World Scientific Research Journal (WSRJ) ISSN: 2472-3703
www.wsr-j.org

Use Linear Regression to Find Major Factors Influencing Air Quality

QingYu Mo*

International school of WuXi Big Bridge Academy, WuXi 214000, China

*xieqichen@126.com

Abstract: Statistics is a useful tool to extract information and help us find the solution to problems. In this paper, I would like to use linear regression technique to find out the major factor influencing the seriousness of air pollution in China. I used Air Quality Index (AQI) as an indicator of how serious the air pollution is and analyzed several factors to determine which factor was the major reason that deteriorated the air quality in China.
Keywords: Use Linear, Regression, Air Quality

1. BACKGROUND & MOTIVATION

In recent years, air pollution has become a really serious problem in China. In some particular months, the haze continues to rage throughout the entire country. Breathing the fresh air and seeing the blue sky have become extravagant hopes for Chinese people. Airborne particulate matter is widely believed to be the most harmful form of air pollution. It is the major cause of several respiratory diseases (lung cancer, etc) and shortens people's life expectancy [1]. Many Chinese people are concerned about the air pollution and are eager to know about the cause of air pollution. In the mainstream media, many experts appear on TV to explain the reasons. Most of them believe that automobile exhaust gas is to blame. Chinese government has taken measures to restrain the severeness of air pollution based on this conclusion. For instance, in some cities, cars are prohibited to be on the road on certain days based on their license plate number. Large vehicles that use unclean fuels are denied entry in some areas. People are not allowed to buy cars under certain conditions. However, if automobile exhaust gas is the main culprit, why is Shanghai, a city that owns more cars than most of other cities, not a city with the worst air quality? Why is the air quality not the worst during peak hours (7 am to 10 am, 4 pm to 6 pm)? We want to find the real reasons so that officials can take actions to better improve the life quality of Chinese people.

2. METHOD

2.1 Source of Data

To obtain the air quality index (AQI) value, we wrote java code to extract and download real-time AQI value from the website.[1] We analyzed 10 different (Beijing, Shijiazhuang, Chengdu, Shenzhen, Shanghai, Guangzhou, Tianjing, Hangzhou, Suzhou, Zhengzhou) cities all across the country and collected AQI data at an interval of two hours for 10 days.

2.2 Choice of Factors

X1: number of vehicles owned by each city. I use this as a factor because experts claimed that automobile exhaust gas was the biggest factor. To obtain the number of vehicles owned by each city, I found a ranking from the internet.

X2: magnitude of industry scale of each city, estimated by the industry income. I use this factor because factories emit waste gas, which is a great source of air pollution. Factories also need power. The process of generating electricity also produces a lot of air pollution. To obtain the magnitude of industry scale of each city, I used the annual industry income of each province to approximate the value.

X3: people's daily major activities that cause air pollution. I use this factor to indicate the influence caused by residents' daily life (i.e. cooking, heating, using A.C, etc). To obtain the coal consumption of each city, I got a table showing the coal consumption (subtracted the amount used by industry) of each city from 1990 to 2010 and I made a prediction based on the past data.

X4: climate and geography. I choose this factor because low pressure is disadvantageous to air circulation. In addition, some cities are close to desert areas, so their dust/sand concentration is higher than other regions. I estimate this value based on the geographical location of each city and the magnitude of air pressure.

2.3 Statistical Analysis

Y-value generating:

Y: the average PM2.5 AQI

X-value generating:

X1: number of vehicles owned by each city[2]

[1] The link to the website is:
http://www.pm25x.com/

[2] Data obtained from this website:
http://www.askci.com/news/dxf/20160920/09224163472.shtml

X2: magnitude of industry scale of each city, estimated by the industry income[3]

X3: people's daily major activities that cause air pollution, predicted by the coal consumption amount from 1990-2010[4]

X4: climate and geography[5]

2.4 Least Squares Regression

In order to make a good prediction using the Ordinary Least Squares Regression, there are several assumptions needed to be satisfied: (Draper, N.R. et,al 1998) [2]

1) constant variance: all y values have the same variance of error
2) independence of error: the error of y value is independent of each other
3) lack of multicollinearity: don't have two or more perfectly correlated explanatory variables

These assumptions need not to be strictly satisfied. Some of the constraints could be relaxed to some degree. Usually, poor performance of ordinary least squares regression may be caused by outliers, cross-correlations or over-fitting. Correlated explanatory variables and collinear effects are major concerns in ordinary least squares regression. (Mason, C. et. al 1991) [3] So before doing the regression analysis, it is necessary to do the correlation analysis of each explanatory variables.

