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

Michael T. Moen

2024-11-27

Contents

1 Abstract

Write last once background, methodology, and results are all established.

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.

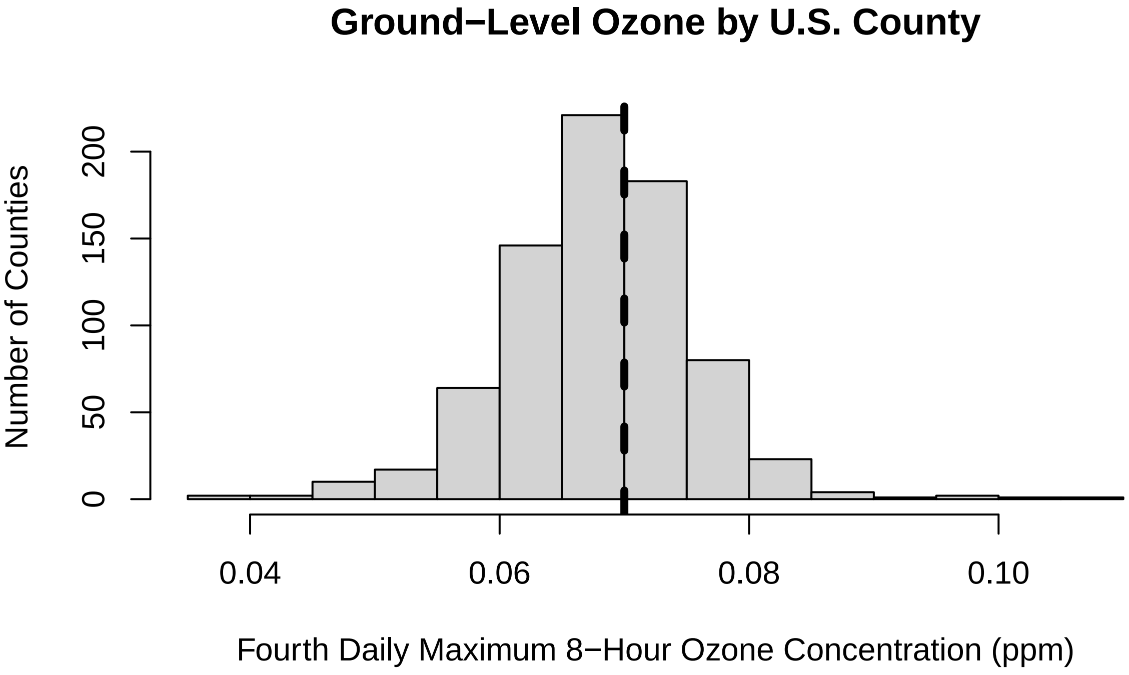
2.1 Data Sources

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.

