Modeling the Relationship Between PM2.5 and County Mortality Rate

Zhenkun Xu, Shucheng Liu and Ruocheng Sun

December 4, 2025

1 Introduction

Exposure to fine particulate matter (PM2.5) is a leading environmental risk factor for global cardiovascular health concern (Feng et al., 2023). The public health burden is significant, as extensive epidemiological studies have established a strong association between long-term exposure to PM2.5 and increased cardiovascular mortality (CMR). The biological mechanisms driving this association are complex and multifaceted, involving pathways such as systemic inflammation, oxidative stress, and autonomic nervous system dysfunction (Feng et al., 2023). Given these established health risks, quantifying the population-level impact of air quality changes becomes a critical task for public policy and environmental regulation. Foundational work in this area by Wyatt et al. (2020) demonstrated that historical reductions in PM2.5 from 1990 to 2010 were associated with significant reductions in CMR across thousands of U.S. counties, confirming the public health benefits of improved air quality (Wyatt et al., 2020).

There are differing views regarding the statistical relationship between CMR and PM2.5 concentrations. Some researchers argue that this association is essentially linear across commonly observed environmental ranges, with evidence showing an approximately linear concentration–response relationship down to very low exposure levels (Schwartz et al., 2002). In contrast, other studies report diminishing marginal effects at higher concentrations, suggesting that the slope of the concentration–response curve decreases as PM2.5 levels rise (Vodonos et al., 2018; Li et al., 2018). A third perspective highlights that the association may become stronger once PM2.5 concentrations surpass certain low-exposure thresholds, with evidence that long-term exposures at or above 6 µg/m^3 are associated with larger mortality effects than exposures below that level (Shi et al., 2016). Thus, using statistical

models to investigate the shape and magnitude of the PM2.5-CMR relationship remains highly meaningful.

Our project is inspired by this body of work and utilizes a similarly structured dataset. Our data is comprised of two main files. The first file, County_annual_PM25_CMR.csv, provides the primary treatment and outcome variables. The treatment is the annual mean PM2.5 concentration, and the outcome is the annual CMR (deaths per 100,000 person-years). This panel data spans 21 years from 1990 to 2010 across 2,132 U.S. counties. The second file, County_RAW_variables.csv, provides a rich set of 9 covariates. These include socioeconomic, housing, and demographic characteristics from U.S. Census data at decadal intervals (1990, 2000, and 2010), such as median income, unemployment rates, and educational attainment. While the link between PM2.5 and CMR is known, this relationship is heavily confounded by these complex socioeconomic factors. Therefore, the primary statistical challenge, and the goal of our project, is to answer the following research question: How can we best model the nonlinear dose-response relationship between PM2.5 and CMR, while properly adjusting for this high-dimensional set of socioeconomic confounders.

2   Related Work

The statistical analysis of the health impacts of PM2.5 is motivated by a large body of scientific literature. Extensive reviews have detailed the complex pathophysiological and molecular mechanisms through which exposure to fine particulate matter can lead to cardiovascular disease (Feng et al., 2023). Given that our response variable, cardiovascular mortality (CMR), represents count or rate data, the analytical foundation for this work is the Generalized Linear Model (GLM) framework (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). This project builds directly upon observational studies that use this framework to quantify the population-level association. A key precedent is the work by Wyatt et al. (2020), which is highly relevant to our project's data and objectives (Wyatt et al., 2020). They analyzed CMR across 2,132 U.S. counties from 1990-2010, using mixed-effect regression models (a form of GLM) to estimate the impact of PM2.5 reductions while accounting for socioeconomic deprivation (Wyatt et al., 2020).

