STAT 156 Re-Analysis

Samuel Gao and Ritvik Iyer

12/2/2021

I. Introduction

In the previous portions of our paper, we have attempted to replicate a paper that analyzes the causal effects of environmental factors on the number of confirmed COVID cases in various regions of China. The authors utilized a pipeline which included machine learning methods (XGBoost and Orthogonal Random Forests) and to make causal inferences. In this section, we will attempt to re-analyze the data using different methods in an attempt to address the same questions.

To begin, we acknowledge that there are numerous challenges that exist when analyzing the data. The biggest challenge in inferring causal effects is that the treatment variable is not well defined for our data set which makes the causal question difficult to formulate. Our "treatment" variables consists of different levels of environmental factors which cannot easily be binarized into a 0 and 1. Fundamentally, the challenge for the question we are addressing is that it is almost impossible to consider an experimental setting where we can change one of the variables while keeping the other covariates constant. Further, given the different levels of environmental factor that we are observing, it is also difficult to state which level should be considered the "treatment" and which level is the "control". Thus the causal question at hand is extremely difficult to formulate in the context of Causal Inference.

II. Method 1- Multiple Regression Model

In re-analyzing the data we have, we first propose the multiple regression model. In this model, we are attempting to measure the direct effects of the different environmental factors on the spread of COVID 19 in China. The model that we are using can be expressed as the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} ......\beta_k X_{ki} + u_i, i = 1...n$$

where Y_i is the outcome of the dependent variable, X_1 can be notated as the "treatment" variable that we are interested in, while the other X's are the different covariates that we are controlling for, and u_i is an error term for entry i. In our case, the different covariates are the economical and demographic data on this data set, such as GDP, population, hospital beds, etc. while the outcome is the number of newly confirmed COVID cases.

Interpreting the results, the coefficient β_1 can be interpreted as the how the number of newly confirmed COVID cases change given changes in the "treatment" variable while controlling for the different covariates. The results of the regressions are shown below in the following page:

Table 1: Estimation of Environmental Impact on COVID Cases- Simple Multivariate Regression Model

	Dependent variable:									
	Case									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TEMP	0.101^{***}									
	(0.031)									
WSPD		0.403^{*}								
		(0.235)								
PRES			0.033^{***}							
			(0.003)							
HUM				0.077^{***}						
				(0.014)						
O3					0.018					
					(0.012)					
NO2						-0.043^{**}				
						(0.018)				
CO							4.253^{***}			
							(0.644)			
SO2								-0.044^{**}		
								(0.022)		
PM2.5									0.023^{***}	
									(0.008)	
PM10										0.015^{**}
										(0.007)
pop	0.001^{**}	0.001^{**}	0.0005^{*}	0.001^{**}	0.001^{**}	0.001^{***}	0.001^{**}	0.001^{**}	0.001^{**}	0.001^{**}
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
GDP	0.066^{***}	0.070^{***}	0.071^{***}	0.068^{***}	0.065^{***}	0.064^{***}	0.061^{***}	0.066^{***}	0.065^{***}	0.065^{***}
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
PRIM	0.177	0.304	0.060	0.083	0.239	0.183	0.251	0.204	0.241	0.244
	(0.223)	(0.221)	(0.223)	(0.223)	(0.221)	(0.219)	(0.222)	(0.221)	(0.222)	(0.222)
SEC	-0.209^{***}	-0.210^{***}	-0.208^{***}	-0.210^{***}	-0.207^{***}	-0.208^{***}	-0.204^{***}	-0.205^{***}	-0.202^{***}	-0.203^{***}
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
A60	0.161^{***}	0.131^{***}	0.077^{*}	0.118^{***}	0.149^{***}	0.142^{***}	0.169^{***}	0.143^{***}	0.156^{***}	0.163^{***}
	(0.043)	(0.044)	(0.043)	(0.041)	(0.042)	(0.042)	(0.043)	(0.042)	(0.042)	(0.043)
BED	-0.784^{***}	-0.794^{***}	-0.658^{***}	-0.808^{***}	-0.834^{***}	-0.773^{***}	-0.967^{***}	-0.807^{***}	-0.942^{***}	-0.936^{***}
	(0.199)	(0.210)	(0.206)	(0.200)	(0.202)	(0.201)	(0.208)	(0.199)	(0.203)	(0.207)
DOC	-5.179^{***}	-5.338^{***}	-5.306^{***}	-4.735^{***}	-5.527^{***}	-5.387^{***}	-5.776^{***}	-5.381^{***}	-5.587^{***}	-5.629^{***}
	(0.527)	(0.542)	(0.540)	(0.503)	(0.548)	(0.547)	(0.572)	(0.538)	(0.551)	(0.567)
NRS	2.037^{***}	2.138^{***}	2.773^{***}	2.131^{***}	2.345^{***}	2.269^{***}	2.543^{***}	2.242^{***}	2.475^{***}	2.469^{***}
	(0.341)	(0.376)	(0.366)	(0.340)	(0.356)	(0.356)	(0.369)	(0.352)	(0.356)	(0.366)
AREA	-0.00005	-0.0001^{**}	0.00004	-0.00002	-0.0001^{*}	-0.0001^{*}	-0.0001^{*}	-0.0001	-0.0001	-0.0001
	(0.00003)	(0.00003)	(0.00004)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
POPDEN	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GDPPC	0.794^{***}	0.780^{***}	0.737^{***}	0.745^{***}	0.810^{***}	0.810^{***}	0.883^{***}	0.782^{***}	0.838^{***}	0.845^{***}
	(0.143)	(0.137)	(0.143)	(0.144)	(0.144)	(0.145)	(0.145)	(0.149)	(0.144)	(0.145)
PRPER	1.039^{***}	1.048^{***}	1.050^{***}	1.070^{***}	1.059^{***}	1.039^{***}	1.065^{***}	1.041^{***}	1.057^{***}	1.064^{***}
	(0.106)	(0.108)	(0.109)	(0.109)	(0.109)	(0.111)	(0.109)	(0.108)	(0.109)	(0.109)
SECPER	0.343^{***}	0.355^{***}	0.359^{***}	0.361^{***}	0.350^{***}	0.355^{***}	0.336^{***}	0.348^{***}	0.341^{***}	0.343^{***}
	(0.032)	(0.034)	(0.033)	(0.033)	(0.033)	(0.033)	(0.032)	(0.033)	(0.033)	(0.033)
ACTV	-6.622^{***}	-6.453^{***}	-6.382^{***}	-6.310^{***}	-6.505^{***}	-6.276^{***}	-6.301^{***}	-6.441^{***}	-6.423^{***}	-6.525^{***}
	(0.383)	(0.357)	(0.356)	(0.344)	(0.361)	(0.372)	(0.345)	(0.357)	(0.356)	(0.366)
Constant	5.800^{*}	5.266	-27.771^{***}	-1.110	4.475	5.844^{*}	1.955	6.708^{*}	5.104	5.197
	(3.262)	(3.357)	(5.311)	(3.081)	(3.371)	(3.292)	(3.196)	(3.534)	(3.260)	(3.272)
Observations	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870
R^2	0.086	0.086	0.087	0.087	0.086	0.086	0.087	0.086	0.086	0.086
Adjusted R^2	0.085	0.085	0.086	0.086	0.085	0.085	0.086	0.085	0.085	0.085
Residual Std. Error (df = 12854)	32.869	32.869	32.851	32.848	32.871	32.870	32.842	32.872	32.866	32.869