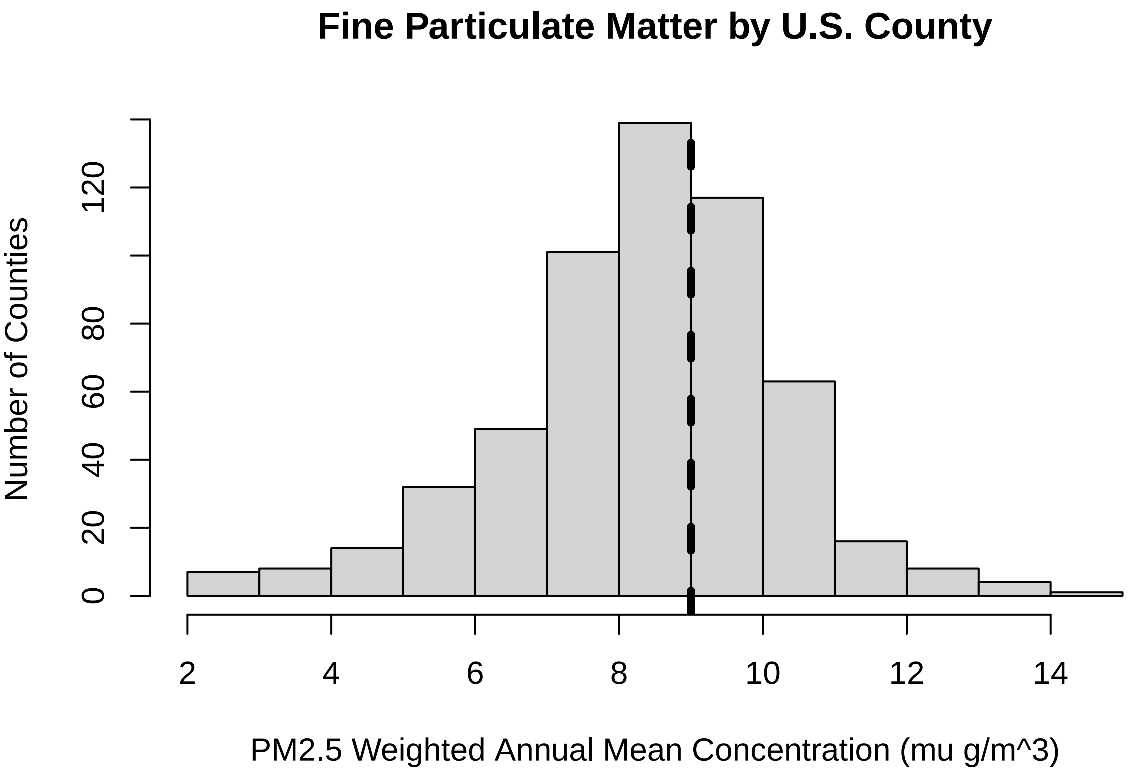


The histogram shows the severity of the ground-level ozone recorded by the EPA in 2023 for each county. Note that the EPA only recorded this data in 761 of the 3143 counties and county-equivalents (parishes in Louisiana, borroughs in Alaska, etc.) in the United States.

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, borroughs in Alaska, etc.) in the United States.

The NAAQS sets the threshold for the fine particulate matter at 9.0 $\mu g/m^3$, 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 file were downloaded and joined using Python's pandas library using the 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 Due to Chronic Respiratory Disease Using Air Quality

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 multilinear 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_{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.

4 Predicting Mortality Due to Cardiovascular Disease

Air quality has been shown to have an impact on cardiovascular health. This section examines this relationship by creating a multilinear 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.

5 Conclusion

Summarize findings and explain importance of findings. Write once finished with other sections

