Modeling the Relationship Between $PM_{2.5}$ and County Mortality Rate

Zhenkun Xu, Shucheng Liu and Ruocheng Sun

December 4, 2025

1 Introduction

Exposure to fine particulate matter ($PM_{2.5}$) is a leading environmental risk factor for global cardiovascular health concern (1). The public health burden is significant, as extensive epidemiological studies have established a strong association between long-term exposure to $PM_{2.5}$ 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 (1). 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 (5) demonstrated that historical reductions in $PM_{2.5}$ 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 (5).

There are differing views regarding the statistical relationship between CMR and $PM_{2.5}$ 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 (10). In contrast, other studies report diminishing marginal effects at higher concentrations, suggesting that the slope of the concentration–response curve decreases as $PM_{2.5}$ levels rise (7; 8). A third perspective highlights that the association may become stronger once $PM_{2.5}$ concentrations surpass certain low-exposure thresholds, with evidence that long-term exposures at or above 6 $\mu g/m^3$ are associated with larger mortality effects than exposures below that level (9). Thus, using statistical models to investigate the shape and magnitude of the $PM_{2.5}$-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 $PM_{2.5}$ 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 $PM_{2.5}$ 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 $PM_{2.5}$ and CMR, while properly adjusting for this high-dimensional set of socioeconomic confounders.

2 Related Work

The statistical analysis of the health impacts of $PM_{2.5}$ 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 (1). 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 (3; 2). 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 (5). 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 $PM_{2.5}$ reductions while accounting for socioeconomic deprivation (5).

More advanced methodologies frame this problem as one of causal inference: estimating the causal dose-response curve (DRC) for a continuous treatment ($PM_{2.5}$ exposure) while adjusting for a set of covariates (such as the socioeconomic data in `County_RAW_variables.csv`) (4; 6). 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 (4). 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 (6).

3 Data Preprocessing

Two preprocessing steps are required to prepare the data for modeling. The first key challenge is the temporal mismatch between the annual $PM_{2.5}$ and crude mortality rate (CMR) data and the decadal covariates from the U.S. Census. To construct a unified cross-sectional dataset and minimize temporal inconsistencies, we restrict our analysis to data from the year 2000. This alignment ensures that the 2000 treatment and outcome variables are paired with the complete set of covariates from the 2000 U.S. Census.

Table 1: Summary Statistics for the 2000 Cross-Sectional Dataset

Variable	Mean	SD	Min	Q1	Median	Q3	Max
PM2.5	7.96	2.09	2.34	6.84	8.34	9.38	13.02
CMR	355.91	64.28	151.22	311.08	351.71	395.91	658.01
civil_unemploy	3.49	1.34	0.50	2.70	3.30	4.10	28.00
median_HH_inc	36799.98	9265.55	16271.00	30758.00	35255.50	40917.00	82929.00
femaleHH_ns_pct	10.97	3.75	3.70	8.60	10.20	12.30	36.40
vacant_HHunit	12.21	8.36	1.50	6.90	10.00	14.60	73.20
owner_occ_pct	73.24	7.86	19.60	69.30	74.60	78.40	89.90
eduattain_HS	34.33	6.78	11.70	30.20	34.40	38.70	53.00
pctfam_pover	10.18	5.52	1.00	6.40	9.00	12.70	47.40
population	125634.43	348365.97	493.00	22951.75	40904.00	95119.50	9519338.00

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. A summary of the aligned dataset is provided below.

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 $PM_{2.5}$ concentrations and the CMR, we remove observations with missing values in either $PM_{2.5}$ or CMR. Because the raw scatterplot exhibits substantial variability, we sort the data by $PM_{2.5}$ 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-$PM_{2.5}$ relationship.

Figure 1 presents the resulting scatterplot of CMR versus $PM_{2.5}$, along with the smoothed trend. The moving average indicates a clear overall upward trajectory: CMR tends to increase as $PM_{2.5}$ rises. More importantly, the rate of increase is not constant. At lower concentrations (approximately below 6 $\mu g/m^3$), the curve remains relatively flat. Between 6 and 8 $\mu g/m^3$, the slope increases noticeably. However, beyond 8 $\mu g/m^3$, the trend appears to level off. These visible changes in curvature provide preliminary evidence of potential threshold effects.

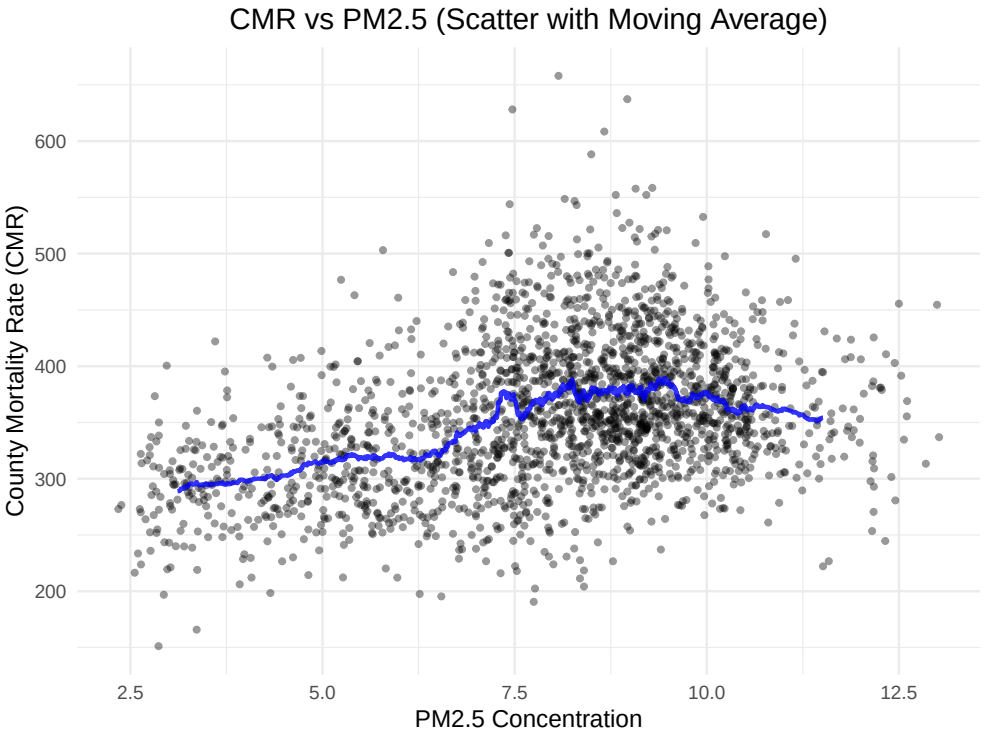

Figure 1: Scatterplot of CMR vs. $PM_{2.5}$ with 100-point centered moving average (blue), used to visualize the underlying trend in the presence of noise.

Based on the patterns observed in the smoothed trend, the changes in curvature motivate the use of hinge-squared terms at 6 and 8 $\mu g/m^3$. Incorporating the nonlinear components $(x-6)_+^2$ and $(x-8)_+^2$ enables the fitted curve to bend upward more sharply once $PM_{2.5}$ rises above 6 $\mu g/m^3$, and to potentially flatten after 8 $\mu g/m^3$, 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 ex-

pected, 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 (pctfam_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.

