ~ 03.8.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)    40.0845    16.3636   2.450  0.0147 *
## 03.8.hr..ppm.   810.7725   201.2295   4.029 6.60e-05 ***
## Adult.Smoking.Percentage  6.9045     0.4627  14.922 < 2e-16 ***
## Percent.Poverty  308.0533    39.6852   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
```