Regression Analysis on PM2.5 Levels for Beijing, China

Team 8: Chris Givens, Dinh Nguyen, Sean Pili, Rongxuan Wang

December 15, 2017

Executive Summary

We did our research project on $PM_{2.5}$ in Beijing. The reason that $PM_{2.5}$'s are dangerous is because that since they are so small, they can easily assimilate into the human body. We were interested in this data set because as a group we wanted to do something in the environmental field. The data set we found [1] had an abundance of observations, and was set in Beijing, which is historically known around the world for having poor air quality. The main research question we had was what factors increase $PM_2.5$ levels in Beijing. The model we will create won't be used to necessarily decrease the $PM_{2.5}$ in Beijing, but to help the local government know when to prepare and expect for $PM_{2.5}$ to be high. This is useful because the local government could be prepared in advance and issue warnings out to the local community so they can take the necessary precautions.

We started off the project by looking at all the variables at hand, and made a hypothesis on which variables will be useful. These variables are *SEASON* (Season of data), *TEMP* (Temperature Celsius), *PRES* (Pressure hPa), *DEWP* (Dew Point Celsius), and *cbwd* (Combined Wind Direction). We then added the Day of the Week and Work Time (work hours, weekend, nonwork hours). We believed that there should also be a time effect involved in our model as well, because the plot of $PM_{2.5}$ indicated to us that we should run a time series analysis. We fitted many models, but each one we fit had normality of error issues. Even after the transformations we did, we were still having issues. A few reasons for that would be that there were so many outliers, and we did not have a reason to get rid of them. When we looked at the Q-Q plots and histograms, the middle area of the plots looked great, but the tails were off-skewed. We ran a BIC to find the best model compared to the other ones we made. We came up with our best models with square root transformations and log transformations, and then compared them by looking at the mean squared prediction error. In the end, the square root model had the lowest mean prediction squared error, and we chose that as our best model. Below is the formula for our best model:

$$\sqrt{PM2.5} = \sqrt{PM2.5_{n-1}} + sin(\frac{2*\pi*t}{24}) + cos(\frac{2*\pi*t}{24}) + \sqrt{PRESSURE}$$
$$+ \sqrt{TEMP} + \sqrt{DEWP} + \sqrt{TEMP} * WTime + NE + NW$$

In conclusion, even though the normality assumptions weren't met, our model can still be used to make useful predictions. We feel that our normality assumption was not met because some of the most important factors that explain $PM_{2.5}$'s variance such as population, number of businesses in the area, atmospheric Nitric Oxide (NO), and atmospheric CO_2 were not present in our model. We believe that if we could add that information to our regression, we would be able to satisfy our normality assumption.

Problem Context

The main focus of our research was on $PM_{2.5}$. $PM_{2.5}$ refers to atmospheric particulate matter (PM) that has a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair. These particulates are extremely hazardous because when they are small, they linger in the air longer. This increases the chances for the particulates to be inhaled by humans and animals. Particles smaller than 2.5 micrometers are able to bypass the nose and throat and penetrate deep into the lungs and some may even enter the circulatory system. As a result, particulates are capable of causing DNA mutations, heart attacks, and premature death. They are also carcinogenic and can attribute to lung cancer. Being able to predict the $PM_{2.5}$ levels and take the necessary precautionary measures to reduce exposure to these particulates will improve the population's health. Our goal is to create a sufficient model for the city of Beijing that can provide insight to when the $PM_{2.5}$ levels are high throughout the year.

The data set was obtained from https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+data. The data set had 43824 observations, and 13 variables. [1] Some of the issues we had with the data was that there were missing values, which was an issue when we performed time series analysis. In order to resolve this issue, we used linear interpolation to fill in the missing data gaps. We then included character variables that helped distinguish weekdays, weekends, and whether it was during a time of high travel. There was also potential issues with performing log transforms on the predictors, temperature and dew point. This dilemma was fixed by creating a new data column which contained the converted values of Celsius to Kelvin for temperature, and translating all the data values in dew point so that the lowest value is one. Another prevalent problem we encountered was the fact that we could not perform an analysis on our data set from the year 2010 to 2014 due to the sheer size of the number of data entries. Because of this, we had to subset our data and perform our analysis on a single year instead.

