Statistics 333 - Applied Regression Analysis Project:

Predicting PM10 with Other Air Quality Measures

Meizi Liu, Ethan Spalding and Qihong Lu

University of Wisconsin-Madison

Predicting PM10 with Other Air Quality Measures

There have been many major cities in developing countries suffered from air pollution, such as Beijing, China (Kan, Chen & Hong, 2009). Importantly, it is well understood that air pollution particulates, such as coarse particulate matter (PM10) and fine particulate matter (PM2.5) are detrimental to people's health. According to United States Environmental Protection Agency (2013), PM refers to a small particles found in the air, consisting of dust, dirt, soot, smoke, and other various solid and liquid chemicals. PM10 are particulate matters with size ranges smaller than 10 micrometers in diameter; PM2.5 refers to fine particulate matters that are 2.5 micrometers in diameter and smaller.

Fine particles are primarily from combustion sources like vehicles, diesel engines, power plants and other industrial facilities. Since the size of the particles is so small, they can pose great health problems to human by penetrating deeply into our lungs and even bloodstream. Numerous research has shown that exposure to fine particles can aggravate lung disease such as asthma, reduce lung function and even cause premature death in people with heart or lung disease (EPA, 2013). A recent report from World Health Organization (2014) suggests that reducing PM10 from 70 to 20 micrograms per cubic meter can potentially reduce 15% of air pollution related deaths, demonstrating the importance of modeling its levels in the air.

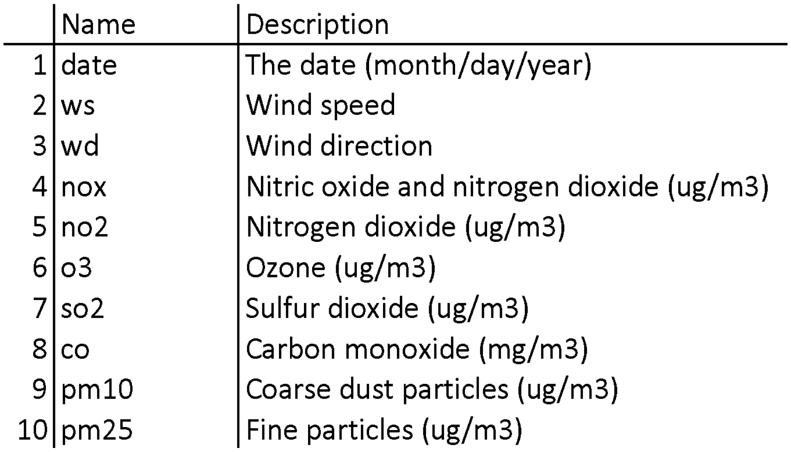
The Clean Air Act requires EPA to set National Ambient Air Quality Standards for six common air pollutants, and particle matter is one of these. The other five air pollutants are ground-level ozone (O3), carbon monoxide(CO), sulfur oxides(SOx), nitrogen oxides(NOx) and lead. Multiple air pollutants tend to be emitted from by the combustion sources simultaneously. Moreover, studies have shown that fine particles can be generated as a product of reactions between nitrogen oxides, sulfur oxides and ammonia in the air (Nevada Division of Environmental Protection, 2009). The objective of the present research is to investigate the relation between concentration of particulate matters and other common air quality measures, including nitric oxide and nitrogen dioxide, ozone, sulfur dioxide and carbon monoxide.

2. Preliminary Data Analysis

2.1 Description of the Database

A dataset was retrieved from The Open Air project for this project. This dataset was collected in London from May 1998 to June 2005. It consists eight independent variables and two responses variables of interest: PM10. All measurements were made 24 times a day for a interval of 1 hour. It has 42892 observations (after eliminating missing data) arranged in a chronological order.

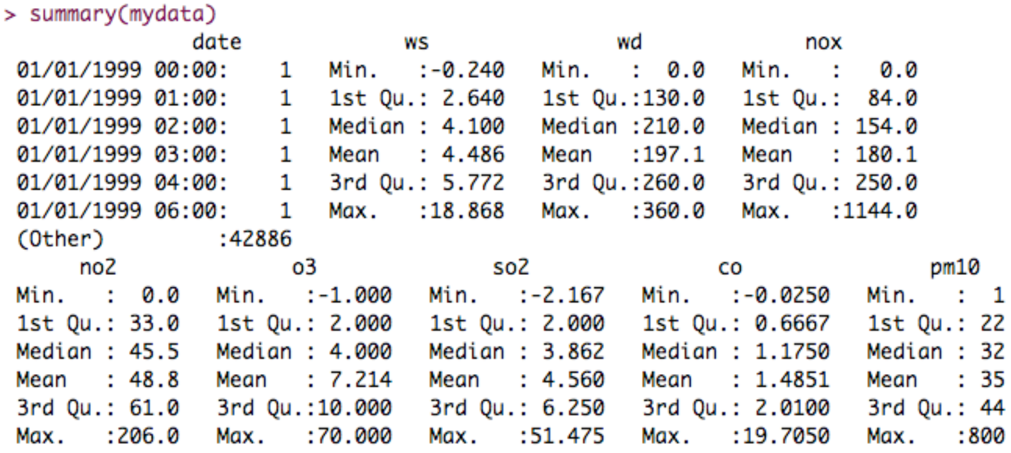
Variables list:



