# Predicting PM2.5 Levels in Beijing Based on Weather Data

Edward Huber
Rochester Institute of Technology

Stat 641
April 2020

## About The Data

In this analysis, I will use data taken from the UCI Machine Learning Repository entitled the "Beijing PM2.5 Data Data Set".[1] This data was recorded between January 1, 2010 and December 31, 2014. There are missing values in the dataset and I will eliminate all rows with missing values. Exploratory analysis shows that the amount of data lost is negligible and will not affect modeling significantly, reducing the data from 43,824 observations to 41,757 observations. The highlight of this dataset and the response variable for this analysis is Beijing's PM2.5 concentration on the hour over the aforementioned time period which will now simply be referred to as 'pm2.5'. It is a measure of atmospheric particulate matter that have a diameter of 2.5 micrometers in micrograms per cubic meter, and is an important indicator of air quality, as these pollutants are such that they cause haziness in the air when levels are high, and at these levels humans and animals are prone to ingesting or breathing them in, causing potential respiratory problems as a they are able to penetrate the lungs and and cause corrosion on the alveolar wall, thus causing extreme damages to the respiratory system.

The data includes eleven variables that I will use in my analysis. The first is the year that the observation was taken, subsequently followed by the month, day, and hour in that year. These variables are self-explanatory. On the variables that need explanation, they are as follows: Dewpoint (dewp), Temperature (temp), Pressure (hPa), Combined Wind Direction (cbwd), Cumulated Wind Speed in meters/second (iws), Cumulated Hours of Snow (is), and Cumulated Hours of Rain (ir). Let's take a closer look at the definition of these variables:

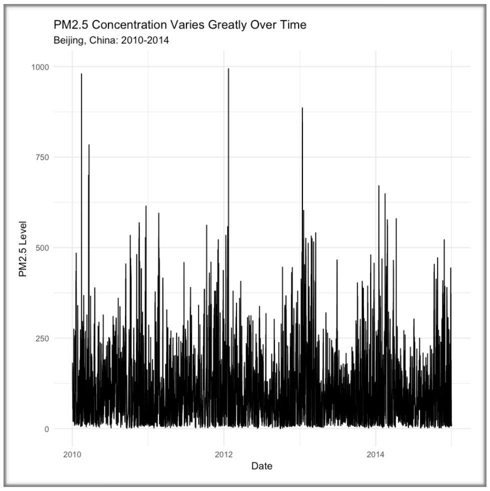
- Dewpoint refers to the temperature that the air needs to be cooled to at constant pressure in order to achieve a relative humidity of 100 percent.
- Temperature is a measure of hourly temperature in degree celsius
- Pressure is measured in hPa, or hectopascal pressure unit. It is the measure of atmospheric pressure at each hour.
- Combined Wind Direction is the reading of combined 'u' and 'v' wind component vectors on the given hour (Northwest, East, etc.)
- Cumulated Hours of Snow refers to the total time of measurable snowfall on the day.
- Cumulated Hours of Rain refers to the total time of measurable rainfall on the day.

Given these variables and the pm2.5 readings, I will use Linear Regression techniques to explore whether or not one can confidently predict pm2.5 levels in Beijing using general hourly weather data. The model will be fit using in the ordinary least squares method which yields a general equation in the format of $Y = \beta_0 + \beta_1 X_1 + \ldots \beta_j X_j$ where least squares estimators will be calculated in matrix format where b is the vector of coefficient estimates, $b = (X'X)^{-1}X'Y$. The model will also follow 5 assumptions required of linear modeling: 1) The relationship between the response and predictor variables is linear, 2) Error has a mean of zero, 3) Error has constant variance, 4) Error terms are independent, 5) Error terms are normally distributed.

