The Data Analysis of Shanghai Air Quality Index Based on Linear Regression Analysis

Wenrong Jiang*
School of Computer and Information, Shanghai Polytechnic University, Shanghai, China

*Corresponding author email: wrjiang@sspu.edu.cn

Abstract. In this paper, the actual data related to air quality in Shanghai is selected, and six factors such as PM2.5, PM10, SO2, NO2, O3 and CO concentration are used as the alternative factors affecting AQI. The Pearson correlation coefficient between these six factors and the air quality index was calculated separately, and the correlation between the variables was analyzed. This paper considers the multicollinearity between the six factors, in order to reduce the influence of multicollinearity on the model, comprehensively consider the elimination of some alternative factors. Then use spss software to analyze in depth, take AQI as the dependent variable, and the other factors as the independent variable to carry out multiple linear regression analysis. Finally, the multiple linear regression analysis equation is obtained and predicted based on this model. The results show that the model is for calculating air. The quality index is more accurate. This model is simpler and more efficient than the original AQI calculation model.

Keywords: Air quality index; Multiple Linear Regression; Multicollinearity; Variable Transformation; Prediction.

1. Introduction
In recent years, China's air pollution problem is still a big problem in the environment. The air quality in Shanghai in the past two years still has a great impact on some special people. Especially in the rapid development of China's economy after the 1990s, major environmental pollution incidents are obviously increasing, and people's lives are also affected more or less. There are many domestic and international reports on the impact of air pollution on people's health. The Air Quality Index (AQI) is a dimensionless index that quantitatively describes the state of air quality.\[1\] The air quality sub-index is also specified for individual pollutants. The main pollutants involved in air quality assessment are fine particles, respirable particulate matter, sulfur dioxide, nitrogen dioxide, ozone, and carbon monoxide.

| Number | The level of AQI | Evaluation       |
|--------|------------------|------------------|
| 0~50   | I                | Excellent        |
| 51~100 | II               | Good             |
| 101~150| III              | Lightly polluted |
| 151~200| IV               | Moderately polluted |
| 201~300| V                | Heavily polluted |
| >300   | VI               | Severely polluted |
In 2000, Neha Khanna proposed the use of the Air Pollution Index (API) system to evaluate air quality; in the first half of 2012, the Air Quality Index (AQI) was used to replace the original Air Pollution Index (API). The air quality index is divided into six levels, from first-grade, second-class, third-class mild pollution, four-level moderate pollution, to five-level heavy pollution, and six-level serious pollution. The higher the level, the more serious the pollution situation. The health hazard to the human body is greater (Table 1)\cite{4}\cite{5}. The study of air quality index has long attracted the attention of scholars at home and abroad.

2. Research Method

2.1. Air Quality Index Original Calculation Model

There are many types of ambient air pollutants, such as nitrogen dioxide (NO2), carbon monoxide (CO), sulfur dioxide (SO2), ozone (O3) and suspended particulate matter, which is always the same as the air quality index. It refers to one of the suspended particles. The Urban Air Quality Index (AQI) is the daily average of several pollutants measured, and the maximum value of the hourly average air quality sub-index is used as the index of daily, hourly air quality and primary pollutants. Therefore, To know the AQI value, the sub-index of each pollutant (IAQI) must be calculated. First, by measuring the degree of fitness of the multiple linear regression, similar to the decision coefficient in the linear regression, multiple decision coefficients are used, which are defined as:

\[
IAQIp = \frac{IAQIHI - IAQILD}{BPHI - BPLD} (Cp - BPL0) + IAQIL0
\]  

(1)

IAQI, the air quality sub-index of the pollutant item P;
Cp - Contaminant item P mass concentration value;
BP - the high value of the contaminant concentration limit similar to Cp;
BPL - the low value of the contaminant concentration limit similar to Cp;
IAQIHI - the air quality sub-index corresponding to BP;
IAQIL - Air quality sub-index corresponding to BPL;
After calculating the sub-index of each monitoring index, the maximum value of the sub-index is taken as the final air quality index (AQI).

2.2. Principle of Multiple Linear Regression Model

In real life, the Multivariable Linear Regression Model refers to a variable that is often affected by multiple variables\cite{3}. For example, the wages people receive are affected by factors such as age, expenditure, working hours, and the company, in addition to the type of occupation. The general form of a multiple linear regression model is

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i \quad i = 1, 2, ..., n
\]  

(2)

Where k is the number of explanatory variables, and \(\beta_j\) (j = 1, 2, ..., k) is called Regression Coefficient. The above formula (2) is also called a random expression of the population regression function. Its non-random expression is

\[
E(Y|X_{i1}, X_{i2}, ..., X_{ik}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}
\]  

(3)

\(\beta_j\) is called Partial Regression Coefficient.
The \(R^2\) becomes the negative correlation coefficient (R), that is, the multiple correlation coefficient. He indicates that the degree of linear correlation between the dependent variable y and all the independent variables actually reflects the degree of correlation between the sample data and the predicted data\cite{6}.

\[
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y-y)^2}{\sum(y-y)^2}
\]  

(4)