More advanced methodologies frame this problem as one of causal inference: estimating the causal dose-response curve (DRC) for a continuous treatment (PM2.5 exposure) while adjusting for a set of covariates (such as the socioeconomic data in County_RAW_variables.csv) (Takatsu & Westling, 2025; Zhang et al., 2025). Recent work in nonparametric statistics offers sophisticated tools for this challenge. Takatsu & Westling (2025) propose a debiased local linear estimator for a "covariate-adjusted regression function," which corresponds to

the DRC under certain conditions (Takatsu & Westling, 2025). A further challenge is the "positivity condition" (i.e., that all individuals have some chance of receiving any exposure level), which may be violated in observational data. Zhang et al. (2025) develop identification and estimation theories for DRCs that do not rely on this assumption, proposing an integral estimator to mitigate this potential bias (Zhang et al., 2025).

3 Data Preprocessing

Two preprocessing steps are required to prepare the data for modeling.

First, a key challenge is the temporal mismatch between the annual PM2.5 and CMR data and the decadal covariates. To create a unified, cross-sectional dataset for our analysis, we will focus our study exclusively on the data from the year 2000. This allows us to align the 2000-era treatment and outcome measures with the complete set of covariates from the 2000 U.S. Census.

Second, as observed in our exploratory analysis, the covariates have vastly different units and scales (e.g., dollar amounts, percentages, counts). To ensure all features are on a comparable scale and to improve the numerical stability of our regression models, we will normalize the covariate features. This process involves transforming each covariate to have a zero mean and unit variance. This final, standardized cross-sectional dataset from the year 2000 will form the basis for our statistical analysis.

4 Exploratory Data Visualization

Before constructing a formal statistical model, we begin with exploratory data visualizations.

First, to examine the empirical relationship between county-level PM2.5 concentrations and the CMR, we remove observations with missing values in either PM2.5 or CMR. Because the raw scatterplot exhibits substantial variability, we sort the data by PM2.5 and compute a 100-point centered moving average of CMR. This smoothing procedure reduces high-frequency noise while preserving the large-scale shape of the CMR-PM2.5 relationship.

Figure 1 presents the resulting scatterplot of CMR versus PM2.5, along with the smoothed trend. The moving average indicates a clear overall upward trajectory: CMR tends to increase as PM2.5 rises. More importantly, the rate of increase is not constant. At lower concentrations (approximately below 6 μg/m^3), the curve remains relatively flat. Between 6 and 8 μg/m^3, the slope increases noticeably. However, beyond 8 μg/m^3, the trend appears to level off. These visible changes in curvature provide preliminary evidence of potential threshold effects.
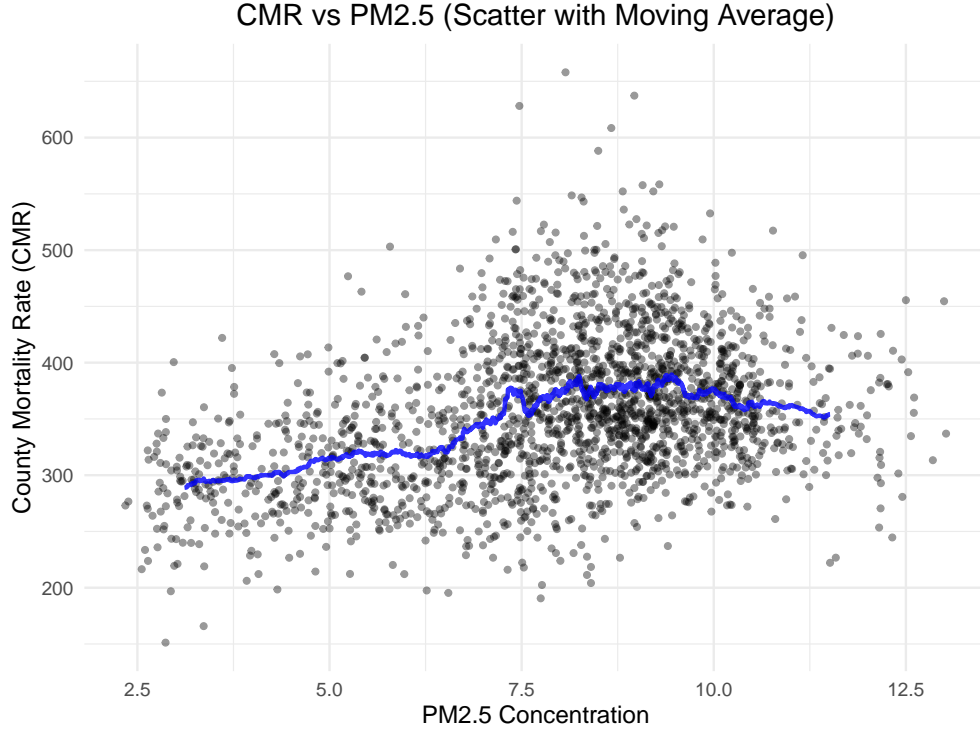
Figure 1: Scatterplot of CMR vs. PM2.5 with 100-point centered moving average (blue), used to visualize the underlying trend in the presence of noise.