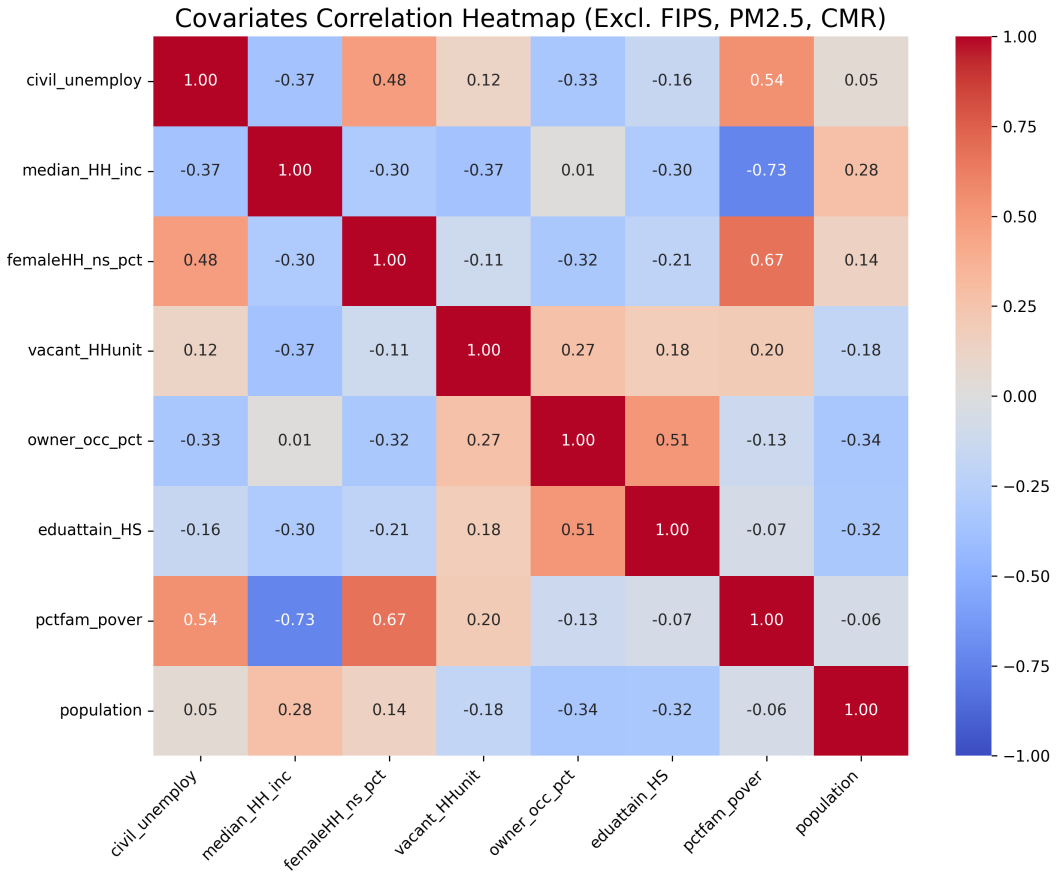

Figure 2: Correlation heatmap of socioeconomic and demographic covariates. The strong blue and red blocks indicate high multicollinearity, motivating the need for dimensionality reduction.

5 Preliminary Analysis

We develop a regression model that allows the association between $PM_{2.5}$ and CMR to vary across different ranges of pollution exposure. The smoothed trend in Figure 1 suggests distinct changes around 6 and 8 $\mu g/m^3$ (subject to change), indicating that the marginal effect of $PM_{2.5}$ 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. In this section, we first examine the performance of a direct hinge model to assess its ability to capture the observed pattern. A more rigorous statistical investigation is presented later in Section 6.

Let x_i denote the $PM_{2.5}$ 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:

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

We fit the preliminary linear model that incorporates hinge-squared terms at 6 and 8 $\mu g/m^3$. 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 $PM_{2.5}$ affects mortality differently across pollution ranges.

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.

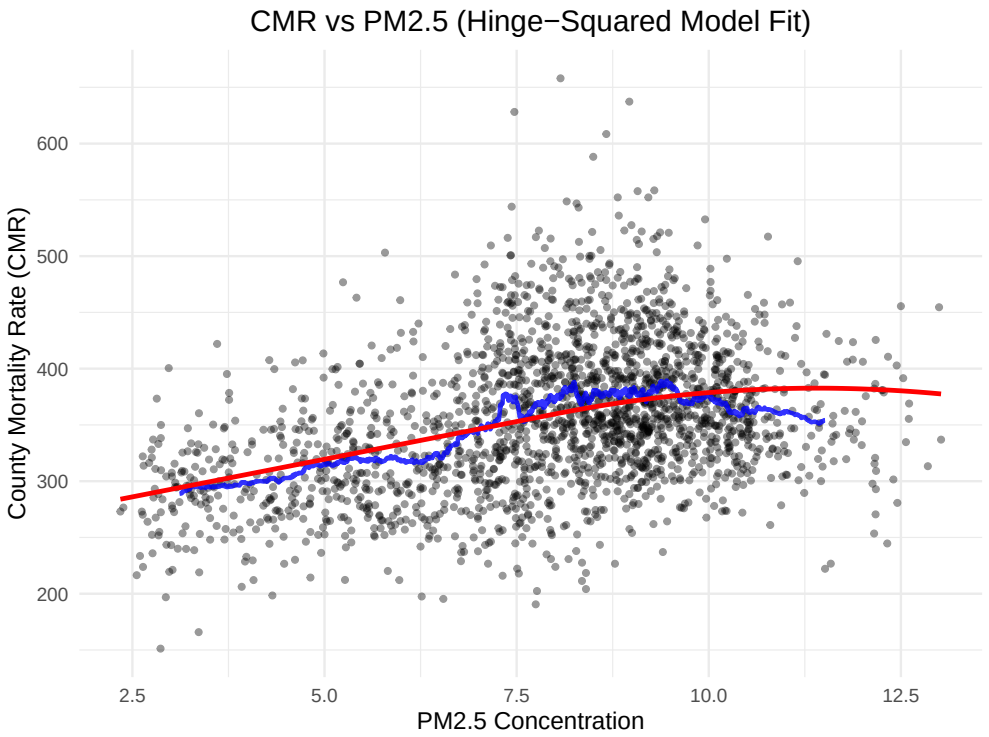

Figure 3: SPreliminary hinge-model fit using fixed breakpoints at 6 and 8 $\mu g/m^3$.

6 Hinge-Based Modeling Framework

In this section, we formally conduct the statistical analysis using a hinge-modeling framework, incorporating several additional components, including a detailed justification for adopting hinge models, a full model-selection analysis, and hypothesis testing procedures. To characterize the association between long-term exposure to ambient $PM_{2.5}$ and CMR, we develop a comprehensive modeling strategy that combines nonlinear exposure modeling, information-theoretic model selection, penalized covariate regularization, and assessment of effect modification. This multi-component framework is designed to achieve three key objectives: (i) providing sufficient flexibility to capture potential nonlinearities in the $PM_{2.5}$–CMR relationship, (ii) ensuring parsimony and robustness against overfitting, and (iii) maintaining interpretability for epidemiological inference.

