How to Understand PM2.5

Bowen Wei

School of North China Electric Power University, Baoding 071000, China

wenuf@sina.com

Keyword: PM2.5, multiple linear regression, principal component analysis

Abstract. In view of the factors that affect PM2.5, we set up two models. Multiple linear regression models were used in the establishment of internal factors. Consider that there are too many external factors, we adopt the dimension reduction method, using the principal component analysis method. Through the internal analysis, we found that CO is the main factor that affects the value of PM2.5. Through the external analysis, we found that rainfall is the main factor affecting the value of PM2.5.

When it comes to the relationship between PM2.5 and the haze, we build model three. We learn that AQI is a major index reflecting the weather conditions, so we transform the study of relationship between PM2.5 and the haze into a study of the relationship between PM2.5 and AQI. After a scatter plot analysis, we use a linear regression model. After residual analysis, we find that the fitting degree of the model is very high. From this, we come to the conclusion that PM2.5 caused by the haze, while the haze reflects the high value of PM2.5

Our models fit the reality well, which can help people understand PM2.5 more scientifically and intuitively. These analyses are also helpful in the management of air pollution. In the future, we can further refine the model from the angle of time and space.

1. Introduction

Haze is the general expression about excessive levels of suspended particulate matter content in atmosphere. With the deterioration of air quality, fog and haze phenomenon increased. In many areas of China, the haze weather phenomenon is incorporated into the fog as a disaster weather warning. The haze has gained wide attention since it appeared in 2013.According to data published by the Ministry of environmental protection in China, AQI of Baoding often exceeds 400 in a few days in a row, that is, "serious pollution", which caused a serious impact on local people's daily lives.

In the study of the main factors affecting the haze, main factors was analyzed affecting on the national large-scale haze phenomenon in November 2013 by using two kinds of statistical analysis methods: non parametric statistical binding method of multivariate regression factor analysis and correspondence analysis method in multivariate statistical analysis. Based on results of the above two kinds of methods and combined with the western developed countries' governance experience on haze pollution, effective methods were proposed to alleviate the haze in China [1].

In the study of the relationship between PM2.5 and haze, a research team used PM2.5/AQI ratio to measure the possibility scale of PM2.5 in haze weather. This ratio can directly characterize PM2.5 in air (haze) influence, arriving at the conclusion that PM2.5 is closely related to haze weather [2].

2. Assumptions and Notations

To facilitate the construction of the model, we make the following assumptions:

- The data in the Baoding Environmental Protection Bureau and other authoritative website is accurate.
- There is no strict divided boundaries between PM2.5/AQI. That is to say, the values of

PM2.5/AQI is at or near the limit, even can be thought in any interval.

- Sunny and thunderstorms weather will not appear fog and haze.
- The error of the estimated weather condition from the picture can be ignored.

We use a list of symbols (cf. Table 1) for simplification of expression.

Table 1. Notations (in the order of model)

Symbol	Definition	Notes
	Model One	
AQI	air quality index	Table 2
y	dependent variable	Equation (1)
x	independent variable	Equation (1)
β	partial regression coefficient	Equation (1)
ε	residual	Equation (1)
	Model Two	
CV	characteristic value	Table 5
CR	contribution rate	Table 5
CCR	cumulative contribution rate	Table 5
	Model Three	
	The same as Model One	

3. The model

Under certain conditions,SO_2, NO, NO_2 in AQI detection indicators is the main gaseous objects before PM2.5 formed [3].

3.1 Model One

We build model one to analyze the internal factors affecting PM2.5. Model one is called *multiple linear regression model*.

We collect data from some professional web [4]. The detection data range from 2015 January 1 to 2015 December 31 in Baoding. We use Sampling method and monthly selected No. 5, 15 and 25 as samples, monitoring the primary pollutants for PM10, SO_2,CO, NO_2,O_3. The initial data is as follows:

