EXPLORING THE CAUSAL EFFECTS BETWEEN CO AND PM2.5: A STUDY ON HARMFUL GAS POLLUTION

Mao Keyu
School of Data Science
Fudan University
Shanghai, China
20307130241.fudan.edu.cn

June 10, 2023

ABSTRACT

In this article, we will explore the causal effects of common air pollutants such as CO and NO2 on the level of PM2.5 pollution. We will begin by providing an overview of the dataset to understand its distribution and make decisions regarding data selection and preprocessing (e.g., encoding object types into numerical values). Based on prior knowledge and empirical assumptions, we will construct a DAG (Directed Acyclic Graph) to depict the causal relationships. Using the information provided by the Causal DAG, we will initially perform simple regression analysis. Subsequently, under specific assumptions, we will employ propensity score modeling for causal inference, which involves techniques such as matching, IP weighting, and stratification. We will conduct sensitivity analysis to assess the contribution of each covariate to the interpretability of the model. Finally, guided by the reasonable assumptions made during the construction of the DAG, we will select an appropriate instrumental variable and perform causal inference using an IV (Instrumental Variable) model. We will compare the interpretability of the different models employed. By the end of our study, we aim to provide insights into the causal effects between the categorization of the harmful gas CO pollution and the level of PM2.5 pollution. We believe that our analysis results are robust to a certain extent, based on the reasonable assumptions initially incorporated into the DAG.

Keywords Causal Inference · Air Pollution · Regression · Propensity Score · Instrumental Variable

1 Introduction

Air pollution refers to the presence of harmful substances in the Earth's atmosphere, resulting in the degradation of air quality. It is a significant environmental issue that affects the health of living organisms, damages ecosystems, and contributes to climate change.

To combat air pollution, governments and organizations around the world are implementing various measures. For the government, it is crucial to have a measure of pollution levels. For instance, in daily life, we use the AQI value to assess the degree of air pollution(1). We have observed that the AQI value of PM2.5 plays a vital role in determining the overall pollution level. Therefore, examining the causal relationship between relevant factors such as the emission levels of harmful gases like NO2 and CO, and the concentration of PM2.5, can assist us in gaining a better understanding of global pollution conditions and making significant contributions towards environmental remediation in the future.

2 Dataset Overview and Problem Description

2.1 Dataset Introduction

In this section, we will introduce the dataset we decide to carry out causal analysis. As we mentioned in the last section, we are interested in the air quality and the factors which may have causal relationship with air quality.

Global Air Pollution Dataset provides geolocated information about the several pollutants. This dataset compiles the AQI values of the following harmful gases:

- **Nitrogen Dioxide [NO2]** : Nitrogen Dioxide is one of the several nitrogen oxides. It is introduced into the air by natural phenomena like entry from stratosphere or lighting.
- **Ozone [O3]** : The Ozone molecule is harmful for outdoor air quality (if outside of the ozone layer). At surface level, ozone is created by chemical reactions between oxides of nitrogen and volatile organic compounds .
- **Carbon Monoxide [CO]** : Carbon Monoxide is a colorless and odorless gas. Outdoor, it is emitted in the air above all by cars, trucks and other vehicles or machineries that burn fossil fuels.
- **Particulate Matter [PM2.5]** : Atmospheric Particulate Matter, also known as atmospheric aerosol particles, are complex mixtures of small solid and liquid matter that get into the air.

Among them, the dataset encompasses air quality indicators from over 20,000 cities in 175 countries worldwide. The AQI value, which is gained by considering all the relevant factors, is also available in the dataset. The data was provided by elichens, making it reliable in scientific experiment.

2.2 Data overview

In this part, we will manage to analyze the data and try to do some simple visualization on the distribution of the data. In the dataset, besides the exact value of AQI, the category(a value among Good', 'Moderate', 'Unhealthy', 'Unhealthy for Sensitive Groups', 'Very Unhealthy', 'Hazardous') of each pollutant is also offered. When exploring distribution, we mainly focus on specific value. The density of each value is shown in figure 1. And as shown in figure 2, Box plot by features and categories can help us learn the relationship between category and AQI value.

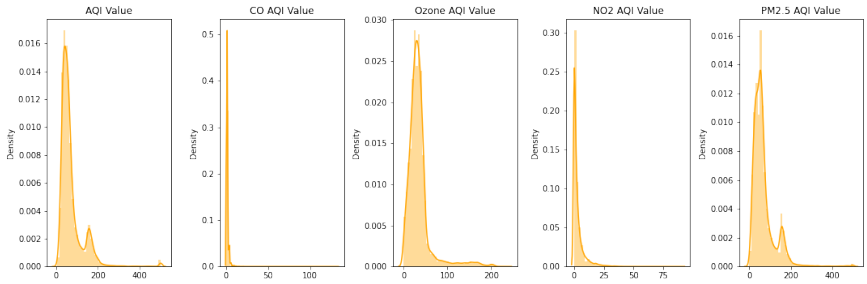

Figure 1: Density figure of AQI

From the box plot, we know that the absolute value of AQI may not make sense when building model, since the standard for category varies among different pollutants.

