# How Sociodemographic Conditions Influence Effect of Air Pollution on Lung Cancer*

### A nationwide county-level analysis of PM2.5 and social equity interactions

Siling Guo

December 10, 2025

This study focuses on exploring the interaction between sociodemographic conditions and air pollution in relation to the lung cancer incidence rates.We apply a multiple regression model with log transformation and conduct a hypothesis t-test. The results show that the sociodemographic conditions have a significant negative impact on the air pollution effect on lung cancer($p<0.001$), and weaken the effect by approximate 1.3%. The model only includes sociodemographic conditions and air pollution as the predictor variables,and focuses on inference rather than prediction, so the results should be read with caution.

## 1 Introduction

Lung cancer is the second most common cancer and the leading cause of cancer death in the United States. Air pollution is one of the risk factors for lung cancer("Lung Cancer Statistics | How Common Is Lung Cancer?" n.d.). In recent years, sociodemographic conditions (social conditions) have also been considered as a risk factor for cancer in many studies. Some research suggests that social conditions play an important role in shaping how strongly air pollution affects lung cancer risk(Acharjee, Das, and S.Stanley 2020). To better understand the interaction between these two factors on lung cancer, we compared different models and selected a multiple regression model with an interaction term and a log transformation for better validity. We also conducted a hypothesis T-test on the interaction term. The model shows that social conditions appear to weaken the effect of air pollution on lung cancer. It provides some new ideas for understanding how social development may be related to cancer incidence in high air pollution areas.

*Project repository available at: https://github.com/silingguo/MATH261A-project2.

First, we introduce the research background and aim, then describe the data used and the selected variables. Next, we introduce the base model and advanced model, as well as their limitations. After presenting the results of the selected model with relevant tables and graphs, we discuss the limitations and future plans.

## 2 Data

The dataset was provided by the research team of (Acharjee, Das, and S.Stanley 2020) and is publicly available from the Harvard Dataverse. It is a county-level dataset that includes information related to lung cancer across 2602 counties in the United States (Kansas, Michigan, Minnesota, and Nevada are excluded due to state legislation and regulations). In our project, we use three variables: Lung Cancer, PM2.5, and Sociod_EQI.

Lung Cancer is a numeric variable which shows the annual lung cancer incidence rate (LC) per 100,000 people from 2010–2014.

Sociod_EQI (Social EQI) is a numeric variable, an index that reflects the social disadvantage, from 2000–2005. It is generated by principal component analysis and includes factors related to sociodemographic conditions(US EPA 2017), such as population density, ethnic diversity, education, crime, and housing(Lobdell et al. 2011). A value of 0 represents the national average; values above 0 indicate below-average-level conditions, and values below 0 indicate above-average-level conditions. The higher the values, the more disadvantaged the social environment(Jagai et al. 2017a).

PM2.5 is a numeric measure of fine particulate matter (µg/m³). According to (Acharjee, Das, and S.Stanley 2020), PM2.5 is classified into high (greater than 10.59 µg/m³) and low (less than 10.59 µg/m³). The original dataset does not provide detailed information on how PM2.5 was collected or which years it represents. Based on the EQI information they provided, the PM2.5 time period seems to be consistent with the Social EQI. This missing information may limit our study's ability to explain.

The dataset was already clean, so we did not do additional data cleaning and transformation.

The summary statistics table shows big differences across counties, meaning that LC, PM2.5 levels, and social conditions vary a lot.

| Column | Mean | Minimun | Maximun | Median | Variance |
|---|---|---|---|---|---|
| LC | 69.17 | 12.90 | 169.90 | 68.60 | 303.39 |
| PM2.5 | 10.125 | 1.70 | 16.91 | 10.61 | 5.48 |
| Social EQI | 0.08 | -4.81 | 3.98 | 0.00 | 0.98 |

The table below compares the counties with the best and worst sociodemographic conditions. It is surprising to see that even though their social conditions are very different, their PM2.5

levels and lung cancer incidence rates are quite similar. The similarity in these outcomes leads to our question: How do social conditions influence the effect of air pollution on lung cancer? We thus use a multiple regression model with interaction terms to investigate.

| County | Social EQI | PM2.5 | PM2.5 Quartile | LC | LC Quartile |
|---|---|---|---|---|---|
| Starr Co | -4.80999 | 6.95 | 0.1283628 | 34.2 | 0.0176787 |
| Douglas | 3.979472 | 6.44 | 0.0903151 | 35.9 | 0.0222905 |

# 3 Methods

The simple linear regression model does not fully address our analysis topic. Therefore, we applied the multiple regression model with an interaction term as our first approach, while also exploring more advanced models.