	Name	Description
1	date	The date (month/day/year)
2	ws	Wind speed
3	wd	Wind direction
4	nox	Nitric oxide and nitrogen dioxide (ug/m3)
5	no2	Nitrogen dioxide (ug/m3)
6	o3	Ozone (ug/m3)
7	so2	Sulfur dioxide (ug/m3)
8	co	Carbon monoxide (mg/m3)
9	pm10	Coarse dust particles (ug/m3)
10	pm25	Fine particles (ug/m3)

Remark: Since PM2.5 concentrations are calculated by PM10 concentration (Smyth, Jiang & Yin, 2006), we decided to fit model for PM10 only.

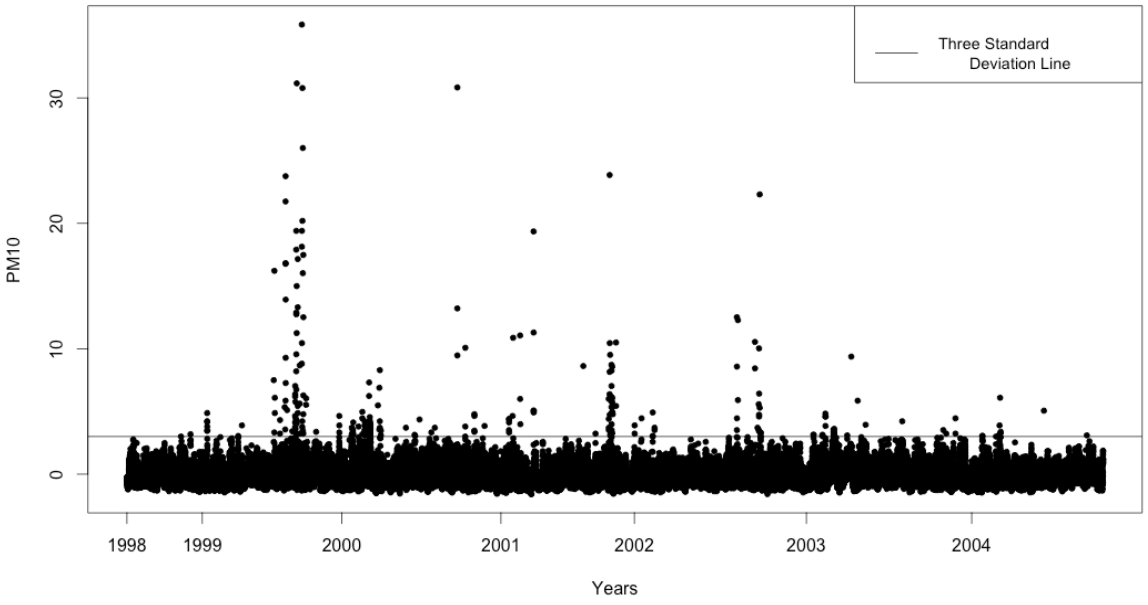
2.2 Overview of the data



```
> summary(mydata)
                  date              ws                wd               nox        
 01/01/1999 00:00:    1   Min.   :-0.240   Min.   :  0.0   Min.   :   0.0  
 01/01/1999 01:00:    1   1st Qu.: 2.640   1st Qu.:130.0   1st Qu.:  84.0  
 01/01/1999 02:00:    1   Median : 4.100   Median :210.0   Median : 154.0  
 01/01/1999 03:00:    1   Mean   : 4.486   Mean   :197.1   Mean   : 180.1  
 01/01/1999 04:00:    1   3rd Qu.: 5.772   3rd Qu.:260.0   3rd Qu.: 250.0  
 01/01/1999 06:00:    1   Max.   :18.868   Max.   :360.0   Max.   :1144.0  
 (Other)          :42886                                                   
      no2              o3               so2               co                pm10    
 Min.   :  0.0   Min.   :-1.000   Min.   :-2.167   Min.   :-0.0250   Min.   :  1  
 1st Qu.: 33.0   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 0.6667   1st Qu.: 22  
 Median : 45.5   Median : 4.000   Median : 3.862   Median : 1.1750   Median : 32  
 Mean   : 48.8   Mean   : 7.214   Mean   : 4.560   Mean   : 1.4851   Mean   : 35  
 3rd Qu.: 61.0   3rd Qu.:10.000   3rd Qu.: 6.250   3rd Qu.: 2.0100   3rd Qu.: 44  
 Max.   :206.0   Max.   :70.000   Max.   :51.475   Max.   :19.7050   Max.   :800  
```