Based on the patterns observed in the smoothed trend, the changes in curvature motivate the use of hinge-squared terms at 6 and 8 μg/m^3. Incorporating the nonlinear components $(x - 6)^2_+$ and $(x - 8)^2_+$ enables the fitted curve to bend upward more sharply once PM2.5 rises above 6 μg/m^3, and to potentially flatten after 8 μg/m^3, consistent with the patterns in Figure 1.

Second, we examine the relationships within our set of socioeconomic covariates from the County_RAW_variables.csv file. Figure 2 displays a correlation heatmap of these variables. The plot immediately reveals the presence of significant multicollinearity. As expected, variables measuring the same construct over time (e.g., median_HH_inc_1990, median_HH_inc_2000, and median_HH_inc_2010) are highly correlated with each other. We also observe strong negative correlations where expected, such as between poverty (pct-fam_pover_1990) and median income (median_HH_inc_1990). This high level of inter-correlation suggests that using all raw covariates directly in a regression model would lead to unstable coefficient estimates.
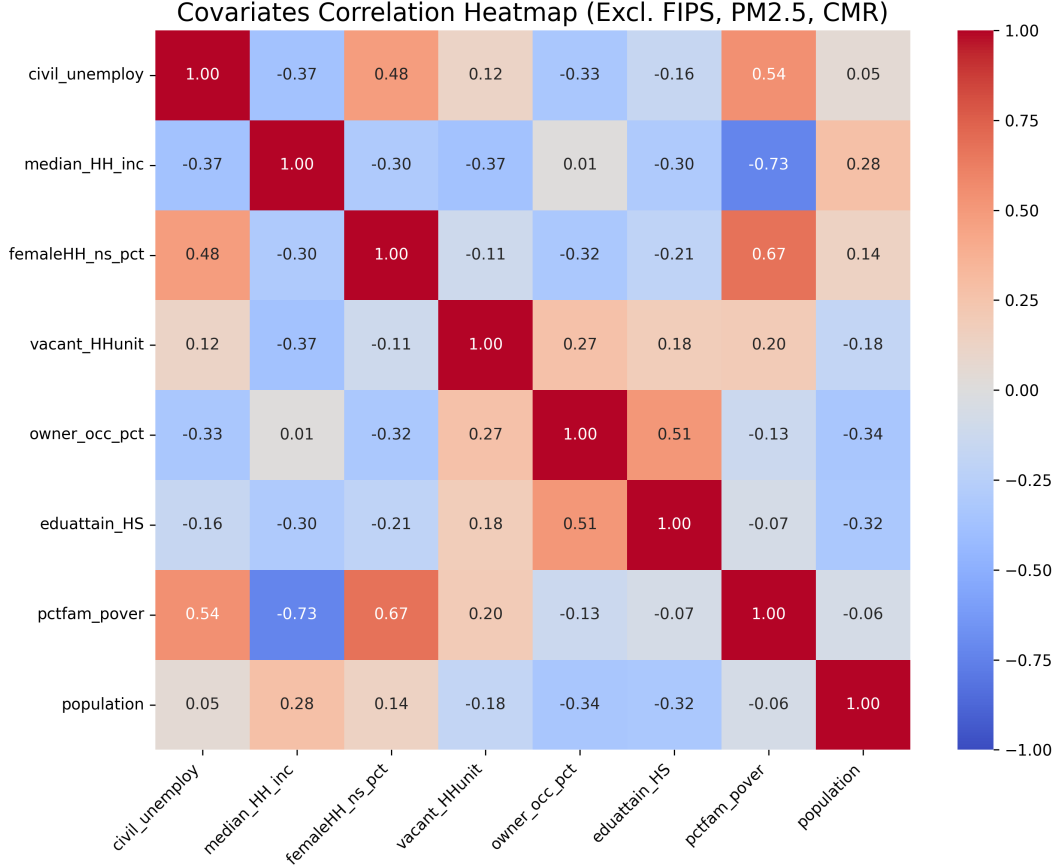
Figure 2: Correlation heatmap of socioeconomic and demographic covariates. The strong blue and red blocks indicate high multicollinearity, motivating the need for dimensionality reduction.