2.3 DAG, Assumptions and Problem definition

In the task, we set the AQI value of PM2.5 as outcome factor, but the choice of treatment factor is confusing so we first consider the DAG. We can build the DAG through prior knowledge. For instance, we can use a certain model like linear regression to evaluate the relation among factors. However, in that case, the DAG will be extremely complex (figure3) since tiny relation will also be detected as relation. Hence, we decide to use some empirical to make assumptions and simplify the DAG:

- Assumption: Specific value of AQI for each pollutant will only directly affect the result of the category. This make Specific value of AQI an additional information of category. This makes sense since most of time, the approximated level of pollution can describe the air level perfectly.
- City will be moved out of the dataset. This makes sense since samples are collected from different city. The function of city equals to index.
- We ignore some relation and pay little attention to integrated AQI value, which is regraded to have strong connection to AQI value of PM2.5.

Then,we can drive the DAG which will be used in this project in figure 4.

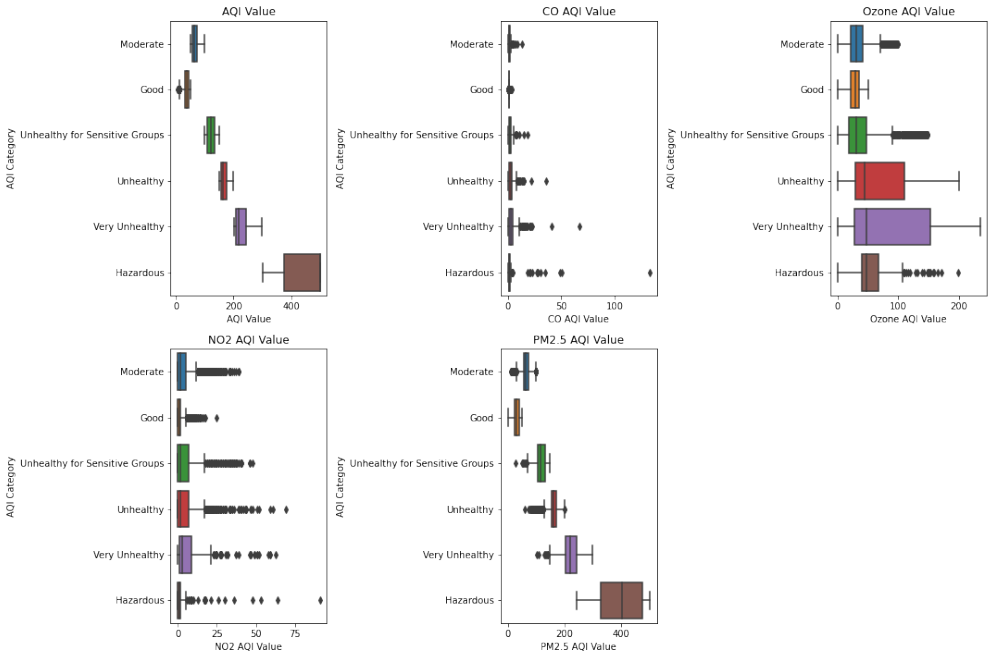

Figure 2: Box plot by features and categories

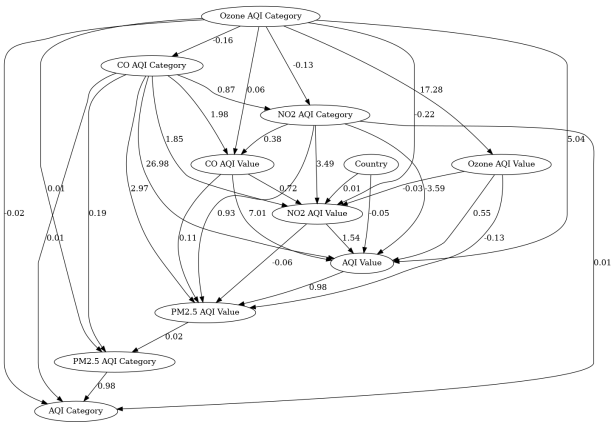

Figure 3: Prior DAG

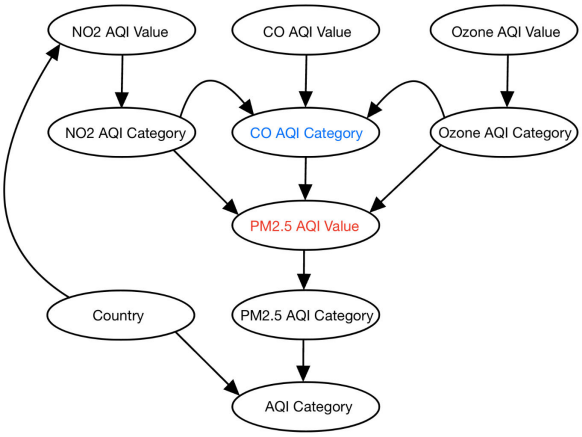

Figure 4: Simplified DAG

For all factors whose data type are object, we manually encode them for further numerical analysis. Specific method is based on the quantile and distribution of the data and can be checked in my coding files.

We notice that CO AQI Category is suitable for treatment variable. Specifically, the value of CO AQI Category doesn't change greatly and about 75% of values stay no more than 1.0. In that case, we can use 1.0 as a boundary and make it a binary value which indicates that 0:slight pollution, 1:severe pollution. Besides, through prior knowledge in figure 3, we know that CO AQI Category has relatively strong relationship between the AQI value of PM2.5.

In conclusion, the problem can be described as:

- **Treatment**: CO AQI Category
- **Outcome**: PM2.5 AQI Value
- **Covariates**: CO AQI Category, NO2 AQI Category, Ozone AQI Category,PM2.5 AQI Category, NO2 AQI Value,Ozone AQI Value, Country

- **Counfounders**: NO2 AQI Category, Ozone AQI Category,NO2 AQI Value,Ozone AQI Value, Country

In the following parts, **we use "val" to denote "AQI Value", and "cate" to denote "AQI Category"**. Furthermore, **we use P,C,N,O,Co to denote PM2.5, CO, NO2, Ozone and Country ID**.