2.3 Selection of a Subset of Data

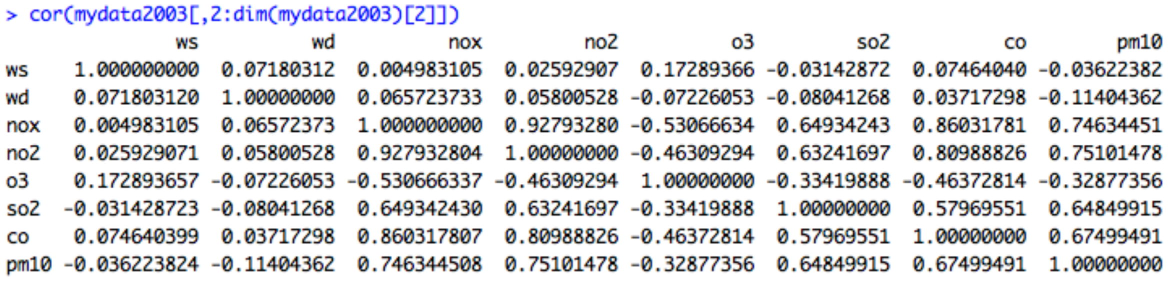
We chose only to use data from 2003 onwards. When plotting our response variable, PM10, against time, there were many clusters of outliers, illustrated with large spikes in some time windows (See the plot below). Since there is no mechanism account for this phenomenon, we chose not to remove them. Instead, we decided to select the data after 2003 onwards, which has a more stable pattern.



2.4 Removal of the Wind Direction Variable

Before fitting our model, we examined each variable by data type, and decided to not attempt to use wind direction in the model. This variable was recorded in zero to 360 degrees so that it is not applicable in the regression setting. More concretely, having degree of 350 or 10 are very different in magnitude, but represents about the same direction.

2.5 Correlation matrix



```
> cor(mydata2003[,2:dim(mydata2003)[2]])
              ws          wd          nox         no2          o3         so2          co        pm10
ws    1.000000000  0.07180312  0.004983105  0.02592907  0.17289366 -0.03142872  0.07464040 -0.03622382
wd    0.071803120  1.00000000  0.065723733  0.05800528 -0.07226053 -0.08041268  0.03717298 -0.11404362
nox   0.004983105  0.06572373  1.000000000  0.92793280 -0.53066634  0.64934243  0.86031781  0.74634451
no2   0.025929071  0.05800528  0.927932804  1.00000000 -0.46309294  0.63241697  0.80988826  0.75101478
o3    0.172893657 -0.07226053 -0.530666337 -0.46309294  1.00000000 -0.33419888 -0.46372814 -0.32877356
so2  -0.031428723 -0.08041268  0.649342430  0.63241697 -0.33419888  1.00000000  0.57969551  0.64849915
co    0.074640399  0.03717298  0.860317807  0.80988826 -0.46372814  0.57969551  1.00000000  0.67499491
pm10 -0.036223824 -0.11404362  0.746344508  0.75101478 -0.32877356  0.64849915  0.67499491  1.00000000
```

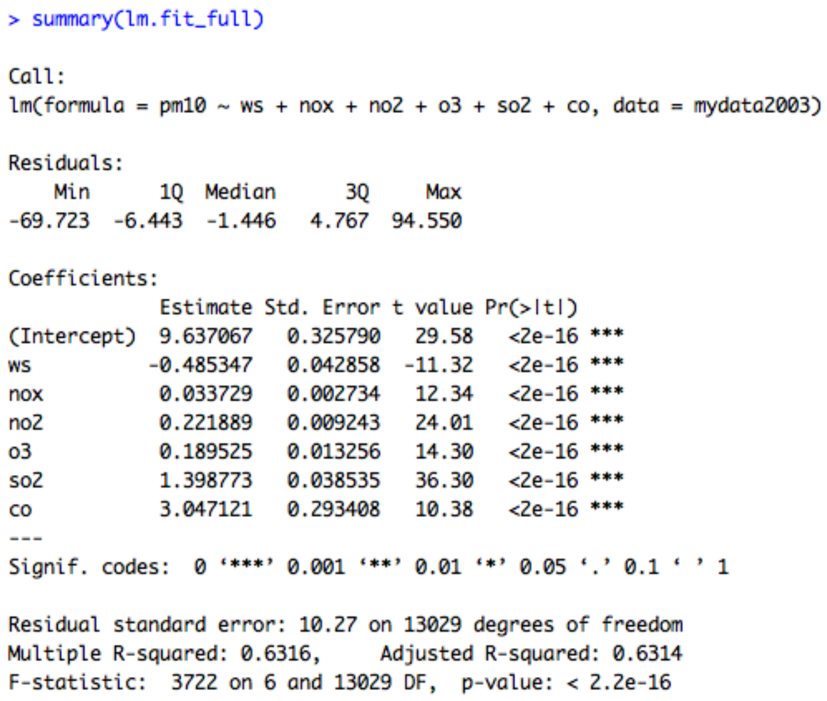
One can see some inter-correlations among predictors (Also see Appendix A for the scatter plot matrix). For example, there are strong associations between Nox and NO2 (r = 0.783), SO2 (r = 0.706), and CO (r = 0.842), which suggest some potential needs for multicollinearity remediation.