[1] http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data
Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257

## Exploring the Data

Again, the question at hand is whether or not one can confidently predict PM2.5 level in Beijing based on basic weather data. First, let's take a look at some summary statistics taken from the data. One can clearly see from the plot below that PM2.5 level vary greatly day to day, but that the overall trend is that the levels have declined over time. Initial impressions say that this data may be hard to model with a simple linear model. See below the plot to take a look at some summary statistics for the variables that will be used in analysis.



```
pm2.5              dewp              temp             pres
 Min.   :  0.00   Min.   :-40.00   Min.   :-19.0   Min.   : 991
 1st Qu.: 29.00   1st Qu.:-10.00   1st Qu.:  2.0   1st Qu.:1008
 Median : 72.00   Median :  2.00   Median : 14.0   Median :1016
 Mean   : 98.61   Mean   :  1.75   Mean   : 12.4   Mean   :1016
 3rd Qu.:137.00   3rd Qu.: 15.00   3rd Qu.: 23.0   3rd Qu.:1025
 Max.   :994.00   Max.   : 28.00   Max.   : 42.0   Max.   :1046
      iws               is                  ir
 Min.   :  0.45   Min.   : 0.00000   Min.   : 0.0000
 1st Qu.:  1.79   1st Qu.: 0.00000   1st Qu.: 0.0000
 Median :  5.37   Median : 0.00000   Median : 0.0000
 Mean   : 23.87   Mean   : 0.05534   Mean   : 0.1949
 3rd Qu.: 21.91   3rd Qu.: 0.00000   3rd Qu.: 0.0000
 Max.   :565.49   Max.   :27.00000   Max.   :36.0000
```

A quick glance at the above summary statistics shows the following:

- pm2.5 takes on a large range of values from 0 to 994, with the majority of values occurring around 100. This is not good news for Beijing, as any reading over 35.4 ug/$m^3$ is considered unsafe. [2]
- The dew point has a median of 2 and a max of 28, which tells us the air was relatively dry for most of the recorded time period. [3]
- Temperature doesn't seem to have any abnormalities, and is consistent with average ranges that tend to be recorded for Beijing year to year. [4]
- The average pressure (hPa) is consistent with observed values over time of hPa, whose yearly range is 980 - 1030 hPa.[5]
- Cumulated wind speed (lws) appears that It may have some outliers, considering its range of values goes from 0.45 to 565, something that may need to taken into account when modeling the data.
- Cumulated hours of snow does not appear to show any points that may have a significant effect on the model.
- The same goes for cumulated hours of rain.

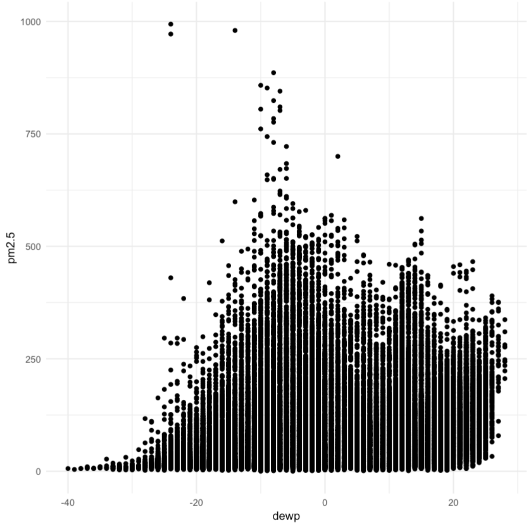
Before calculating a linear model on the data, let's take a look at each variables relationship to pm2.5 values. The plot pm2.5 vs. each respective variable is found on the next page. We see that each variable has a slightly different relationship with the response. Cumulated rain hours cumulated snow hours, cumulated wind speed appear to have a logarithmic relationship with the response. Dew point has a weak-positive relationship with the response, while pressure and temp have a weak-negative relationship with the response. Now that we have an idea of these relationships and the ranges in values that we see in the each variable. Let's fit the initial full model.