3 Regression analysis

3.1 Regression model

In this section, we will apply regression analysis for causal effect estimation with a sensitivity analysis for unmeasured confounding. Determining the powers of the variables in linear regression is a crucial task prior to fitting the data. In order to gain an initial understanding of the approximate polynomial relationship between the regression terms and the outcome variable, it is advisable to commence with the creation of a pair plot and a heatmap(Figure 5, Figure 6).

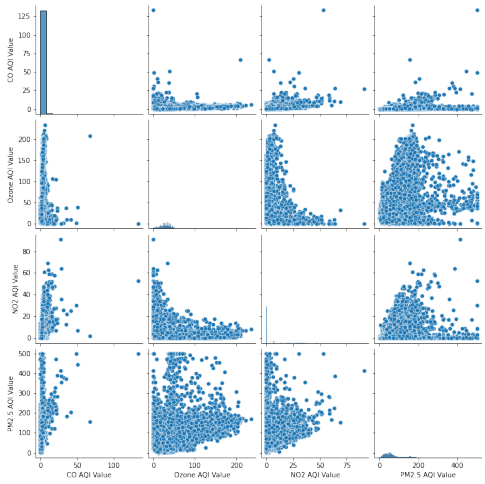

Figure 5: Pair plot

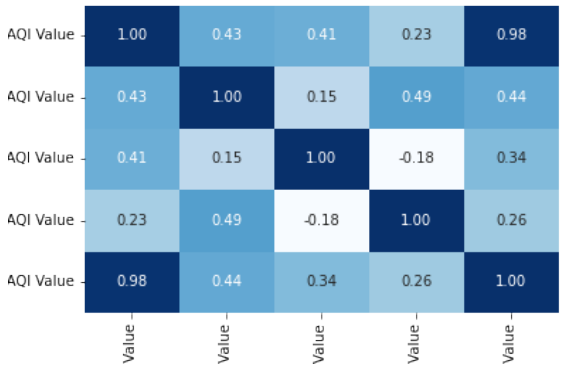

Figure 6: Heat Map

Clearly, the relation of PM2.5-NO2 and PM2.5-CO are hard to be interpreted linearly. Therefore, we introduce their quadratic terms to help describe the model. The final term of our regression model is:

$$\begin{aligned} P_{val} = 40.38C_{cate} + 1.98N_{val} + 0.01V_{val}^2 - 2.21N_{cate} + 0.83O_{val} \\ - 0.001O_{val}^2 + 2.59O_{cate} + 0.08Co + 32.81 \end{aligned} \tag{1}$$

Obviously, in the linear model, C_{cate}(CO AQI Category) plays the most important role, which is the same as our prior knowledge. But it's not enough for us to conclude causal relation. On the other hand, Though 2 quadratic terms and Country ID have rather low coefficient value, this may result from the scale of data and will not affect the status of C_{cate}. The R square of model is 0.2882, indicating simple regression cannot fit the model very well.

3.2 Sensitivity analysis

3.2.1 Conventional Method

The purpose of sensitivity analysis is to understand the robustness and reliability of the model or system by examining the sensitivity of the output to different input scenarios. It helps in identifying which input variables have the most significant influence on the output and how changes in those variables can affect the overall results.

We try following strategies:

1. Remove quadratic terms in the regression model
2. Remove all val term(specifc value) and Country ID
3. Remove all confounders

Coefficients	C_{cate}	N_{val}	O_{val}	N_{cate}	O_{cate}	Co	N_{val}^2	O_{val}^2	R sqaure
Original	40.38	1.98	0.83	-2.21	2.59	-0.08	0.01	-0.001	0.2884
Remove Quadratic	39.87	2.47	0.45	-3.75	5.65	-0.08	0	0	0.287
Remove val and Co	56.41	0	0	3.32	11.3	0	0	0	0.2448
Remove Confounders	57.48	0	0	0	0	0	0	0	0.1965

Table 1: Sensitivity Test

The result is shown in table 1

The coefficient of determination (R-squared) is employed to assess the model's interpretative capacity. In summary, the inclusion of the quadratic term exerts minimal influence on the regression model, albeit resulting in a slight improvement. Notably, the categorical representation of each pollutant can effectively describe the model even after removing precise numerical values. This aligns with the earlier assertion that approximate values suffice for the majority of air level descriptions. Ultimately, even in the absence of confounding variables, the model retains its interpretative capacity, underscoring the significance of the treatment variable, denoted as C_{cate}.

Additionally, propensity score model can help us to further carry out sensitivity test due to its interpretation on treatment and covariates(2). With the help of regression model and propensity model, we can estimate the effect of treatment under a number of different assumptions about the direction and magnitude of unmeasured confounding. We can then generate a sequence of values for α to determine the range of possible magnitude of ignorability violations.