5 Model Specification

Guided by the exploratory analysis in the previous section, we develop a regression model that allows the association between PM2.5 and CMR to vary across different ranges of pollution exposure. The smoothed trend in Figure 1 suggests distinct changes around 6 and 8 μg/m^3, indicating that the marginal effect of PM2.5 may not be constant. To accommodate this pattern within a linear modeling framework, we introduce hinge-squared terms that enable flexible changes in slope and curvature beyond these thresholds.

Let x_i denote the PM2.5 concentration for county i. We define two hinge-squared basis functions as

$$h_{6,i} = (x_i - 6)_+^2, \qquad h_{8,i} = (x_i - 8)_+^2,$$

where $(u)_+ = \max(u, 0)$ denotes the positive-part operator. Each term is equal to zero below

its respective threshold, and increases quadratically once the threshold is crossed.

Based on the first exploratory findings, our preliminary regression model is specified as follows:

$$\begin{aligned}
\text{CMR}_i = \beta_0 &+ \beta_1 x_i + \beta_2 h_{6,i} + \beta_3 h_{8,i} \\
&+ \gamma_1 \text{civil\_unemploy}_i + \gamma_2 \text{median\_HH\_inc}_i + \gamma_3 \text{femaleHH\_ns\_pct}_i \\
&+ \gamma_4 \text{vacant\_HHunit}_i + \gamma_5 \text{owner\_occ\_pct}_i + \gamma_6 \text{eduattain\_HS}_i \\
&+ \gamma_7 \text{pctfam\_pover}_i + \gamma_8 \text{population}_i + \varepsilon_i,
\end{aligned}$$

where ε_i denotes the error term.

This specification preserves the interpretability and tractability of a standard linear regression model, while the hinge-squared terms introduce localized flexibility to accommodate the nonlinear patterns observed around the 6 and 8 $\mu\text{g/m}^3$ thresholds in the exploratory analysis. These terms allow the effect of PM2.5 to change in magnitude or curvature once the pollution level surpasses each threshold, without imposing a fully global nonlinear structure.

The additional socioeconomic and demographic covariates help adjust for potential confounding factors that may otherwise distort the estimated association between PM2.5 exposure and mortality. In the subsequent analysis, we may refine or extend this model to assess whether the hinge-squared structure adequately captures the underlying relationship or whether further nonlinear modeling is warranted.

Interpretation of Variables

- x_i: PM2.5 concentration.

- $h_{6,i}$: nonlinear curvature activated when $x_i > 6$.

- $h_{8,i}$: additional curvature activated when $x_i > 8$.

