Examining Beijing Air Pollution with Meteorological Variables

STAT 425: Applied Regression and Design
Summer 2021

Group: Benny Z
Benny Zhao, bzhao22

Instructor: Professor Lelys Bravo De Guenni

1 Introduction

This project aims to predict the extent of air pollution in Beijing using PM2.5 concentration as a metric for air pollution. PM2.5 concentration refers to the concentration of fine particular matter, specifically the mass in micrograms of airborne particles with aerodynamic diameters of less than 2.5 micrometers observed in a volume of a meter cubed.

In order to predict PM2.5 concentration, both meteorological variables, such as temperature and air pressure, and temporal variables, such as month and year, will be used as predictors. While the data set was taken from the University of California, Irvine, with the following source:

- https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

it was an aggregate of the meteorological variables were measured at the Beijing Capital International Airport (BCIA), from weather.nocrew.org, and the PM2.5 concentration measures recorded at the U.S. Embassy in Beijing; both data series ran from January 1, 2010 up to December 31, 2014. It should be noted that although the airport and embassy are 17 kilometers apart, the two locations have the same weather, according to a 2015 academic paper that used the data set in a journal published by the Royal Society, titled "Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating", which can be found here:

- https://doi.org/10.1098/rspa.2015.0257.

Additionally, given how weather conditions will have high variability and can confound the variables used, it was critical that the time span the data was collected in was of sufficient length and that the frequency of the observations was high enough for air pollution to be modeled accurately.

2 Exploratory Data Analysis

The data set has 43,824 hourly observations, and the main response variable that is to be modeled is:

- `PM2.5`: The concentration of fine particulate matter composed of particles with aerodynamic diameters of less than 2.5 μm ($\mu g/m^3$)

The data set that will be used has the following temporal variables:

- `Year`: Year of the observation
- `Month`: Month of the observation
- `Day`: Day of the observation
- `Hour`: Hour of the observation

The meteorological variables in the data set are the following:

- `DEWP`: Dew point needed to achieve a relative humidity of 100% ($^\circ C$)
- `TEMP`: Temperature ($^\circ C$)
- `PRES`: Pressure (hPa)
- `CBWD`: Combined wind direction
- `LWS`: Accumulated wind speed (m/s)
- `LS`: Accumulated hours of snow
- `LR`: Accumulated hours of rain

However, working at a daily level rather than an hourly level is more ideal since air pollution can be predicted using the day in the year as a predictor rather than the hour.

As a result, the daily observation data set has 1,826 observations, and the non-temporal variables have been aggregated by date into the following variables:

- `Avg_Day_PM2.5`: Average PM2.5 concentration for that day ($\mu g/m^3$)
- `Avg_Day_DEWP`: Average dew point needed to achieve a relative humidity of 100% for that day ($^\circ C$)
- `Avg_Day_TEMP`: Average temperature for that day ($^\circ C$)
- `Avg_Day_PRES`: Average pressure for that day (kPa)
- `Max_Day_CBWD`: Maximum combined wind direction for that day
- `Max_Day_LWS`: Maximum accumulated wind speed for that day (m/s)
- `Max_Day_LS`: Maximum accumulated hours of snow for that day
- `Max_Day_LR`: Maximum accumulated hours of rain for that day