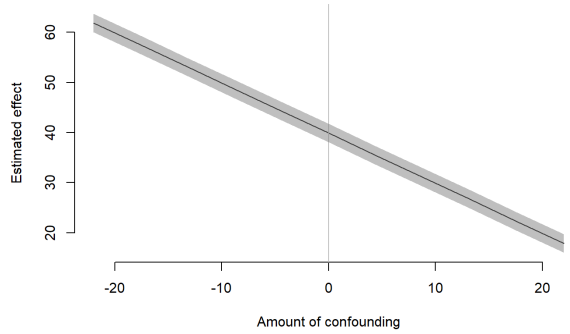

Figure 7: Sensitivity Analysis

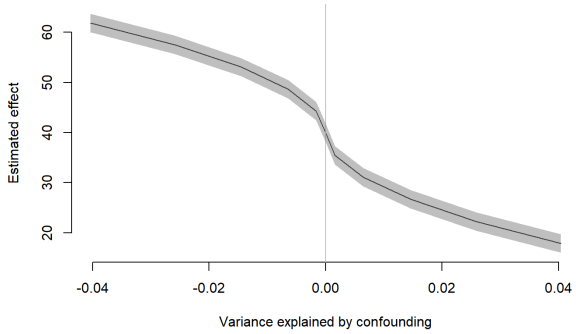

Figure 8: Sensitivity Analysis in terms of variance

From Figure 7 and Figure 8, we can see the difference between the raw confounding in terms of the magnitude of ignorability violations and the sensitivity in terms of variance explained by confounding. In addition, we can compare the strength of confounding against the variance explained by the covariates.

3.2.2 E-Value for Sensitivity Analysis

Conventional sensitivity analysis is subject to subjectivity: researchers can choose sensitivity parameters to make their results more robust. And Evaluation of assumptions required: The essence of sensitivity analysis is to evaluate whether changing the assumptions on which the study is based would lead to changes in the research results. However, conducting sensitivity analysis often requires new assumptions and possesses complexity: many sensitivity analysis methods are highly complex, and many researchers struggle to effectively utilize and articulate the results.

In this subsection, we decide to report E-Value, which is defined as the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured confounding by VanderWeele and Ding in 2016(3). E-Value has several strengths compared to conventional methods including:

- No assumptions required (provides a conservative result).
- Simple calculations
- Reduces subjectivity

In order to calculate E-Value, we first consider observed Risk Ratio(RR):

$$RR_c^{obs} = \frac{P(Outcome = 1|Explosure = 1, C = c)}{P(Outcome = 1|Explosure = 1, C = c)} \tag{2}$$

where C denotes Confounders. Then E-value can be computed by:

$$\text{E-Value} = \begin{cases} RR + \sqrt{RR \times (RR - 1)} & RR > 1 \\ 1/RR + \sqrt{1/RR \times (1/RR - 1)} & RR < 1 \end{cases} \tag{3}$$

The E-value quantifies the sensitivity of the study results to unmeasured confounding factors. A higher E-value indicates more robust results, suggesting that even in the presence of unmeasured confounders, the study conclusion is unlikely to change. Conversely, a lower E-value indicates greater susceptibility to the influence of unmeasured confounding factors, requiring stronger evidence to support the stability of the conclusions. We report our results of regression model in table 2.

Table 2: Point Estimate of Point Estimate

	Point Estimate	Lower Bound(95%)
RR	1.95	1.89
E-Value	3.31	3.19

E-Value cannot directly conclude the causal effect, but it can help us to confirm whether our model is robust enough. In this case, we set the threshold of E-value as 2, indicating that unmeasured confounder(s) would need to double the probability of a subject's having exposure equal to c + delta instead of c, where c is an arbitrary value(4).

4 Estimation with Propensity Score

4.1 Propensity score model

In this section, we will use several methods: Matching, IP Weighting and Stratification to estimate the ATE(Average Treatment Effect). All of these method are based on propensity score model.

Since we have to balance our covariates in this task, we tend to choose covariates with discrete value. And in the last tasks, we have concluded that:"val" has strong connection with "cate"(based on prior knowledge and assumption) and approximated pollutant level(here we use category) can almost describe the model. Hence, in this task, we focus on balance covariates exclude specific value.

First of all, we build a propensity score model based on dataset and visualize it in Figure 9. The distribution of the propensity score ensures that for practically every interval, there are exposure samples and control samples that can be matched.

4.2 Matching

Assumptions in this part(C denote confounders):

1. Consistency: $P_{val} = C_{cate}P^1_{val} + (1 - C_{cate})P^0_{val}$
2. SUTVA: $P_{val,i}(C_{cate,1}, \cdots, C_{cate,n}) = P_{val,i}(C_{cate,i})$
3. NUCA:$C_{cate,i} \perp (P^1_{val}, P^0_{val})|C$
4. Positivity(Overlap): $0 < P(C_{cate,i} = 1|C_i = c) < 1$

4.2.1 Matching Without Replacement

Matching in causal inference is a technique used to reduce or eliminate the effects of confounding variables when estimating the causal impact of a treatment or intervention(5). It involves identifying and pairing treated and control units that are similar in terms of their observed characteristics, thus creating a more comparable comparison group.

We match control group data for treatment group to estimate ATT(average treatment effect for treatment group) and then match treatment data to estimate ATC(average treatment effect for control group). The matched results are shown in 3.