*p<0.1; **p<0.05; ***p<0.01

However, in order to deepen our understanding of the results and critique the conclusions that we have made, we must consider the basic assumptions of the multiple regression model. The assumptions of the multiple regression models are:

- u_i has conditional mean zero given $X_{1i}...X_{ki}$. Or expressed in mathematics:

$$E[u_i|X_{1i}...X_{ki}] = 0$$

- $(X_{1i}...X_{ki}, Y_i)$, i = 1. . . .n are drawn i.i.d. from the joint distributions.
- Large outliers are unlikely: $X_{1i}...X_{ki}$ and Y_i all have non-zero, finite fourth moments.
- There is no perfect multicollinearity, which means that none of the regressors are perfect linear functions of the other regressors.

The first assumption should hold relatively easily as there is an intercept term in the regression that we are performing.

The second assumption is quite strong, but must be imposed in order for us to make any progress on the multiple regression. We must assume that all of the draws are drawn independently from the joint distribution of the covariates and outcome, which is highly unlikely due to the fact that a lot of the covariates are dependent on each other in one way or another (for example, a larger population size tends to correlate with a higher level of GDP.)

The third assumption holds relatively easily as the entries are all finite.

The fourth assumption holds if we make some adjustments to our data. In our data, the only place where multicollinearity happens is related to the different sectors' GDP. The primary sector, secondary sector, and tertiary sectors should sum up to 100%, so in our analysis, we exclude one of the sectors (tertiary sector) in terms of the percentage it accounts for in a city's overall GDP as well as the absolute value of the GDP. By excluding those two variables, we should not have any multicollinearity in our data.

Further, if we assume that the covariates that are included in the data set are complete so that there is no omitted variables bias, and that controlling for all the covariates, unconfoundedness holds then the coefficients may be interpreted with causal framework in answering how does changing the specific environmental factor affect the number of confirmed COVID cases.

However, this assumption, along with the assumption that the samples are drawn i.i.d. from the joint distribution is extremely strong and unlikely to hold. We do not know if there is an omitted variable that may impact the outcome in a significant way, nor can we verify that the samples are drawn independently, which are significant weaknesses of our model. In addition, we are not considering the environmental factors are covariates in this case, as they may very well be covariates for the other environmental factors as well. Controlling for environmental factors in the analysis may help improve the interpretability of our results (we omit this regression as there are 1024 possible regressions to run and the results would be impossible to report here.) We have further assumed that the linear model is true here, which we have not tested and examined in full detail, which may a further area this analysis can be improved upon in the future.

III. Method 2- Multiple Regression Model with Interactions

We here propose an alternative method to analyze the data using multiple regression, by including an interaction term between the "treatment" variable and the covariates. The model can be described as the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}......\beta_k X_{ki} + \beta_{k+1} X_{1i} X_{2i}... + \beta_{2k-1} X_{1i} X_{ki} + u_i, i = 1...n$$

where X_{1i} is the treatment variable and the other X_i terms are the covariates, and u_i is an error term.

Similar assumptions with relation to multiple regression holds here. We also exclude the percentage of GDP in the tertiary sector and the absolute value of GDP in the tertiary sector in the analysis to avoid multicollinearity.

The difference in this approach compared to the simple multiple regression is that we are now considering the different effects the covariates may have on the treatments In the simple multiple regression, we assume that the effects of the treatment are constant given the covariates, where the β_1 term measuring the effect for the treatment X_{1i}. With interaction terms, we no longer assume that the effects of the covariates are constant, and we now also account for the different effects the covariate terms may have on the treatment variable.

The results from this regression is shown in the following page as Table 2. We exclude the interaction terms in the table due to their volume. Only the results for the covariate terms and the treatment terms are reported.

Table 2: Estimation of Environmental Impact on COVID Cases- Multivariate Regression Model with Interaction