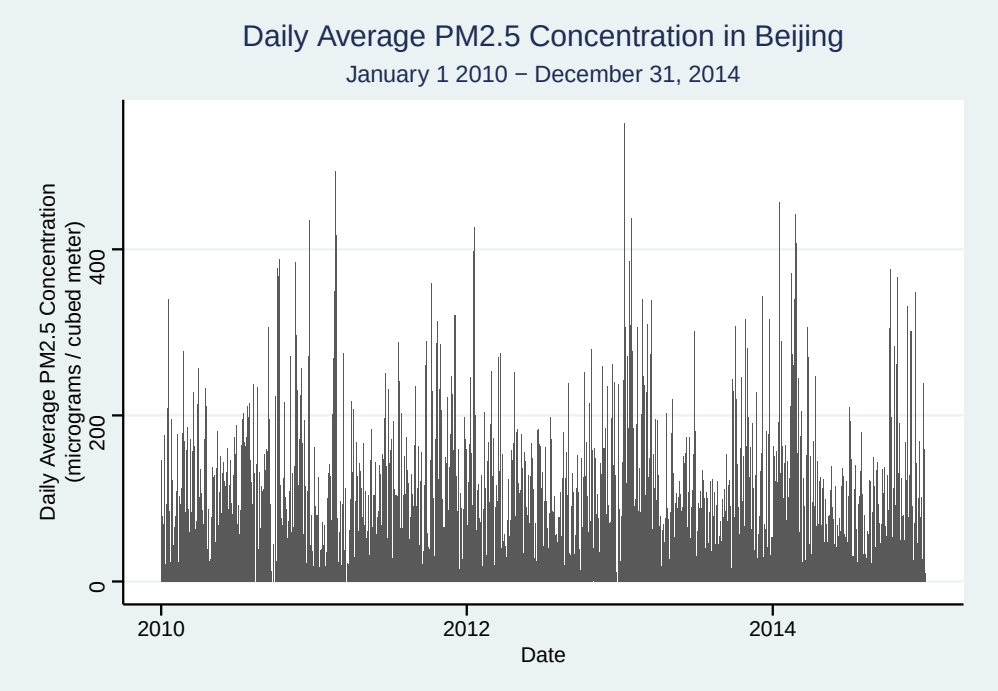

Figure 1: Daily Average PM2.5 Concentration in Beijing

Looking at the histogram of the daily average PM2.5 concentration in Figure 1, it appears that the daily average concentration peaks at the beginning of each year, before decreasing during the middle of the year.

In fact, when sorting the data set that has a daily temporacy frequency in descending average PM2.5 concentration order, the days that have the top 5 average PM2.5 concentrations are:

Day	Average Daily PM2.5 Concentration (micrograms / meters-cubed)
2013-01-12	552.4783
2011-02-21	493.9167
2014-01-16	457.5000
2014-02-25	442.6667
2013-01-29	438.1304

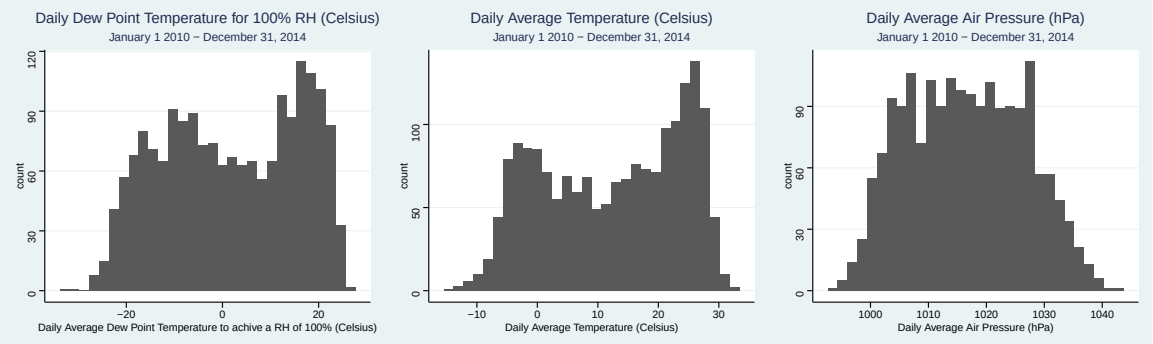

Figure 2: Histograms of Daily Average Meteorological Variables

In Figure 2, the distribution of both the daily average dew point temperatures and the daily average temperatures appear to be bimodal with left skew. As for the daily average air pressure, its disribution appears to be symmetric about its median, and the distribution appears to be unimodal.