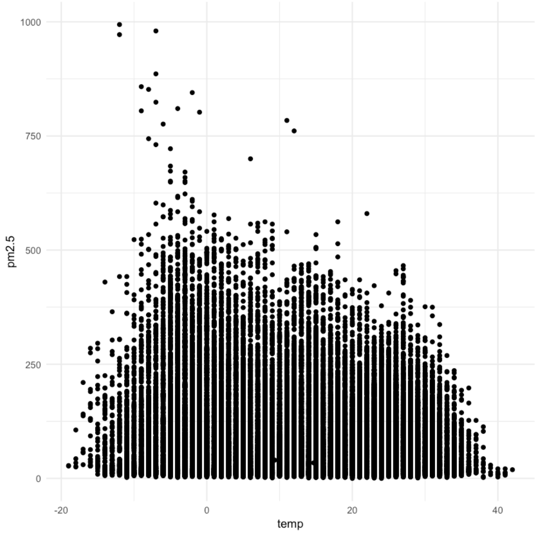
[2] The World Air Quality Index project. (n.d.). Revised PM2.5 AQI breakpoints. Retrieved from https://aqicn.org/faq/2013-09-09/revised-pm25-aqi-breakpoints/

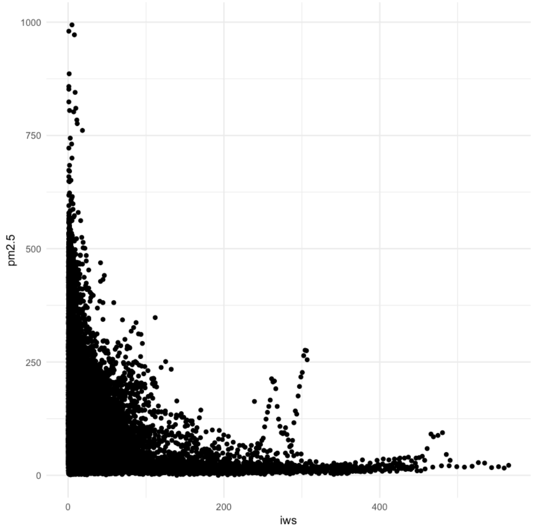
[3] (n.d.). Retrieved from https://www.theweatherprediction.com/habyhints/190/

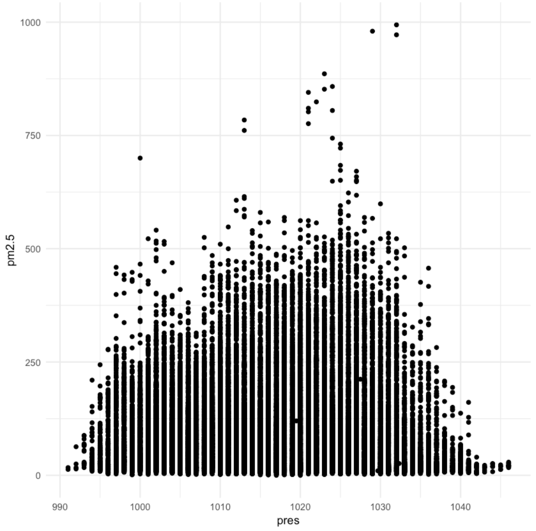
[4] Data.org. (n.d.). Retrieved from https://en.climate-data.org/asia/china/beijing/beijing-134/

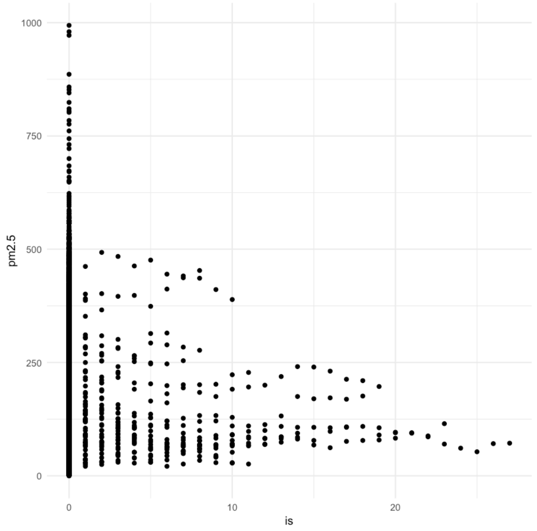
[5] (n.d.). Retrieved from https://blog.metservice.com/node/1021