6.1 Rationale for Using Hinge Models

In Section 2, we noted that a growing body of environmental epidemiology research suggests that the association between $PM_{2.5}$ and mortality may not be linear across the full exposure range. In particular, when $PM_{2.5}$ concentrations are relatively low, the relationship tends to be approximately linear, whereas at higher concentrations the rate of increase becomes more

gradual. Traditional linear specifications may therefore obscure such structural features, potentially resulting in biased estimates of health effects.

In Section 4, we visualized this relationship using a scatter plot, as shown in Figure 1. The empirical pattern aligns well with findings in the existing literature: the CMR initially increases nearly linearly with $PM_{2.5}$, and then the growth begins to flatten at higher concentration levels.

To allow for flexible yet interpretable nonlinearities, we considered hinge-squared functions with two candidate change points. Hinge functions are attractive because they allow abrupt changes in curvature while still producing piecewise smooth exposure–response curves. Compared to high-degree splines, hinge models retain interpretability through explicit breakpoints that can correspond to regulatory standards or biological thresholds.

We included hinge terms of the form

$$(\mathrm{PM}_{2.5} - k)_+^2 ,$$

which allow the curvature to increase smoothly above each breakpoint, capturing potential accelerations in risk. The quadratic form avoids discontinuities in derivatives and therefore reflects plausible dose–response structures in environmental health.

6.2 Primary Variable Model Selection

6.2.1 Selection of Optimal Breakpoints for the Hinge Model

In the preliminary analysis, we identified the values 6 and 8 as potential breakpoints based solely on the visual patterns observed in Figure 1. Although this provides an intuitive indication that structural changes may occur near these values, directly fixing the breakpoints at 6 and 8 would be somewhat hasty. A more rigorous approach is to examine a set of candidate breakpoints and determine the optimal specification through model selection, which improves the robustness of our conclusions.

To implement this procedure, we evaluated a grid of candidate change points:

$$k_1 \in \{5.5, 6.0, 6.5\}, \qquad k_2 \in \{7.5, 8.0, 8.5\},$$

which produced nine hinge models in total. Each of these nine models incorporated the same set of socioeconomic covariates (unemployment rate, median household income, percentage of female single-householder families, number of vacant housing units, owner-occupied housing rate, high-school attainment rate, poverty rate, and population), allowing direct comparability across model specifications.

To determine the best-fitting hinge specification, we computed Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all nine models. AIC prioritizes goodness-of-fit with moderate penalization for model size, while BIC imposes a stronger penalty, favoring simpler models under large sample sizes. Using both metrics provides complementary assessments and helps avoid overfitting.

Table 2: Comparison of Hinge Models with Varying Breakpoints

model	lower_knot	upper_knot	AIC	BIC
hinge_6_7.5	6.00	7.50	22601.38	22675.02
hinge_5.5_7.5	5.50	7.50	22601.39	22675.03
hinge_6.5_7.5	6.50	7.50	22601.45	22675.09
hinge_5.5_8	5.50	8.00	22601.68	22675.32
hinge_6_8	6.00	8.00	22601.68	22675.33
hinge_6.5_8	6.50	8.00	22601.68	22675.32
hinge_6.5_8.5	6.50	8.50	22602.00	22675.64
hinge_6_8.5	6.00	8.50	22602.12	22675.77
hinge_5.5_8.5	5.50	8.50	22602.17	22675.82

As shown in Table 2, the hinge model with breakpoints at $(k_1, k_2) = (6, 7.5)$ achieved the lowest AIC and BIC among all candidate specifications, indicating superior performance under both fit-based and complexity-penalized criteria. Models with nearby breakpoint combinations produced similar values, suggesting a relatively stable optimum around this region of the parameter space. The $(6, 7.5)$ configuration, however, consistently provided the strongest evidence of improved fit without unnecessary model complexity. Given this balance between explanatory power and parsimony, this breakpoint pair was selected as the primary exposure structure for all subsequent analyses.

6.2.2 Comparison Between the Hinge Model and Spline Alternatives

Under the assumption that a hinge specification is appropriate, our model selection procedure identified the optimal pair of breakpoints. However, we were still concerned that the hinge structure might be overly restrictive and that a smoother functional form could potentially capture additional nuances in the exposure–response relationship. To further assess model adequacy, we extended our model selection framework to include comparisons with natural cubic spline models. In particular, we fitted spline specifications with degrees of freedom

$$\text{df} \in \{3, 4, 5, 6\},$$

while keeping the adjustment set identical to that used in the hinge specifications. This ensured a direct and fair comparison between the hinge model with knots at $(6, 7.5)$ and a sequence of increasingly flexible spline models. The model selection results, reported in Table 3, summarize the AIC and BIC values across all candidate specifications.

Table 3: Comparison of Hinge and Spline Models

Model	AIC	BIC
hinge_6_7.5	22601.38	22675.02
spline_df_3	22601.63	22675.28
spline_df_4	22603.35	22682.65
spline_df_5	22605.22	22690.19
spline_df_6	22607.05	22697.69

By evaluating both AIC and BIC across all models, we found that none of the spline specifications outperformed the hinge model with knots at $(6, 7.5)$. Although spline models introduce additional smoothness and flexibility to the exposure–response curve, their AIC and BIC values were consistently slightly higher, indicating no improvement in model fit.

These results suggest that the hinge model provides an adequate and parsimonious characterization of the association, capturing the overall pattern of an initial increase followed by a more gradual rise. Moreover, its performance is comparable to, and in some cases slightly better than, that of smoother spline alternatives. Taken together, this evidence supports the hinge specification as a reasonable and empirically justified modeling framework in our analysis.

6.3 LASSO-Based Covariate Selection

To obtain a stable and interpretable model for estimating the association between $PM_{2.5}$ exposure and CMR, we applied the LASSO (Least Absolute Shrinkage and Selection Operator) to select socioeconomic covariates while preserving the full exposure–response structure. The design matrix included $PM_{2.5}$, two hinge terms at 6 and 7.5 $\mu g/m^3$ to capture potential nonlinearities, and a set of socioeconomic indicators such as unemployment, income, education, poverty, and housing characteristics. These SES variables are often correlated, and including all of them without regularization may introduce multicollinearity, inflate variance, and reduce the stability of the estimated exposure effect.

LASSO addresses this issue by shrinking coefficients of less relevant predictors toward zero, thereby removing redundant covariates and mitigating multicollinearity among socioeconomic variables. Because our analytic goal is to estimate the $PM_{2.5}$ exposure–response

function without distortion, the penalty factor for $PM_{2.5}$ and the two hinge terms was set to zero so that these exposure terms were never penalized or excluded. Only the socioeconomic covariates were subject to LASSO selection. We used ten-fold cross-validation to determine the optimal tuning parameter $\lambda_{\min}$ and then refit the model at this value to obtain the final coefficients.