	Dependent variable:									
	Case									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TEMP	−1.048***									
	(0.317)									
WSPD		−9.450***								
		(2.954)								
PRES			−0.466***							
			(0.073)							
HUM				0.176						
				(0.212)						
O3					0.220					
					(0.175)					
NO2						−0.871**				
						(0.361)				
CO							13.454			
							(13.252)			
SO2								−1.534***		
								(0.314)		
PM2.5									−0.136	
									(0.109)	
PM10										0.038
										(0.067)
pop	0.002***	−0.001	−0.123***	−0.004***	0.002	0.002*	−0.001	0.001*	0.001***	0.001**
	(0.001)	(0.001)	(0.013)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0005)	(0.001)
GDP	0.040*	0.074***	0.456	0.091*	0.107*	0.088*	0.028	0.009	0.032	0.057**
	(0.023)	(0.028)	(0.701)	(0.050)	(0.059)	(0.047)	(0.031)	(0.027)	(0.021)	(0.028)
PRIM	0.735**	0.481	46.245***	0.719	1.868*	1.191*	−0.486	−0.321	−0.882**	−0.389
	(0.370)	(0.486)	(6.541)	(0.936)	(1.010)	(0.698)	(0.531)	(0.347)	(0.344)	(0.396)
SEC	−0.209***	0.213**	18.090***	0.299***	−0.768***	−0.345***	0.317***	−0.051	−0.144***	−0.237***
	(0.053)	(0.094)	(2.200)	(0.111)	(0.139)	(0.079)	(0.072)	(0.042)	(0.045)	(0.058)
A60	−0.360***	−0.015	−3.031***	−0.243	−0.156	0.320**	0.187*	0.053	0.406***	0.295***
	(0.067)	(0.087)	(1.010)	(0.164)	(0.190)	(0.159)	(0.109)	(0.072)	(0.072)	(0.084)
BED	−0.982***	−2.747***	−1.821	2.433***	0.119	−2.195***	−0.167	−1.293***	−0.293	0.563
	(0.303)	(0.468)	(5.774)	(0.797)	(0.786)	(0.805)	(0.544)	(0.394)	(0.311)	(0.368)
DOC	−11.141***	−2.236*	171.943***	12.477***	−13.787***	−18.679***	3.454*	−7.346***	−4.815***	−7.725***
	(1.001)	(1.229)	(19.616)	(1.994)	(2.368)	(2.229)	(1.813)	(1.036)	(1.031)	(1.042)
NRS	3.963***	3.396***	−90.127***	−8.547***	4.394***	9.622***	1.873*	3.099***	3.053***	3.175***
	(0.502)	(0.742)	(12.667)	(1.214)	(1.470)	(1.451)	(1.111)	(0.651)	(0.599)	(0.655)
AREA	−0.0003***	−0.0003***	0.001	0.0005***	−0.00003	0.00003	0.0004***	0.00005	−0.00002	−0.00003
	(0.0001)	(0.0001)	(0.001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
POPDEN	−0.010***	−0.007***	0.097**	0.007**	0.004**	−0.002	0.007***	−0.002*	0.003***	0.001
	(0.002)	(0.001)	(0.041)	(0.004)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
GDPPC	1.579***	−0.711*	−28.448***	−0.729	2.532***	1.467***	−1.202***	0.885***	−0.206	0.335
	(0.283)	(0.411)	(3.554)	(0.554)	(0.834)	(0.396)	(0.390)	(0.178)	(0.233)	(0.286)
PRPER	1.483***	1.194***	−32.666***	−2.084***	1.868***	2.435***	0.463*	1.752***	0.491***	1.192***
	(0.145)	(0.212)	(3.204)	(0.307)	(0.459)	(0.295)	(0.239)	(0.191)	(0.159)	(0.205)
SECPER	0.384***	0.133**	−6.730***	−0.354***	0.744***	0.457***	−0.036	0.328***	0.160***	0.387***
	(0.051)	(0.065)	(0.808)	(0.097)	(0.155)	(0.097)	(0.091)	(0.051)	(0.056)	(0.075)
ACTV	−7.057***	−6.001***	46.176***	7.277***	−9.883***	−14.160***	−3.300***	−9.114***	−5.860***	−8.416***
	(0.462)	(0.684)	(4.004)	(1.347)	(1.645)	(0.987)	(0.523)	(0.610)	(0.499)	(0.575)
Constant	18.279***	30.247***	459.998***	−15.148	−12.132	24.190**	−6.230	16.347***	10.195**	3.338
	(5.285)	(7.372)	(70.770)	(12.627)	(16.590)	(11.898)	(8.984)	(5.486)	(4.979)	(6.277)
Observations	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870
R^2	0.093	0.090	0.113	0.117	0.090	0.120	0.098	0.093	0.093	0.090
Adjusted R^2	0.091	0.088	0.111	0.115	0.088	0.118	0.096	0.091	0.091	0.088
Residual Std. Error (df = 12840)	32.757	32.809	32.386	32.326	32.815	32.274	32.661	32.759	32.764	32.817

*p<0.1; **p<0.05; ***p<0.01

IV. Method 3- Fixed Effects Regression

What the multiple regression model presented earlier does not account for is the characteristics of the cities that are fixed over time. For example, in our data set, demographic and economic indicators are fixed over time. Even though factors such as population, GDP, hospital beds, etc. are fluid over time, government agencies tend to record these data for a given time period given their difficulty in data agglomeration and estimation. The data that this data set contains is from government reports and statistical yearbooks for 2019, and we will assume that the covariate data is constant across the time.

With the above mentioned assumptions, an analysis considering the effects of time may be more applicable than a multiple regression model. Hence we here conduct our analysis again but as panel data with fixed effects regression, and we consider the following model:

$$Y_{it} = \beta_1 X_{1,it} + ...\beta_k X_{k,it} + \alpha_i + u_{it}$$

Where Y_{it} is the outcome for entity i and time t. For this data set, the increments of t that we are observing is days, and the entity we are identifying are the various cities in China. $X_{k,it}$ describe the value of the regressors for entity i at time period t. In the fixed effects regression that we run, we include a regressor for the specific pollutant level for city i at time t, and also a variable for the activity level of the city, which is a covariate that is varying in time. The remainder of the covariates are assumed to be constant in time, and are notated as α_i in the above equation. The components of α_i includes the various economic and demographic factors that we have mentioned earlier that are assumed to be constant over time. Lastly, u_{it} is the error for entity i at time t.

How the programs run fixed effect regressions is done in two steps. In the first step, entity specific averages are subtracted out for each of the variables. In this manner, all of the fixed effects that are identified will be differenced out. More specifically, expressed in mathematics for two variables that vary across the time as we are analyzing for our current data set, the procedure is:

$$Y_{it} - \bar{Y} = \beta_1 (X_{1,it} - \bar{X_{1,i}}) + \beta_2 (X_{2,it} - \bar{X_{2,i}}) + (u_{it} - \bar{u_i})$$

With the above values, OLS regression is ran and β_1 is estimated. The results of the fixed effects regression for our data set is displayed in the following page as Table 3.