## 3.1 Base Model

$$LC = \beta_0 + \beta_1 \,\text{PM2.5} + \beta_2 \,\text{Social EQI} + \beta_3 \,(\text{PM2.5} \times \text{Social EQI}) + \varepsilon_i$$

Here, $LC$ is the response variable. $PM2.5$ is the first predictor variable, and $SocialEQI$ is the second variable. PM2.5 $\cdot$ Social EQI,the interaction term, is the third predictor. The intercept, $\beta_0$, represents LC when both Social EQI are equal to zero. However, in pratice, this value may not be meaningful because PM2.5 is rarely zero. The slope, $\beta_1$, represents the main effect of PM2.5, indicating the average change in LC for each one-unit increase in PM2.5 when Social EQI is zero (national average level). The slope, $\beta_2$, represents the main effect of Social EQI, indicating that the average change in LC for a one-unit increase in Social EQI when PM2.5 is zero. The slope of the interaction term, $\beta_3$, represents how the Social EQI affects the relationship between PM2.5 and LC. In other words, $\beta_3$ measures how much the effect of PM2.5 on lung cancer changes at different levels of Social EQI. The error, $\varepsilon_i$, represents the error term. The ordinary least squares (OLS) method was used to estimate the parameters.

To make sure the model is valid, we check several assumptions using plots. We use a scatter plot to check whether the predictor and the response look linear, a QQ plot to check the normality of the error term, a residual-fitted value plot to check the homoscedasticity of the error term. We assume the observations (each county) are independent because the data represent different counties and there is no clear dependence between them. The diagnostic plots showed slightly unequal variance, but the other model assumptions are reasonably satisfied.

Previous studies suggest that PM2.5 and social conditions are related to LC(Jagai et al. (2017b)). To determine whether Social EQI changes the strength of the PM2.5 and LC relationship, we apply the t-test on interaction term. The null hypothesis, $H_0$: $\beta_3 = 0$, represents

social EQI has no impact on on the PM2.5 effect. The alternative hypothesis, $H_a$: $\beta_3 \neq 0$, represents that social conditions have an impact on the PM2.5 effect. We use a significance level of $\alpha$=0.05.

### 3.2 Advanced Model

The residuals vs. fitted value plot of the base model shows slightly unequal variance. Therefore, we apply a log transformation to the response variable to improve the validity.

$$\log(LC) = \beta_0 + \beta_1 \,\mathrm{PM}_{2.5} + \beta_2 \,\text{Social EQI} + \beta_3 \,(\mathrm{PM}_{2.5} \times \text{Social EQI}) + \varepsilon_i$$

The structure of the advanced model is the same as the base model. After log transformation, the interpretation of the coefficients is in percentage changes rather than absolute changes, which is more meaningful in explaining disease incidence rate.

Here, $\log(LC)$ is the response variable. The slope, $\beta_1$, represents the main effect of PM2.5, indicating the approximate percentage change in LC for a one-unit increase in PM2.5 when Social EQI is zero (national average level). The slope, $\beta_2$, represents the main effect of Social EQI, indicating that the approximate percentage change in LC for a one-unit increase in Social EQI when PM2.5 is zero. The slope of the interaction term, $\beta_3$, represents how Social EQI affects the relationship between PM2.5 and LC. In other words, $\beta_3$ measures how much the percentage effect of PM2.5 on lung cancer changes at different levels of Social EQI. The error, $\varepsilon_i$, represents the error term.

The process of hypothesis testing and assumption checking stays the same, but the interpretation changes.