Similar to the unary linear regression is 0<<1 and the closer to 1, then the regression equation will have a higher degree of fit, and vice versa. In fact, the number of independent variables will also be
affected. From (4) we can see that the SSR increases with the increase of SSR. However, this trend has nothing to do with the degree of fitting, so it is different. In the regression equation of the number, it is necessary to correct the fitting when comparing the fitting degrees. The adjustment is as follows:

\[ R^2 = 1 - \frac{\text{SSR}}{\text{SST}(n-1)} = 1 - (1-R^2) \frac{n-1}{n-k-1} \]  

The collinearity problem refers to the existence of a linear relationship or an approximately linear relationship between independent variables when fitting multiple linear regression. The method of collinearity diagnosis is based on the analysis of the matrix \(X'X\) of the observation data of the independent variables, and various indicators reflecting the correlation between the independent variables are used. Common statistics for collinearity diagnosis include variance expansion factor VIF (or tolerance TOL), conditional index and variance ratio.\(^5\) The variance expansion factor VIF refers to a relative measure of the variance of the regression coefficient due to the collinearity of the independent variables. For the ith regression coefficient, its variance expansion factor is defined as

\[ \text{VIF}_i = \frac{1}{1-R_i^2} = \frac{1}{\text{TOL}_i} \]  

Among them is the R-square of the independent regression linear regression model of the independent variable pair model. The reciprocal is also called Tolerance.

3. Analysis and Evaluation of AQI Impact Factors

3.1. Data Source Description

The research data of this paper comes from the SODA competition. The source is the Shanghai Environmental Protection Bureau. The data collected is hourly data, from July 1st, 2014 to 00:00-February 28th, 2015, 23:00, lasting 5 months. The forecast data is from March 2015 to April 2, 2015. 2014 to April 2015, and visualizes it to make it simple.

3.2. Data Preprocessing

3.2.1. Preprocessing. The research data of this paper comes from the SODA competition. The source is the Shanghai Environmental Protection Bureau. The data collected is hourly data, from July 1st, 2014 to 00:00-February 28th, 23:00, lasting 5 months. The forecast data is from March 2015 to April 2, 2015, 2014 to April 2015, and visualizes it to make it simple.

3.2.2. Visualization.

**Figure 1.** The bar chart of air quality level.
Figure 2. The histogram of Shanghai AQI distribution.

It can be seen from the above figure that the weather quality in Shanghai is good during this period, and most of the time AQI is within the range of 30-100.\cite{2}

3.3. Analysis and evaluation of AQI influencing factors

3.3.1. Correlation analysis between AQI and its influencing factors. Analyze the correlation between variables, see Table 2.

| Table 2. Correlation coefficient table. |
|----------------------------------------|
| Pearson correlation coefficient        | PM2.5 Daily average | PM10 Daily average | O3 Daily average | NO2 Daily average | SO2 Daily average | CO Daily average | AQI Daily average |
|----------------------------------------|---------------------|---------------------|-----------------|------------------|------------------|-----------------|------------------|
| PM2.5 Daily average                    | 1                   |                     |                 |                  |                  |                 |                  |
| PM10 Daily average                     | - .260              |                     |                 |                  |                  |                 |                  |
| O3 Daily average                       | -.089               | .891                |                 |                  |                  |                 |                  |
| NO2 Daily average                      | .721                | -.485               | -.302           |                  |                  |                 |                  |
| SO2 Daily average                      | .769                | -.468               | -.345           | .779             | 1                |                 |                  |
| CO Daily average                       | .907                | -.405               | -.240           | .785             | .791             |                 |                  |
| AQI Daily average                      | .990                | -.246               | -.088           | .710             | .775             | .893            |                  |

It can be seen from the above table that the correlation coefficient between AQI and PM2.5 concentration and CO concentration and AQI concentration is greater than 0.9, and the correlation
The coefficient between PM10 concentration and O3 concentration is close to 0.9, and the correlation degree is also high. It was found that AQI was significantly correlated with NO2 concentration, SO2 concentration and CO concentration. The correlation coefficient between NO2 concentration, SO2 concentration and CO concentration was close to 0.8, so the correlation between the three was also high.

3.3.2. Filter variable
By observing the above correlation coefficients, a scatter plot of partial variables is made to more intuitively observe the relationship between the variables.

From Figure 3, we can see the obvious linear relationship between PM2.5 and AQI. From Figure 4, we can see the obvious linear relationship between PM2.5 and CO.

A clear linear relationship between O3 and PM10 can also be seen from Figure 5. In order to weaken the multicollinearity between independent variables and eliminate the factors that have less influence on the dependent variables, the method of screening variables is as follows: the daily average of CO concentration is highly correlated with PM2.5 and AQI, but the concentration of CO is strong. Compared with the concentration of PM2.5, the influence of the dependent variable AQI is small, and the correlation between CO and SO2 and NO2 is also strong. Therefore, the daily average value of CO concentration is removed, and the correlation between PM10 and O3 is also strong. Similarly, O3 The daily mean value of the concentration relative to the concentration of PM10 has a relatively small effect on AQI, so the daily mean value of O3 concentration is excluded.