To more fully interpret the results of the fixed effects regression model, we must again understand the assumptions the model is making. The assumptions are quite close to the assumptions made under the multiple regression model, and are stated below:

- u_i has conditional mean zero given $X_{1,it}, X_{2,it}...X_{k,it}, \alpha_i$. Or expressed in mathematics:

$$E[u_{it}|X_{1,it}, X_{2,it}...X_{k,it}, \alpha_i] = 0$$

- $(X_{1,it}, X_{2,it}...X_{k,it}, u_{i1}, u_{i2}...u_{iT})$, i = 1. . . .n are drawn i.i.d. from the joint distributions.
- Large outliers are unlikely: X_{it} and u_{it} all have non-zero, finite fourth moments.
- There is no perfect multicollinearity.

Table 3: Fixed Effects Regression Analysis

	Dependent variable:									
	Case									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TEMP	-0.266^{***}									
	(0.089)									
WSPD		-0.327								
		(0.324)								
PRES			0.192^{***}							
			(0.074)							
HUM				-0.015						
				(0.022)						
O3					-0.028					
					(0.021)					
NO2						0.170^{*}				
						(0.094)				
CO							-0.771			
							(1.298)			
SO2								-0.011		
								(0.055)		
PM2.5									0.048^{***}	
									(0.015)	
PM10										0.037^{***}
										(0.013)
ACTV	-4.420^{***}	-4.954^{***}	-4.599^{***}	-4.992^{***}	-4.868^{***}	-5.719^{***}	-4.983^{***}	-4.960^{***}	-4.906^{***}	-5.170^{***}
	(0.810)	(0.935)	(0.831)	(0.920)	(0.887)	(1.300)	(0.961)	(0.935)	(0.925)	(0.996)
Observations	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870	12,870
R^2	0.041	0.040	0.041	0.040	0.041	0.042	0.040	0.040	0.042	0.042
Adjusted R^2	0.029	0.028	0.029	0.028	0.028	0.029	0.028	0.028	0.030	0.029
F Statistic (df = 2; 12704)	273.297^{***}	266.923^{***}	271.607^{***}	266.645^{***}	268.635^{***}	277.490^{***}	266.700^{***}	266.358^{***}	279.138^{***}	275.528^{***}

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The first, third, and fourth assumptions hold more easily while the second reasoning is more challenging to hold for our data set for similar reasoning as explained in the section about multiple regression. However, advantages are present with the fixed effects regression model compared to the multiple regression models. Given that our data consists of data from different cities across China and across multiple time periods, this model of analysis is the most appropriate panel data analysis is meant specifically for this type of analysis.

With this model, there are also obvious disadvantages and the assumptions that must be considered also. In this model, we assume that the "fixed" covariates do not have an effect on the outcome, but this may contribute to omitted variables bias if the covariates actually have an impact on the outcome. For example, if there were actually some effect of GDP on the newly confirmed cases in a city, this effect cannot be captured as the GDP values have all been differenced out and cannot be observed. In order for the fixed effects regrssion to be more causal in the interpretation, an additional assumption should be said that the fixed effects from the different entities have no effects on the outcomes, or else the estimates we received from the regression will be biased.

In addition, the assumption that all of the variables are fixed in time is also a very strong assumption, and may affect the interpretation of our results. For example, we have assumed in the original data set that the number of hospital beds, nurses, and doctors are constant in the city across this time period. However, due to the news of the COVID-19 pandemic, these numbers may fluctuate wildly from what were originally reported. If we assume that these factors do not impact the spread of COVID-19 in a particular city, then by differencing them out in the fixed regression the analysis will be sensible. However, if the truth is, for a heuristic example, that if a city that obtains more hospital beds to treat patients that are infected by the virus it may be more successful in containing the virus later on, then differencing out these factors and assuming that they are constant would bias the results that were obtained by the fixed effects regression. In order to interpret the data better, we must further assume that the fixed effects that we have identified are in fact, fixed, which is likely not true for our data set.

Further, even if the above two assumption holds perfectly, we must always consider the case where there may be other variables that vary in time that we have failed to consider, that also has an impact on the outcome. With the assumptions of the fixed effects regression model, variables that do not vary in time do not contribute to the outcome but variables that vary in time may still contribute to the outcome. In our data set, we identified the activity level of a city as a variable that is varying in time. However, other obvious variables have not been identified that may significantly contribute to the outcome. For example, the number of COVID tests that were given in a particular day may significantly influence the number of confirmed cases within a particular city. If no COVID tests were given, 0 cases will be identified, while if the testing capacities increased, more cases may become confirmed.

If the four assumptions for the fixed effects regression model hold, plus the additional three assumptions that we present above also hold, then the results derived from the fixed regression model may be interpreted causally. However, as previously discussed, the assumptions are extremely unlikely to hold and more analysis and a more complete data may be needed in order to obtain a more robust result.

V. Comparison with Methods in original Paper

In our re-analysis of the data, we propose 3 models to estimate the causal effects of the treatment on COVID-19 cases. This approach is quite different than the methods used by the authors in the original paper. Instead, the authors use more of a "toolbox"-type approach in two stages. First, the authors don't define their set of treatment variables a priori. Rather, they fit various machine learning models coupled with interpretation tools (e.g. Feature importance, Permutation scores, SHAP, etc.) which yield the features which are highly predictive of the outcome. Whereas, in our approach, we clearly define which treatment variables we assess at the beginning of the process. Secondly, the authors assume a particular causal graph structure for the basis of their analysis, and try out several different machine learning methods to estimate causal effects within this graph structure. This approach may be problematic because if the causal graph is mis-specified, then the model estimation is obsolete. In addition, the author do not clearly outlay the assumptions behind most of their analysis, where we attempt to identify as clearly as possible the assumptions we are imposing on each of our analysis.

There are other notable differences in the execution of the analysis as well. The original authors separated the dataset into three clusters using Machine Learning methods, while we analyzed the data without considering the data in stratas. Instead, we believe that the multiple regression models and the time series model take into consideration already the variations that may occur due to the different covariates present, and the focus is more put on whether the covariates have been properly controlled for and whether unconfoundedness holds.

The original authors further analyzed the data by parsing the data into two phases, where they consider one phase to be the "Spreading" phase while the second as the "Postpeak" phase. In our analysis, we also did not take into consideration this factor. We believe that this decision was arbitrary and if there were any true differences across periods, the time fixed effects regression model should account for the differences by design.

