Application of Multiple Regression in Mathematical Model of Air Quality Calibration

Xiaohua Que¹, Hong Li ², *

¹School of Mathematics and Information Science, Guangxi College of Education, Nanning, Guangxi, China
²Nanning Normal University, Nanning, China

*Corresponding author e-mail: 248241080@qq.com

Abstract. Based on the existing self-built point data, state-controlled point data and wind speed, pressure, precipitation, temperature, humidity and other data, a multivariate regression mathematical model is established in this paper to optimize the self-built point data. The model is optimized via data to find out the factors that have great influence on the data of self-built points and get more accurate regression equation model. After verification, it can be seen that the calibration degree of the model to the self-built point data is high.

Keywords: Air Quality, Data Calibration, Mathematical Model

1. Introductions

With the rapid development of economy, all countries begin to pay attention to the problem of environmental pollution, especially how to protect air quality. In 2017, 239 out of 338 cities at or above the prefectural level exceeded the standard in the urban environment air quality [1]. The first step to improve air quality is to improve the method of air quality detection and improve the accuracy of air quality detection. Air quality test data should be accurate, real-time and intelligent [2]. A wireless sensor network air quality detection system- CASPP prediction system has been designed by the researchers [3, 4]. Since 1996, China has formulated the quality standard of air detection, and began to build automatic detection system. China has built 1436 air quality inspection stations in 338 prefecture-level cities and above [5]. At present, due to the limitation of objective conditions, it is impossible to ensure that the national air quality monitoring and control stations can be distributed in all regions, so it is difficult to meet the actual needs. But self-built sites can be used to detect air quality. Self-built sites are less expensive, but less accurate, so calibration is required. In the existing research, there is a lack of calibration research on air quality detection self-built point data. Therefore, based on the state control point and self-built point data of air quality, the regression mathematical model is established to calibrate the self-built point data, so that it has high accuracy.
2. Data and methods

2.1. Data
The data in this paper are the data of each hour collected by a national control point in a certain period of time and the corresponding self-built point data. Table 1 shows some of the data. For complete information, please refer to question D of the 2019 national college students mathematical modeling competition. The data are explained as follows:

(1) Self-built point data can affect the accuracy of mathematical models. In order to explore the data relationship between self-built point and state control point and achieve the purpose of calibration, we assume that the temperature, humidity, wind speed, air pressure and precipitation data monitored by self-built point are accurate values.

(2) Air quality may also be affected by other factors, but in this problem we only consider the environmental factors (temperature, humidity, wind speed, air pressure, precipitation) given by the topic, and the other factors are ignored.

Table 1. Question D of the 2019 national college students mathematical modeling competition

| PM 2.5 | PM10 | CO   | NO₂  | SO₂  | O₃   | Time           |
|--------|------|------|------|------|------|----------------|
| 33     | 71   | 0.756| 9    | 25   | 80   | 2018/11/14 10:00 |
| 32     | 69   | 0.736| 9    | 22   | 86   | 2018/11/14 11:00 |
| 33     | 64   | 0.804| 9    | 26   | 88   | 2018/11/14 12:00 |
| 35     | 63   | 0.75 | 8    | 26   | 88   | 2018/11/14 13:00 |
| 30     | 69   | 0.855| 8    | 34   | 88   | 2018/11/14 14:00 |

2.2. Modeling
Regression models are established with wind speed, pressure, precipitation, temperature and humidity, respectively, using the differences between state-controlled point data and self-built point data PM 2.5, PM10, CO, NO₂, SO₂ and O₃. Supposing the wind speed, pressure, precipitation, temperature and humidity as variables which affect the data of self-built points X₁, X₂, X₃, X₄ and X₅, PM 2.5 PM10, CO, NO₂, SO₂ and O₃ are Y₁, Y₂, Y₃, Y₄, Y₅ and Y₆ respectively. The following regression relationship can be obtained:

\[ Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \nu \]

\[ Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \nu \]

\[ Y_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \nu \]

\[ Y_4 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \nu \]

\[ Y_5 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \nu \]

\[ Y_6 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \nu \]
3. Results and inspection