Analysis

The preliminary approach to performing an analysis on this data set was to first create a data column that calculated what day of the week the data entries were. From there we took into consideration of the active hours throughout a weekday and denoted the hours between 6 AM to 10 PM as "Work Time", 11 PM to 5 am as "Non Work Time", and the days Saturday and Sunday as "Weekend". After creating all of the variables for our analysis, we initially tried to use stepwise regression to find the best model for our data, but we realized that the data was too large, and it could not run the stepwise regression. This encouraged us to just use 2013 data to create our models. Furthermore, after selecting our best model for the 2013 data, we could use it to predict the 2014 $PM_{2.5}$. After consulting with our professor, we realized that we did not need to start with stepwise regression; we could use our intuition to select a few candidate models and build from there. After plotting our response and its auto-correlation function, we believed that we needed to add a time series component to our model because (as shown in Figure 2) the ACF decayed over many lags, meaning it stayed significant, and the plot of our response (shown in Figure 1) appeared to show periodicity. Due to this, we made sure to add an AR(1) term to all of our regressions and created indicator variables for month and season. In addition, we created daily and weekly sine and cosine variables in case the previously listed indicator variables fail to fully capture the periodicity exhibited by the data.

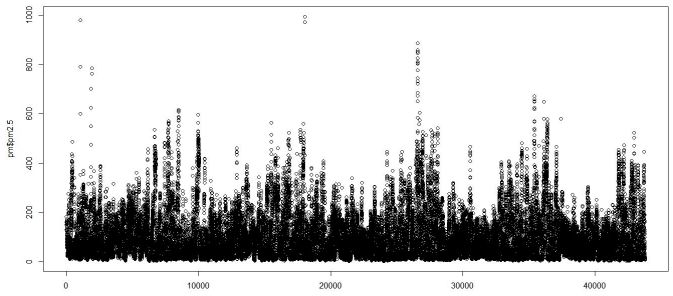

Figure 1: PM2.5 plot

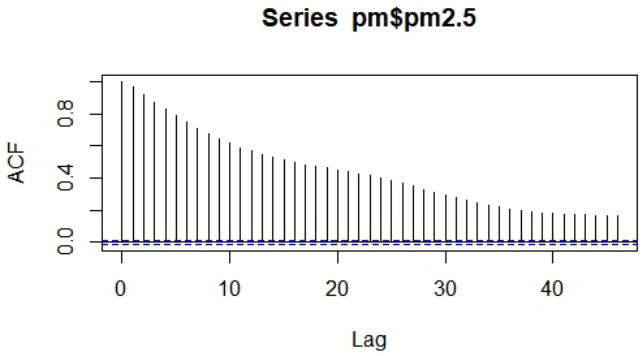

Figure 2: PM2.5 ACF

Our first models included all of our original features plus an AR(1) component and various periodicity capturing variables, but we realized it we might need to tweak it for these reasons; our normality of errors assumption was definitively violated (as shown in Figure 3), our constant variance assumption was most likely violated (as shown in figure 4), there was too many influential outliers (as shown in figure 5) and there appeared to be strong linear relationship between three of our predictors: dew point, pressure and temperature (as shown in Figure 6.)

