Predicting COPD Prevalence and Health Spendings from Air Quality Data

Sarah Abouchleih, Kaydence Lin, and Rishpiath Satter

DS 3000: Foundations of Data Science

November 29, 2023

Abstract

This project explores the connections between air quality, healthcare spending, and the prevalence of Chronic Obstructive Pulmonary Disease (COPD) across U.S. states. We used air quality data obtained from wisevoter.com and the openaq API, along with COPD prevalence data and healthcare spending in each U.S. state from the CDC and GHDE. Our project leverages a multiple linear regression model to predict the prevalence of COPD and potential healthcare costs from the percentage of good, moderate, and unhealthy air quality days. As found in our Results, it revealed that states with better air quality tend to exhibit lower COPD prevalence rates and spend less on healthcare. Specifically, a one-unit increase in the percentage of good air quality days corresponds to a 10.04% reduction in COPD prevalence, while the same increase in moderate and unhealthy days leads to decreases of 7.9% and 3.30%. In healthcare spending, an increase in the percentage of good, moderate, and unhealthy air quality days corresponds to decreases of $8393.06, $7655.86, and $1924.53.

Introduction

Air quality significantly impacts public health, potentially leading to various respiratory and cardiovascular issues. This project aimed to explore the relationship between air quality and public health indicators in the United States, specifically focusing on the correlation between air

quality and healthcare costs and the prevalence of Chronic Obstructive Pulmonary Disease (COPD).

Poor air quality has been linked to adverse health effects, raising concerns about its broader implications for public health. According to the World Health Organization (WHO), exposure to air pollution contributes to approximately 7 million premature deaths worldwide each year (WHO). The United States, with its diverse environmental conditions, provides an interesting context to examine the correlation between air quality and public health indicators. Our leading questions of interest were: Is there a significant correlation between air quality and healthcare costs between different states in the US and is there a significant correlation between air quality and the prevalence of COPD? Using these questions to guide us, we explored the economic impact of air quality on healthcare costs to find insights into the potential financial burden associated with poor air quality. We investigated the relationship between air quality and COPD prevalence to find valuable information about the respiratory health implications of varying air quality levels.

This project aims to contribute valuable insights, through multiple linear regressions, into the intersection of air quality and public health, providing a foundation for potential interventions and policy recommendations to mitigate the health effects of poor air quality.

Data Description

We collected air quality data from two main sources: wisevoter.com for Air Quality Index (AQI) data and the open API for additional international air quality metrics. In addition, we are utilizing CSV datasets from the Centers for Disease Control and Prevention (CDC) and the Global Health Data Exchange (GHDE) for COPD prevalence and healthcare costs, respectively. The cleaned dataset includes information on the percentage of good, moderate, and unhealthy air

quality days from the AQI, average percent COPD prevalence from the CDC, and healthcare spending per capita for each state in the US from the GHDE. The reliability of these datasets is bolstered by the credibility of the respective sources. However, challenges in data cleaning may arise due to potential inconsistencies in formatting, missing values, and variations in measurement methodologies between sources.

The cleaning process handled missing values, standardized data formats, mapped state names for consistency, converted percentage values to numeric formats, and merged the datasets based on a common 'State' column. This meticulous cleaning approach aims to ensure a cohesive and accurate dataset, setting the stage for a robust exploratory data analysis of the relationships between air quality, COPD prevalence, and healthcare spending.

Method

This project utilized a multiple linear regression model to predict the percentage of a state's COPD prevalence and a state's health spending ($) based on the percentage of good, moderate, and unhealthy air quality Days. A regression model was created for each feature we wanted to predict. We chose a multiple linear regression model because it could take into account all types of air quality days.

We implemented a single-fold cross-validation with a 70-30 split for both models. We chose this method of cross-validation because it helps provide a reliable performance estimate and helps reduce overfitting or underfitting. We chose a 70-30 split because our dataset is relatively small. It is also easy to implement and analyze the model's performance

The assumptions in the regression model to predict COPD were not fully met. The independence assumption was met because there are no patterns and the plots are randomly

scattered. The constant variance assumption was not fully met, the concern was mainly the unhealthy air quality days. The plot of unhealthy air quality days against the residuals seems like there is a little bit of a dependency as most of the plots are scattered like a vertical line towards the left side of the plot with some outliers. The good air quality days may have a dependency when plotted against the residuals, but it is not as heavy or apparent as the unhealthy air quality days. The moderate air quality days seem to pass the constant variance assumption when plotted against the residuals. The normality assumption was met because the plots are evenly distributed along the red line.