3.1. Solution of model

By substituting the corresponding values into the regression equation, the following table can be obtained by excel calculation:

**Table 2.** The parameter values of the regression equation corresponding to $PM_{2.5}$

| Parameters | Parameter estimates | Confidence interval |
|------------|---------------------|---------------------|
| $\beta_0$  | 123                 | 123                 |
| $\beta_1$  | 644                 | 644                 |
| $\beta_2$  | 649                 | 649                 |

With the estimates of regression coefficients $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ and $\beta_5$ in Table 2, the equation with each factor is obtained:

$$Y = -427.483 + 3.3240X_1 + 0.4060X_2 + 0.0106X_3 + 0.0500X_4 + 0.3631X_5 + 18.7784$$

Similarly, according to the data in the table (supporting materials 1, 2), the equations between $PM_{10}$, $CO$, $NO_2$, $SO_2$, $O_3$ and wind speed, air pressure, precipitation, temperature and humidity can be obtained.

By substituting the corresponding values of $PM_{10}$ into the regression equation, the following table can be obtained:

**Table 3.** The parameter values of the regression equation corresponding to $PM_{10}$

| Parameters | Parameter estimates | Confidence interval |
|------------|---------------------|---------------------|
| $\beta_0$  | -2945.56            | [-3304.98, -2586.14]|
| $\beta_1$  | -2.0647             | [-4.99061, 0.861203]|
| $\beta_2$  | 2.801961            | [2.456057, 3.147865]|
| $\beta_3$  | 0.04789             | [0.032545, 0.063236]|
| $\beta_4$  | 1.22928             | [0.832208, 1.626351]|
| $\beta_5$  | 1.606707            | [1.522871, 1.690544]|

$$Y_2 = -2945.56 - 2.0647X_1 + 2.8020X_2 + 0.0479X_3 + 1.2293X_4 + 1.6067X_5 + 41.6963$$

By substituting the corresponding values of $CO$ into the regression equation, the following table can be obtained:

**Table 4.** The parameter values of the regression equation corresponding to $CO$

| Parameters | Parameter estimates | Confidence interval |
|------------|---------------------|---------------------|
| $\beta_0$  | -36.7802            | [-40.4927, -33.0677]|
| $\beta_1$  | 0.13624             | [0.106018, 0.166462]|
| $\beta_2$  | 0.034939            | [0.031366, 0.038512]|
| $\beta_3$  | -0.00038            | [-0.00054, -0.00022]|
| $\beta_4$  | 0.041732            | [0.037631, 0.045834]|
| $\beta_5$  | 0.001996            | [0.001113, 0.002862]|

By substituting the corresponding values of $PM_{10}$ into the regression equation, the following table can be obtained:
\[ Y_3 = -36.780 + 0.1362X_1 + 0.0349X_2 - 0.0004X_3 + 0.0417X_4 + 0.0020X_5 + 0.4306 \]

(4) By substituting the corresponding values of NO\textsubscript{2} into the regression equation, the following table can be obtained by excel calculation:

**Table 5.** The parameter values of the regression equation corresponding to NO\textsubscript{2}

| Parameters | Parameter estimates | Confidence interval |
|------------|---------------------|---------------------|
| \( \beta_0 \) | -1821.78 | [-2026.34, -1617.22] |
| \( \beta_1 \) | 0.218427 | [-1.44681, 1.883677] |
| \( \beta_2 \) | 1.720344 | [1.523478, 1.917211] |
| \( \beta_3 \) | 0.1162 | [0.107466, 0.124933] |
| \( \beta_4 \) | 2.185918 | [1.959931, 2.411906] |
| \( \beta_5 \) | 0.767048 | [0.719334, 0.814762] |

\[ Y_4 = -1821.78 + 0.2184X_1 + 1.7203X_2 + 0.1162X_3 + 2.1859X_4 + 0.7670X_5 + 23.7308 \]