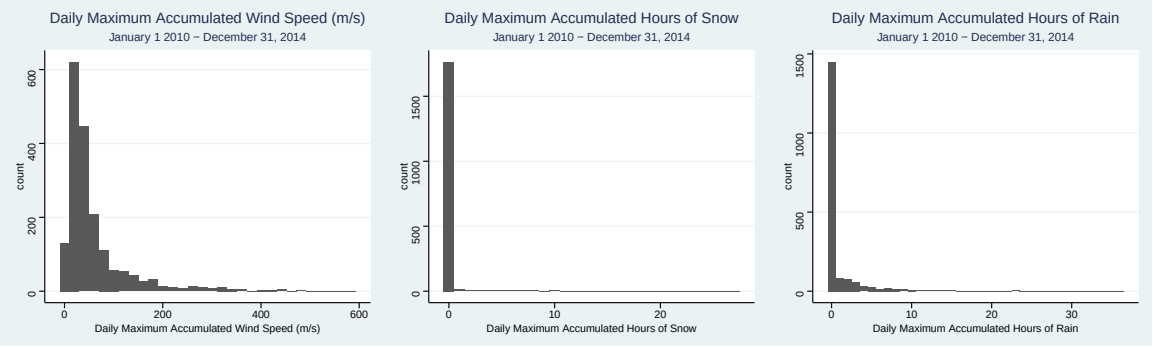

Figure 3: Histograms of Daily Maximum Meteorological Variables

Lastly in Figure 3, since the distributions for the daily accumulated wind speed, accumulated hours of snow, and accumulated hours of rain appear to have heavy right skew, using a log transformation or inverse transformation on these variables might prove to be fruitful when fitting model later on.

3 Methods

Before making the preliminary model, we will separate the data set which had a daily temporal frequency into training and testing sets. We will randomly select $\frac{1}{10}$ of the data set without replacement to be the testing set, and the remaining $\frac{9}{10}$ of the data set will be the training set.

3.1 Model Fitting

The first preliminary model I made was a multiple linear regression model that included all of the meteorological predictors along with the date that each observation was recorded.

Based on the Residuals vs Fitted plot in the above figure, it is apparent that the assumption of the residuals have constant variance has been violated as the spread of the residuals appears to increase as the magnitude of the fitted values increase. This is also apparent in the Scale-Location plot where the spread of the residuals begins to increase dramatically when fittd values are greater than 50.

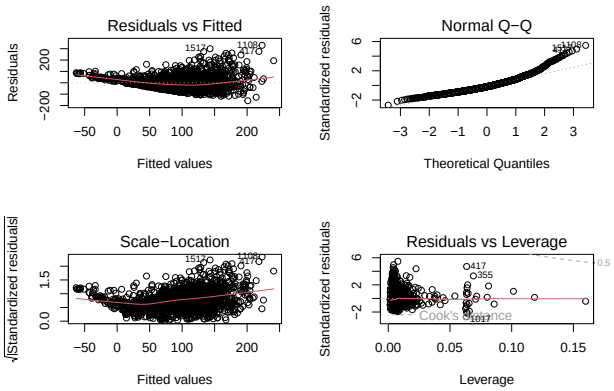

Figure 4: MLR model diagnostic plots

Based on the Normal-QQ plot, there also appears to be violations of the assumption of the residuals following a normal distribution as there are deviations from the 1:1 line when the theoretical quantiles are greater than 1.5.

Based on the Residuals vs Leverage plot, there does not appear to be any highly influential points as there were not any observations that were in the threshold where the Cook's distance was at least 1.

Next, I used the following formal tests to test for the respective model assumptions:

- Breusch-Pagan test for homoscedasticity
- Shapiro-Wilks test for the residuals following a normal distribution
- Durbin-Watson test for non-correlated errors

Formal test	p-value
Breusch-Pagan	0
Shapiro-Wilks	0
Durbin-Watson	0

As the p-value of the three tests are all less than 0.05, all three assumptions of the residuals having constant variance, the residuals following a normal distribution, and non-correlated errors, are violated.

Next, I checked the multiple linear regression model for unusual observations.

When checking for high-leverage points, there were a total of 151 high leverage points, meaning there were 151 observations that were far away from the center of the data.

As for outliers, there were 6 outliers according to outlier test run with a Bonferroni correction which had a critical value of $\frac{0.05}{2 * n}$, with $n = 1643$.

Using the Cook's distance of the observations to check for highly influential points, there were not any that had a Cook's distance of at least 1. Hence, there were not any highly influential points.

Due to the presence of 6 outliers according to outlier test, I made the decision to make a copy of the training set that omitted those observations.

As the multiple linear regression model had violations in the assumptions for homoscedasticity in the residuals, residuals following a normal distribution, and errors not being correlated, I tried to use a Boxcox transformation to try to deal with the violation with the normality assumption.