Table 4: LASSO coefficients at $\lambda_{\min}$

Variable	**Coefficient**
(Intercept)	250.88
$PM_{2.5}$	13.51
$(PM_{2.5} - 6)_+^2$	0.54
$(PM_{2.5} - 7.5)_+^2$	-2.53
civil_unemploy	0.00
median_HH_inc	-14.00
femaleHH_ns_pct	14.19
vacant_HHunit	0.00
owner_occ_pct	3.61
eduattain_HS	6.62
pctfam_pover	10.23
population	-1.10

Covariates with nonzero coefficients were retained, forming the final adjustment set $\mathcal{S}$. With these selected variables, the final model takes the form

$$\text{CMR} = \beta_0 + \beta_1 \text{PM}_{2.5} + \beta_2 (\text{PM}_{2.5} - 6)_+^2 + \beta_3 (\text{PM}_{2.5} - 7.5)_+^2 + \sum_{X_j \in \mathcal{S}} \gamma_j X_j.$$

This procedure yields a more stable and interpretable confounder set by removing socioeconomic variables whose contributions to explaining variation in CMR were negligible after accounting for the remaining predictors. In particular, `civil_unemploy` and `vacant_HHunit` received coefficients of exactly zero. Both variables are highly correlated with stronger SES indicators already included in the model—for example, poverty rate, income, and education tend to capture the underlying socioeconomic environment more directly and with less measurement noise. After these more informative covariates were accounted for, unemployment and housing vacancy contributed little additional independent variability in CMR, leading LASSO to shrink their coefficients to zero. Excluding such redundant predictors reduces multicollinearity and avoids overadjustment while preserving the integrity of the $PM_{2.5}$ exposure–response relationship.

Interestingly, some pairs of covariates that are often considered strongly correlated, such

as poverty rate and median income, were both retained. This suggests that each variable may capture a distinct aspect of socioeconomic disadvantage. For instance, poverty rate reflects concentrated deprivation at the lower tail of the income distribution, whereas median income provides a more continuous measure of overall economic resources. Their joint retention indicates that these measures provide complementary information relevant for confounding control rather than fully redundant signals.

6.4 Model Diagnostics

We assessed the adequacy of the LASSO-selected model using the standard suite of regression diagnostics, shown in Figure 4. The residual–fitted plot (top left) displays a cloud of residuals centered around zero with no discernible systematic pattern, suggesting that the linear functional form is appropriate and that no major model misspecification is present. Although the smoothing curve shows a slight downward trend at higher fitted values, the deviation is small relative to the overall residual scale and does not indicate meaningful nonlinearity.

The normal Q–Q plot (top right) shows that the standardized residuals closely follow the theoretical normal line over most of the distribution, with moderate deviations in the upper tail. These departures are limited to a small number of observations and do not substantially undermine the approximate normality assumption used for inference. Overall, the distribution of residuals is reasonably compatible with Gaussian errors.

The scale–location plot (bottom left) shows that the square root of the standardized residuals remains roughly constant across the range of fitted values. While a mild upward trend is visible, the spread of the residuals is broadly stable, providing no strong evidence of heteroskedasticity. This supports the assumption of constant variance across levels of the predictors included after LASSO selection.

Finally, the residuals–leverage plot (bottom right) reveals that all observations fall well below the Cook's distance thresholds, indicating the absence of points exerting disproportionate influence on the fitted model. A few data points exhibit moderate leverage, but their standardized residuals remain small, implying that they do not materially affect the estimated coefficients.

Taken together, these diagnostics provide strong support for the adequacy of the linear modeling assumptions in the specification obtained after LASSO covariate selection.

We further evaluated multicollinearity among the retained covariates using variance inflation factors (VIFs). The two hinge-based $PM_{2.5}$ terms show high VIF values, which is expected given their overlapping construction and does not indicate problematic multi-

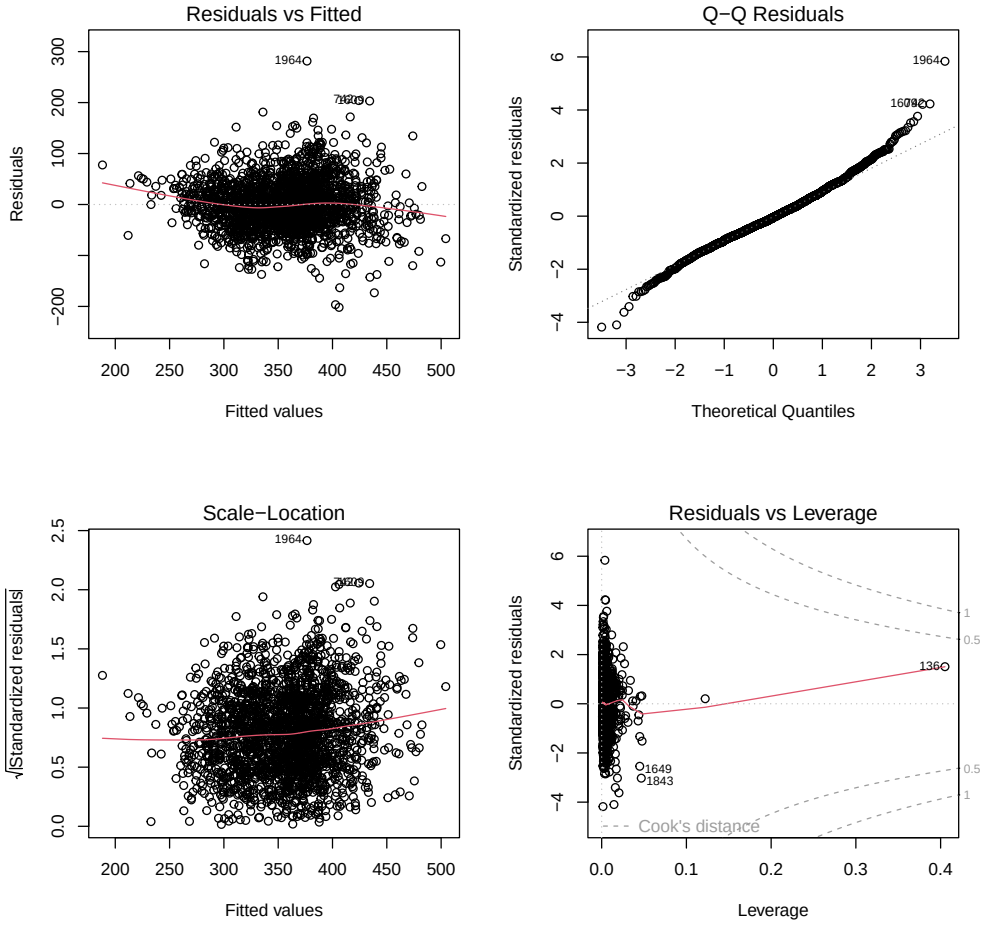

Figure 4: Residual diagnostic plots for the LASSO-selected linear model

collinearity. Focusing on the remaining socioeconomic covariates, we observed moderately elevated VIFs for median household income and the poverty rate. This is intuitive, as income and poverty are naturally related measures of socioeconomic status. However, their VIF values remain within an acceptable range and the two variables capture distinct aspects of community conditions; therefore, both were retained in the final specification. All other covariates exhibit low VIF values, confirming that the overall model does not suffer from concerning multicollinearity.

6.5 Analysis of Post-LASSO Hinge Model

Having established through diagnostic analysis that the post-LASSO hinge model reasonably satisfies the standard assumptions of linear regression, we now examine the fitted exposure–response relationship between $PM_{2.5}$ and CMR. The fitted curve is shown in Figure 5.