(5) By substituting the corresponding values of SO\textsubscript{2} into the regression equation, the following table can be obtained by excel calculation:

**Table 6.** The parameter values of the regression equation corresponding to SO\textsubscript{2}

| Parameters | Parameter estimates | Confidence interval |
|------------|---------------------|---------------------|
| \( \beta_0 \) | 1441.893 | [1191.943, 1691.843] |
| \( \beta_1 \) | 5.278797 | [3.244045, 7.313548] |
| \( \beta_2 \) | -1.38458 | [-1.62513, -1.14403] |
| \( \beta_3 \) | -0.05993 | [-0.0706, -0.04926] |
| \( \beta_4 \) | -1.48078 | [-1.75691, -1.20464] |
| \( \beta_5 \) | -0.22007 | [-0.27837, -0.16176] |

\[ Y_5 = 1441.893 + 5.2788X_1 - 1.3846X_2 - 0.0599X_3 - 1.4808X_4 - 0.2201X_5 + 28.9967 \]

(6) By substituting the corresponding values of O\textsubscript{2} into the regression equation, the following table can be obtained:

**Table 7.** The parameter values of the regression equation corresponding to O\textsubscript{2}

| Parameters | Parameter estimates | Confidence interval |
|------------|---------------------|---------------------|
| \( \beta_0 \) | -852.478 | [-1163.87, -541.085] |
| \( \beta_1 \) | -23.3023 | [-25.8377, -20.7669] |
| \( \beta_2 \) | 0.828216 | [0.528534, 1.127899] |
| \( \beta_3 \) | 0.123725 | [0.110428, 0.137022] |
| \( \beta_4 \) | -0.89543 | [-1.23944, -0.55141] |
| \( \beta_5 \) | 0.474315 | [0.401669, 0.546961] |
\[ Y_6 = -852.478 - 23.3023X_1 + 0.8282X_2 + 0.1237X_3 - 0.8954X_4 + 0.4743X_5 + 36.1329 \]

By calculating the difference between the state-controlled point and the self-built point and the multiple linear regression, we obtain the multivariate linear regression equation. We analyze the results and find that there are different degrees of correlation among these factors. So we also need to eliminate the small factors and do further optimization.

### 3.2. Model inspection

In the original data, the representative is substituted into the above equations to test, and the following results are obtained:

\[ Y_1 = -427.483 + 3.3240 \times 1.25 + 0.4060 \times 1017.35 + 0.3631 \times 53 + 18.7784 \]
\[ Y_2 = -2945.56 + 2.8020 \times 1017.35 + 1.2293 + 1.6067 \times 53 + 41.6963 \]
\[ Y_3 = -36.780 + 0.1362 \times 1.25 + 0.0349 \times 1017.35 + 0.0417 \times 17 + 0.4306 \]
\[ Y_4 = -1821.78 + 1.7203 \times 1017.35 + 2.1859 \times 17 + 0.7670 \times 53 + 23.7308 \]
\[ Y_5 = 1441.893 + 5.2788 \times 1.25 - 1.3846 \times 1017.35 - 1.4808 \times 17 + 28.9967 \]
\[ Y_6 = -852.478 - 23.3023 \times 1.25 + 0.8282 \times 17 + 0.1237 \times 53 + 36.1329 \]

**Table 8. The results of the first data validation**

| Verification | \( Y_1 \) | \( Y_2 \) | \( Y_3 \) | \( Y_4 \) | \( Y_5 \) | \( Y_6 \) |
|--------------|-----------|-----------|-----------|-----------|-----------|-----------|
| Standard     | 33        | 76        | 0.763     | 8         | 37        | 89        |

It can be seen from the table that the results of the verification and the standard have different degrees of correlation, and the correlation degree between and is quite different, so further optimization is needed.