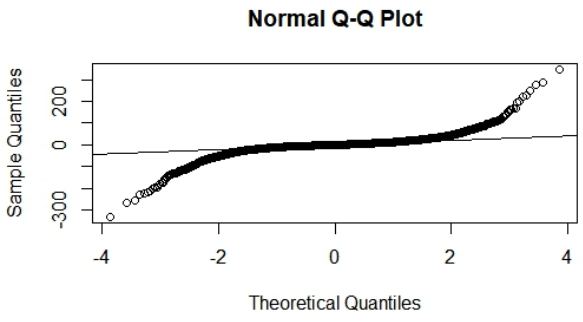

Figure 3: Q-Q Plot

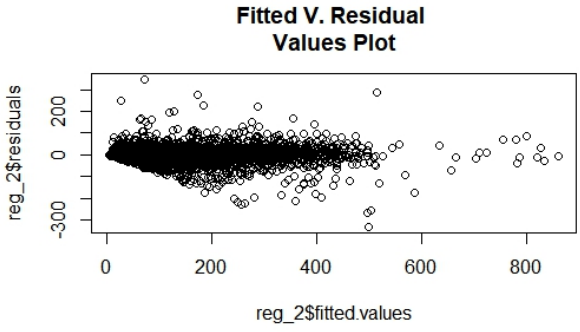

Figure 4: Fitted Values

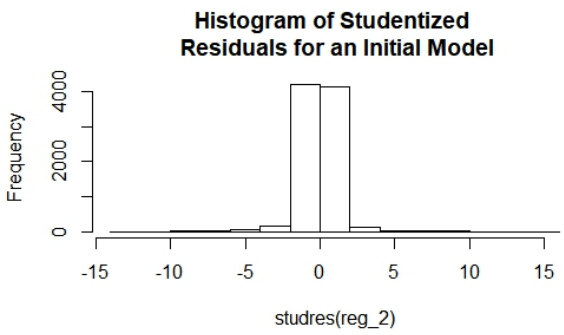

Figure 5: Studentized Residuals

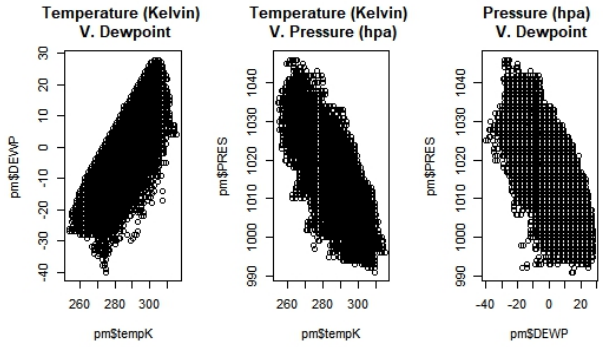

Figure 6: Interactions

To deal with the colinearity in our data, we tried fitting models that did not include all three of our correlated predictors, as well as models that contained their interaction terms. We transformed our response and most of our numerical predictors using the natural log and square root functions in an attempt to satisfy our normality of errors and constant variance assumptions (in addition to hopefully making our outliers less extreme.) In addition, we tried plotting each of our continuous predictors against the residuals from each of our models to see if our model needed an extra polynomial term, such as temperature squared, but none of our continuous predictors appeared to show a polynomial, or other, pattern when plotted against any of our models' residuals. Unfortunately, we were never able to fully satisfy our constant variance or normality of errors assumptions with any of our models so we ultimately decided to use BIC and Mean Squared Prediction Error as the criterion to find our best model. When creating these models, we altered our predictors such as: transforming our continuous predictors with the log or square root functions, subtracting one, two, or all three of our correlated predictors, adding one or more interaction terms between our correlated predictors and removing non-significant levels of indicator variables (or the variable altogether), and choosing different periodicity variables. Interestingly, we found that February was the only significant level of our month feature in both our square root and natural log response models. One of our group members is from Beijing and he mentioned that February is the coldest month there. He also mentioned that most buildings in Beijing used coal-powered central heating, which generates pollutants into the air. However, according to a boxplot of monthly residuals, we may be drawing a premature conclusion.

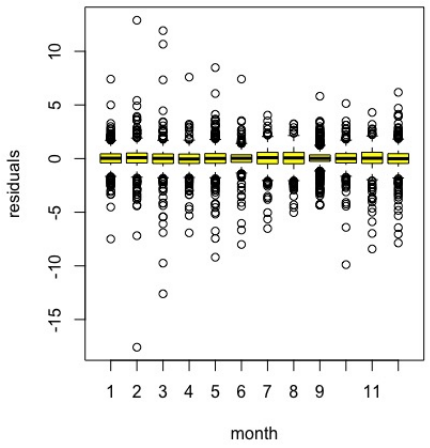

Figure 7: Month Boxplots

There does not appear to be a difference in the means of the residuals for any of the months, so we believed that February being considered a significant predictor was a Type I error. We used BIC to make sure of this and we were correct. Both our log and square root response models with the lowest BIC's did not include a February or month predictor. However, BIC did and this helped assure us that only the North East (NE) and North West (NW) levels of the Combined Wind Direction variable were useful, as they were the only significant levels of that variable in all of the models we fit.

After creating a slew of potential candidate models, we used BIC to determine the final models for each response transformation as we cannot compare the BIC of a logged response model to the BIC of a square rooted response model.