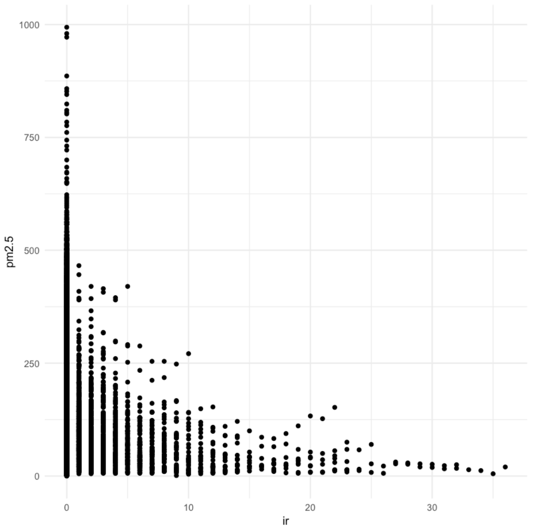

pm2.5
1000
750
500
250
0
-40
-20
0
20
dewp


pm2.5
1000
750
500
250
0
-20
0
20
40
temp


pm2.5
1000
750
500
250
0
0
200
400
iws


pm2.5
1000
750
500
250
0
990
1000
1010
1020
1030
1040
pres


pm2.5
1000
750
500
250
0
0
10
20
is


pm2.5
1000
750
500
250
0
0
10
20
30
ir

## Modeling and Testing

First, I will fit a basic multiple linear regression model to PM2.5 in response to a linear combination of the variables DEWP, TEMP, PRES, IWS, IS, and IR. In testing significance and confidence, I will use a standard of 95% throughout the analysis. Fitting the regression model gives the equation and summary output:

$$E(pm2.5) = 1.728 + 4.282 * dewp - 6.068 * temp - 1.529 * pres - 0.262 * iws - 2.267 * is - 7.206 * ir$$