In the conclusion reached by the authors, the found virtually no significant effect on any of the treatments, with only 1 passing all of the refutation tests, for which the authors still did not fully express their willingness to confirm it as a causal effect. In our analysis, we share this skepticism with the authors and do not wish any statements that are explicitly causal, as there are many assumptions that need to hold in order for the effects to be actually determined causal. There are many data that we simply do not have access to which may affect the outcomes significantly. So while our analysis did turn out some statistically significant results (5 for the simple multiple regression model, 4 for the multiple regression model with interactions, and 4 for the fixed effects regressions), we do caution the reader from making any explicit causal inferences from such results.

In comparing the models that we have so far presented, the fixed effects regression holds the greatest promise as the assumptions are explicitly stated with the shortfalls identified more clearly, which makes the interpretation of the results more straightforward and potentially more causal compared to the other models presented. As discussed in Section IV on the fixed effects regression model, though, there are clear assumptions that are violated with this data set as we simply do not have all the data required to accurately perform a full, well vetted fixed effects regression. However, we do believe that this method, given the data set that we have, is the most appropriate model for the purposes of analytics, despite its shortfalls, and may be improved upon the most easily if additional data resources are available at different researchers' disposal.

Appendix of Code

```
time1_overall = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-rep

time1_spreading = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-r

time1_post = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-replic

time2_overall = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-rep

time2_spreading = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-r

time2_post = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-replic

time3_overall = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-rep

time3_spreading = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-r

time3_post = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-replic

cleaned = read.csv("~/Documents/UC BERKELEY/STATISTICS/STAT 156/Final Project/GitHub/EnvCausal-replicat

all_dat = rbind(time1_overall, time2_overall, time3_overall)

test = cbind(all_dat, cleaned)

cleaned$City = as.character(cleaned$City)
all_dat$City = as.character(all_dat$City)

all_dat$pop = c(1)
all_dat$GDP = c(0)
all_dat$PRIM = c(0)
all_dat$SEC = c(0)
all_dat$TERT = c(0)
all_dat$A60 = c(0)
all_dat$BED = c(0)
all_dat$DOC = c(0)
all_dat$NRS = c(0)
all_dat$AREA = c(0)
all_dat$POPDEN = c(0)
all_dat$GDPPC = c(0)
all_dat$PRPER = c(0)
all_dat$SECPER = c(0)
all_dat$TERPEC = c(0)


cleaned$City[111] = "Suzhou_1"
all_dat$City[6085:6162] = "Suzhou_1"
cleaned$City[98] = "Taizhou_1"
all_dat$City[6787:6864] = "Taizhou_1"
cleaned$City[41] = "Lvliang"
cleaned$City[103] = "Bangbu"
all_dat$City[7723:7800] = "Chaoyang"
```

```r
cleaned = cleaned[-79, ]
cleaned = cleaned[-67, ]
all_dat = all_dat[-(9283:9360), ]


for(i in 1:164){
all_dat[all_dat$City == cleaned$City[i], ]$pop = cleaned$POP[i]
all_dat[all_dat$City == cleaned$City[i], ]$GDP = cleaned$GDP[i]
all_dat[all_dat$City == cleaned$City[i], ]$PRIM = cleaned$PRIM[i]
all_dat[all_dat$City == cleaned$City[i], ]$SEC = cleaned$SEC[i]
all_dat[all_dat$City == cleaned$City[i], ]$TERT = cleaned$TERT[i]
all_dat[all_dat$City == cleaned$City[i], ]$A60 = cleaned$X.60yr.[i]
all_dat[all_dat$City == cleaned$City[i], ]$BED = cleaned$BED[i]
all_dat[all_dat$City == cleaned$City[i], ]$DOC = cleaned$DOC[i]
all_dat[all_dat$City == cleaned$City[i], ]$NRS = cleaned$NRS[i]
all_dat[all_dat$City == cleaned$City[i], ]$AREA = cleaned$Area[i]
all_dat[all_dat$City == cleaned$City[i], ]$POPDEN = cleaned$POPDENS[i]
all_dat[all_dat$City == cleaned$City[i], ]$GDPPC = cleaned$GDPpc[i]
all_dat[all_dat$City == cleaned$City[i], ]$PRPER = cleaned$Prim.[i]
all_dat[all_dat$City == cleaned$City[i], ]$SECPER = cleaned$Sec.[i]
all_dat[all_dat$City == cleaned$City[i], ]$TERPEC = cleaned$Tert.[i]
}
```

```r
results1 = lm(Case ~ TEMP + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results1)$coeff[2, 4]
```

```
## [1] 0.06685475
```

```r
results2 = lm(Case ~ WSPD + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results2)$coeff[2, 4]
```

```
## [1] 0.08028935
```

```r
results3 = lm(Case ~ PRES + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results3)$coeff[2, 4]
```

```
## [1] 3.163902e-05
```

```r
results4 = lm(Case ~ HUM + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results4)$coeff[2, 4]
```

```
## [1] 1.02808e-05
```

```r
results5 = lm(Case ~ O3 + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results5)$coeff[2, 4]
```

```
## [1] 0.1738511
```

```r
results6 = lm(Case ~ NO2 + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results6)$coeff[2, 4]
```

```
## [1] 0.1660672
```

```
results7 = lm(Case ~ CO + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results7)$coeff[2, 4]
```

```
## [1] 8.972021e-07
```

```
results8 = lm(Case ~ SO2 + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results8)$coeff[2, 4]
```

```
## [1] 0.3133944
```

```
results9 = lm(Case ~ PM2.5 + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results9)$coeff[2, 4]
```

```
## [1] 0.02100447
```

```
results10 = lm(Case ~ PM10 + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV , dat = all_dat)
summary(results10)$coeff[2, 4]
```

```
## [1] 0.06323277
```

```
rob_se_1 <- list(sqrt(diag(vcovHC(results1, type = "HC1"))),
               sqrt(diag(vcovHC(results2, type = "HC1"))),
               sqrt(diag(vcovHC(results3, type = "HC1"))),
               sqrt(diag(vcovHC(results4, type = "HC1"))),
               sqrt(diag(vcovHC(results5, type = "HC1"))),
               sqrt(diag(vcovHC(results6, type = "HC1"))),
               sqrt(diag(vcovHC(results7, type = "HC1"))),
               sqrt(diag(vcovHC(results8, type = "HC1"))),
               sqrt(diag(vcovHC(results9, type = "HC1"))),
               sqrt(diag(vcovHC(results10, type = "HC1"))))