Here is the formula for our logged response model with the lowest BIC (-23453.79, 65.81) percent probability of being correct given other models:

$$\log PM2.5 \;=\; \log PM2.5_{n-1} \;+\; sin(\frac{2*\pi*t}{24}) \;+\; cos(\frac{2*\pi*t}{24}) \;+\; \log TEMP$$
$$+\; \log DEWP \;+\; \log TEMP * \log DEWP \;+\; NE \;+\; NW$$

Below are the assumption plots for our final log response model. The Normality of errors assumption is still violated, likely because there are too many outliers (shown in Figure 11). The residuals appear to satisfy the normality assumption for the first two standard deviations, but are too large/small past that point. If we had less outliers, our data might be normal. If we used moving averages (over days instead of hours) our data might be more normal, but we did not have the time to consider this. The constant variance of errors and identically distributed errors assumptions both appear to be violated as shown in the fitted versus residual plot below.

Here is the formula for our square rooted response model with the lowest BIC(1064.461, 91.57) percent probability of being correct given other models:

$$\sqrt{PM2.5} \;=\; \sqrt{PM2.5_{n-1}} \;+\; sin(\frac{2*\pi*t}{24}) \;+\; cos(\frac{2*\pi*t}{24}) \;+\; \sqrt{PRESSURE}$$
$$+\; \sqrt{TEMP} \;+\; \sqrt{DEWP} \;+\; \sqrt{TEMP} * WTime \;+\; NE \;+\; NW$$

Below are the assumption plots for our final square root response model. The Normality of errors assumption is still violated, likely because there are too many outliers, as is shown in the studentized residual histogram

below. The residuals appear to satisfy the normality assumption for the first two standard deviations, but are too large/small past that point. If we had less outliers, our data might be normal. If we used moving averages, such as over days instead of hours, our data might be more normal, but we did not have the time to consider this. The constant variance of errors and identically distributed errors assumptions both appear to be violated as shown in the fitted versus residual plot below.

We decided to include the interaction between temperature and Work Time because we believed that it would generally be warmer during business hours v. non-business hours.

Since we could not use BIC to compare our two final models, we decided to use Mean Squared Prediction Error to decide on our final model, especially because our multiple linear regression assumptions were not met. Regardless, if our assumptions were met, we can still use our models for prediction, but not inference, and if they are genuinely useful in predicting our response, we feel that it would still be worthwhile to complete this project.

The Mean Squared Prediction error of our Square root response model, 1811.69, was roughly ten times lower than the Mean Squared Prediction error of our logged response model, 12663.51, making it a clear winner.

Although we cannot perform inference on our model, we decided to include our final model's coefficients below.

```
Call:
lm(formula = pm25sqrt[2:8760] ~ pm25sqrt[1:8759] + YX$t + sin24h +
    cos24h + TEMPsqrt[1:8759] + DEWPsqrt[1:8759] + TEMPsqrt[1:8759] *
    YX$wtime[1:8759] + PRESsqrt[1:8759] + NE[1:8759] + NW[1:8759])

Residuals:
     Min       1Q   Median       3Q      Max
-10.3767  -0.4041   0.0620   0.4723  12.2001

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    1.432e+01  5.635e+00   2.542   0.0111 *
pm25sqrt[1:8759]                               9.461e-01  3.427e-03 276.024  < 2e-16 ***
YX$t                                           4.724e-07  4.952e-06   0.095   0.9240
sin24h                                        -2.509e-01  2.891e-02  -8.680  < 2e-16 ***
cos24h                                         9.574e-02  1.996e-02   4.797 1.64e-06 ***
TEMPsqrt[1:8759]                              -6.055e-01  1.189e-01  -5.091 3.63e-07 ***
DEWPsqrt[1:8759]                               2.181e-01  2.365e-02   9.222  < 2e-16 ***
YX$wtime[1:8759]Weekends                      -2.114e+00  2.336e+00  -0.905   0.3655
YX$wtime[1:8759]WorkHours                      6.171e+00  1.402e+00   4.402 1.08e-05 ***
PRESsqrt[1:8759]                              -1.584e-01  1.389e-01  -1.140   0.2542
NE[1:8759]1                                   -8.787e-02  3.558e-02  -2.469   0.0136 *
NW[1:8759]1                                    8.369e-03  2.700e-02   0.310   0.7566
TEMPsqrt[1:8759]:YX$wtime[1:8759]Weekends      1.317e-01  1.390e-01   0.947   0.3435
TEMPsqrt[1:8759]:YX$wtime[1:8759]WorkHours    -3.602e-01  8.348e-02  -4.315 1.61e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.073 on 8745 degrees of freedom
Multiple R-squared:  0.9397,    Adjusted R-squared:  0.9396
F-statistic: 1.049e+04 on 13 and 8745 DF,  p-value: < 2.2e-16
```