In order to pass the constant variance assumption, we tried transforming the y features using different methods. When we did a log transformation, there was no significant change in the graphs with the different air quality days plotted against the residuals. A similar result also happened when we did a square root transformation. When we tried to do a box cox transformation with the help of the scipy.stats library, the plots did change. The good and moderate air quality days could pass constant variance, however, there are multiple outliers in the graph. The unhealthy air quality days changed, however, it still did not pass constant variance because the plots were scattered around the left side of the plot. Despite our different methods, the constant variance assumption was still not met.

The assumptions in the regression model to predict health spending were not fully met. The independence assumption was met because there are no patterns and the plots are randomly scattered. The constant variance assumption was not fully met, the concern was mainly the unhealthy air quality days. The plot of unhealthy air quality days against the residuals seems like there is a dependency as most of the plots are scattered like a vertical line towards the left side of the plot with some outliers. The good air quality days seem to pass the constant variance

assumption, however, there are some outliers. The moderate air quality days seem to pass the constant variance assumption when plotted against the residuals. The normality assumption was met because the plots are evenly distributed along the red line.

To pass the constant variance assumption, we tried transforming the y features using different methods. When we did a log transformation, the good air quality days had an obvious dependency as the plots were scattered along a negative linear slope. The moderate air quality days had an obvious dependency as the plots were scattered along a positive linear slope. There was no significant change in the unhealthy air quality days plotted against the new residuals. A similar result also happened when we did a square root and boxcox transformation. Despite our different methods, the constant variance assumption was still not met.

Overall, out of the machine learning models we have learned in class, we believed that the multiple regression model worked best with the data we gathered and the goals of this project. With air quality data, we can utilize that to predict a state's prevalence of COPD or health spending.

Results

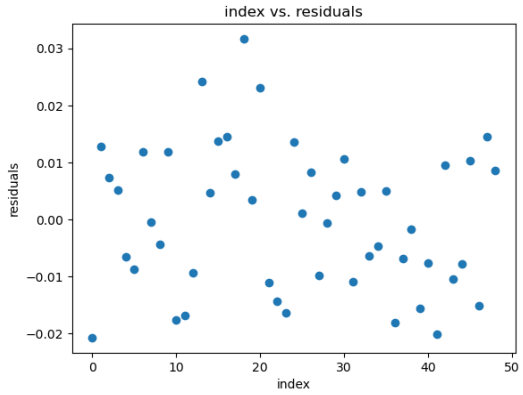
For our multiple linear regression model to predict COPD, our results show that COPD prevalence may decrease as air quality improves, as indicated by the negative coefficients (-0.1004, -0.0790, -0.0330). When the percentage of good air quality days increases by one, the prevalence of COPD decreases by 10.04%. When the percentage of moderate air quality days increases by one, the prevalence of COPD decreases by 7.9%. When the percentage of unhealthy air quality days increases by one, the prevalence of COPD decreases by 3.30%. The intercept was 1.5079, which means that when the percentage of air quality days is 0, the prevalence of

COPD is 150.795%.

For the cross validated MSE, we got 0.0002, which is a good value because it is very low. For the cross-validated R^2, we got 0.0222, which indicates that the model explains only approximately 2.22% of the variability, which is an acceptable value because it is between 0 and 1. These values show that our regression model is somewhat acceptable.