This model meets the assumptions, and the unequal variance issue of the base model is mostly corrected. In the Results section, we will explain this in detail.

Our analysis also has some clear limitations: the data are at the county level, so the results only reflect county patterns and not individual risk. Besides, only PM2.5 and Social EQI are included in the model, while other important factors of lung cancer, such as smoking and genetics, are not included, which limits the predictive ability. Moreover, the Social EQI score is a rough combined measure, so it doesn't really capture the details of social conditions.

We use the base R's lm() function to implement the linear model(R Core Team 2024)

## 4 Results

Here are the scatter plots of the base model and the advanced model. Essentially, rather than changing the variables in the model, the log transformation mainly rescales the response by compressing large values, so the scale of the plots is different. Figure 1 shows that the two

predictor variables have a linear relationship with the response variable. PM2.5 is positively associated with LC, while Social EQI is negatively associated with cancer incidence. This association is counterintuitive. Why are better social conditions associated with a higher lung cancer incidence rate? It is reasonable that better social conditions usually mean better access to medical support and more accurate screening of cancer, so more lung cancer cases can be diagnosed("Theory of Fundamental Causes" 2025). Besides, other factors such as genetics, aging, and local industry structure may also play roles in LC. The scatter plot is based on the original data, without controlling for other factors.

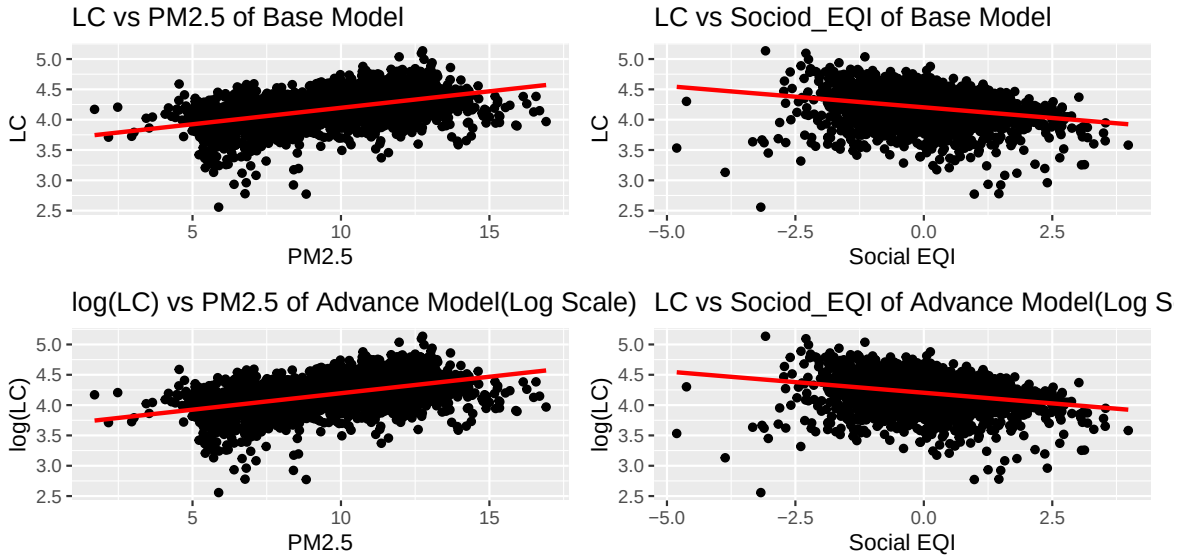


Figure 1: Scatter Plot

In the Residuals vs. Fitted Value plot of base model(Figure 2), the residuals are spread out along with the higher fitted value, which indicates the unequal variance issue. After the transformation, the residuals are closer to zero, which indicates that the residuals have equal variance.

From Figure 3, we can see that the majority of the points follow the reference line in the base model. The left tail falls below the reference line, and the right tail falls above it, indicating that some observations have more extreme residuals than expected. After transformation, the normality issue is improved, and the right tail follows the reference line better, but the left tail still falls below it.