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
      (Intercept)  1.728e+03  7.299e+01  23.680  < 2e-16 ***
      dewp         4.282e+00  5.346e-02  80.109  < 2e-16 ***
      temp        -6.068e+00  6.836e-02 -88.764  < 2e-16 ***
      pres        -1.529e+00  7.135e-02 -21.431  < 2e-16 ***
      iws         -2.616e-01  8.436e-03 -31.015  < 2e-16 ***
      is          -2.267e+00  5.097e-01  -4.448  8.7e-06 ***
      ir          -7.206e+00  2.816e-01 -25.593  < 2e-16 ***
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      Residual standard error: 80.46 on 41750 degrees of freedom
      Multiple R-squared:  0.2361,     Adjusted R-squared:  0.236
      F-statistic:  2151 on 6 and 41750 DF,  p-value: < 2.2e-16
```

Now that we have a model, let's consider the variables as a whole and determine if the group is significant in predicting pm2.5 levels using the hypotheses:

$H_0 : \beta_1 = \beta_2 \ldots = \beta_j = 0$
$H_A : \beta_j \neq 0$, for at least one j, j = 1 …, p

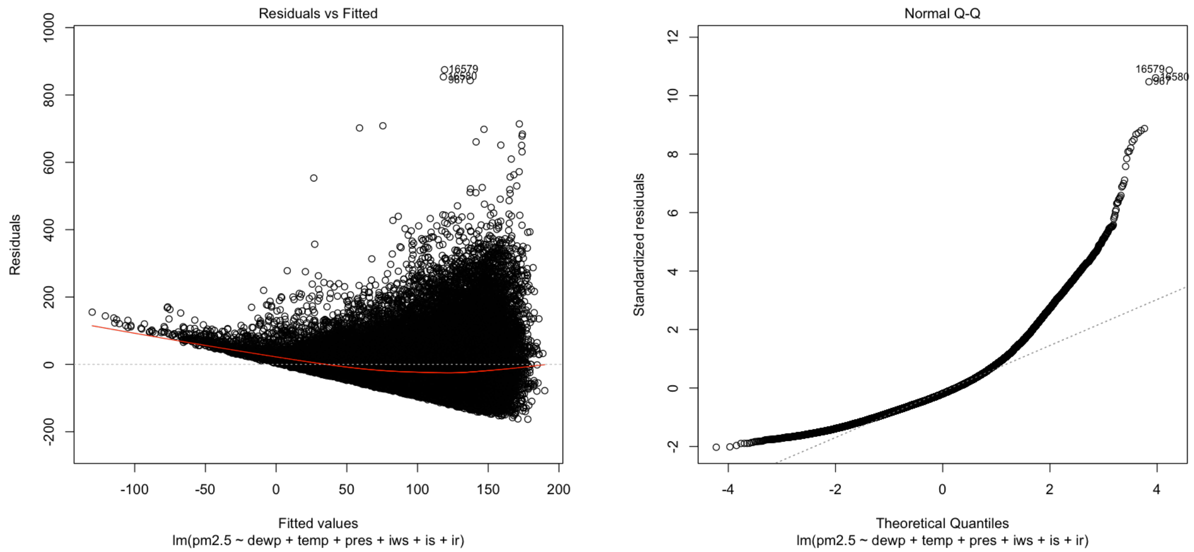
For this generalization using the general F-test in conjunction with the summary output of the model, we see that:

$F_0 = 2151 > F_{6,\, 41750,\, 0.95} = 2.09$

Thus, we reject the null hypothesis in favor of the alternative that at least one Beta is not equal to zero. The next natural question is: which, if not all Betas, are significant to the model? In order to accomplish this, I will calculate the critical t-value for the data on 7 and 41757 degrees of freedom and compare that to the absolute t-value for each beta. By converting the data to matrices in R, I have calculated a vector of absolute t-values for each Beta as well as the critical t-value for this data. Using the hypotheses

$H_0 : \beta_j = 0$, for j = 1…, p
$H_A : \beta_j \neq 0$, for j = 1…, p