Figure 1

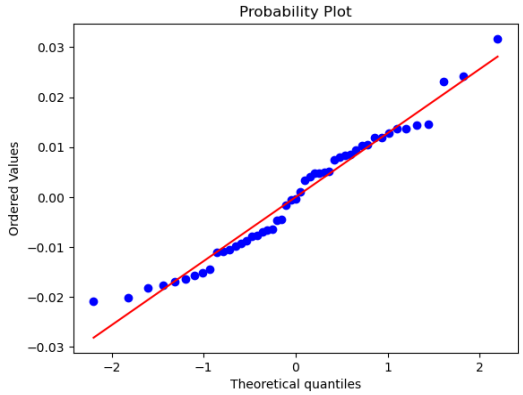
Predict COPD: Index vs Residuals



Note: Plot to check for independence assumption

Figure 2

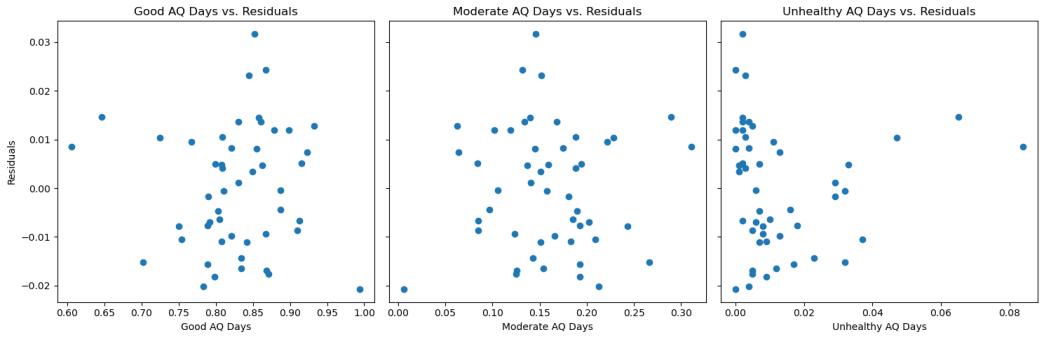
Predict COPD: QQ Plot



Note: Plot to check for normality assumption

Figure 3

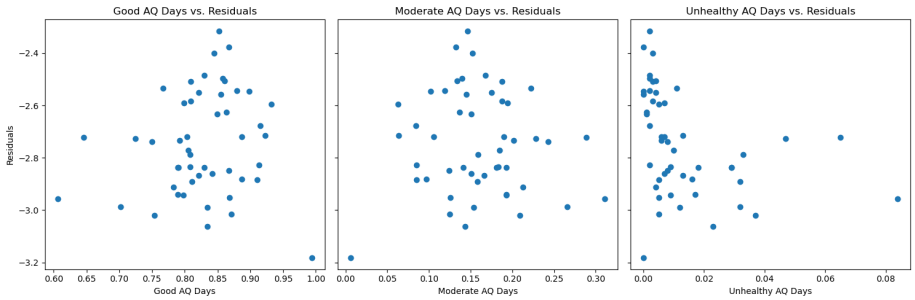
Predict COPD: X features vs. Residuals



Note: Plots to check for constant variance assumption

Figure 4

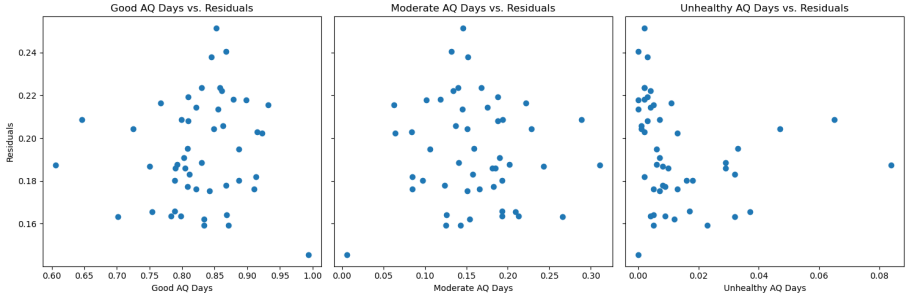
Predict COPD: X features vs. Residuals



Note: Plots to check for constant variance assumption with a log transformation

Figure 5

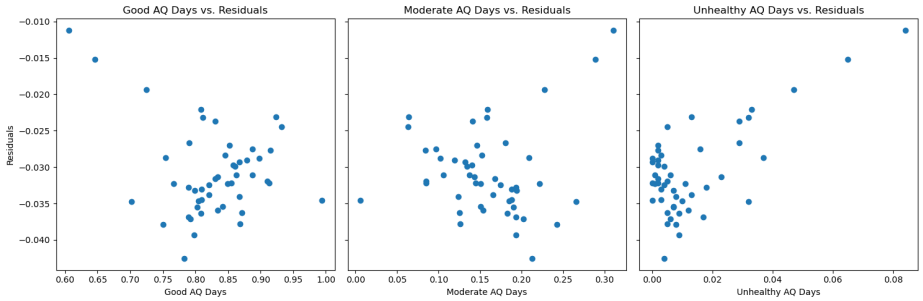
Predict COPD: X features vs. Residuals



Note: Plots to check for constant variance assumption with a square root transformation

Figure 6

Predict COPD: X features vs. Residuals



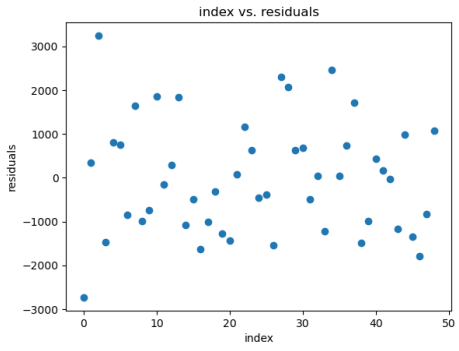
Note: Plots to check for constant variance assumption with a box cox transformation