Table 5: VIFs for covariates in the LASSO-selected model

Covariate	**VIF**
$PM_{2.5}$	8.68
$(PM_{2.5} - 6)_+^2$	66.99
$(PM_{2.5} - 7.5)_+^2$	40.08
median_HH_inc	4.45
femaleHH_ns_pct	2.96
owner_occ_pct	1.76
eduattain_HS	2.22
pctfam_pover	5.33
population	1.30

The estimated exposure–response curve exhibits an approximately linear increase in CMR at lower $PM_{2.5}$ concentrations. As $PM_{2.5}$ continues to rise, however, the slope of the curve gradually decreases, indicating a clear pattern of diminishing marginal effects at higher concentration levels, consistent with our expectation.

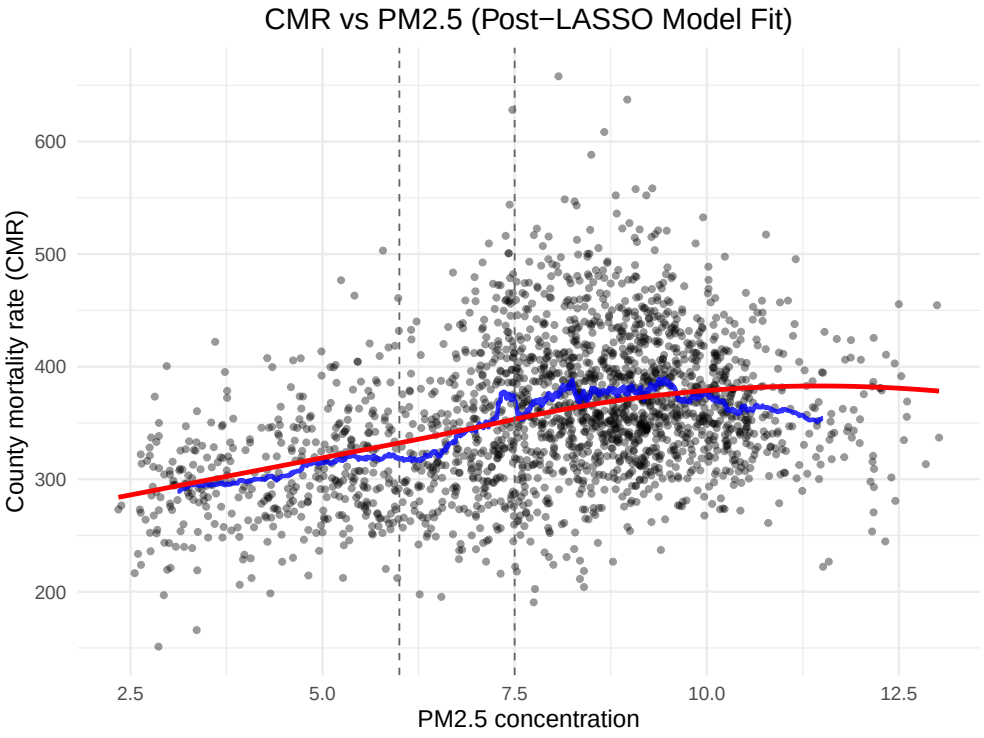

Figure 5: CMR vs. $PM_{2.5}$ with smoothed trend and post-LASSO fitted curve

6.5.1 Individual Coefficient Hypothesis Tests

Next, we evaluate the statistical significance of the covariates retained in the final specification. For each predictor X_j, we test the null hypothesis $H_0 : \beta_j = 0$ against the two-sided alternative $H_1 : \beta_j \neq 0$, and report the estimated coefficients, standard errors, p-values, and 95% confidence intervals. The results of these individual coefficient tests are presented in Table 6.

Table 6: Individual Coefficient Hypothesis Tests for the post-LASSO hinge model

Term	Estimate	Std. Error	CI Lower	CI Upper	p-value	Significant
$PM_{2.5}$	13.21	1.47	10.32	16.10	0.00	Yes
$(PM_{2.5} - 6)_+^2$	0.56	1.04	-1.48	2.61	0.59	No
$(PM_{2.5} - 7.5)_+^2$	-2.44	1.55	-5.47	0.59	0.11	No
median_HH_inc	-14.25	2.21	-18.58	-9.92	0.00	Yes
femaleHH_ns_pct	15.15	1.80	11.62	18.68	0.00	Yes
owner_occ_pct	4.21	1.39	1.49	6.93	0.00	Yes
eduattain_HS	6.95	1.56	3.89	10.01	0.00	Yes
pctfam_pover	10.02	2.42	5.28	14.76	0.00	Yes
population	-1.43	1.19	-3.77	0.91	0.23	No

The results in Table 6 reveal clear contrasts in the statistical relevance of the predictors. The linear $PM_{2.5}$ term shows a large and precisely estimated positive effect, with a narrow confidence interval and a p-value well below conventional thresholds, indicating that $PM_{2.5}$ levels are strongly associated with the outcome even after accounting for socioeconomic controls. In comparison, the two nonlinear $PM_{2.5}$ terms fail to reach significance, as their confidence intervals encompass zero and the associated p-values remain relatively large. This suggests that, conditional on the linear component, there is little evidence supporting additional curvature or threshold behavior at the chosen cut-off points.

Socioeconomic characteristics also show systematic and interpretable relationships with CMR. Median household income exhibits a significant negative association, consistent with the idea that higher-income communities typically benefit from better access to healthcare, healthier living environments, and reduced exposure to chronic stressors that are known to mitigate cardiovascular risks.

In contrast, factors such as the share of female-headed households without a spouse, owner-occupancy rates, educational attainment, and poverty levels all display significant positive associations with CMR. These patterns may reflect different underlying mechanisms: a higher prevalence of single-parent households may indicate greater socioeconomic vulnerability and elevated chronic stress; higher owner-occupancy rates may correlate with demographic compositions, such as older or long-term resident populations, who face in-

herently higher cardiovascular risks; higher educational attainment may capture community structures or occupational profiles associated with sedentary lifestyles or other risk-related behaviors; and elevated poverty rates often correspond to reduced healthcare access, higher exposure to environmental hazards, and cumulative disadvantages that aggravate cardiovascular outcomes.

By contrast, population size does not show a significant effect, suggesting that once socioeconomic and environmental factors are accounted for, simple differences in scale do not systematically relate to CMR.

6.5.2 Joint Hypothesis Test for Hinge Terms

To assess whether the two hinge terms jointly improve model fit, we compared the full hinge model with a reduced model excluding both hinge components while retaining all socioeconomic covariates. The F-test results are summarized in Table 7.

Table 7: Joint F-test for nonlinear hinge terms in the $PM_{2.5}$ model

Model	df	RSS	Δdf	ΔRSS	F	p-value
Reduced model (linear $PM_{2.5}$ only)	2124	4,998,109.52				
Full hinge model (two hinge terms)	2122	4,955,406.32	2	42,703.20	9.14	0.00011

The hinge model achieves a reduction in the residual sum of squares of approximately 42703 with two additional degrees of freedom. This result is statistically significant ($F(2, 2122) = 9.14$, $p = 0.00011$), indicating that the hinge components collectively capture nonlinear structure in the relationship between $PM_{2.5}$ and CMR. Although each hinge term is individually significant, the joint F-test provides a stricter evaluation and confirms that these nonlinear effects remain important when considered together. This finding suggests that a purely linear model cannot adequately describe the association.