```

```
table1 = stargazer(results1, results2, results3,
          results4, results5, results6, results7, results8, results9, results10,
          digits = 3,
          header = FALSE,
          type = "latex",
          se = rob_se_1,
          title = "Estimation of Environmental Impact on COVID Cases- Simple Multivariate Regression Mo
          model.numbers = FALSE,
          column.labels = c("(1)", "(2)", "(3)", "(4)", "(5)", "(6)", "(7)", "(8)", "(9)", "(10)"))
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{Estimation of Environmental Impact on COVID Cases- Simple Multivariate Regression Model}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcccccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{10}{c}{\textit{Dependent variable:}} \\
## \cline{2-11}
## \\[-1.8ex] & \multicolumn{10}{c}{Case} \\
##  & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\
## \hline \\[-1.8ex]
```

```
##  TEMP & 0.101$^{***}$ &  &  &  &  &  &  &  &  &  \\
##   & (0.031) &  &  &  &  &  &  &  &  &  \\
##   & & & & & & & & & & \\
##  WSPD &  & 0.403$^{*}$ &  &  &  &  &  &  &  &  \\
##   &  & (0.235) &  &  &  &  &  &  &  &  \\
##   & & & & & & & & & & \\
##  PRES &  &  & 0.033$^{***}$ &  &  &  &  &  &  &  \\
##   &  &  & (0.003) &  &  &  &  &  &  &  \\
##   & & & & & & & & & & \\
##  HUM &  &  &  & 0.077$^{***}$ &  &  &  &  &  &  \\
##   &  &  &  & (0.014) &  &  &  &  &  &  \\
##   & & & & & & & & & & \\
##  O3 &  &  &  &  & 0.018 &  &  &  &  &  \\
##   &  &  &  &  & (0.012) &  &  &  &  &  \\
##   & & & & & & & & & & \\
##  NO2 &  &  &  &  &  & $-$0.043$^{**}$ &  &  &  &  \\
##   &  &  &  &  &  & (0.018) &  &  &  &  \\
##   & & & & & & & & & & \\
##  CO &  &  &  &  &  &  & 4.253$^{***}$ &  &  &  \\
##   &  &  &  &  &  &  & (0.644) &  &  &  \\
##   & & & & & & & & & & \\
##  SO2 &  &  &  &  &  &  &  & $-$0.044$^{**}$ &  &  \\
##   &  &  &  &  &  &  &  & (0.022) &  &  \\
##   & & & & & & & & & & \\
##  PM2.5 &  &  &  &  &  &  &  &  & 0.023$^{***}$ &  \\
##   &  &  &  &  &  &  &  &  & (0.008) &  \\
##   & & & & & & & & & & \\
##  PM10 &  &  &  &  &  &  &  &  &  & 0.015$^{**}$ \\
##   &  &  &  &  &  &  &  &  &  & (0.007) \\
##   & & & & & & & & & & \\
##  pop & 0.001$^{**}$ & 0.001$^{**}$ & 0.0005$^{*}$ & 0.001$^{**}$ & 0.001$^{**}$ & 0.001$^{***}$ & 0.0
##   & (0.0003) & (0.0003) & (0.0003) & (0.0003) & (0.0003) & (0.0003) & (0.0003) & (0.0003) & (0.0003)
##   & & & & & & & & & & \\
##  GDP & 0.066$^{***}$ & 0.070$^{***}$ & 0.071$^{***}$ & 0.068$^{***}$ & 0.065$^{***}$ & 0.064$^{***}$
##   & (0.014) & (0.014) & (0.014) & (0.014) & (0.014) & (0.014) & (0.014) & (0.014) & (0.014) & (0.014
##   & & & & & & & & & & \\
##  PRIM & 0.177 & 0.304 & 0.060 & 0.083 & 0.239 & 0.183 & 0.251 & 0.204 & 0.241 & 0.244 \\
##   & (0.223) & (0.221) & (0.223) & (0.223) & (0.221) & (0.219) & (0.222) & (0.221) & (0.222) & (0.222
##   & & & & & & & & & & \\
##  SEC & $-$0.209$^{***}$ & $-$0.210$^{***}$ & $-$0.208$^{***}$ & $-$0.210$^{***}$ & $-$0.207$^{***}$ &
##   & (0.026) & (0.026) & (0.026) & (0.026) & (0.026) & (0.026) & (0.026) & (0.026) & (0.026) & (0.026
##   & & & & & & & & & & \\
##  A60 & 0.161$^{***}$ & 0.131$^{***}$ & 0.077$^{*}$ & 0.118$^{***}$ & 0.149$^{***}$ & 0.142$^{***}$ &
##   & (0.043) & (0.044) & (0.043) & (0.041) & (0.042) & (0.042) & (0.043) & (0.042) & (0.042) & (0.043
##   & & & & & & & & & & \\
##  BED & $-$0.784$^{***}$ & $-$0.794$^{***}$ & $-$0.658$^{***}$ & $-$0.808$^{***}$ & $-$0.834$^{***}$ &
##   & (0.199) & (0.210) & (0.206) & (0.200) & (0.202) & (0.201) & (0.208) & (0.199) & (0.203) & (0.207
##   & & & & & & & & & & \\
##  DOC & $-$5.179$^{***}$ & $-$5.338$^{***}$ & $-$5.306$^{***}$ & $-$4.735$^{***}$ & $-$5.527$^{***}$ &
##   & (0.527) & (0.542) & (0.540) & (0.503) & (0.548) & (0.547) & (0.572) & (0.538) & (0.551) & (0.567
##   & & & & & & & & & & \\
##  NRS & 2.037$^{***}$ & 2.138$^{***}$ & 2.773$^{***}$ & 2.131$^{***}$ & 2.345$^{***}$ & 2.269$^{***}$
##   & (0.341) & (0.376) & (0.366) & (0.340) & (0.356) & (0.356) & (0.369) & (0.352) & (0.356) & (0.366
##   & & & & & & & & & & \\
```

```
##  AREA & $-$0.00005 & $-$0.0001$^{**}$ & 0.00004 & $-$0.00002 & $-$0.0001$^{*}$ & $-$0.0001$^{*}$ & $-
##   & (0.00003) & (0.00003) & (0.00004) & (0.00003) & (0.00003) & (0.00003) & (0.00003) & (0.00003) &
##   & & & & & & & & & & \\
##  POPDEN & $-$0.002$^{***}$ & $-$0.002$^{***}$ & $-$0.002$^{***}$ & $-$0.002$^{***}$ & $-$0.002$^{***
##   & (0.001) & (0.001) & (0.001) & (0.001) & (0.001) & (0.001) & (0.001) & (0.001) & (0.001) & (0.001
##   & & & & & & & & & & \\
##  GDPPC & 0.794$^{***}$ & 0.780$^{***}$ & 0.737$^{***}$ & 0.745$^{***}$ & 0.810$^{***}$ & 0.810$^{***
##   & (0.143) & (0.137) & (0.143) & (0.144) & (0.144) & (0.145) & (0.145) & (0.149) & (0.144) & (0.145
##   & & & & & & & & & & \\
##  PRPER & 1.039$^{***}$ & 1.048$^{***}$ & 1.050$^{***}$ & 1.070$^{***}$ & 1.059$^{***}$ & 1.039$^{***
##   & (0.106) & (0.108) & (0.109) & (0.109) & (0.109) & (0.111) & (0.109) & (0.108) & (0.109) & (0.109
##   & & & & & & & & & & \\
##  SECPER & 0.343$^{***}$ & 0.355$^{***}$ & 0.359$^{***}$ & 0.361$^{***}$ & 0.350$^{***}$ & 0.355$^{**
##   & (0.032) & (0.034) & (0.033) & (0.033) & (0.033) & (0.033) & (0.032) & (0.033) & (0.033) & (0.033
##   & & & & & & & & & & \\
##  ACTV & $-$6.622$^{***}$ & $-$6.453$^{***}$ & $-$6.382$^{***}$ & $-$6.310$^{***}$ & $-$6.505$^{***}$
##   & (0.383) & (0.357) & (0.356) & (0.344) & (0.361) & (0.372) & (0.345) & (0.357) & (0.356) & (0.366
##   & & & & & & & & & & \\
##  Constant & 5.800$^{*}$ & 5.266 & $-$27.771$^{***}$ & $-$1.110 & 4.475 & 5.844$^{*}$ & 1.955 & 6.708
##   & (3.262) & (3.357) & (5.311) & (3.081) & (3.371) & (3.292) & (3.196) & (3.534) & (3.260) & (3.272
##   & & & & & & & & & & \\
## \hline \\[-1.8ex]
## Observations & 12,870 & 12,870 & 12,870 & 12,870 & 12,870 & 12,870 & 12,870 & 12,870 & 12,870 & 12,8
## R$^{2}$ & 0.086 & 0.086 & 0.087 & 0.087 & 0.086 & 0.086 & 0.087 & 0.086 & 0.086 & 0.086 \\
## Adjusted R$^{2}$ & 0.085 & 0.085 & 0.086 & 0.086 & 0.085 & 0.085 & 0.086 & 0.085 & 0.085 & 0.085 \\
## Residual Std. Error (df = 12854) & 32.869 & 32.869 & 32.851 & 32.848 & 32.871 & 32.870 & 32.842 & 32
## F Statistic (df = 15; 12854) & 80.450$^{***}$ & 80.428$^{***}$ & 81.468$^{***}$ & 81.625$^{***}$ & 8
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{10}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