6 Preliminary Analysis

We fit the preliminary linear model that incorporates hinge-squared terms at 6 and 8 $\mu\text{g/m}^3$. The fitted curve, displayed in Figure 3, captures the accelerated increase in CMR between these two thresholds and the subsequent flattening at higher concentrations. This model-based trend is consistent with the exploratory findings and provides additional support for the hypothesis that $\text{PM}_{2.5}$ affects mortality differently across pollution ranges.
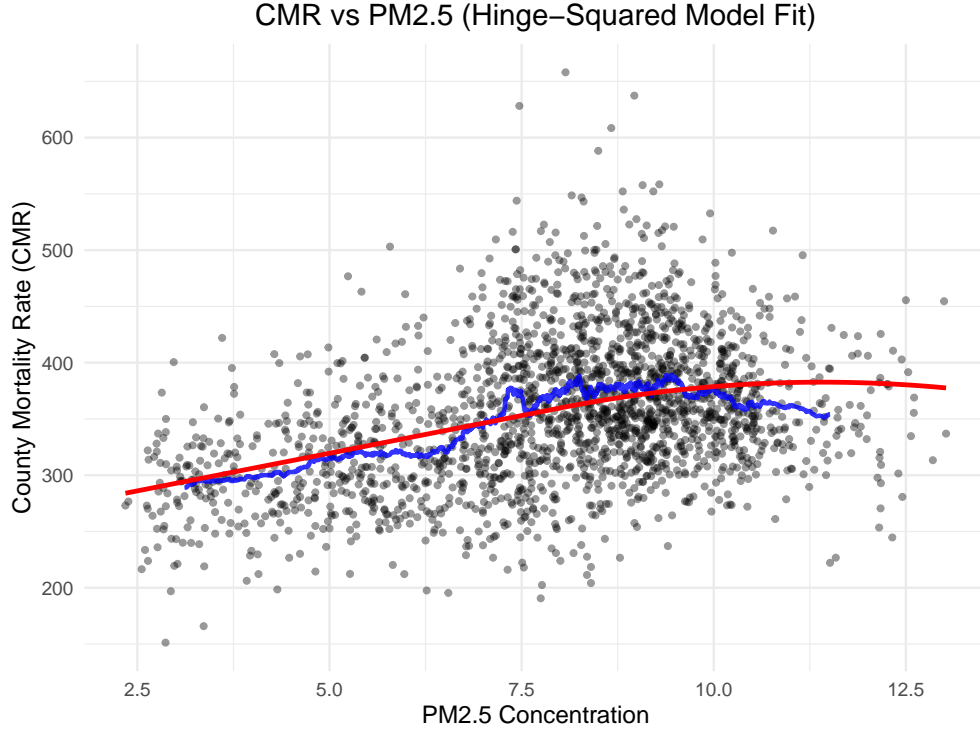
Figure 3: Scatterplot of CMR vs. PM2.5 with 100-point centered moving average (blue), used to visualize the underlying trend in the presence of noise.

Taken together, both the moving-average exploration and the hinge-squared model fit indicate that (i) the association between $PM_{2.5}$ and mortality is positive overall, and (ii) its strength varies across distinct concentration ranges. These preliminary results motivate the development of formal research questions.

At the same time, we also observe that the fitted hinge–squared model does not perfectly capture the underlying trend, suggesting that the adequacy of the nonlinear specification requires further investigation. For example, although the increase in CMR appears somewhat faster in the 6–8 $\mu g/m^3$ range, the slope is not dramatically steeper, implying that the practical and statistical significance of the hinge–quadratic terms should be examined more carefully.

7 Research Questions and Corresponding Hypotheses

This section outlines the research questions guiding our analysis, together with the corresponding null hypotheses.

RQ1. Overall Association between PM$_{2.5}$ and CMR

Research Question: Is county-level PM$_{2.5}$ concentration positively associated with the CMR after adjusting for covariates?

Null Hypothesis H1 (Overall Effect):

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, \qquad H_A : \text{At least one of } \beta_1, \beta_2, \beta_3 \neq 0.$$

RQ2. Nonlinearity of the Exposure-Response Curve

Research Question: Does the PM$_{2.5}$-CMR association deviate from linearity?