Although the joint test does not rule out the possibility that the hinge specification may not be the optimal nonlinear representation, the presence of nonlinear structure indicates that other nonlinear models, such as smooth functions or flexible nonpolynomial basis representations, may also be considered to provide a more complete characterization of the relationship between $PM_{2.5}$ and CMR.

6.6 Effect-Modification Assessment

Having established the primary exposure–response structure and the LASSO-selected adjustment set, we next evaluated whether the association between $PM_{2.5}$ exposure and CMR

varies across socioeconomic contexts. Understanding heterogeneity in pollution effects is essential for identifying vulnerable subpopulations and ensuring equitable environmental health policies.

For each SES variable Z selected by the LASSO procedure, we assessed effect modification of the association between $PM_{2.5}$ and county-level mortality by fitting the following interaction model:

$$\text{CMR} = \beta_0 + \beta_1 \, \text{PM}_{2.5} + \beta_2 Z + \delta \, (\text{PM}_{2.5} \times Z) + \text{hinge terms} + \text{other SES covariates}.$$

The coefficient δ represents the modification of the marginal effect of $PM_{2.5}$ across levels of Z. A positive value of δ indicates that the adverse effect of $PM_{2.5}$ becomes stronger in counties with higher levels of the SES indicator. Conversely, a negative value suggests enhanced vulnerability in counties with lower SES levels. By adjusting for all non-focal SES covariates in each model, the estimated interaction effect reflects the independent moderating role of Z, rather than confounding by correlated socioeconomic variables.

For each fitted model, we extracted the point estimate of the interaction term, its confidence interval, and the corresponding p-value. SES indicators with statistically significant interaction effects ($p < 0.05$) were considered potential modifiers of the $PM_{2.5}$–CMR relationship. The results are summarized in Table 8.

Table 8: Interaction estimates for $PM_{2.5}$ and SES indicators

SES indicator	δ	**CI lower**	**CI upper**	p**-value**
median_HH_inc	−0.92	−2.03	0.19	0.10
femaleHH_ns_pct	1.78	0.69	2.87	0.00
owner_occ_pct	1.82	0.91	2.74	0.00
eduattain_HS	−0.74	−1.73	0.25	0.14
pctfam_pover	3.28	2.18	4.38	0.00
population	−1.40	−2.34	−0.46	0.00

Significant positive interaction terms were observed for the percentage of female-headed households, the proportion of owner-occupied housing, and the poverty rate. These patterns suggest that $PM_{2.5}$-related mortality effects become stronger in counties with higher levels of these SES characteristics. Counties with more female-headed households may include populations that are more vulnerable to environmental stressors, making them more susceptible to the health impacts of pollution. Higher owner-occupied housing rates may reflect more stable, long-term residency, resulting in greater cumulative exposure to local air quality conditions. The strong positive interaction with poverty rate is consistent with environmental justice evidence: communities with higher poverty burdens often have fewer health resources

and poorer baseline health, amplifying the adverse effects of pollution.

In contrast, the interaction term for population size was significantly negative, indicating that the mortality effects of PM2.5 are relatively weaker in more populous counties, which are typically urban or more highly urbanized areas. This pattern may reflect the fact that counties with larger populations are often located in urban environments with better-developed infrastructure, including more abundant medical resources, easier access to healthcare, and more comprehensive public health systems, which collectively help buffer the health risks associated with air pollution. In comparison, sparsely populated areas tend to have fewer medical facilities, longer travel distances for care, and an older population structure, making them more vulnerable when exposed to PM2.5. Other SES indicators showed certain interaction patterns but did not reach statistical significance, providing only limited evidence of effect modification.

To illustrate effect modification more concretely, we plotted model-predicted CMR across the observed $PM_{2.5}$ range at the 10th, 50th, and 90th percentiles of each SES variable, holding all other covariates at their sample means so that differences across curves reflect only the SES modifier. The resulting patterns are shown in Fig. 6.

Across the socioeconomic indicators examined, the clearest evidence of effect modification appears for the proportion of female-headed households and the proportion of families in poverty. For both indicators, the differences between SES levels widen substantially as $PM_{2.5}$ increases. Areas at the 90th percentile exhibit a notably steeper rise in CMR, reaching higher peak values and doing so at slightly larger $PM_{2.5}$ concentrations. In contrast, the 10th-percentile areas display a flatter upper tail, with predicted CMR leveling off or declining earlier.

For the percentage of owner-occupied housing, the degree of effect modification is weaker but still detectable. The three curves remain relatively close, yet they differ in both shape and magnitude. Areas with lower owner- occupancy tend to have slightly higher CMR at lower $PM_{2.5}$ levels, whereas at moderate to higher $PM_{2.5}$ concentrations the curve for high owner-occupancy areas becomes marginally higher.

For population size, the curves almost completely overlap. Both the slopes and overall shapes are highly similar across SES levels, indicating minimal effect modification. However, this may partly reflect the distribution of population size: although the full range spans from very small to extremely large populations, the 10th, 50th, and 90th percentiles used for visualization are relatively concentrated and do not capture the most extreme contrasts, reducing the apparent separation between the curves.