Figure 8: Final Model Coefficients

Below are plots of our our fitted values overlaid on our response (figure 10) and our predicted response values for 2014 (figure 9), 1 year out, overlaid on our 2014 response. We believe that our predictions were good as they appeared to follow the general pattern of the data one year out. Therefore, we believe that our final model was useful in predicting our response.

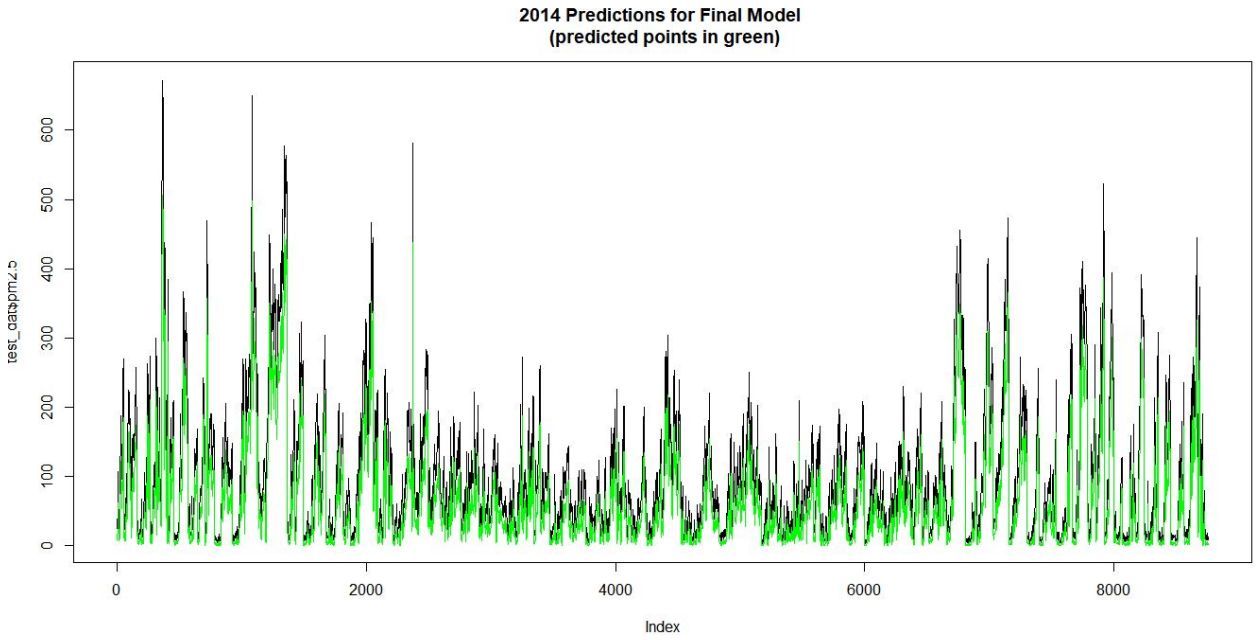

Figure 9: Final Model Predictions

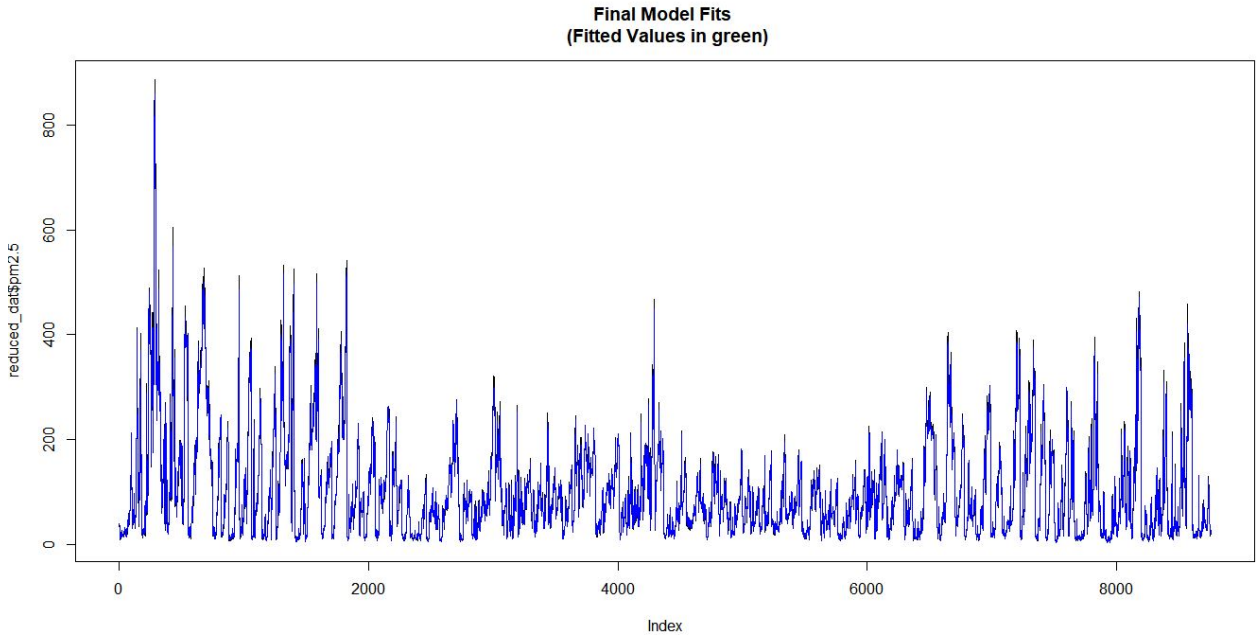

Figure 10: Final Model Fits

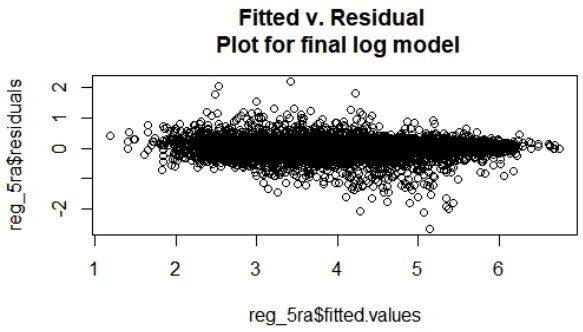

Figure 11: Log Fitted Values

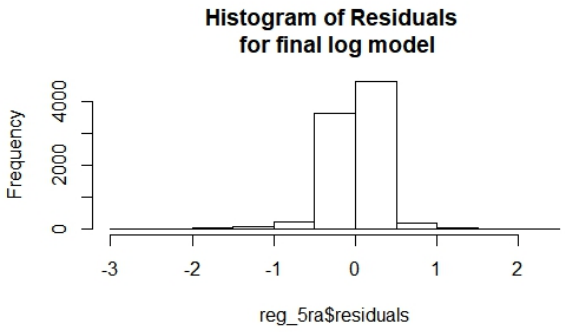

Figure 12: Log Histogram

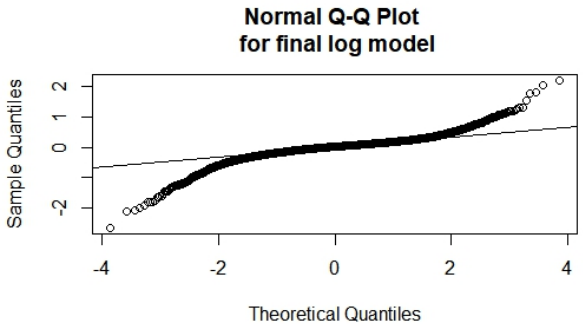

Figure 13: Log Q-Q Plot

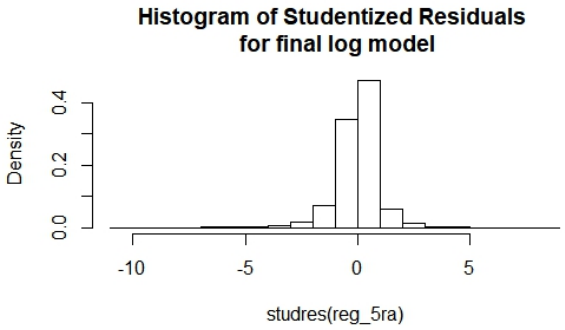

Figure 14: Log Studentized Residuals

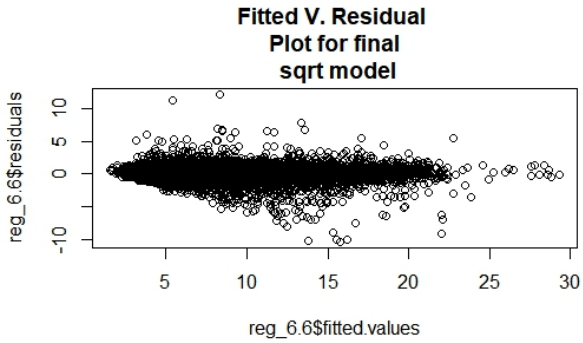