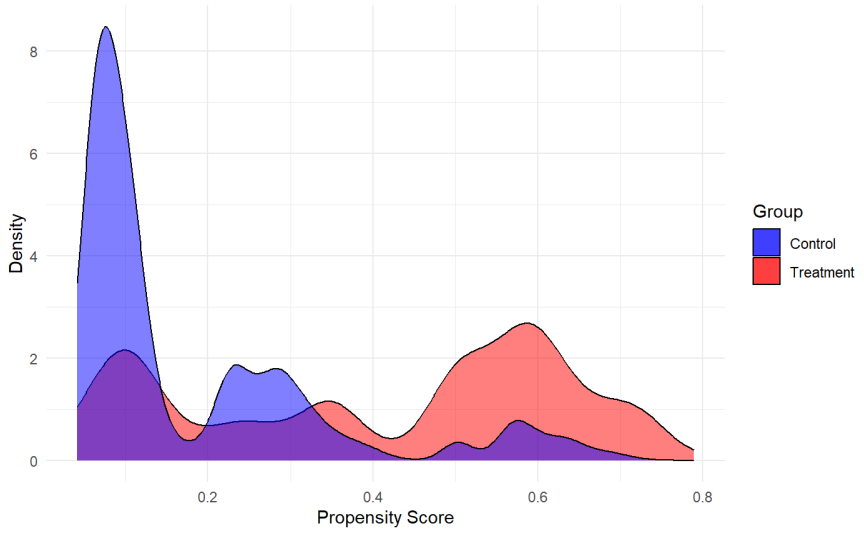

Figure 9: Propensity score distribution

Control	Treated	
All	17601	5434
Matched	5434	5434
Unmatched	12167	0

	SMD(O_{cate})	SMD(N_{cate})	SMD(Co)	ATT/ATC
ATT	0.015	0.006	0.004	44.95
ATC	0.002	0.032	0.025	69.15

Table 3: Matching Results

From table 3, we can find that the smd of each covariate is lower than 0.1, indicating we balance covariates successfully. And 5434 pairs of data are used in the test, which is enough for a splendid results. By the eqation below we compute sample ATE:

$$\begin{aligned} \hat{ATE} &= \frac{N_{treat}}{N}\hat{ATT} + (1 - \frac{N_{treat}}{N})\hat{ATC} \\ &= 63.44 \end{aligned} \tag{4}$$

4.2.2 Matching With Replacement

In the experiment, matching with replacement refers to the process of pairing individuals from a treated group with individuals from a control group, allowing for the possibility of matching the same control individual with multiple treated individuals. This means that control group individuals can be used as matches for multiple treated individuals.

Additionally, a caliper of 0.1 is employed in the matching process. This means that the difference in propensity scores between the matched pairs should not exceed 0.1 standard deviations of the propensity score distribution. We simply follow the procedure of Matching Without Replacement, and results are shown in Table 4

Control	Treated	
All	17601	5434
Matched	1585	5403
Unmatched	16016	31

	SMD(O_{cate})	SMD(N_{cate})	SMD(Co)	ATT/ATC
ATT	0.028	0.024	0.008	27.66
ATC	0.004	0.018	0.004	39.62

Table 4: Matching Results

And the estimated ATE can be computed as well:

$$\begin{aligned} \hat{ATE} &= \frac{N_{treat}}{N}\hat{ATT} + (1 - \frac{N_{treat}}{N})\hat{ATC} \\ &= 36.8 \end{aligned} \tag{5}$$

There are several reasons that may lead to the difference between 2 matching methods. Main factors include availability of suitable matches, bias introduced by replacement, differences in the distribution of covariates and methodological considerations like the set of caliper size.

4.3 IP Weighting

Assumptions in this part(C denote confounders):

1. NUCA:$C_{cate,i} \perp (P_{val}^1, P_{val}^0)|C$
2. Positivity: $0 < P(C_{cate,i} = 1|C_i = c) < 1$

Inverse probability weighting is a method used to address confounding and estimate causal effects(6). It involves assigning weights to observations based on their propensity scores, which are the predicted probabilities of receiving a particular treatment given the observed covariates.

The purpose of inverse probability weighting is to create a pseudo-population where the distribution of the covariates is balanced between the treated and control groups. In the task, we use the inverse of propensity scores to reweight the population and compute the sample ATE based on new population. Then we have:

$$\begin{aligned} \hat{ATE} &= \frac{\sum_{i=1}^{n} \text{Treatment}_i \times \text{Outcome}_i \times \text{IPW weight}_i}{\sum_{i=1}^{n} \text{Treatment}_i \times \text{IPW weight}_i} - \frac{\sum_{i=1}^{n} (1 - \text{Treatment}_i) \times \text{Outcome}_i \times \text{IPW weight}_i}{\sum_{i=1}^{n} (1 - \text{Treatment}_i) \times \text{IPW weight}_i} \\ &= 49.54 \end{aligned} \tag{6}$$

4.4 Stratification

1. NUCA(conditional exchangeability):$C_{cate,i} \perp (P_{val}^1, P_{val}^0)|C$
2. Positivity: $0 < P(C_{cate,i} = 1|C_i = c) < 1$

conditional exchangeability Stratification involves dividing the study population into distinct strata or groups based on the values of a specific covariate or a combination of covariates(7).

The purpose of stratification is to create homogeneous subgroups within the population, where individuals within each stratum share similar characteristics. The procedure of Stratification is similar to Matching. We have:

$$\begin{aligned} \hat{ATE}_s &= \frac{1}{n_t} \sum_{i=1}^{n_t} Y_{t,i} - \frac{1}{n_c} \sum_{i=1}^{n_c} Y_{c,i} \\ \hat{ATE} &= \sum_s (\hat{ATE}_s \times \frac{N_s}{N}) \\ &= 55.67 \end{aligned} \tag{7}$$

where ATE stratum ATE_s represents the estimated average treatment effect within a specific stratum. n_t, n_c represent the sample sizes of the treated and control groups within the stratum, respectively. $Y_{t,i}, Y_{c,i}$ represent the outcome values for the treated and control units, respectively, within the specific stratum. In conclusion, the results in this section is shown in table 5. We can find that sample ATE values estimated by 3 methods is close, indicating we get rather reasonable result.