Table 2. The Initial Data of Five Main Factors

Date	AQI	PM2.5	PM10	SO_2	CO	NO_2	O_3
1.05	327	265.7	450	149.9	5.512	113.3	49
1.15	365	299.7	462.8	97.5	4.861	82.7	13
1.25	163	126.3	181.6	75.5	2.603	47	28
2.05	112	79.5	137.3	115.1	2.162	47.7	82
2.15	264	206.8	364.9	125.9	3.967	70.6	51
2.25	184	139.2	253.5	64.4	2.213	36	82
3.05	141	106.6	175	108.5	1.596	57.4	85
3.15	198	132.3	289.3	99.2	2.166	70.6	110
3.25	174	128.3	243.8	50.4	2.166	70.5	114
4.05	112	69.2	130.1	17.1	1.017	34	108
4.15	194	114.4	252.8	63.7	1.517	57.8	125
4.25	126	90.9	175	52.1	1.254	46.9	233
5.05	116	84.2	151.7	47.8	1.092	49.7	190
5.15	72	39.6	85.6	20.7	0.567	34.3	159
5.25	159	105.3	162.8	56.7	1.164	40.9	281
6.05	90	51.9	127.3	42.6	0.729	45.5	216
6.15	121	84.4	165.4	47.3	1.267	59.5	266

6.25	97	71	113.3	22.6	0.81	29.8	194
7.05	95	56.1	90	19.9	0.796	30.5	225
7.15	105	76.5	134.7	26.9	0.927	51	146
7.25	121	91.5	155.3	24.2	1.097	35.9	238
8.05	132	98.8	169.8	18.7	1.426	30.5	172
8.15	116	80	140.2	17.6	0.683	48.9	165
8.25	49	23.3	53.5	10.1	0.485	26.5	126
9.05	51	33.6	52.6	9.7	0.501	17.1	102
9.15	128	96.9	168	18.8	0.99	64.2	178
9.25	60	32	71.6	17.7	0.503	38	122
10.05	228	179.7	280.2	32.6	1.781	97.8	217
10.15	231	167.9	287	40.9	2.32	89.6	266
10.25	59	26.3	68.9	47	0.872	67.3	35
11.05	149	100.8	176.6	34.2	1.084	73.1	46
11.15	202	153.8	225.5	59.5	3.793	98.9	33
11.25	79	57.3	70.4	58.1	1.676	41.8	60
12.05	284	237.8	323.3	118.2	5.189	99.8	11
12.15	48	21.5	51.3	28.4	0.737	23.5	69
12.25	307	250.8	375.8	61.6	6.698	110.2	23

Then we established a multivariate linear regression model based on PM2.5 as the dependent variable and SO2 CO NO2 O3 PM10 as the independent variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \quad (1)$$

The parameters of the β, from β_0 to β_5 are regression coefficients remained to be estimated, ε is random error (roughly follows a mean zero normal distribution) .

Table 3. Results of Linear Regression Model

Subscript Value of β	β	Confidence Interval of β			
0	-6.0001	-22.3165 -- 10.3163			
1	0.4719	0.3748 -- 0.5689			
2	-0.1259	-0.3081 -- 0.0563			
3	15.5195	8.2927 -- 22.7463			
4	0.0209	-0.2791 -- 0.3210			
5	0.0254	-0.0490 -- 0.0997			
Parameter	R^2	F		P	S^2
Value	0.9726	213.1029		1.7303	159.4828

From the table we can see that the R^2=0.9726 shows 97.26% of variable Y (standard of PM2.5 levels) can be determined by the established model. The value of F exceeds the critical value of the F test and P is much smaller than that of α = 0.05, so the model from the overall look is available. So we can get a rough linear model:

$$y = -6.0001 + 0.4719x_1 - 0.1259x_2 + 15.5195x_3 + 0.0209x_4 + 0.0254x_5 \quad (2)$$

The confidence interval of Beta 0, beta 2, beta 4, beta 5 contains zero, which means the impact of X2, X4, X5, on variable Y is not significant, so we focus on the consideration of X1 and X3. From the estimation of parameter values we can find that PM2.5 was affected mostly by X3 , farther more, they have a positive correlation, indicating that carbon dioxide produced by burning coal is the main internal effects of PM2.5 which means the more coal burned, the bigger the PM2.5 value is. This also explains why PM2.5 is so serious in the heating season. X1 means the content of PM10, which is positively correlated with the content of PM2.5.

Both of X1 and X2 are consistent with the actual situation, which shows that the linear fitting equation is significant.