We can reject the null hypotheses for all j values for this data as the vector of absolute t-values for Betas 1 to 7 is (23.68, 80.11, 88.76, 21.43, 31.01, 4.45, 25.59), all of which are greater than the 0.67, the critical t-value. According to this model, then, all variables contribute significantly to the expectation of pm2.5 levels. An initial point of concern is that we see an $R^2$ of 23.6%, which means that this model accounts for very little of the variance in the data. As this is a multiple regression analysis with many data points and 6 regressor variables, it is important to



consider the model in other ways. Let's start by looking at the Residuals vs. Fitted  and Normal QQ plots (above) for the model. These plots show a story different than we might think from the original significance tests. The normal plot shows that the data is heavily right skewed while the Residuals vs. Fitted plot shows a right facing cone, indicating that a logarithmic transformation on some or all of the variables and or response should be considered.

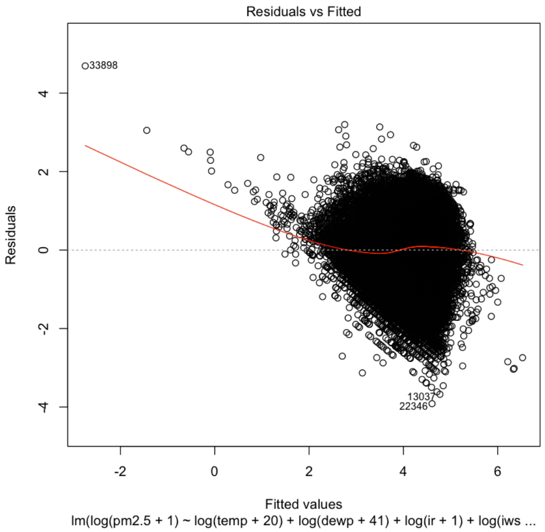
## Transformation

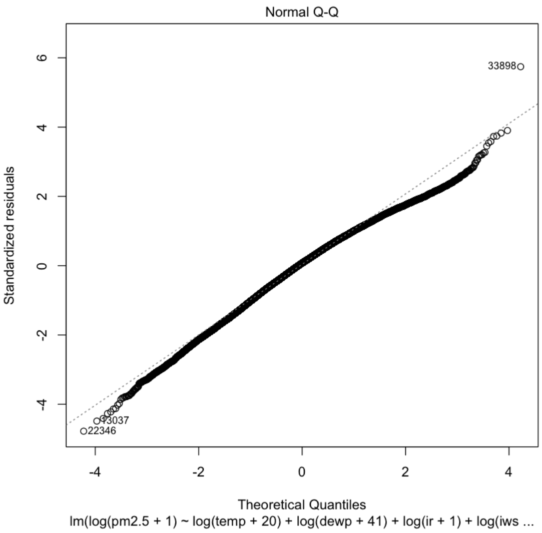
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    66.246346   4.878310  13.580   <2e-16 ***
log(temp + 20) -1.364319   0.017341 -78.677   <2e-16 ***
log(dewp + 41)  1.863805   0.019710  94.561   <2e-16 ***
log(ir + 1)    -0.317148   0.012597 -25.176   <2e-16 ***
log(iws)       -0.137644   0.002805 -49.069   <2e-16 ***
log(is + 1)     0.024739   0.023314   1.061    0.289    
log(pres)      -9.250335   0.697051 -13.271   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8196 on 41750 degrees of freedom
Multiple R-squared:  0.3358,     Adjusted R-squared:  0.3357 
F-statistic:  3518 on 6 and 41750 DF,  p-value: < 2.2e-16
```