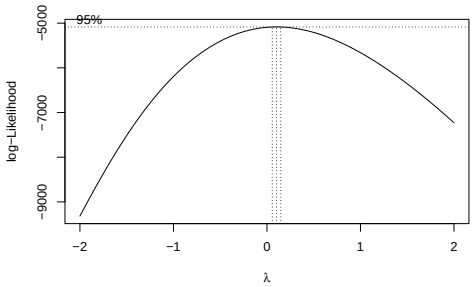

Figure 5: MLR Model Boxcox Plot

Since Figure 4 displays the log-likeihood being at its maximum near 0, I chose to only transform the reponse variable, the daily average PM2.5 concentration, using a log transformation

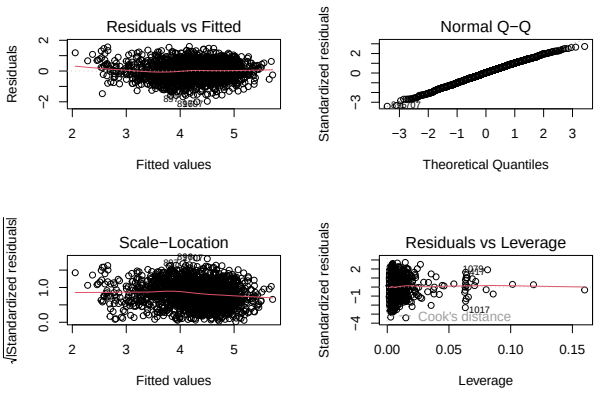

Figure 6: Log Model Diagnostic Plots

Looking at Figure 6's Residuals vs Fitted plot, it appears does not appear to be any issues with non-linearity in the recently made log model.

There also does not appear to be any violations of the normality assumption either, as there does not appear to be any deviation from the 1:1 line in the Normal Q-Q plot.

There apperas to be slight issues variance of the residuals in the Scale-Location plot as the spread appears to increase when fitted values approach 4, but start to decrease when increasing greater than 4.

As for the Residuals vs Leverage plot, there does not appear to be any observations that are highly influential as there are not any points in regions where the Cook's distance is at least 1.

When running the three formal tests that were I also ran on the first multiple regression model, the three tests produced the following p-values:

Formal test	p-value
Breusch-Pagan	0.0000000
Shapiro-Wilks	0.0213391

Formal test	p-value
Durbin-Watson	0.0000000

While there appears to be a slight improvement in the normality assumption, as the p-value of the Shapiro-Wilks test on the multiple linear regression model was substantially closer to zero, there does not appear to be any improvement in the constant variance assumption, nor the non-correlated errors assumption for the log model.

As I previously observed how the variables for the maximum daily accumulated wind speed, maximum daily hours of accumulated snow, and maximum daily hours of accumulated rain had heavy right skew, I tried using the log transformations on those variables. Since the maximum daily accumulated hours of snow and the maximum daily accumulated hours of rain had minimums of zero, I added one to both variables in the observations in the data set.

Now, for Figure 7, the same things said for the first log model can be said for the second model when viewing the second model's diagnostic plots, where there are not any indications of non-linearity based on the Residuals vs Fitted plot, there are not any deviations from the 1:1 line in the Normal Q-Q plot, there are some indications of the variance of the residuals changing when fitted values increase, and there are not any unusual observations that would be considered highly influence in the Residuals vs Leverage plot.

Looking at the p-values produced from the formal tests, they are as follows:

Formal test	p-value
Breusch-Pagan	0.0000000
Shapiro-Wilks	0.0245004
Durbin-Watson	0.0000000

While there appeared to be minimal improvements in the normality assumption, the assumptions of the residuals having constant variance and non-correlated errors do not appear to have any improvements.

I then tried including the interaction between the maximum combined wind direction and the maximum accumulated wind speed, as both variables are related to wind.

In Figure 8, there does not appear to be any issues with the model assumptions when viewing the interaction model's diagnostic plots

Formal test	p-value
Breusch-Pagan	0.0000000
Shapiro-Wilks	0.1466757
Durbin-Watson	0.0000000

While there did not appear to be any improvements based off the Bruesh-Pagan and Durbin-Watson test still returning p-values that are very close to zero, the Shapiro-Wilks test did not reject the null hypothesis of normally distributed residuals.