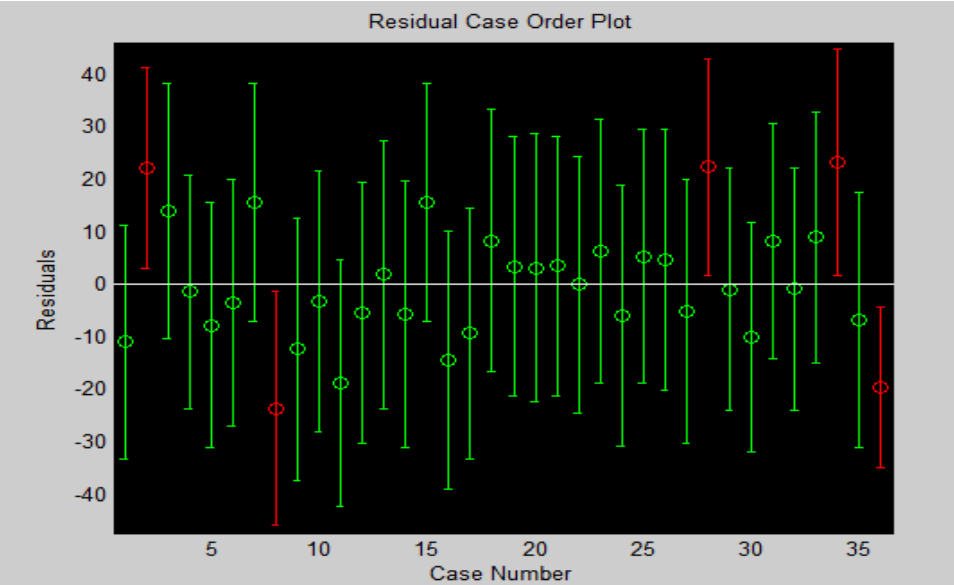

Figure 1. Residual Plot of Model One

From residual plot, we can see the distance between data residuals and zero. When the residual confidence intervals include zero, indicating that the regression model can fit the original data well, or could be regarded as outliers. From the figure, we can see only five outliers, means the regression model is fit to the actual situation.

3.2 Model Two

Model two is also used to analyze the factors affecting PM2.5. Model two is considered from the external causes. We use *principal component analysis* to build our model.

We collect related data from professional web [5]. We collected meteorological data of Baoding in 2015. The data shows as follows:

Table 4. Meteorological Data

Month	Temperature	Dew Point	Pressure	Wind Speed	Rainfall
1	-1	-12	1029	7	0.8
2	2	-10	1025	7	4.3
3	9	-5	1021	8	10.2
4	15	6	1015	9	32.2
5	20	11	1008	9	42.9
6	25	15	1004	9	11.8
7	27	20	1005	8	194.5
8	26	21	1008	6	29.7
9	21	16	1015	7	64
10	15	7	1020	7	11.1
11	4	1	1027	6	35.5
12	0	-5	1029	6	0.7

We used MATLAB, calculated the characteristics of numerical matrix of the five factors of twelve months' values were 0.0121, 0.0372, 0.5461, 0.8981, 3.5065, so we can calculate the contribution rate was 0.24%, 0.74%, 10.92%, 17.96%, 70.13%.

Table 5. The Result of Influence Degree Determination

Factors	CV	CR (%)	CCR (%)
Rainfall	3.5065	70.13	70.13
Wind Speed	0.8981	17.96	88.09
Pressure	0.5461	10.92	99.01
Dew point	0.0372	0.74	99.76
Temperature	0.0121	0.24	100

These data are intuitive to show that rainfall, wind speed and pressure decided effect is big in these five factors affecting PM2.5. In addition, the rainfall is bigger relative to the role of wind speed and pressure, their contribution rate and 99.01%. Thus, we can think that these three factors are the main factors influencing PM2.5. The principal component analysis was significant.

3.3 Model Three

We need to look at the relationship between PM2.5 and haze. Whether PM2.5 is the main factor which influence air quality (we use AQI to measure the main factors). The data we collected is shown in Table 2 as above.