After running the pairwise correlation analysis, we get the following result:

	X1	X2	X3	X4
X1	1	-0.12	-0.60	-0.35
X2		1	0.50	-0.61
X3			1	-0.18
X4				1

[3] Data obtained from these websites:
http://www.stats.gov.cn/tjsj/zxfb/201701/t20170126_1458143.html
http://www.360doc.com/content/16/0321/20/502486_544127491.shtml

[4] Data obtained from this website:
http://wenku.baidu.com/link?url=ancVowr-oZA0bcRW7TBZMZuP1QqnELDbIXmnmgJ_hbRn99KI25jcbqsmmI9ZxfuXQkajLhz1YrZk9Jlbx9lassBR9UIHtmGwKiL88JT4pF_

[5] Some data obtained from this website
http://wenku.baidu.com/link?url=8ErjrNqfVrjOPnPzkek1Tdmqj8_Gqgm-ixtKSI9oWBFt6rEFtzfGBmSyQnRrjTB1wieJhw0tQ1Kc0S1fss1FzBZNFP8ZKeavpBznx3Y8iOy

Some correlation are relatively high (larger than 0.60). Consequently, the assumption of multicollinearity may not hold and the high-correlation may influence the prediction. We should keep that in mind when doing the analysis.

We use multivariable linear regression to build a model to predict the AQI value. We normalize all the numerical data. We select the K-Best features (K=1,2, 3, 4) based on the F-values between label/feature for regression tasks. The feature that are significantly different from zero and has great influence in y value is given higher priority and will be selected first.

K	Attributes	Coefficients	R^2
1	{x2}	{0.602}	0.84
2	{x2,x4}	{0.50,-0.61}	0.863
3	{x2,x3,x4}	{0.61,-0.60,-0.46}	0.919
4	{x1,x2,x3,x4}	{-0.15,0.59,-0.72,-0.65}	0.923

3. CONCLUSION

As we can see from the table, when there are 4 attributes, the R^2 value is the largest, which means 92% of the variation could be explained by the linear regression and that is the best linear regression we can do. The linear regression formula is y=-0.15X1+0.59X2-0.72X3-0.65X4+837. All coefficients are significantly different from 0. However, the number of automobiles (X1) is not the biggest contributor. On the contrary, the magnitude of industry scale is the most influential one (X2). When we modify the K (number of attributes), X2 is always selected. According to this result, we can say that industry is the most significant contributor to the air pollution. So, if we really want to improve the air condition, we need officials to allocate enough resources so that factories are constantly under regulation and inspection. If factories produce unnecessary wastes because they are unwilling to upgrade their recycling and cleaning system, they should be shut down. Government should also appropriate enough research funding to support institutes and researchers to develop better technologies. Concentrating on automobile regulation may work, but industry regulation is more urgent when dealing with serious air pollution.

4. FUTURE WORK

Due to the time and resource limit, this paper can still be further improved. The data size is not large enough. In the future, I want to increase the sample size so that the data points can be representative enough. When the sample size is large enough, we can use 10-fold cross-validation for better accuracy analysis. Many other factors that may cause the air

pollution haven’t been taken into consideration when doing the regression analysis. Also, I only analyzed the AQI value for ten days. I want to include values from other months.

REFERENCES

[1] Hamra, Ghassan B., Neela Guha, Aaron Cohen, Francine Laden, Ole Raaschou-Nielsen, Jonathan M. Samet, Paolo Vineis, Francesco Forastiere, Paulo Saldiva, Takashi Yorifuji, and Dana Loomis. "Outdoor Particulate Matter Exposure and Lung Cancer: A Systematic Review and Meta-Analysis." Environmental Health Perspectives (2014): n. pag. Web.

[2] Draper, N.R.; Smith, H. (1998). “Applied Regression Analysis (3rd Ed.)”. John Wiley. ISBN 0-471-17082-8.

[3] Mason, Charlotte H., and William D. Perreault. “Collinearity, Power, and Interpretation of Multiple Regression Analysis”. Journal of Marketing Research 28.3 (1991): 268–280. Web