Method	Matching w/o	IP Weighting	Stratification
Sample ATE	63.44/36.8	49.54	55.67

Table 5: Estimated results

The results suggests that exposure to the treatment variable (in this case, the Carbon Monoxide Air Quality Index, CO AQI) leads to an average increase of $\hat{ATE}$ units in the outcome variable (Particulate Matter 2.5 Air Quality Index, PM2.5 AQI).

This means that, on average, individuals or areas exposed to higher levels of CO AQI experience a $\hat{ATE}$-unit increase in PM2.5 AQI compared to those with lower CO AQI levels. The magnitude of the effect may vary depending on the measurement scale and units used for the AQI indices.

4.5 Sensitivity Analysis

4.5.1 Conventional Methods

In this part we will do sensitivity analysis to test three models above. Our method is to evaluate the importance of each covariate by removing it and comparing. However, in this task, we don't have a ground truth of causal effect and we don't have criteria like R square for the regression model. As a consequence, we will first change the propensity score(which will lead to the change of all three models), and then have a comparison of the following results and the distribution of new propensity score.

We mainly focus on 2 strategies:

1. Remove covariate Co. According to our conclusion in the sensitivity analysis part of regression analysis, Co is supposed to contribute little to the model.
2. Follow strategy 1, then remove N_{cate}.
3. Follow strategy 1, then remove O_{cate}.

For 3 strategies, we fit our propensity model as shown in Figure 10,Figure 11 and Figure 12.

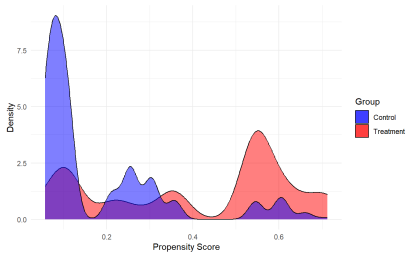

Figure 10: Propensity score-strategy 1

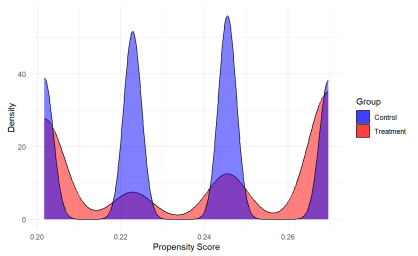

Figure 11: Propensity score-strategy 2

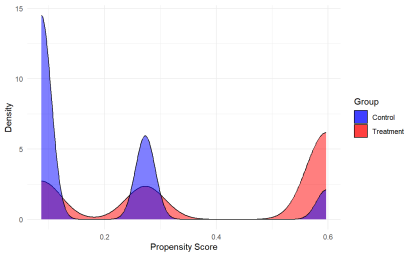

Figure 12: Propensity score-strategy 3

From new propensity score models and their distribution, we can find that the covariate Co has little influence on the propensity score model. Both strategy 2 and 3 will change the distribution of propensity score greatly. However, the density curves of propensity scores for the exposure group and control group, showing overlapping peaks and valleys in their positions under both strategies, provide favorable conditions for utilizing propensity score models in subsequent experiments, such as matching and stratification. The analysis results are shown in table 8

Table 6: Senstivity Analysis

Method	Matching	IP Weighting	Stratification
Original	63.44	49.54	55.6
Strategy 1	63.39	50.56	54.26
Strategy 2	46.4	59.93	55.96
Strategy 3	55.58	57.14	59

The fluctuations in ACE estimates can be attributed to the removal of covariates during sensitivity analysis. Covariates play a crucial role in adjusting for confounding variables and ensuring accurate causal inference. In the case of matching, the ACE estimates are more sensitive to the removal of covariates compared to weighting and stratification. Matching relies heavily on identifying similar units between the treatment and control groups, and the exclusion of certain covariates may disrupt the balance achieved through matching, resulting in more noticeable changes in the estimated ACE.

Meantime, we can conclude that covariate N_{cate} and O_{cate} contribute more to interpreting the model than covariate Co, indicating that regional difference is less significant than some pollutant levels.

4.5.2 E-Value For Sensitivity Analysis

Similar to Regression Analysis part, we also report the E-Value of our model in this section. The threshold of E-Value is set to 2, which is same as last part.

Table 7 demonstrates our results. Clearly, all the E-Value even the lower bound of E-Value(95%CI) is higher than our threshlold, meaning that the research results are relatively robust against unmeasured confounding factors. In other

Table 7: Results of E-Value For Sensitivity Analysis

Method	RR	RR Lower Bound	E-Value	E-Value Lower Bound
Matching With Rep	2.24	2.17	3.92	3.76
Matching Without Rep	1.47	1.49	2.3	2.14
IP Weighting	4.8	4.69	9.07	8.85
Stratifcation	2.52	2.46	4.49	4.35

words, even in the presence of unmeasured confounders, the study conclusion is unlikely to change significantly. This also indicates a certain level of reliability in the research findings and a higher level of stability in the conclusions.