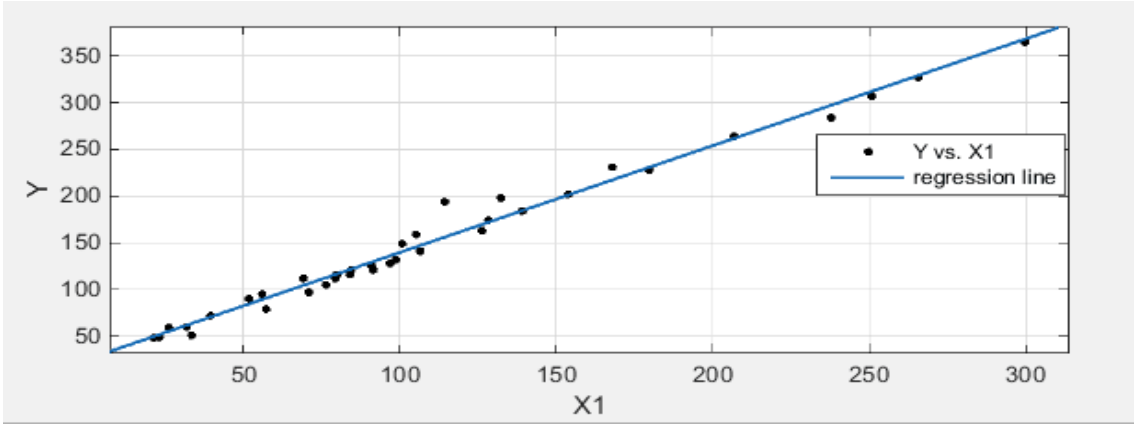

Figure2. Sample Regression Line

From the point of scatter diagram, it can be seen that PM2.5 and AQI are closely related, and there is a linear positive correlation between them. Based on this, we construct *linear regression model* to analyze it.

Then we establish a linear regression model.

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (3)$$

By solving the equation with MATLAB, the regression equation of the model is

$$y = 25.1108 + 1.1445 x_1 \quad (4)$$

Table 6. Results of Linear Regression Model

Parameter	R^2	F	P	S^2
Value	0.9850	2229.6191	1.3791	102.6446

Assuming a given α=0.05, through the calculation and look-up table we can obtain a result that F>>F0.05 (1, 34), so there is a very significant linear relationship between PM2.5 and AQI.

In order to further verify the above regression model's fit, we use related data for residual analysis. We can see that there are only two outliers, which means that the regression model is in line with the actual situation. That is a good model fit, PM2.5 is indeed the main factors affecting the haze. PM2.5 caused by the fog and haze, while the fog and haze reflects the high value of PM2.5.

4. Conclusion

From the above analysis, we can see that the PM2.5 of Baoding is mainly influenced by PM10. Coal combustion, vehicle exhaust, construction site dust and is the main source of PM2.5. Time

changes from the point of view, the pollutants in the city has obvious time variation characteristics and relative higher during heating and no heating period is relatively low. From regional distribution point of view, due to the business district is located in the downtown area, the environment is more complex, in addition the business district is located in the old city, central heating rate and clean energy utilization rate is relatively low, and the concentration of pollutants is relatively high.

Air polluted in Baoding is mainly coal-burning pollution. The particulate matter is the primary air pollutants in urban areas. Coal-fired flue gas and automobile exhaust gas and the dust from the ground is the main source of PM2.5 in Baoding. Downtown by the heat and power plant centralized heating and gas heating ratio accounted for only 30%. The rest mainly is scattered coal-fired boiler. Automobile exhaust pollution is becoming more and more serious. So it is still difficult to fundamentally improve the air quality in Baoding.

Reference

[1] Shaorong Feng. factors haze influence and measures based on statistical analysis methods [M]

[2] Qingchun Li, Weihua Cao. The Characteristics of the Haze and The Analysis of Influencing Factors in Baoding. China Meteorological Administration[J]. Baoding Institute of Urban Meteorology, August 28, 2013

[3] Feng Liu, Li Yin, Xing Zhang. Part of the Linear Model in the Practice and Understanding of the Air Quality Index of Fine Particulate Matter PM 2.5 analysis. [J]．Applied Mathematics44, 2014 (9):130 – 134

[4] PM2.5 China [N].

[5] Weather Underground [N].

[6] Baidu Encyclopedia [N].