As I did not consider there to be other models to attempt, I began to using backwards variable selection on the interaction model, using the BIC as the selection criterion.

From backwards variable selection, I was able to generate the following model:

$$\begin{gathered}\log(y_{\text{Avg. Daily PM2.5}}) = \beta_0 + \beta_1 x_{\text{Avg. Dew Point}} + \beta_2 x_{\text{Avg. Day Temp}} + \beta_3 x_{\text{Avg. Day Pressure}} + \\ + \beta_4 d_{\text{Max Day CBWD, NE}} + \beta_5 d_{\text{Max Day CBWD, NW}} + \beta_6 d_{\text{Max Day CBWD, SE}} + \\ + \beta_7 \log(x_{\text{Max Day Wind Spd}}) + \beta_8 \log(x_{\text{Max Day Rain}}) + \beta_9 \log(x_{\text{Max Day Snow}}) + \\ + \beta_{10} x_{\text{Max Day Wind Spd}} + \beta_{11} \left(x_{\text{Max Day Wind Spd}}, d_{\text{Max Day CBWD, NE}}\right) + \\ + \beta_{12} \left(x_{\text{Max Day Wind Spd}}, d_{\text{Max Day CBWD, NW}}\right) + \beta_{13} \left(x_{\text{Max Day Wind Spd}}, d_{\text{Max Day CBWD, SE}}\right)\end{gathered}$$

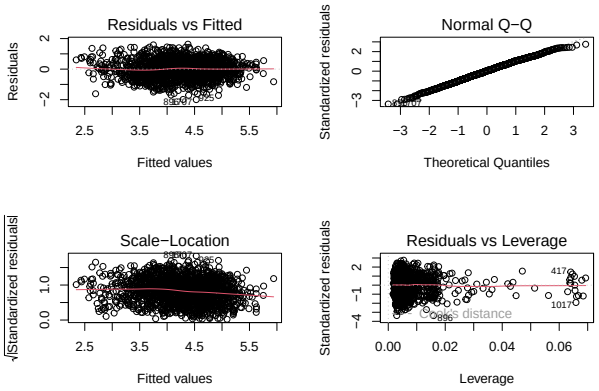

Figure 7: Second Log Model Diagnostic Plots

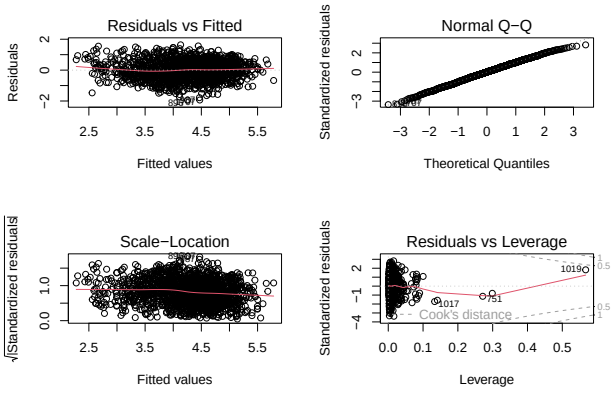

Figure 8: Second Log Model with Interaction Diagnostic Plots

The coefficients for the predictors are as follows:

β_i	Value
0	36.2705421
1	0.0642724
2	-0.0998289
3	-0.0294314
4	-0.3349571
5	-0.2463578
6	-0.1409296
7	-0.2074820
8	-0.2076164
9	-0.1603069
10	-0.0052686
11	0.0042511
12	0.0044572
13	0.0108519

Next, I checked the model diagnostics for the BIC model.

