Under the Dome -- Beijing PM 2.5 Air Pollution Analysis Using Time Series Analysis and Linear Regression

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
DEPARTMENT OF STATISTICS

Xinge Jia: xj2221
Nikita Tourani: nrt2117
Xinyi Hu: xh2383
Project Mentor: Professor Banu Baydil

Department of Statistics, Columbia University, New York, NY 10027, USA

Background

China has faced a severe and unsolved air pollution problem for years, with PM2.5 being the main pollutant. A 2012 study shows an estimated 8,572 premature deaths occurred in four major Chinese cities due to high PM2.5 levels. The report also said severe air pollution in Shanghai, Guangzhou, Xi'an and Beijing has led to a total economic loss of 6.8 billion yuan ($1.09 billion).

Aims

Our dataset covers PM2.5 readings and other weather data from 2010-2014. We will examine the data and quantify the severity of the problem with a statistical approach.

In the first section, some descriptive statistics will be provided.

Followed by the time series analysis section, we focus on how PM2.5 values vary under different frequencies. And then by fitting suitable models and doing model diagnostics, we try to find the pattern of PM 2.5.

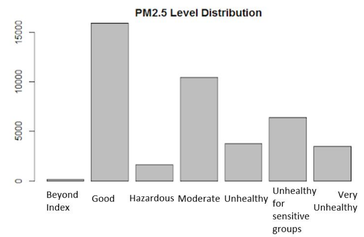
Finally, the goal of the linear regression analysis section is to select the model consisting of variables that influence PM2.5 in Beijing significantly, which can help us to identify the relevant factors and take actions to prevent the high levels of PM2.5 in the future.

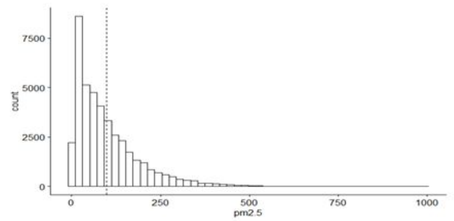
Beijing CBD

Summary Statistics

According to the World Health Organization's findings, the highest PM2.5 level for acceptable air quality is 25 micrograms per cubic meter.

Our data set shows that the median PM2.5 level in Beijing registers approximately 3 times the recommended level. (72 mg/m^3)





Seasonal distribution: winter/autumn months tend to have a much higher average PM2.5 level than the summer/spring months

24-hour distribution: night levels tend to be higher than day levels

Time Series Analysis

In this section, only the PM 2.5 values were analyzed. The PM 2.5 data were processed into 2 types: daily PM 2.5 data and monthly PM 2.5 data, using the averaged values. Then logarithm was used on both data sets due to the existence of changing variability.

Model selection:

An MA(2) model for daily PM 2.5 data was chosen as the ACF plot of the data showed that the sample covariance cut off at lag 2 and converged after lag 2. As for the monthly PM 2.5 data, first an MA(6) model and an AR(6) model were tested, and then MA(6) was chosen because it has smaller AIC value.

Diagnostics:

Hypothesis tests in terms of randomness and normality were tested on the fitted residuals of two models.

Fitting daily PM 2.5 data by MA(2) model, we obtained non-Gaussian but independently distributed noise. Fitting monthly PM 2.5 data by MA(6) model, the fitted residuals are normally and independently distributed.

Conclusion:

We want the fitted residuals to behave like Gaussian white noise, thusly we consider that the fitted model MA(6) using monthly PM 2.5 data might be a good fit.

Checking the AICs and RMSEs of the two models, we found that MA(2) model for daily PM 2.5 data has much larger AIC and RMSE. Hence, to build a better time series model for daily PM 2.5 data, one might need to consider more complicated time series models.

```
##                      AIC  RMSE
## Daily PM 2.5    4034.693 0.746
## Monthly PM 2.5    12.830 0.221
```

Linear Regression Analysis

In this part, we select an appropriate model by making transformations of predictor variables and response variable, including interactions of predictor variables and using criteria for model selection.

Model selection:

Classifying the type of predictor variables as qualitative and quantitative predictors.

Making log transformation of response variable(PM2.5) in order to improve the fitness of our model.

Using some statistical criteria like ridge regression and Lasso to select variables. The result shows that we should keep the seven variables.

Making transformations of predictor variables to find whether the response variable has a statistical interaction between the polynomial of quantitative variables and the interaction between qualitative variables and quantitative variables respectively as well as the interaction between quantitative variables themselves. After comparing each model's AIC and R-adjusted square, we decide not to add any interaction terms in the model.

Conclusion:

The final model consists of the seven original predictor variables and the log transform of response variable. The quantities of the final model can be concluded as:

AIC	R squared	MSPR
99005	0.413	0.622

Discussion

Using data exploring, we found a seasonal and daily distribution. The data showed a strong positive skew and 12.5% of all levels registered within the "Very Unhealthy / Hazardous / Beyond Index" categories.

In the time series analysis and linear regression, we found that using parametric methods can't give us a good fitted model. This suggests that the Beijing PM2.5 nature may be captured with a nonparametric model.

In the linear regression, the model hasn't been improved much by making any transformations of predictor variables or adding any interaction terms. This suggests that we may collect other variables like the number of cars to see the deep link between PM2.5 and our life.