Null Hypothesis H2 (Linearity Test):

$$H_0 : \beta_2 = \beta_3 = 0, \qquad H_A : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0.$$

RQ3. Threshold at 6 μg/m^3

Research Question: Does the slope or curvature change when PM$_{2.5}$ exceeds 6 μg/m^3?

Null Hypothesis H3 (6 μg/m^3 Threshold):

$$H_0 : \beta_2 = 0, \qquad H_A : \beta_2 \neq 0.$$

RQ4. Threshold at 8 μg/m^3

Research Question: Does the rate of increase in CMR change after 8 μg/m^3?

Null Hypothesis H4 (8 μg/m^3 Threshold):

$$H_0 : \beta_3 = 0, \qquad H_A : \beta_3 \neq 0.$$

RQ5. Distributional Adequacy of the Error Term

Research Question:
Does the error term in the hinge-squared regression model satisfy the distributional assumptions required for valid inference, in particular, approximate normality and homoscedasticity?

Null Hypothesis H5 (Error Normality and Homoscedasticity):

$$H_0 : \text{The residuals are approximately normally distributed and homoscedastic.}$$

H_A : The residuals deviate significantly from normality or exhibit heteroscedasticity.

RQ6. Optimal Breakpoint Location

Research Question: Are 6 μg/m^3 and 8 μg/m^3 the most appropriate breakpoints, or do alternative locations better capture the nonlinear pattern?

Null Hypothesis H6 (Breakpoint Optimality):

$$H_0 : 6 \text{ and } 8 \ \mu\text{g/m}^3 \text{ are optimal breakpoints.}$$

$$H_A : \text{Alternative breakpoints yield a significantly better fit.}$$

RQ7. Adequacy of the Hinge-Squared Model

Research Question: Does the hinge-squared specification adequately capture the nonlinear pattern, or do more flexible models provide a significantly better fit?

Null Hypothesis H7 (Model Adequacy):

$$H_0 : \text{The hinge-squared model provides an adequate fit.}$$

$$H_A : \text{More flexible models (e.g., splines) significantly improve the fit.}$$

RQ8. Significance of Socioeconomic Covariates

Research Question: Do socioeconomic and demographic covariates significantly explain variation in CMR beyond PM$_{2.5}$?

Null Hypothesis H8 (Covariate Effects):

$$H_0 : \gamma_j = 0 \quad \text{for each covariate } j, \qquad H_A : \gamma_j \neq 0.$$

RQ9. Effect Modification by Socioeconomic Factors

Research Question: Do socioeconomic factors like poverty, income, or education modify the association between PM$_{2.5}$ and CMR? That is, does the PM$_{2.5}$ slope vary across socioeconomic profiles?

Null Hypothesis H9 (No Effect Modification):

$$H_0 : \delta_k = 0 \quad \text{for all interaction terms } k,$$

$$H_A : \text{At least one interaction coefficient } \delta_k \neq 0.$$

Here, each interaction term k corresponds to $\mathrm{PM_{2.5}}$ multiplied by a selected socioeconomic variable, and the coefficients δ_k are obtained by re-estimating the model with these interaction terms included.

8 Analysis Plan

Building upon the preliminary hinge-squared regression in Section 6, the next stage of our study is to conduct a more rigorous and systematic examination of the estimated relationship between PM2.5 and CMR. Although the initial OLS fit indicates that the association may vary across different pollution ranges, particularly near the 6 and 8 μg/m^3 thresholds, these observations remain exploratory. A more formal analysis is therefore required to determine whether the patterns suggested in the preliminary fit are statistically meaningful and whether the hinge-squared structure provides an adequate representation of the exposure-response curve.