### 3.3. Model optimization

Considering that some factors are not high, they are eliminated directly. The following optimized regression model is obtained.

\[ Y_1 = -427.483 + 3.3240 \times 0.75 + 0.4060 \times 1019.55 + 0.3631 \times 94 + 18.7784 \]
\[ Y_2 = -2945.56 + 2.8020 \times 1019.55 + 1.2293 \times 12 + 1.6067 \times 94 + 41.6963 \]
\[ Y_3 = -36.780 + 0.1362 \times 0.75 + 0.0349 \times 1019.55 + 0.0417 \times 12 + 0.4306 \]
\[ Y_4 = -1821.78 + 1.7203 \times 1019.55 + 2.1859 \times 12 + 23.7308 \]
\[ Y_5 = 1441.893 + 5.2788 \times 0.75 - 1.3846 \times 1019.55 - 1.4808 \times 12 + 28.9967 \]
\[ Y_6 = -852.478 - 23.3023 \times 0.75 - 0.8954 \times 12 + 36.1329 \]

In the original data, the representative is taken into the above equations to test, and the following results are obtained:

\[ Y_1 = -427.483 + 3.3240 \times 0.75 + 0.4060 \times 1019.55 + 0.3631 \times 94 + 18.7784 \]
\[ Y_2 = -2945.56 + 2.8020 \times 1019.55 + 1.2293 \times 12 + 1.6067 \times 94 + 41.6963 \]
\[ Y_3 = -36.780 + 0.1362 \times 0.75 + 0.0349 \times 1019.55 + 0.0417 \times 12 + 0.4306 \]
\[ Y_4 = -1821.78 + 1.7203 \times 1019.55 + 2.1859 \times 12 + 23.7308 \]
\[ Y_5 = 1441.893 + 5.2788 \times 0.75 - 1.3846 \times 1019.55 - 1.4808 \times 12 + 28.9967 \]
\[ Y_6 = -852.478 - 23.3023 \times 0.75 - 0.8954 \times 12 + 36.1329 \]

**Table 9. The second data validation results**

| Verification | \( Y_1 \) | \( Y_2 \) | \( Y_3 \) | \( Y_4 \) | \( Y_5 \) | \( Y_6 \) |
|--------------|-----------|-----------|-----------|-----------|-----------|-----------|
| Standard     | 35        | 29        | 0.68      | 6         | 33        | 42        |
4. Conclusion
Because the self-built point will be affected by many factors in the detection of air quality, it is difficult to describe it very accurately by mathematical model. However, based on a large number of data, this paper establishes a relatively simple regression model, which makes it have certain application value. The calibrated self-built point data is not different from the state-controlled point data, and the model has high accuracy. In addition, the linear regression model is adopted uniformly when considering the influencing factors. Through data calibration, it is found that there are some errors in the difference between them, which may have a nonlinear relationship with them. It is the direction worthy of further study.

Acknowledgments
This work was financially supported by the Basic Ability Improvement Projects for Young and Middle-aged Teachers from Colleges of Guangxi in 2019 (NO.2019KY1679).

References
[1] Qu Changyu. Study on air quality assessment and influencing factors of major cities in China , D. Jilin University of Finance and Economics, 2019.
[2] Lyu Zhen. Design and implementation of urban air quality detection system, D. Zhengzhou University, 2014.
[3] Zhu Rong, Xu Dahai, Meng Yanjun, Cheng Conglan, Shao Demin. Urban air pollution numerical prediction system CAPPS and its application, J. Journal of Applied Meteorological Science, No.03 (2001) 267-278.
[4] Yang Yang, Zhu Shan'an. Design and implementation of environmental monitoring system based on wireless sensor network, J. Industrial Control Computer, Vol.020 No.009 (2007) 6-8.
[5] Zhou Fei. Discussion on quality control of automatic monitoring of ambient air, J. China Resources Comprehensive Utilization, Vol.36 No.11 (2018) 166-168.