Figure 15: Square Root Fitted Values

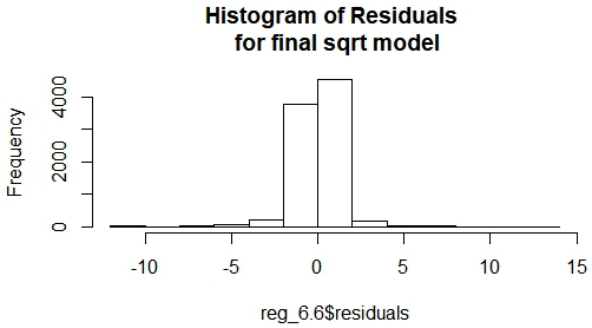

Figure 16: Square Root Histogram

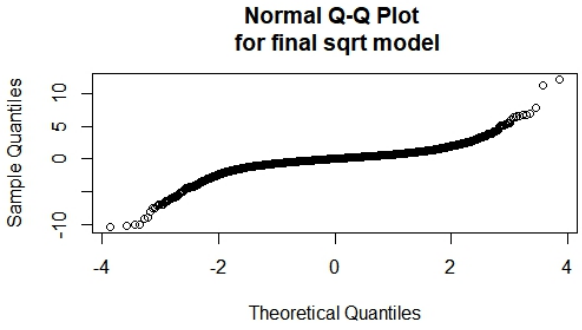

Figure 17: Square Root Q-Q Plot

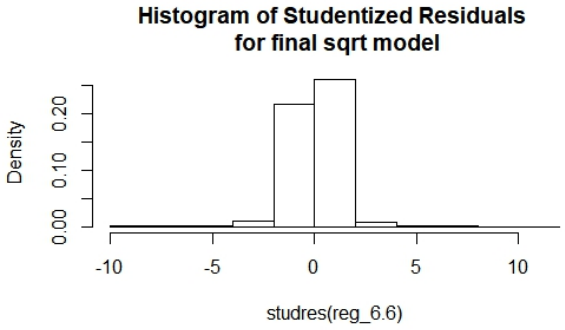

Figure 18: Square Root Studentized Residuals

Conclusion

When looking back at the basic question of interest, we focused on what predictors in the data would help to predict the $PM_{2.5}$ levels. Our model can be used to predict the $PM_{2.5}$ levels from knowing the temperature, pressure, dew point, work time, and wind direction. This is important because it can be used as a precautionary tool for the Beijing community; during circumstances when predictors reach certain levels such that $PM_{2.5}$ levels are considered to be extremely dangerous, the local government could issue warnings to its citizens.

If we could improve our model, we would look into consolidating the data from hours into days by averaging the $PM_{2.5}$ values of 24 hour spans so that we could look at the data through the years. We would also try and gather more data on things like nearby businesses, construction, and car activity [2]. This would hopefully make our normality assumptions look a lot nicer assuming that our model does not include all of the important factors in predicting $PM_{2.5}$. However, as mentioned before, we can still use our model for prediction because it still shows a satisfactory predictive ability. Since the model has some accuracy in predicting $PM_{2.5}$, we believe that this model is still useful.

References

[1] Song Xi Chen. *Beijing PM2.5 Data Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data`.

[2] Department for Environment Food Rural Affairs. *Public Health: Sources and Effects of PM2.5*. URL: `https://laqm.defra.gov.uk/public-health/pm25.html`.