Our first step is to validate the initial findings using formal hypothesis testing procedures. We will apply an overall F-test to evaluate whether PM2.5 and its nonlinear hinge terms jointly contribute to explaining variation in CMR. This test evaluates the null hypothesis that the coefficients associated with the linear component and the two hinge-squared components are simultaneously equal to zero. In addition, we will conduct t-tests on each individual PM2.5-related coefficient to assess whether the linear trend, the change in curvature around 6 μg/m^3, and the further curvature beyond 8 μg/m^3 are statistically significant. These tests will clarify whether the visual patterns identified in the exploratory analysis reflect genuine structural changes or simply arise from sampling variability.

After assessing the statistical significance of the model coefficients, we will perform a series of diagnostic checks to evaluate the robustness of the fitted model. The distribution of residuals will be examined through QQ-plots and formal normality tests to determine whether the standard OLS assumptions are reasonable. If evidence of non-constant variance is detected, further adjustments to the model will be required.

Beyond this OLS validation, our primary objective is to shift from correlation to a more rigorous estimation of the causal Average Dose-Response Function (ADRF). This allows us to quantify the causal relationship between PM2.5 (as a continuous treatment) and CMR, but requires addressing the significant confounding bias from our 9 covariates. We will adopt a two-stage approach. The first stage will focus on covariate balancing by computing non-parametric optimal weights. The goal is to create a pseudo-population where the covariate

distributions are independent of the PM2.5 exposure level, thus controlling for confounding without relying on a specific model form.

In the second stage, we will estimate the ADRF by fitting a flexible, spline-based regression model on this newly weighted pseudo-population. This approach is superior to the fixed-knot hinge model, as it allows the data to determine the complex, nonlinear shape of the response curve. This framework also allows us to investigate potential effect modification by estimating stratified ADRFs for different socioeconomic groups (e.g., by poverty level) to see if the causal effect of PM2.5 varies across these contexts. This two-stage strategy offers greater robustness by separating the task of controlling for confounding from the task of modeling the outcome.

9 Acknowledgements

References

Feng, S., Huang, F., Zhang, Y., Feng, Y., Zhang, Y., Cao, Y., & Wang, X. (2023). The pathophysiological and molecular mechanisms of atmospheric PM2.5 affecting cardiovascular health: A review. *Ecotoxicology and Environmental Safety, 249*, 114444.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall/CRC.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370–384.

Takatsu, K., & Westling, T. (2025). Debiased inference for a covariate-adjusted regression function. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 87*, 33-55.

Wyatt, L. H., Peterson, G. C. L., Wade, T. J., Neas, L. M., & Rappold, A. G. (2020). The contribution of improved air quality to reduced cardiovascular mortality: Declines in socioeconomic differences over time. *Environment International, 136*, 105430.

Zhang, Y., Chen, Y., & Giessing, A. (2025). Nonparametric Inference on Dose-Response Curves Without the Positivity Condition. *arXiv:2405.09003v2 [stat.ME]*.

Vodonos, A., Awad, Y. A., & Schwartz, J. (2018). The concentration-response between long-term PM2.5 exposure and mortality; A meta-regression approach. *Environmental Research*, 166, 677–689. doi:10.1016/j.envres.2018.06.021.

Li, T., Zhang, Y., Wang, J., Xu, D., Yin, Z., Chen, H., Lv, Y., Luo, J., Zeng, Y., Liu, Y., Kinney, P. L., & Shi, X. (2018). All-cause mortality risk associated with long-term exposure to ambient PM2.5 in China: a cohort study. *The Lancet Public Health*, 3(10), e470–e477.

Shi, L., Zanobetti, A., Kloog, I., Coull, B. A., Koutrakis, P., Melly, S. J., & Schwartz, J. D. (2016). Low-Concentration PM2.5 and Mortality: Estimating Acute and Chronic Effects in a Population-Based Study. *Environmental Health Perspectives*, 124(1), 46–52. doi:10.1289/ehp.1409111.

Schwartz, J., Laden, F., & Zanobetti, A. (2002). The concentration-response relation between PM2.5 and daily deaths. *Environmental Health Perspectives*, 110(10), 1025–1029.