6 Appendices

6.1 Apendix 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.

```python
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 multilinear 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 multilinear 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
```

```
##
## Call:
## lm(formula = Mortality.Rate..2014. ~ O3.8.hr..ppm. + PM2.5.Wtd.AM..mu.g.m3. +
##     Adult.Smoking.Percentage + Percent.Area.Developed + Median.Household.Income +
##     Percent.Poverty + Percent.Unemployed + Percent.Uninsured,
##     data = filtered_resp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.314  -6.980  -0.577   6.233  52.795
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.201e+01  9.104e+00   3.516 0.000483 ***
## O3.8.hr..ppm.             7.624e+01  7.508e+01   1.015 0.310442
## PM2.5.Wtd.AM..mu.g.m3.   -7.531e-01  3.466e-01  -2.173 0.030331 *
## Adult.Smoking.Percentage  2.180e+00  1.839e-01  11.854  < 2e-16 ***
## Percent.Area.Developed   -1.164e+01  3.323e+00  -3.503 0.000507 ***
## Median.Household.Income  -1.350e-04  5.953e-05  -2.268 0.023840 *
## Percent.Poverty          -9.360e+00  1.851e+01  -0.506 0.613314
## Percent.Unemployed       -4.595e+01  3.765e+01  -1.220 0.222967
## Percent.Uninsured         1.780e+01  1.328e+01   1.340 0.180796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.79 on 435 degrees of freedom
## Multiple R-squared:  0.5186, Adjusted R-squared:  0.5098
## F-statistic: 58.58 on 8 and 435 DF,  p-value: < 2.2e-16
```

```
## Start:  AIC=2121.06
## Mortality.Rate..2014. ~ O3.8.hr..ppm. + PM2.5.Wtd.AM..mu.g.m3. +
##     Adult.Smoking.Percentage + Percent.Area.Developed + Median.Household.Income +
##     Percent.Poverty + Percent.Unemployed + Percent.Uninsured
##
##                            Df Sum of Sq   RSS    AIC
## - Percent.Poverty           1      29.8 50667 2119.3
## - O3.8.hr..ppm.             1     120.0 50757 2120.1
## - Percent.Unemployed        1     173.4 50810 2120.6
## - Percent.Uninsured         1     209.2 50846 2120.9
## <none>                                  50637 2121.1
## - PM2.5.Wtd.AM..mu.g.m3.    1     549.6 51186 2123.8
## - Median.Household.Income   1     598.6 51235 2124.3
## - Percent.Area.Developed    1    1428.8 52066 2131.4
## - Adult.Smoking.Percentage  1   16356.8 66994 2243.3
##
## Step:  AIC=2119.32
## Mortality.Rate..2014. ~ O3.8.hr..ppm. + PM2.5.Wtd.AM..mu.g.m3. +
##     Adult.Smoking.Percentage + Percent.Area.Developed + Median.Household.Income +
##     Percent.Unemployed + Percent.Uninsured
##
##                            Df Sum of Sq   RSS    AIC
## - O3.8.hr..ppm.             1     130.7 50797 2118.5
## - Percent.Uninsured         1     187.1 50854 2118.9
## <none>                                  50667 2119.3
## - Percent.Unemployed        1     251.1 50918 2119.5
## - PM2.5.Wtd.AM..mu.g.m3.    1     570.7 51237 2122.3
## - Median.Household.Income   1     669.6 51336 2123.2
## - Percent.Area.Developed    1    1561.8 52228 2130.8
## - Adult.Smoking.Percentage  1   16332.7 66999 2241.4
##
## Step:  AIC=2118.46
## Mortality.Rate..2014. ~ PM2.5.Wtd.AM..mu.g.m3. + Adult.Smoking.Percentage +
##     Percent.Area.Developed + Median.Household.Income + Percent.Unemployed +
##     Percent.Uninsured
##
##                            Df Sum of Sq   RSS    AIC
## - Percent.Uninsured         1     175.0 50972 2118.0
## <none>                                  50797 2118.5
## - Percent.Unemployed        1     231.0 51028 2118.5
## - PM2.5.Wtd.AM..mu.g.m3.    1     441.2 51239 2120.3
## - Median.Household.Income   1     651.7 51449 2122.1
## - Percent.Area.Developed    1    1534.9 52332 2129.7
## - Adult.Smoking.Percentage  1   16218.1 67015 2239.5
##
## Step:  AIC=2117.99
## Mortality.Rate..2014. ~ PM2.5.Wtd.AM..mu.g.m3. + Adult.Smoking.Percentage +
##     Percent.Area.Developed + Median.Household.Income + Percent.Unemployed
##
##                            Df Sum of Sq   RSS    AIC
## <none>                                 50972 2118.0
## - Percent.Unemployed        1     247.8 51220 2118.1
## - PM2.5.Wtd.AM..mu.g.m3.    1     513.0 51485 2120.4
## - Median.Household.Income   1    1118.9 52091 2125.6
## - Percent.Area.Developed    1    1441.7 52414 2128.4
## - Adult.Smoking.Percentage  1   16114.5 67087 2238.0
```

```
## 
## Call:
## lm(formula = Mortality.Rate..2014. ~ PM2.5.Wtd.AM..mu.g.m3. + 
##     Adult.Smoking.Percentage + Percent.Area.Developed + Median.Household.Income + 
##     Percent.Unemployed, data = filtered_resp_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -28.718  -6.985  -0.672   6.339  51.269 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               3.863e+01  6.303e+00   6.129 1.98e-09 ***
## PM2.5.Wtd.AM..mu.g.m3.   -6.486e-01  3.089e-01  -2.100 0.036339 *  
## Adult.Smoking.Percentage  2.117e+00  1.799e-01  11.767  < 2e-16 ***
## Percent.Area.Developed   -1.142e+01  3.243e+00  -3.520 0.000477 ***
## Median.Household.Income  -1.402e-04  4.521e-05  -3.101 0.002055 ** 
## Percent.Unemployed       -5.174e+01  3.546e+01  -1.459 0.145240    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.79 on 438 degrees of freedom
## Multiple R-squared:  0.5154, Adjusted R-squared:  0.5099 
## F-statistic: 93.17 on 5 and 438 DF,  p-value: < 2.2e-16
```

```
## 
## Call:
## lm(formula = Mortality.Rate..2014. ~ Adult.Smoking.Percentage + 
##     Percent.Area.Developed + Median.Household.Income + Percent.Unemployed, 
##     data = filtered_resp_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -29.629  -6.990  -0.554   6.234  51.529 
```

```
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.468e+01  6.040e+00   5.742 1.74e-08 ***
## Adult.Smoking.Percentage  2.071e+00  1.793e-01  11.554  < 2e-16 ***
## Percent.Area.Developed   -1.448e+01  2.907e+00  -4.983 9.01e-07 ***
## Median.Household.Income  -1.357e-04  4.534e-05  -2.993  0.00292 **
## Percent.Unemployed       -6.273e+01  3.521e+01  -1.782  0.07548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.83 on 439 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.5061
## F-statistic: 114.5 on 4 and 439 DF,  p-value: < 2.2e-16
```