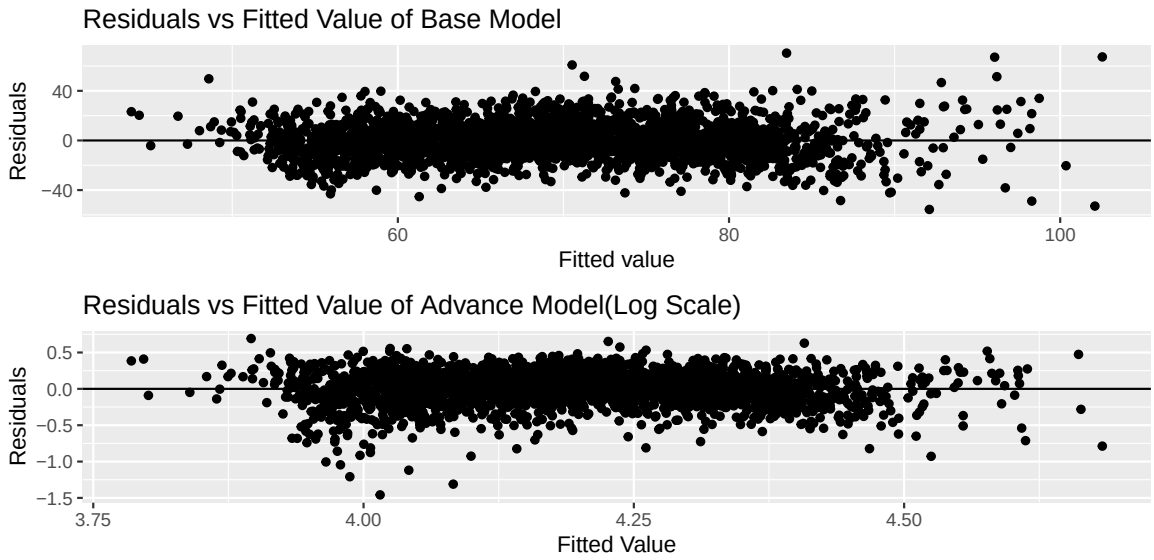


Figure 2: Residuals vs Fitted Value Plot

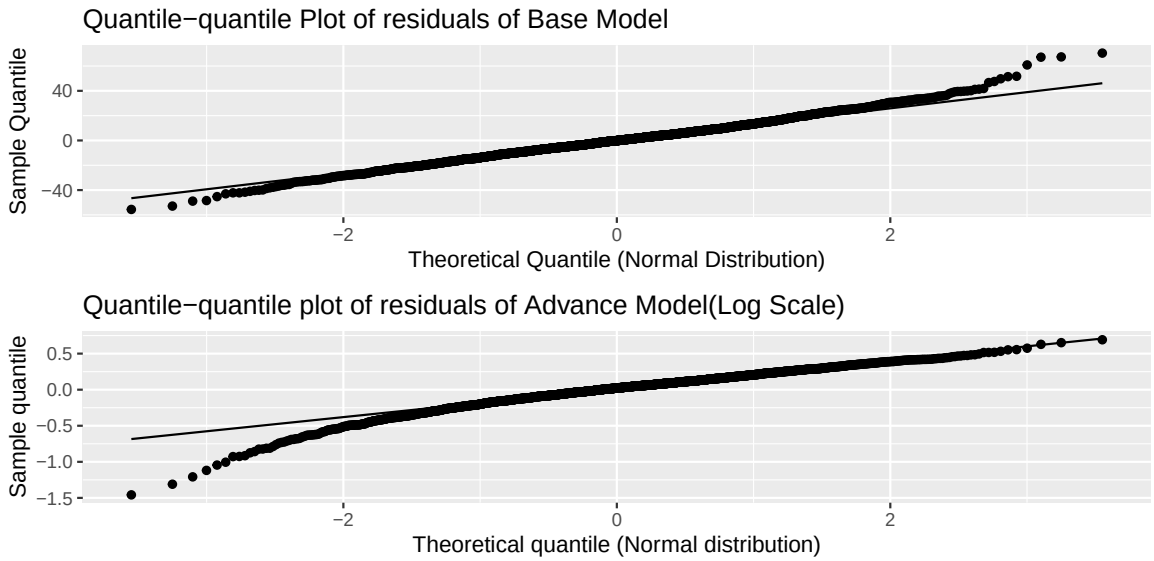


Figure 3: QQ Plot

| Term/Statistic | Base Model | Advanced Model log |
|---|---|---|
| Intercept | 35.132 | 3.659 |
| PM2.5 | 3.3634 | 0.0537 |
| Social EQI | 6.2154 | 0.0646 |
| PM2.5 · Social EQI | -1.1118 | -0.0131 |
| Residual Std. Error | 14.4173 | 0.2243 |
| R-squared | 0.316 | 0.302 |
| Adjusted R- squared | 0.315 | 0.301 |
| F-statistic | 399.46 | 374.79 |
| P-value | $< 0.001$ | $< 0.001$ |

The table above shows the data of the two models. It shows that the two models give similar conclusions and explain the variation in the data in a similar way. Considering the model validity and interpretability, we choose the log-transformed model as our final model. Here are the analyses:

The data above reveal the relationships among three variables: There is a 5.4% change in LC for a one-unit increase in PM2.5 when holding Social EQI constant, and PM2.5 is positively related to LC ($p < 0.05$).

There is a 6.5% change in LC for a one-unit increase in Social EQI when holding PM2.5 constant, and Social EQI is positively related to LC ($p < 0.05$). The scatter plot above shows a negative relationship between them. One possible reason is that the scatter plot shows an overall relationship, which is influenced by other factors. However, the model helps reveal a clearer relationship by controlling the PM2.5 effect.

There is a -1.3% change in the effect of PM2.5 on LC at different levels of Social EQI. For each one-unit increase in Social EQI, the effect of PM2.5 on LC will be weakened by approximately 1.3%. In contrast, for each one-unit decrease in Social EQI, the effect of PM2.5 on LC will be strengthened by approximately 1.3%. In other words, when social conditions are worse, the effect of PM2.5 on lung cancer becomes weaker. Besides, from Figure 4, we can see that as Social EQI increases, the slope becomes less steep, showing LC increases more slowly with PM2.5.

We applied a t-test on the interaction terms. According to the result($t = -6.959$, p-value $< 0.001$), we then reject $H_0$ and conclude that the social conditions have a significant negative impact on the PM2.5 effect on LC.

R-squared is around 30%, meaning that our regression model can explain 30% variation of the LC. Around 70% of the variation in LC is still unexplained since the model only includes two factors and the interaction effect between them, and does not include all major risk factors of LC. This level of explanatory power is considered acceptable. Adjusted R-squared is slightly smaller than R-squared, meaning that the model is not overfitted. The p-value of the F-test