In summary, the remaining independent variables are PM2.5 daily average, NO2 daily average, SO2 daily average, PM10 daily average, and the dependent variable is the AQI daily average.

4. Analyze and Prediction

4.1. Prepare before Analyzing
Since linear regression models require variables to satisfy the assumption that each variable obeys the sum of normal distributions,
\[ E(X_i|\mathbf{X}) \approx \alpha_0 + \alpha X_i \tag{7} \]

Therefore, the variables are first transformed, and generally become logarithmic or cubic. After filtering, it is decided to transform each variable into

\[ \frac{1}{\sqrt{A}} \cdot P_1^{0.1} \cdot \sqrt{N} \cdot S^{0.05} \cdot \log P_2 \]

The air quality index is A, PM2.5 is P1, NO2 is N, SO2 is S, and P2 is PM10.

### 4.2. Using SPSS to Analyze Backwards Stepwise Regression

The analysis results are as follows:

| Model | R     | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|-------|-------|----------|-------------------|-----------------------------|---------------|
| 1     | .970  | .940     | .939              | .007650931                  | 1.154         |
| 2     | .969  | .940     | .939              | .007642562                  | 1.154         |

**Figure 6.** Model summary diagram.

It can be seen from the above model 1 and 2 that the Adjusted R Square is 0.939, indicating that the model has a good fit.

| Model | Sum of Squares | df | Mean Square | F | Sig. |
|-------|----------------|----|-------------|---|------|
| 1 Regression | .218 | 4  | .053 | 931.043 | .0005 |
| Residual | .014 | 238 | .000 | | |
| Total | .232 | 242 | | | |
| 2 Regression | .218 | 3  | .073 | 1243.959 | .0001 |
| Residual | .014 | 239 | .000 | | |
| Total | .232 | 242 | | | |

**Figure 7.** Discrete analysis.

**Figure 8.** Regression equation coefficients.

It can be seen from the above figure that the F value of Model 2 is large, indicating that the model is significant. Observing these regression values is statistically significant, and the multiple linear regression equations can be derived as:

\[ \frac{1}{\sqrt{A}} = 0.587 - 0.245P_1^{0.1} - 0.052S^{0.05} - 0.009 \log P_2 \tag{8} \]

It can be seen from the above Figures 8 and 9 that the residual obeys the normal distribution. Explain that the model is reasonable.

**Figure 9.** Residual normal probability map.

**Figure 10.** Residual normal probability map.
5. Conclusion and Suggestion

5.1. Conclusion
Firstly, multiple monitoring factors are used as the factors to influence the AQI. Firstly, the correlation analysis is performed on these variables, and the Pearson coefficient is calculated to determine the relationship between each independent variable and the relationship between the independent variable and the dependent variable. Multi-collinearity, according to a reasonable method to screen out the variables, using spss software to carry out stepwise regression analysis, determine multiple regression coefficients, and finally obtain multiple regression equations, using the resulting model calculations, not only improve efficiency, but also enhance AQI Interpretability and accuracy solve the problem that the original model only considers the monitoring index with the highest degree of pollution as the evaluation index of AQI, and the result is one-sided and large. The variable selection after multivariate regression model not only fully considers all The impact of pollutants on AQI also greatly reduces the complexity of the model.

5.2. Suggestion
The proposal is more focused on calculating the concentration of several monitoring indicators retained in the model and increasing the environmental governance according to the degree of influence, finding more efficient solutions, improving the level of environmental governance, and improving our air as soon as possible.

Acknowledgments
This work is supported by the Key Disciplines of Computer Science and Technology of Shanghai Polytechnic University under Grant No. XXKZD1604.

References
[1] Han Lida. Analysis of O3 Pollution Characteristics in Downtown Area of Dalian[J]. Journal of Environmental Management College of China, 2009, 19(2):76–77, 84.
[2] Research on EXCEL Implementation of Ambient Air Quality Index (AQI) [J]. Environmental Science and Management. 2013, 5:140.
[3] Tao Jia, Qi Li, Wenzhong Shi. Estimation and analysis of emissions from on-road vehicles in Mainland China for the period 2011–2015[J]. Atmospheric Environment, 2018, 191.
[4] Jong Seong Khim, Seongjin Hong, Seo Joon Yoon, Jungho Nam, Jongsooeng Ryu, Seong-Gil Kang. A comparative review and analysis of tentative ecological quality objectives (EcoQOs) for protection of marine environments in Korea and China[J]. Environmental Pollution, 2018.
[5] JIAO Dong-Fang, SUN Zhi-Hua. Regression analysis of air quality index[J]. Periodical of Ocean University of China, 2018, 48(Sup.II): 228-234.
[6] Wenrong Jiang, Yulei Yang. Provide data analysis ideas for small travel agencies[J]. 2019 2nd International Conference on Informatics, Control and Automation (ICA2019), Hangzhou, China, DEStech Publications, Inc. May 26-27, 2019. 107-111.
[7] Wenrong Jiang, Wenting Shao. Analysis and Evaluation of Development Degree of the Special Township in a Prefecture-level City[J]. 2018 2nd International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA2018), Sanya, China, ATLANTIS Press. May 28, 2018, 265-270.