5 Instrumental Variables

5.1 Choice of Instrumental Variable

Review the whole article, there's a covariate never uses in all models: C_{val}. Due to our assumption, the specific value of AQI will only directly have causal effect to it's own specific category. In that case, we consider the essential assumptions for instrumental variable:

1. Relevance: The instrumental variable(s) must be correlated with the treatment variable of interest (CO AQI in this case). This means that the instrumental variable(s) should have a significant association with the treatment variable and should affect the likelihood of receiving the treatment.
2. Exclusion: The instrumental variable(s) must not have a direct effect on the outcome variable (PM2.5 AQI) other than through its influence on the treatment variable. In other words, the instrumental variable(s) should only affect the outcome variable through its impact on the treatment variable and should not have any direct causal relationship with the outcome.
3. Independence: The instrumental variable(s) should be independent of any confounding factors that may affect the outcome variable.
4. Monotonicity: This assumption states that there are no "defiers" in the population—individuals who would receive the treatment if assigned to the control group and vice versa.

The assumption 1 and 2 are perfectly satisfied in our case(please refer to Figure 4, and Figure 13 shows the only causal path from C_{val} to P_{val}), and we make assumption 3 and 4 in the following experiment. Then, C_{val} is a splendid choice of instrumental variable.

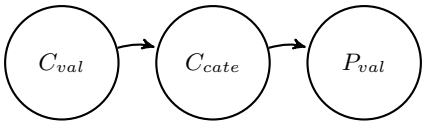

Figure 13: Causal path from C_{val} to P_{val}

5.2 IV models

5.2.1 OLS and 2SLS

The Two-Stage Least Squares (2SLS) model is a statistical technique used in econometrics to estimate the causal effect between an endogenous treatment variable and an outcome variable in the presence of endogeneity(8). The 2SLS model is often employed when the treatment variable is affected by omitted variables or measurement errors, leading to biased estimates in a simple regression framework.

In general, 2SLS consists of 2 stages. In the first stage, instrumental variables are used to estimate the endogenous treatment variable. The first stage involves regressing the endogenous treatment variable C_{cate} on the instrumental variables C_{val} to obtain the predicted values of the treatment variable, known as the "fitted values."

$$\hat{C}_{cate} = \alpha_0 + \alpha_1 C_{val} + \alpha_2^T C \tag{8}$$

where we use C to denote the rest of confounders.

In the second stage, the predicted values of the treatment variable from the first stage are used as the explanatory variable in the outcome equation. The outcome equation captures the relationship between the predicted treatment variable and the outcome variable of interest (here is P_{val}). Besides, we choose log function as our link functions to establish the relationship between endogenous treatment variable and outcome variable, namely:

$$\log(P_{val}) = \mu_0 + \mu_1 \hat{C}_{cate} + \mu_2^T C \tag{9}$$

Transform the regression result of μ_1 and we can get our ATE estimation.

Ordinary Least Squares (OLS) regression can be seen as a special case of the Two-Stage Least Squares (2SLS) model when there are no endogeneity concerns or instrumental variables involved.In OLS, the focus is on estimating the coefficients of the explanatory variables directly from the observed data. In this case, OLS model will be used for comparison.

5.2.2 GMM

In the context of instrumental variable analysis, the GMM(Generalized Method of Moments) model extends the two-stage least squares (2SLS) approach by incorporating moment conditions that capture the relationship between the instrumental variables, the endogenous treatment variable, and the outcome variable(9).

The GMM model involves the following steps:

1. Moment Conditions: Moment conditions are specified based on the instrumental variables and the relationship between the treatment variable and the outcome variable.
2. Objective Function: An objective function is defined, which measures the distance between the sample moments and the population moments.
3. Estimation: The GMM estimator iteratively adjusts the parameter estimates to minimize the objective function.
4. Inference: After estimating the model parameters, statistical inference can be performed to assess the significance and precision of the estimates.

Besides, we also test CUE-GMM, The Generalized Method of Moments estimation with continuously updating, which is an extension of the traditional GMM estimation method. CUE-GMM addresses the issue of parameter instability by incorporating the idea of recursively updating the parameter estimates as new data becomes available.

5.2.3 Results

After we fit those 3 models on our dataset, we can take a look at the final results. All three model using same instrumental variable reaches the same estimated ATE as we can see in figure 14. The fianl coefficients for each covariate and treatment variable are similar in 3 models, though there are some differences on 95% confidence interval each coefficient. After transformation, we can get our estimated ATE through IV method:

$$\hat{ATE}_{iv} = 32.94 \tag{10}$$

When both 2SLS and GMM estimators yield the same ATE estimate with a single IV, it suggests that the instruments used in both methods have a similar impact on the endogenous variable and, subsequently, on the outcome variable. This alignment of estimates between the two methods provides confidence in the consistency and robustness of the estimated ATE.

Additionally, When using the instrumental variable (IV) method to estimate the average treatment effect (ATE), the resulting estimates may be slightly smaller than those obtained using other methods such as matching. This can be attributed to the different characteristics of the instrumental variable method and matching methods in addressing causal inference problems. The instrumental variable method addresses endogeneity, which refers to the underlying interdependence between the treatment variable and the outcome variable, by utilizing an instrumental variable. In contrast, matching methods create a comparison group similar to a randomized experiment by selecting and treating individuals with similar characteristics.

We are also interested in the interpretative capability of IV models, so we compare them with OLS(this result aligns to the model in Regression Analysis part which removes quadratic terms), which can be treated as a special case of 2SLS. The results are shown in table 8.