For our multiple linear regression model to predict health spending, our results show that healthcare costs per capita may decrease as air quality improves, as the coefficients indicate (-8393.06, -7655.86, -1924.53). When the percentage of good air quality days increases by one, the healthcare costs per capita decrease by $8393.06. When the percentage of moderate air quality days increases by one, the healthcare costs per capita decrease by $7655.86. When the percentage of unhealthy air quality days increases by one, the healthcare costs per capita decrease by $1924.53. The intercept was 132584.2791, which means that when the percentage of air quality days is 0, the health care spending is $132584.28.

For the cross validated MSE, we got 2522751.3325, which is a bad value because it is very high. For the cross-validated R^2, we got 0.2271, which indicates that the model explains only approximately 22.71% of the variability, which is an acceptable value because it is between 0 and 1. These values show that our regression model is somewhat acceptable.

Figure 7

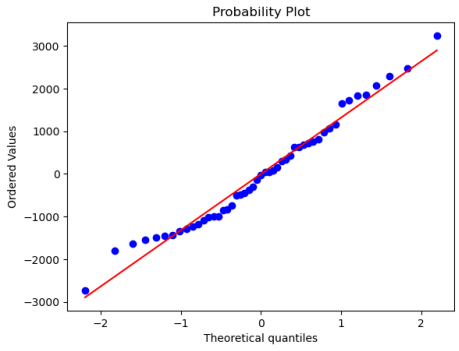
Predict Health Spendings: Index vs Residuals



Note: Plot to check for independence assumption

Figure 8

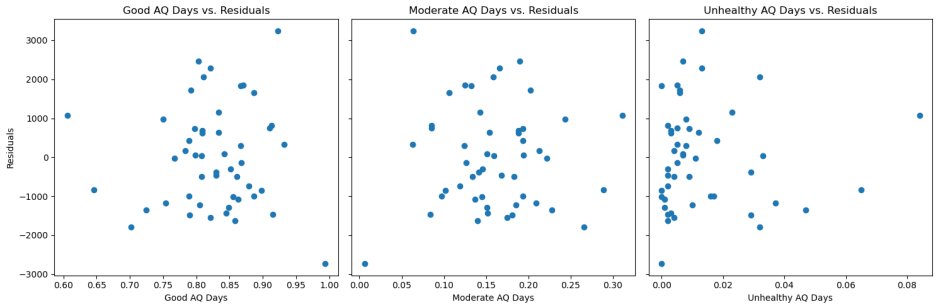
Predict Health Spendings: QQ Plot



Note: Plot to check for normality assumption

Figure 9

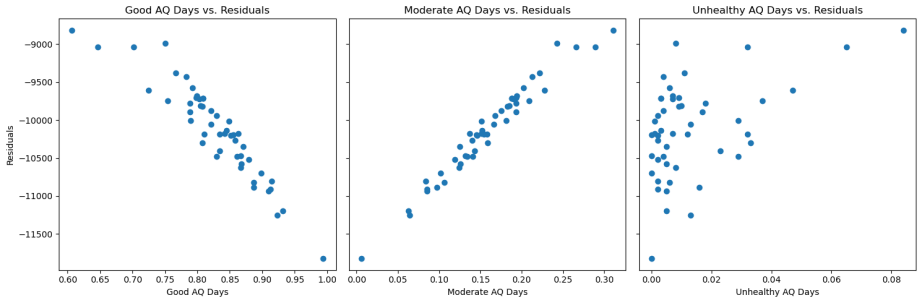
Predict Health Spendings: X features vs. Residuals



Note: Plots to check for constant variance assumption

Figure 10

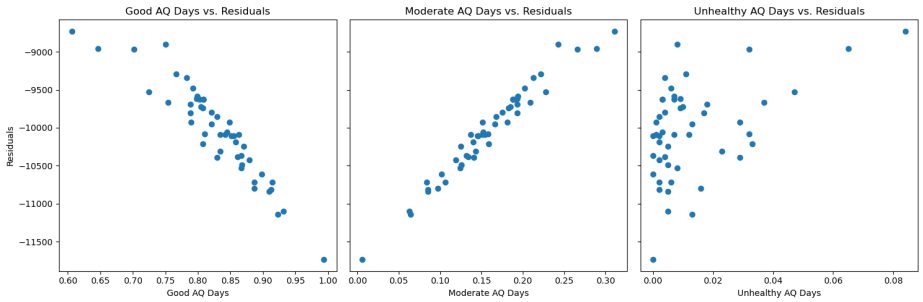
Predict Health Spendings: X features vs. Residuals