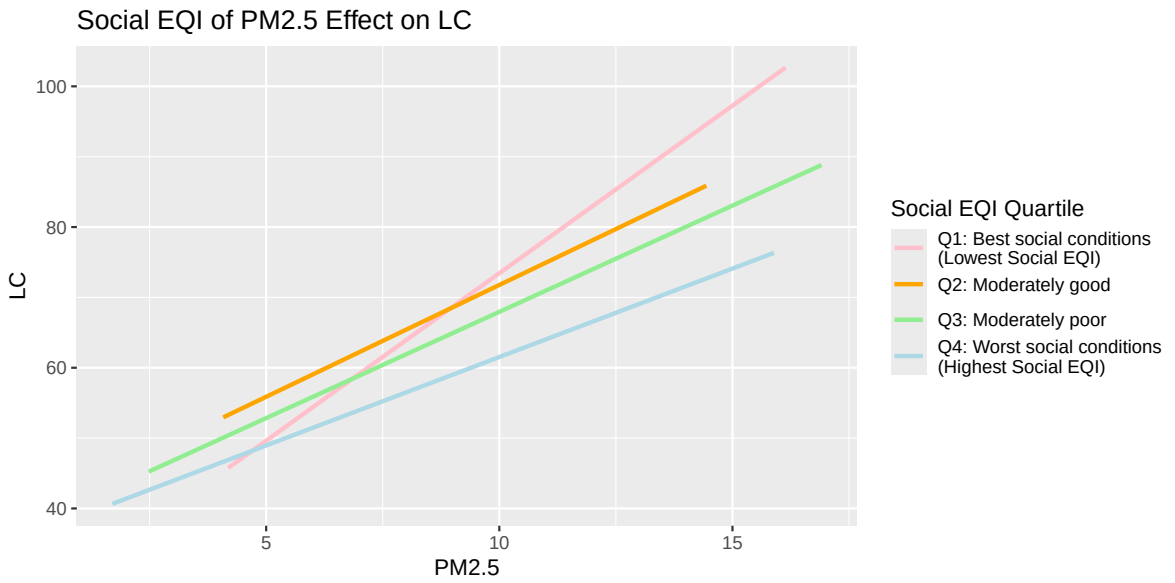


Figure 4: Interaction Plot

is less than 0.05, meaning that the overall model is highly significant, and PM2.5 and Social EQI can somehow explain the LC.

## 5 Discussion

In this study, we fit a multiple regression model with an interaction term to answer how social conditions (Social EQI) affect the air pollution (PM2.5) impact on the lung cancer incidence rate across counties. To improve the model's validity, we applied a log transformation. Air pollution and social conditions are positively associated with the lung cancer incidence rate. Moreover, social conditions appear to weaken the effect of air pollution on lung cancer: when social conditions are better, the impact of air pollution on lung cancer is weaker.

Counties with better social conditions are able to reduce the effect of air pollution on the lung cancer incidence rate with better medical support, stronger community support, and better public awareness of the disease. In contrast, counties with poorer social conditions may have worse industrial environments, poorer food access, and worse medical support, which can make the impact of air pollution on lung cancer more severe.

Although the model helps us to understand the interaction between air pollution and social conditions, it only includes two predictor variables, making it harder to fully explain and accurately predict the lung cancer incidence rate. In the future, we plan to go beyond the inference and explore the prediction. We will try splitting 70% of the data as training set and apply regularized regression models, such as ridge and LASSO.LASSO would be better for

this dataset because there are many variables and some of them are highly correlated. Using Lasso can help us to select important variables while controlling overfitting. to find the add more variables and improve the prediction performance.

## References


Acharjee, Mithun, Kumer Pial Das, and Young S.Stanley. 2020. "Air Quality-Lung Cancer Data." Harvard Dataverse. https://doi.org/10.7910/DVN/HMOEJO.

Hystad, Perry, Richard M. Carpiano, Paul A. Demers, Kenneth C. Johnson, and Michael Brauer. 2013. "Neighbourhood Socioeconomic Status and Individual Lung Cancer Risk: Evaluating Long-Term Exposure Measures and Mediating Mechanisms." *Social Science & Medicine (1982)* 97 (November): 95–103. https://doi.org/10.1016/j.socscimed.2013.08.005.

Jagai, JS, LC Messer, KM Rappazzo, CL Gray, SC Grabich, and DT Lobdell. 2017b. "County-Level Cumulative Environmental Quality Associated with Cancer Incidence." *Cancer* 123 (15): 2901–8. https://doi.org/10.1002/cncr.30709.

———. 2017a. "County-Level Cumulative Environmental Quality Associated with Cancer Incidence." *Cancer* 123 (15): 2901–8. https://doi.org/10.1002/cncr.30709.

Lobdell, Danelle T., Jyotsna S. Jagai, Kristen Rappazzo, and Lynne C. Messer. 2011. "Data Sources for an Environmental Quality Index: Availability, Quality, and Utility." *American Journal of Public Health* 101 (Suppl 1): S277–85. https://doi.org/10.2105/AJPH.2011.300184.

"Lung Cancer Statistics | How Common Is Lung Cancer?" n.d. Accessed December 7, 2025. https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." https://www.R-project.org/.

"Theory of Fundamental Causes." 2025. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Theory_of_fundamental_causes&oldid=1286993629.

US EPA, ORD. 2017. "Environmental Quality Index (EQI)." Overviews and {Factsheets}. https://www.epa.gov/healthresearch/environmental-quality-index-eqi.