First I have applied a logarithmic transformation to all variables and the response. In order to avoid taking Log(0) or Log of a negative, I have added the following to create a minimum of 1 for each predictor: [pm2.5 + 1, temp + 20, dewp + 41, ir + 1, is + 1]. Looking at the summary output of this model gives some better insight as to what variables are significant or not. Recall that in the basic model output we saw the highest p-value in cumulated snow hours, and while still significant in that model, its p-value was much higher than the rest of the predictors. Using the logarithmic transformation we see a very high p-value, so let's consider the model eliminating cumulated snow hours:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.046737   4.874689   13.55   <2e-16 ***
log(temp + 20) -1.366840   0.017177  -79.57   <2e-16 ***
log(dewp + 41)  1.866705   0.019520   95.63   <2e-16 ***
log(ir + 1)    -0.317493   0.012593  -25.21   <2e-16 ***
log(iws)       -0.137392   0.002795  -49.16   <2e-16 ***
log(pres)      -9.221829   0.696534  -13.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8196 on 41751 degrees of freedom
Multiple R-squared:  0.3358,	Adjusted R-squared:  0.3357 
F-statistic:  4221 on 5 and 41751 DF,  p-value: < 2.2e-16
```

By applying the logarithmic transformation to the model we see that now all predictor variables are considered significant to the model based on a significance level of 95%, each having a p-value very close to zero. In both transformations we see that the $R^2$ value is essentially the same, coming in at about 34%. While it is generally desired to have a higher $R^2$ value, this does not necessarily mean that the model is not adequate. Let's take a look at the normal and residual vs. fitted plots to see if they are any different than those from the simple model:





Comparing these plots to the plots from the original model, we see a great improvement. While not perfect, the Normal plot shows slight left-skewness but a much better fit than the previous model, and residuals no longer show that a transformation is necessary. As we have seen low $R^2$ values let's consider the PRESS statistic for the models, as it is generally accepted as a better indicator of a model's predictive power and accuracy.[6] Calculating the PRESS statistic on the original model and the transformed model, we see that the logarithmic transformation

[6] Mitsa, T. (n.d.). Use PRESS, not R squared to judge predictive power of regression. Retrieved from https://www.analyticbridge.datasciencecentral.com/profiles/blogs/use-press-not-r-squared-to-judge-predictive-power-of-regression

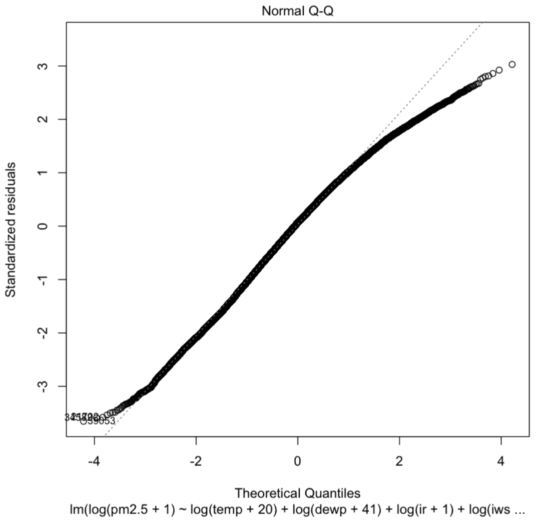
sans cumulated snow hours has significantly better predictive accuracy than the original model, their scores being:

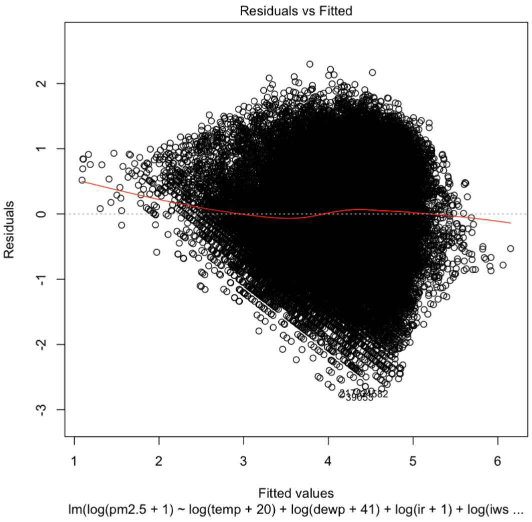
$\text{PRESS}_{\text{original}}$ = 270338852.26 and $\text{PRESS}_{\text{logMod}}$ = 28052.50

Now that we know that the logarithmic transformation on the partial model provides better prediction accuracy, let's improve on it by removing outliers. I will do this by using Cook's Distance to determine what points are influential points and removing them (1,694 points) from the data. Doing so results in the following model output:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    69.915960   4.717281   14.82   <2e-16 ***
log(temp + 20) -1.567131   0.017143  -91.42   <2e-16 ***
log(dewp + 41)  2.061474   0.019344  106.57   <2e-16 ***
log(ir + 1)    -0.283592   0.015382  -18.44   <2e-16 ***
log(iws)       -0.133139   0.002675  -49.77   <2e-16 ***
log(pres)      -9.783992   0.673881  -14.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7584 on 40057 degrees of freedom
Multiple R-squared:  0.3887,     Adjusted R-squared:  0.3886 
F-statistic:  5094 on 5 and 40057 DF,  p-value: < 2.2e-16
```