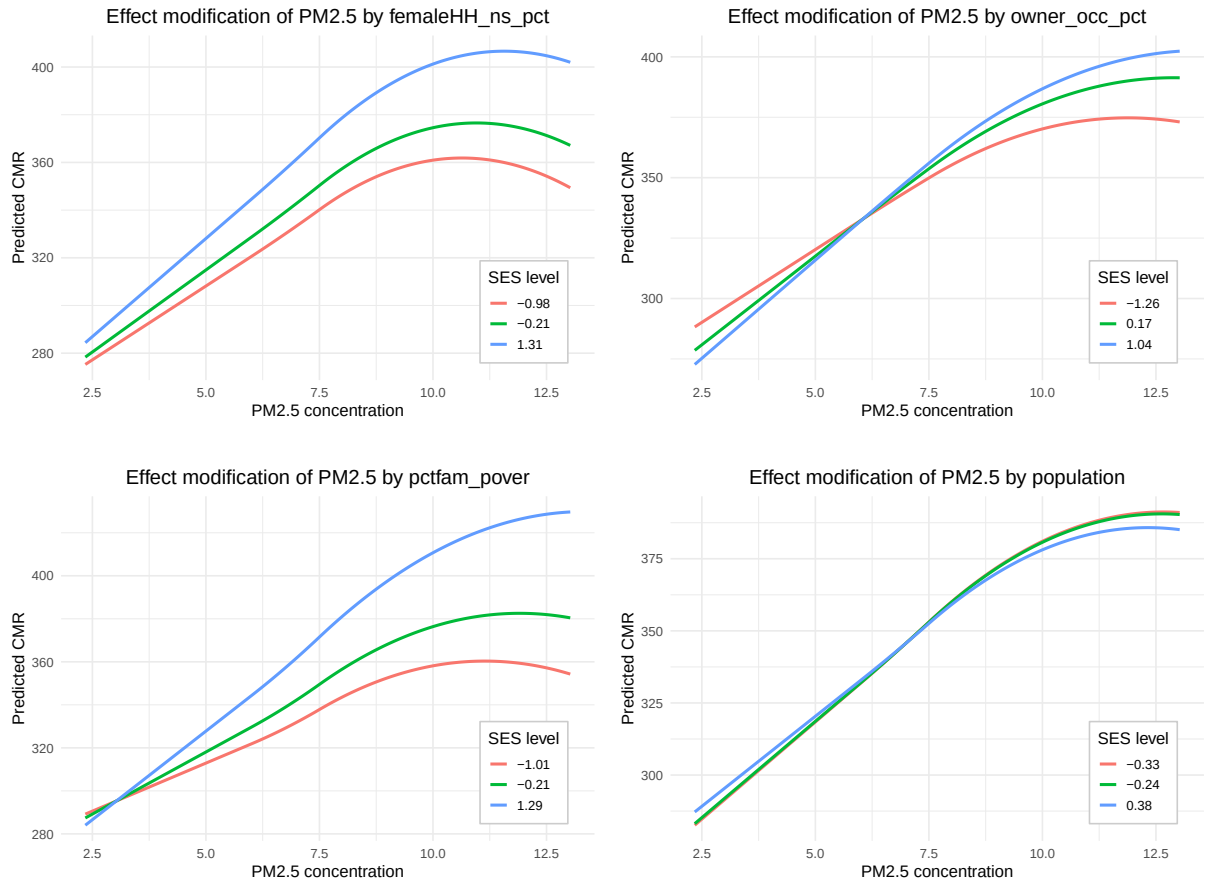

Figure 6: $PM_{2.5}$–CMR by SES Levels

7 Causal Effect Estimation using VCNet

7.1 Motivation: From Association to Causality

The statistical modeling framework presented in Section 1, including the hinge-squared regression and LASSO covariate selection, effectively characterizes the conditional expectation of the CMR given $PM_{2.5}$ exposure and socioeconomic factors. However, standard regression methods primarily capture statistical associations and do not necessarily account for the confounding bias inherent in observational data. To support policy interventions, we aim to estimate the causal effect of long-term $PM_{2.5}$ exposure on mortality, formally defined as the Average Dose-Response Function (ADRF).

Unlike discrete treatments, $PM_{2.5}$ concentration is a continuous variable. Traditional methods often discretize the continuous treatment into bins, estimating effects separately for each interval. This approach, however, often leads to discontinuous ADRF estimates and loss of information within bins. To address these limitations and estimate a smooth, continuous dose-response curve, we employ the Varying Coefficient Neural Network (VCNet)

(11).

7.2 Notation and Assumptions

We adopt the potential outcomes framework for continuous treatments. Let $T \in \mathcal{T} = [0, 1]$ denote the continuous treatment variable (normalized $PM_{2.5}$ concentration), $Y \in \mathbb{R}$ denote the outcome (CMR), and $X \in \mathcal{X} \subset \mathbb{R}^d$ denote the vector of observed confounding covariates (e.g., unemployment rate, income, educational attainment).

We define the potential outcome $Y(t)$ as the mortality rate that would be observed if a county were exposed to $PM_{2.5}$ level t. Our target estimand is the Average Dose-Response Function (ADRF), defined as:

$$\psi(t) = \mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid X, T = t]]. \tag{1}$$

To identify $\psi(t)$ from observational data, we rely on two standard assumptions:

1. **Unconfoundedness:** The treatment assignment is independent of potential outcomes given the covariates, i.e., $Y(t) \perp T \mid X$ for all $t \in \mathcal{T}$. The rich set of socioeconomic and housing characteristics selected in our previous step supports the validity of this assumption.

2. **Overlap (Positivity):** The generalized propensity score $\pi(t|x)$ is bounded away from zero, meaning that for any covariate profile x, there is a positive probability density of receiving treatment level t.

7.3 The VCNet Methodology

We utilize VCNet to model the generalized propensity score $\pi(t|X)$ and the outcome regression function $\mu(t, x) = \mathbb{E}[Y \mid T = t, X = x]$. A key innovation of VCNet is its ability to preserve the continuity of the ADRF without discretizing the treatment.

As illustrated in the network structure (see Figure 7), VCNet employs a "varying coefficient" head for the outcome prediction. Instead of treating the treatment T as a simple input feature which might be lost in high-dimensional representations, VCNet defines the weights of the neural network as functions of T. Specifically, the outcome prediction head is defined as $f_{\theta(t)}(Z)$, where Z is a representation of covariates X, and the parameters $\theta(t)$ are modeled using B-splines:

$$\theta(t) = \sum_{k=1}^{K} \beta_k \phi_k(t), \tag{2}$$

where $\phi_k(\cdot)$ are spline basis functions. This structure ensures that the estimated causal effect changes smoothly with respect to the $PM_{2.5}$ concentration, yielding a continuous and differentiable ADRF.

Training Objective

To learn the parameters effectively, VCNet employs a joint training strategy. The network is trained by minimizing a composite loss function that balances outcome prediction accuracy with the estimation of the generalized propensity score. The objective function is defined as:

$$\mathcal{L}[\mu^{NN}, \pi^{NN}] = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu^{NN}(t_i, x_i))^2 - \frac{\alpha}{n}\sum_{i=1}^{n}\log(\pi^{NN}(t_i|x_i)), \tag{3}$$

where the first term represents the mean squared error (MSE) for the outcome prediction μ^{NN}, and the second term is the negative log-likelihood for the conditional density estimator π^{NN}. The hyperparameter α controls the trade-off between these two objectives.

This joint optimization is crucial because the conditional density estimator π^{NN} serves a dual purpose: it estimates the propensity score required for bias correction and helps extract informative latent features Z from the covariates X. These shared features are then utilized by the outcome prediction head μ^{NN}, preventing the loss of treatment information in high-dimensional latent spaces and improving the estimation of the causal effect.

Estimation of ADRF

Once the model is trained, the Average Dose-Response Function (ADRF) is estimated by averaging the predicted conditional expected outcomes over the empirical distribution of the covariates. Specifically, for any treatment level $t \in \mathcal{T}$, the estimated ADRF $\hat{\psi}(t)$ is given by:

$$\hat{\psi}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}^{NN}(t, x_i), \tag{4}$$

where $\hat{\mu}^{NN}(t, x_i)$ is the predicted outcome for unit i if they were assigned treatment t, based on their observed covariates x_i. This estimator effectively marginalizes over the confounding variables X to recover the population-level dose-response curve.

7.4 Empirical Results and Analysis

We applied the VCNet method to the U.S. county-level dataset. To ensure temporal alignment with the available Census covariates (unemployment rates, income, housing conditions,

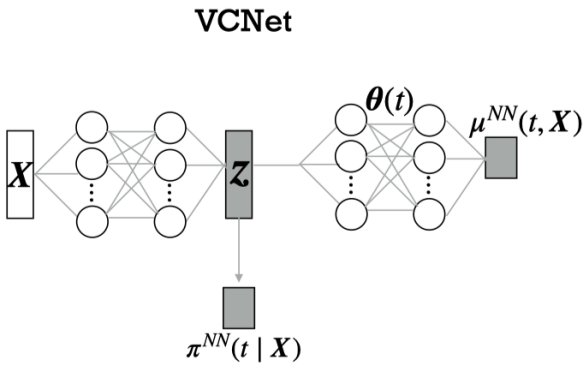

Figure 7: The Varying Coefficient Neural Network (VCNet) architecture. The model ensures continuity in the estimated dose-response curve by allowing the network weights $\theta(t)$ to vary smoothly as a function of the treatment t.

etc.), we focused on treatment and outcome data from the year 2000. The data was randomly split into training and testing sets with a 2:1 ratio.