```r
results11 = lm(Case ~ TEMP + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV + TEMP*pop +
                 TEMP*GDP + TEMP*PRIM + TEMP*SEC + TEMP*A60 + TEMP*BED + TEMP*DOC +
                 TEMP*NRS + TEMP*AREA + TEMP*POPDEN + TEMP*GDPPC + TEMP*PRPER +
                 TEMP*SECPER + TEMP*ACTV  , dat = all_dat)
summary(results11)$coeff[2, 4]
```

```
## [1] 0.1972296
```

```r
results12 = lm(Case ~ WSPD + pop + GDP + PRIM + SEC  + A60 + BED + DOC + NRS +
                AREA + POPDEN + GDPPC + PRPER + SECPER + ACTV +
                 WSPD*pop + WSPD*GDP + WSPD*PRIM + WSPD*SEC + WSPD*A60 +
                 WSPD*BED + WSPD*DOC + WSPD*NRS + WSPD*AREA + WSPD*POPDEN +
                 WSPD*GDPPC + WSPD*PRPER + WSPD*SECPER + WSPD*ACTV , dat = all_dat)
summary(results12)$coeff[2, 4]
```

```
## [1] 0.05641236
```

```r
results13 = lm(Case ~ PRES + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                 AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV +
                 PRES*pop + PRES*GDP + PRES*PRIM + PRES*SEC + PRES*A60 + PRES*BED +
                 PRES*DOC + PRES*NRS + PRES*AREA + PRES*POPDEN + PRES*GDPPC +
                 PRES*PRPER + PRES*SECPER + PRES*ACTV, dat = all_dat)
```

```r
summary(results13)$coeff[2, 4]
```

```
## [1] 0.001115243
```

```r
results14 = lm(Case ~ HUM + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                 AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV +
                 HUM*pop + HUM*GDP + HUM*PRIM + HUM*SEC + HUM*A60 + HUM*BED +
                 HUM*DOC + HUM*NRS + HUM*AREA + HUM*POPDEN + HUM*GDPPC +
                 HUM*PRPER + HUM*SECPER  + HUM*ACTV , dat = all_dat)
summary(results14)$coeff[2, 4]
```

```
## [1] 0.5543436
```

```r
results15 = lm(Case ~ O3 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                 AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV +
                 O3*pop + O3*GDP + O3*PRIM + O3*SEC + O3*A60 + O3*BED +
                 O3*DOC + O3*NRS + O3*AREA + O3*POPDEN + O3*GDPPC + O3*PRPER +
                 O3*SECPER+ O3*ACTV , dat = all_dat)
summary(results15)$coeff[2, 4]
```