Notice that all regressor variables are still significant per the p-value approach, and that the $R^2$ value has gone up to about 39%. Calculating the PRESS statistic on this model gives the value 23044.904, a significant improvement from the original transformation. Looking at the Normal and Residual vs. Fitted plots for this model (below) also shows some improvement. Although the normal plot still shows some skewness, the Residuals vs. Fitted plot shows more evenly dispersed points about the zero line. That being said, it appears as though this is the best linear model to use for purposes of the question at hand, whether or not we can predict pm2.5 levels using a set of weather data variables.





Now that we have decided on the model, it is important to check validity and significance of the model and Beta values using standard hypothesis tests.

$H_0 : \beta_1 = \beta_{2 \ldots} = \beta_j = 0$
$H_A : \beta_j \neq 0$, for at least one j, j = 1 …, p

For this generalization using the general F-test in conjunction with the summary output of the model, we see that:

$F_0 = 5094 > F_{5,\ 40057,\ 0.95} = 2.21$

Thus, we reject the null hypothesis in favor of the alternative that at least one Beta is not equal to zero. The next natural question is: which, if not all Betas, are significant to the model? In order to accomplish this, I will calculate the critical t-value for the data on 6 and 40057 degrees of freedom and compare that to the absolute t-value for each beta. By converting the data to matrices in R, I have calculated a vector of absolute t-values for each Beta as well as the critical t-value for this data. Using the hypotheses

$H_0 : \beta_j = 0$, for j = 1…, p
$H_A : \beta_j \neq 0$, for j = 1…, p

We can reject the null hypotheses for all j values for this data as the vector of absolute t-values for Betas 1 to 6 is (22.47, 83.59, 91.55, 20.17, 32.11, 21.35), all of which are greater than the 0.67, the critical t-value. According to this model, then, all variables contribute significantly to the expectation of pm2.5 levels. As a final step, let’s take a look at the 95% confidence intervals for each regressor in this model:

- $60.70 \leq \beta_1$ (intercept) $\leq 79.16$
- $-1.6 \leq \beta_2$ (temperature) $\leq -1.53$
- $2.02 \leq \beta_3$ (dew point) $\leq 2.10$
- $-0.31 \leq \beta_4$ (cumulated rain hours) $\leq -0.25$
- $-0.14 \leq \beta_5$ (cumulated wind speed) $\leq -0.13$
- $-11.1 \leq \beta_6$ (pressure) $\leq -8.46$

The range on the 95% confidence interval for each Beta is relatively small, and we can conclude that based on this analysis the appropriate model for predicting pm2.5 levels in Beijing at a given time and given weather factors of temperature, dew point, cumulated rain hours, cumulated wind speed, and hPa pressure is:

$$E(PM2.5) = 69.96 - 1.57 * log(temp + 20) + 2.06 * log(dewp + 41) - 0.28 * log(ir + 1) - 0.13 * log(iws) - 9.78 * log(pres)$$

### Conclusion

In summary, it appears that it is possible to predict Beijing’s PM2.5 levels based on certain weather data provided. As noted in the beginning of the analysis, a time-series plot shows that PM2.5 levels fluctuate greatly in short time periods, but appear to be going down in general over time. The initial look at the summary statistics of the variables used in predicting PM2.5 levels shows that there are no irregularities in the data that would contribute to an unreliable model based on weather normalcies. Modeling of the data shows that there is not an exact linear relationship between the weather variables and PM2.5 levels and that is perhaps more complicated than a cut and dry relationship. By applying linear transformation techniques to the data, I have developed a model that we can say is accurate to predicting pm2.5 levels in Beijing at a 95% confidence level. The best use for this model is not to predict the exact value of pm2.5 levels, but more to predict whether or not those levels will be in a safe or harmful range. As mentioned earlier, any level of pm2.5 over 35.4 ug/m$^3$ is considered unsafe. Therefore, if we can be 95% confident in the prediction, it is equivalent to say that we can be almost certain that we can predict whether air conditions will be hazardous or not during a given time period based on any weather data for variables used in this analysis.

## References

(n.d.). Retrieved from https://www.theweatherprediction.com/habyhints/190/

(n.d.). Retrieved from https://blog.metservice.com/node/1021

Data.org. (n.d.). Retrieved from https://en.climate-data.org/asia/china/beijing/beijing-134/

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257. <http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>

Mitsa, T. (n.d.). Use PRESS, not R squared to judge predictive power of regression. Retrieved from https://

www.analyticbridge.datasciencecentral.com/profiles/blogs/use-press-not-r-squared-to-judge-predictive-power-of-regression

The World Air Quality Index project. (n.d.). Revised PM2.5 AQI breakpoints. Retrieved from https://aqicn.org/faq/2013-09-09/revised-pm25-aqi-breakpoints/