The estimated ADRF curve representing the causal relationship between $PM_{2.5}$ concentration and Cardiovascular Mortality Rate (CMR) is presented in Figure 8.

Analysis of the Dose-Response Relationship

The estimated ADRF reveals a distinct non-monotonic relationship, characterized by an "inverted-U" pattern, which offers nuanced insights compared to simple linear associations:

- **Low-to-Moderate Exposure:** In the lower range of $PM_{2.5}$ concentrations, the CMR increases steadily with exposure. This aligns with standard epidemiological evidence suggesting that clean air is protective and that initial increases in particulate matter correlate with higher cardiovascular risk.

- **Peak Risk:** The mortality risk appears to peak at a concentration around 10 $\mu g/m^3$. This suggests that the marginal harm of $PM_{2.5}$ is most pronounced leading up to this threshold.

- **High Exposure and Saturation:** Beyond 10 $\mu g/m^3$, the curve begins to decline gradually. This counter-intuitive trend in highly exposed regions may be attributed to a "saturation effect," where the marginal biological harm of additional pollution diminishes at high levels.

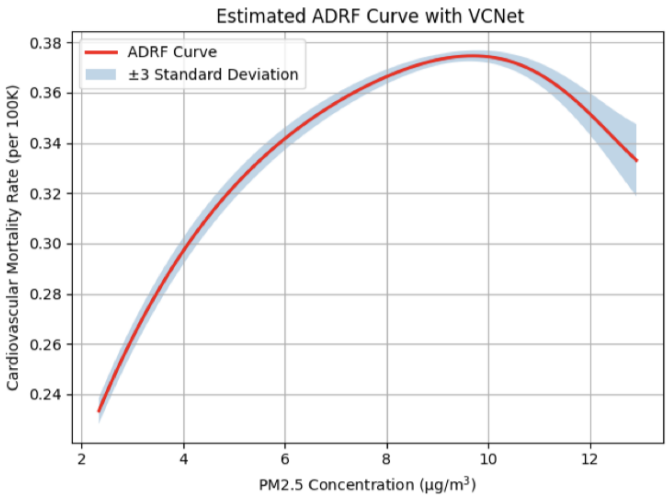

Figure 8: Estimated Average Dose-Response Function (ADRF) using VCNet. The curve illustrates the causal effect of $PM_{2.5}$ concentration ($\mu g/m^3$) on Cardiovascular Mortality Rate (per 1,000 individuals).

Potential Mechanisms

The observed decline at high concentrations warrants careful interpretation. One plausible explanation is the presence of compensatory mechanisms. Counties with higher pollution levels are often more urbanized and may possess more robust healthcare infrastructure or higher socioeconomic status compared to rural, lower-pollution areas. These factors could buffer the adverse health impacts of pollution, leading to lower observed mortality rates despite high exposure.

Alternatively, this trend may be partially driven by unmeasured confounders that vary regionally and influence both pollution exposure and health outcomes. However, by employing VCNet, which provides a flexible, bias-corrected estimator capable of modeling complex nonlinear dependencies, we avoid the oversimplifications of parametric models that force monotonic effects. This highlights the necessity of using advanced causal inference methods to capture the true structural nature of environmental health risks.

8 Future Directions and Potential Extensions

There remain several meaningful opportunities to extend and refine this study. First, because our analysis focuses on the year 2000 to align treatment, outcome, and covariates, future work can incorporate the full longitudinal structure of the 1990 to 2010 dataset. This would

permit mixed effects modeling, temporal causal inference, and dynamic estimation of dose response relationships, providing a more complete view of long term exposure effects.

Second, although hinge based modeling performs well and captures core nonlinear patterns, more flexible approaches such as Gaussian process regression, Bayesian nonparametric methods, or shape constrained estimators can be explored to validate the stability of the inferred exposure response structure. These alternatives could reveal additional smoothness or curvature not captured by the current specification.

Finally, the effect modification results suggest notable heterogeneity across socioeconomic contexts. Future research can examine this heterogeneity more deeply through varying coefficient models or subgroup specific dose response analyses. In addition, expanding the covariate set to include environmental, behavioral, and healthcare access measures may further strengthen causal identification and improve the robustness of the estimated causal effect of $PM_{2.5}$ on cardiovascular mortality.

Conclusion

This project examines the relationship between long term $PM_{2.5}$ exposure and cardiovascular mortality using hinge based regression, covariate selection, and causal inference. Through model comparison using AIC and BIC, the hinge specification with breakpoints at 6 and 7.5 provides the best balance between fit and interpretability, performing as well as or better than spline alternatives. LASSO regularization identifies a focused set of socioeconomic covariates, reduces multicollinearity, and yields a stable post selection model.

The final regression results show a strong positive association between $PM_{2.5}$ and mortality, with the linear component highly significant. Although the individual hinge terms are not significant, the joint F test indicates that nonlinear structure still contributes meaningfully to model performance. Several socioeconomic variables also modify pollution effects, suggesting that vulnerability varies across communities and that environmental risks are not evenly distributed.

To estimate causal effects, we applied VCNet and obtained a smooth Average Dose Response Function. The estimated curve shows an inverted U shape, with mortality increasing at lower and moderate $PM_{2.5}$ levels and flattening or declining at higher exposures. This highlights the complexity of pollution effects and the importance of flexible causal methods.

Acknowledgements

We used Chatgpt and gemini to help us refine our code in R and LaTeX. We discussed together for ideas and experiments. Zhenkun Xu and Shucheng Liu contributed to the data, experiments, and report writing. Ruocheng Sun contributed to the slides and presentation. We all contributed to the github branches.

References

[1] Feng, S., Huang, F., Zhang, Y., Feng, Y., Zhang, Y., Cao, Y., & Wang, X. (2023). The pathophysiological and molecular mechanisms of atmospheric $PM_{2.5}$ affecting cardiovascular health: A review. *Ecotoxicology and Environmental Safety, 249*, 114444.

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

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

[4] 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.

[5] 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.

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

[7] Vodonos, A., Awad, Y. A., & Schwartz, J. (2018). The concentration-response between long-term $PM_{2.5}$ exposure and mortality; A meta-regression approach. *Environmental Research*, 166, 677–689. doi:10.1016/j.envres.2018.06.021.

[8] 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 $PM_{2.5}$ in China: a cohort study. *The Lancet Public Health*, 3(10), e470–e477.

[9] Shi, L., Zanobetti, A., Kloog, I., Coull, B. A., Koutrakis, P., Melly, S. J., & Schwartz, J. D. (2016). Low-Concentration $PM_{2.5}$ and Mortality: Estimating Acute and Chronic Effects in a Population-Based Study. *Environmental Health Perspectives*, 124(1), 46–52. doi:10.1289/ehp.1409111.

[10] Schwartz, J., Laden, F., & Zanobetti, A. (2002). The concentration-response relation between $PM_{2.5}$ and daily deaths. *Environmental Health Perspectives*, 110(10), 1025–1029.

[11] Nie, L., Ye, M., Liu, Q., & Nicolae, D. (2021). VCNet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*.