```
## [1] 0.4141118
```

```r
results16 = lm(Case ~ NO2 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                 AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV +
                 NO2*pop + NO2*GDP + NO2*PRIM + NO2*SEC + NO2*A60 + NO2*BED +
                 NO2*DOC + NO2*NRS + NO2*AREA + NO2*POPDEN + NO2*GDPPC +
                 NO2*PRPER + NO2*SECPER + NO2*ACTV , dat = all_dat)
summary(results16)$coeff[2, 4]
```

```
## [1] 0.07433373
```

```r
results17 = lm(Case ~ CO + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                 AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV +
                 CO*pop + CO*GDP + CO*PRIM + CO*SEC + CO*A60 + CO*BED +
                 CO*DOC + CO*NRS + CO*AREA + CO*POPDEN + CO*GDPPC +
                 CO*PRPER + CO*SECPER + CO*ACTV , dat = all_dat)
summary(results17)$coeff[2, 4]
```

```
## [1] 0.3803437
```

```r
results18 = lm(Case ~ SO2 + pop + GDP + PRIM + SEC + A60 + BED + DOC +
                 NRS + AREA + POPDEN + GDPPC + PRPER + SECPER  +
                 ACTV + SO2*pop + SO2*GDP + SO2*PRIM + SO2*SEC + SO2*A60 +
                 SO2*BED + SO2*DOC + SO2*NRS + SO2*AREA + SO2*POPDEN +
                 SO2*GDPPC + SO2*PRPER + SO2*SECPER + SO2*ACTV , dat = all_dat)
summary(results18)$coeff[2, 4]
```

```
## [1] 0.04133779
```

```r
results19 = lm(Case ~ PM2.5 + pop + GDP + PRIM + SEC + A60 + BED + DOC +
                 NRS + AREA + POPDEN + GDPPC + PRPER + SECPER  +
                 ACTV+ PM2.5*pop + PM2.5*GDP + PM2.5*PRIM + PM2.5*SEC +
                 PM2.5*A60 + PM2.5*BED + PM2.5*DOC + PM2.5*NRS + PM2.5*AREA +
                 PM2.5*POPDEN + PM2.5*GDPPC + PM2.5*PRPER + PM2.5*SECPER
                  + PM2.5*ACTV, dat = all_dat)
summary(results19)$coeff[2, 4]
```

```
## [1] 0.45548
```

```r
results20 = lm(Case ~ PM10 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                 AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV +
                 PM10*pop + PM10*GDP + PM10*PRIM + PM10*SEC + PM10*A60 +
                 PM10*BED + PM10*DOC + PM10*NRS + PM10*AREA + PM10*POPDEN +
                 PM10*GDPPC + PM10*PRPER + PM10*SECPER + PM10*ACTV, dat = all_dat)
summary(results20)$coeff[2, 4]
```

```
## [1] 0.7941046
```

```r
rob_se_2 <- list(sqrt(diag(vcovHC(results11, type = "HC1"))),
               sqrt(diag(vcovHC(results12, type = "HC1"))),
               sqrt(diag(vcovHC(results13, type = "HC1"))),
               sqrt(diag(vcovHC(results14, type = "HC1"))),
               sqrt(diag(vcovHC(results15, type = "HC1"))),
               sqrt(diag(vcovHC(results16, type = "HC1"))),
               sqrt(diag(vcovHC(results17, type = "HC1"))),
               sqrt(diag(vcovHC(results18, type = "HC1"))),
               sqrt(diag(vcovHC(results19, type = "HC1"))),
               sqrt(diag(vcovHC(results20, type = "HC1"))))
```

```r
results21 = plm(Case ~ TEMP + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results21, vcov. = vcovHC, type = "HC1")

results22 = plm(Case ~ WSPD + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results22, vcov. = vcovHC, type = "HC1")

results23 = plm(Case ~ PRES + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results23, vcov. = vcovHC, type = "HC1")

results24 = plm(Case ~ HUM + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results24, vcov. = vcovHC, type = "HC1")

results25 = plm(Case ~ O3 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results25, vcov. = vcovHC, type = "HC1")

results26 = plm(Case ~ NO2 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results26, vcov. = vcovHC, type = "HC1")

results27 = plm(Case ~ CO + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results27, vcov. = vcovHC, type = "HC1")
```

```r
results28 = plm(Case ~ SO2 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results28, vcov. = vcovHC, type = "HC1")

results29 = plm(Case ~ PM2.5 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results29, vcov. = vcovHC, type = "HC1")

results30 = plm(Case ~ PM10 + pop + GDP + PRIM + SEC + A60 + BED + DOC + NRS +
                  AREA + POPDEN + GDPPC + PRPER + SECPER  + ACTV ,
                data = all_dat, model = "within", index = c("City"))
coeftest(results30, vcov. = vcovHC, type = "HC1")

rob_se_3 <- list(sqrt(diag(vcovHC(results21, type = "HC1"))),
               sqrt(diag(vcovHC(results22, type = "HC1"))),
               sqrt(diag(vcovHC(results23, type = "HC1"))),
               sqrt(diag(vcovHC(results24, type = "HC1"))),
               sqrt(diag(vcovHC(results25, type = "HC1"))),
               sqrt(diag(vcovHC(results26, type = "HC1"))),
               sqrt(diag(vcovHC(results27, type = "HC1"))),
               sqrt(diag(vcovHC(results28, type = "HC1"))),
               sqrt(diag(vcovHC(results29, type = "HC1"))),
               sqrt(diag(vcovHC(results30, type = "HC1"))))
```

```r
table3 = stargazer(results21, results22, results23,
          results24, results25, results26, results27, results28, results29, results30,
          digits = 3,
          header = FALSE,
          type = "latex",
          se = rob_se_3,
          title = "Fixed Effects Regression Analysis",
          model.numbers = FALSE,
          column.labels = c("(1)", "(2)", "(3)", "(4)", "(5)", "(6)", "(7)", "(8)", "(9)", "(10)"))
```