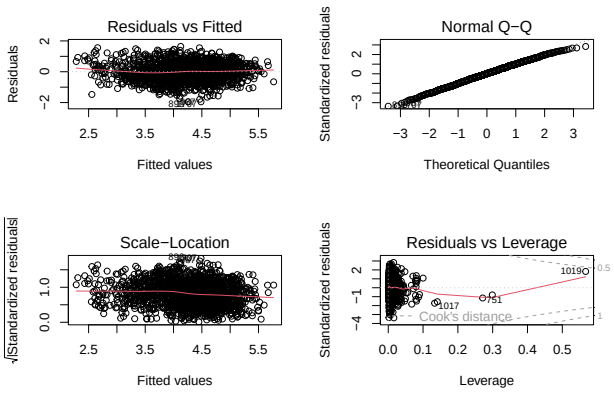

Figure 9: BIC Model Diagnostic Plots

Look at Figure 9's diagnostic plots, the Residuals vs Fitted plot does not show any indication of non-linearity in the model. As for the Normal Q-Q plot, there does not appear to any deviation from the 1:1 line, which would mean the normality assumption would not appear to be violated. As for the Scale-Location plot, the same issue of the variance of the residuals increasing before decreasing is present. As for the Residuals vs Leverage plot, there does not appear to be any highly influential points as all of the points appear to be within the interval where the Cook's distance is less than 1.

Formal test	p-value
Breusch-Pagan	0.0000000
Shapiro-Wilks	0.1219447
Durbin-Watson	0.0000000

Again, while the normality assumption does not appear to be violated since the Shapiro-Wilks test did not reject the null hypothesis of normality, the Breush-Pagan and Durbin-Watson test still reject their null hypotheses, meaning the BIC model violates the assumption of constant variance among the residuals and non-correlated errors.

3.2 Prediction

Next, using the most optimal model from the previous part, which was the interaction log model which had backwards BIC variable selection, I will predict the average daily PM2.5 concentration using the data from the testing set as predictors.

In order to gauge how accurate my most optimal model is, I will use Root Mean Square Error as the metric to report on the model's performance, while removing any observations that do not have a recorded average daily PM2.5 concentration when computing the mean.

The RMSE value was 18.5192015, which would mean that on average, the prediction made by my optimal model would be 18.5192015 micrograms / cubic meter away from the actual observed daily average PM2.5 concentration.

4 Discussion of Results and Conclusions

Using the optimal model I produced at the end, which was the variable selected log model with interaction terms, using BIC as the selection criteron, my predictions of the average daily PM2.5 concentration of the testing data on average were off by 18.5192015 micrograms / cubic meter away. Additionally, the BIC model's Adjusted R^2 value of 0.5084191 means that the 50.8419094 percent of the variance in the logged daily average PM2.5 concentration was explained by the BIC model, when accounting for the addition of additional predictors to the model. While this appears to be relatively accurate, considering how the standard deviation of the average daily PM2.5 reading of 80.2306045 micrograms per cubic meter, this model failed to have the following assumptions satisfied:

- Constant variance among the residuals
- Non-correlated errors

As the observations are from a time series data set, the autocorrelation of the errors was to be expected, and an approach that may help in resolving this would be to use Generalized Linear Regression to make a model.

When it comes to applying the model in the real world, we can look at the magnitudes of the coefficients in the BIC model; it appears that a combined daily maximum wind direction in the Northeast direction would contribute the most to observing a smaller average daily PM2.5 concentration. Additionally, as for the predictor that would contribute the most to observing a bigger average daily PM2.5 concentration, that would be the average daily dew point temperature needed to reach a relative humidity of 100%.

Closing thoughts:

- The final BIC model would not be ideal for interpretability, given how the response variable, the average daily PM2.5 concentration, has a logarithmic scale, and how there are multiple interaction terms as predictors, along with the predictors like the maximum daily accumulated hours of rain, being on a logarithmic scale as well.
- As the data was collected from an observational study rather than an experiment, the model loses much of its utility of the goal was to find ways to reduce the amount of PM2.5 concentration in the air. This is further compounded with the fact that many of the predictor variables are not things humans can control directly, like wind speed and air pressure.
- As this model fails to have the assumptions of the residuals having constant variance and the errors not being correlated, the utility from the model will also begin to diminish when attempting to extrapolate the amount of PM2.5 concentration using extreme values for predictors.

5 Acknowledgements

I thank Burak Ogan Mancarci (University of British Columbia) for the assistance in the formatting of the report and the title page. I adapted the formatting used in the RMarkdown file and the `title.sty` file of their thesis proposal, which can be found at:

- https://github.com/oganm/ThesisProposal