Note: Plots to check for constant variance assumption with a log transformation

Figure 11

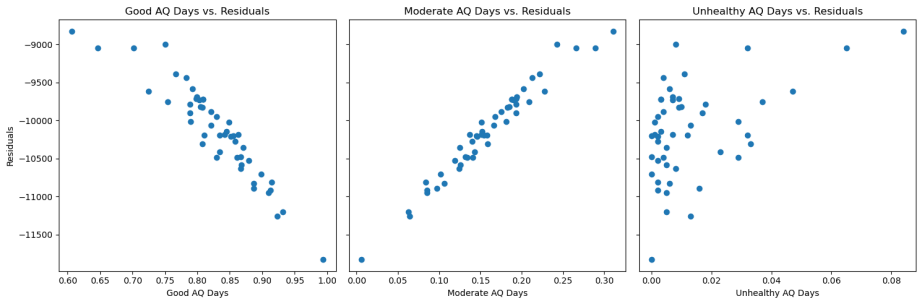
Predict Health Spendings: X features vs. Residuals



Note: Plots to check for constant variance assumption with a square root transformation

Figure 12

Predict Health Spendings: X features vs. Residuals



Note: Plots to check for constant variance assumption with a box cox transformation

Discussion

We used multiple linear regression analysis to investigate the associations between air quality metrics and public health indicators, namely COPD prevalence and healthcare costs. We aimed to determine how the percentages of good, moderate, and unhealthy air quality days relate to these important health outcomes.

From our regression model to predict COPD, the coefficients show that states with better air quality may have lower COPD prevalence rates. However, these correlations are only moderate in strength, suggesting that there are other factors that affect COPD prevalence, which we did not include in our analysis.

From our regression model to predict health spending, the coefficients show that states with better air quality may spend less on healthcare. However, these correlations are also moderate in strength, implying that there are other variables that influence healthcare spending, which we did not account for in our analysis.

We acknowledge the limitations of our linear regression approach. Linear regression assumes a linear relationship between variables, which may not capture the complex dynamics

between air quality and public health. Therefore, our findings may not fully reflect the reality of these relationships. The observed relationships may not be linear in nature and may require more advanced models to describe them better.

We evaluated the assumption of constant variance by plotting the residuals against the predictor variables. However, some of the features had not passed the constant variance assumption. We attempted to transform the response variable (percentage of COPD and healthcare spending) using log, square root, and box cox transformations to address this issue, so the variability of the residuals (the differences between the observed and predicted values) should have remained constant across all levels of the predictors. Despite doing this, the constant variance assumption was not fully met.

Given what we learned from that, in the future we could try out different models instead, such as a polynomial or logistic regression model. With polynomial regression, we could extend the linear regression by including polynomial terms to capture nonlinear relationships, the advantage of this model being that it can handle curved relationships between predictors and the response variable. Other models that could also be used in the future that we have not explored in class are Naive Bayes and RandomForests. We could try using other public health indicators to predict using air quality days, such as asthma or cancer. We could also use other air quality metrics, such as the levels of different pollutants, to predict COPD and health spending. In addition, we can also analyze public health indicators in other countries using the air quality data we had gathered from the OpenAQ AQI.

Overall, our regression models provide a good foundation for predicting public health indicators from air quality data. Although it is not the best model, we can use this to improve it with additional features and data, as well as apply it to new models.

References

Air Quality by state 2023. Wisevoter. (2023, May 15). https://wisevoter.com/state-rankings/air-quality-by-state/

Centers for Disease Control and Prevention. (2022, July 11). *State estimates - chronic obstructive pulmonary disease (COPD)*. Centers for Disease Control and Prevention. https://www.cdc.gov/copd/data-and-statistics/state-estimates.html

Global Health Data Exchange. *United States Health Expenditure by state, payer, and type of care, 2003-2019*. GHDx. (1970, January 1). https://ghdx.healthdata.org/record/ihme-data/united-states-health-spending-by-state-payer-type-service-2003-2019

OpenAQ. (n.d.). *Home*. https://openaq.org/

World Health Organization. (n.d.). *Air Pollution*. World Health Organization. https://www.who.int/health-topics/air-pollution#tab=tab_1.