The R-squared of the instrumental variable (IV) model is significantly greater than that of the ordinary least squares (OLS) model, indicating that the IV model has a much higher explanatory power than the OLS model. In summary, after

	2SLS	GMM	GMM-Con
Dep. Variable	log(PMval+0.01)	log(PMval+0.01)	log(PMval+0.01)
Estimator	IV-2SLS	IV-GMM	IV-GMM
No. Observations	23035	23035	23035
Cov. Est.	unadjusted	unadjusted	robust
R-squared	0.8158	0.8158	0.8158
Adj. R-squared	0.8157	0.8157	0.8157
F-statistic	1.098e+05	1.197e+05	1.366e+05
P-value (F-stat)	0.0000	0.0000	0.0000
Country	0.0188	0.0188	0.0188
	(89.589)	(93.505)	(93.673)
NO2.AQI.Value	-0.0592	-0.0592	-0.0592
	(-12.377)	(-12.918)	(-12.256)
NO2.AQI.Category	0.2251	0.2251	0.2251
	(9.3392)	(9.7475)	(8.1241)
Ozone.AQI.Value	-0.0066	-0.0066	-0.0066
	(-6.7571)	(-7.0525)	(-6.0558)
Ozone.AQI.Category	0.7540	0.7540	0.7540
	(44.689)	(46.643)	(42.402)
CO.AQI.Category	3.4946	3.4946	3.4946
	(31.167)	(32.529)	(27.509)
Instruments	CO.AQI.Value	CO.AQI.Value	CO.AQI.Value

Figure 14: Results of IV method

incorporating instrumental variables into the modeling process, we obtained a highly explanatory model. Furthermore, the consistency of the estimated values across different IV models, including 2SLS and GMM, enhances the credibility of our instrumental variables and the robustness of the model.

Table 8: Interpertative Capability Analysis

Model	R squared	Adj. R-squared	ATE
OLS	0.287	0.2869	39.87
2SLS	0.8158	0.8157	32.94
GMM	0.8158	0.8157	32.94

6 Conclusion and Future work

Table 9: Results for estimated ATE

Model	$\hat{ATE}$
Polynomial Regression	40.38
Linear Regression	39.87
Matching With Rep	63.44
Matching Without Rep	36.8
Weighting	49.54
Stratification	55.67
2SLS	32.94
GMM/CUE-GMM	32.94

In this experiment, we selected a dataset related to air pollution to investigate the factors influencing PM2.5 (using AQI as an indicator). Under reasonable assumptions, we chose the Category of one of the harmful gases, CO, as the treatment variable for our study and transformed it into a binary treatment using an appropriate method. We first constructed a causal Directed Acyclic Graph (DAG) based on prior knowledge and simplified it according to empirical knowledge and assumptions.

During the analysis process, we employed regression analysis, matching, weighting, stratification, and IV modeling to evaluate the causal effects. The results are presented in Table 9. The estimated values for the Average Treatment Effect

(ATE) are concentrated in the range of 35-60, with minimal variation and no inconsistency in the positive or negative direction. This indicates a clear causal relationship between the treatment variable, C_{cate}, and the outcome variable, P_{val}. Specifically, as the air quality related to CO worsens (note that we only need a rough classification of pollution levels), the PM2.5 pollution tends to become more severe.

However, there are several limitations in this experiment. One important assumption is that a specific AQI value will only directly affect its AQI Category, and an approximated pollutant level can adequately describe the overall air condition. This assumption is based on empirical knowledge since we often focus on a categorical representation of air quality in our daily lives rather than specific values. Our IV model is also based on this assumption, and it is necessary to further validate its validity. Additionally, we only have data from a single time period, and ideally, a longitudinal analysis of data across time would be desirable. Finally, a limitation of this study is the limited consideration of factors and covariates. For instance, significant anthropogenic factors were not extensively taken into account, and it would be beneficial to include relevant data in future experiments.

Despite these limitations, this study provides insights into the causal relationship between harmful gases and PM2.5 pollution, specifically regarding the approximate classification of the degree of harmful gas pollution and its causal effect on PM2.5 pollution levels (AQI Value).

References

[1] S. Nigam, B. Rao, N. Kumar, and V. Mhaisalkar, "Air quality index-a comparative study for assessing the status of air quality," *Research Journal of Engineering and Technology*, vol. 6, no. 2, pp. 267–274, 2015.

[2] B. A. Brumback, M. A. Hernán, S. J. Haneuse, and J. M. Robins, "Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures," *Statistics in medicine*, vol. 23, no. 5, pp. 749–767, 2004.

[3] P. Ding and T. J. VanderWeele, "Sensitivity analysis without assumptions," *Epidemiology (Cambridge, Mass.)*, vol. 27, no. 3, p. 368, 2016.

[4] S. Chinn, "A simple method for converting an odds ratio to effect size for use in meta-analysis," *Statistics in medicine*, vol. 19, no. 22, pp. 3127–3131, 2000.

[5] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, p. 1, 2010.

[6] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," *Statistical methods in medical research*, vol. 22, no. 3, pp. 278–295, 2013.

[7] C. E. Frangakis and D. B. Rubin, "Principal stratification in causal inference," *Biometrics*, vol. 58, no. 1, pp. 21–29, 2002.

[8] K. A. Bollen, J. B. Kirby, P. J. Curran, P. M. Paxton, and F. Chen, "Latent variable models under misspecification: Two-stage least squares (2sls) and maximum likelihood (ml) estimators," *Sociological Methods & Research*, vol. 36, no. 1, pp. 48–86, 2007.

[9] C. F. Baum, M. E. Schaffer, and S. Stillman, "Instrumental variables and gmm: Estimation and testing," *The Stata Journal*, vol. 3, no. 1, pp. 1–31, 2003.