On the other hand, the response, PM10, is correlated with many other predictors, such as NOx (r = 0.746), NO2 (r = 0.751), SO2 (r = 0.648), CO (r = 0.675). These results suggest that regression model might be feasible method to predict PM10.

3. Analysis

3.1 Fitting the Full Model

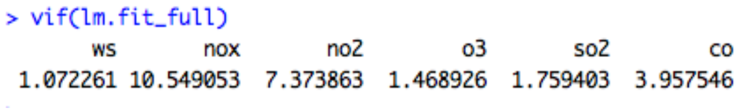
In order to establish a baseline for evaluating model performance, a full model, which has all predictors, was constructed. Here are the summary statistics of the full model.



```
> summary(lm.fit_full)

Call:
lm(formula = pm10 ~ ws + nox + no2 + o3 + so2 + co, data = mydata2003)

Residuals:
    Min      1Q  Median      3Q     Max 
-69.723  -6.443  -1.446   4.767  94.550 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.637067   0.325790   29.58   <2e-16 ***
ws          -0.485347   0.042858  -11.32   <2e-16 ***
nox          0.033729   0.002734   12.34   <2e-16 ***
no2          0.221889   0.009243   24.01   <2e-16 ***
o3           0.189525   0.013256   14.30   <2e-16 ***
so2          1.398773   0.038535   36.30   <2e-16 ***
co           3.047121   0.293408   10.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.27 on 13029 degrees of freedom
Multiple R-squared: 0.6316,	Adjusted R-squared: 0.6314 
F-statistic:  3722 on 6 and 13029 DF,  p-value: < 2.2e-16
```

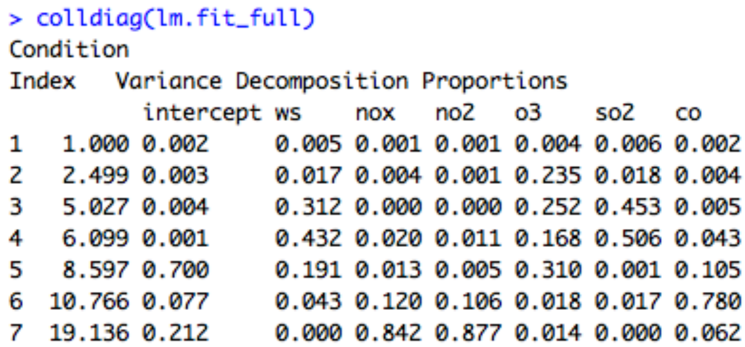
3.2 Multicollinearity diagnostics

The variance inflation factor was computed to inform the multicollinearity diagnostics.



```
> vif(lm.fit_full)
       ws       nox       no2        o3       so2        co 
 1.072261 10.549053  7.373863  1.468926  1.759403  3.957546 
```

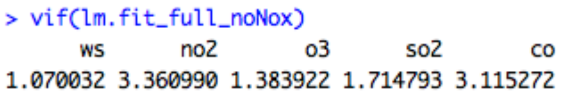
Here are the conditional indices and conditional numbers:



```
> colldiag(lm.fit_full)
Condition
Index   Variance Decomposition Proportions
          intercept ws    nox   no2   o3    so2   co   
1   1.000 0.002     0.005 0.001 0.001 0.004 0.006 0.002
2   2.499 0.003     0.017 0.004 0.001 0.235 0.018 0.004
3   5.027 0.004     0.312 0.000 0.000 0.252 0.453 0.005
4   6.099 0.001     0.432 0.020 0.011 0.168 0.506 0.043
5   8.597 0.700     0.191 0.013 0.005 0.310 0.001 0.105
6  10.766 0.077     0.043 0.120 0.106 0.018 0.017 0.780
7  19.136 0.212     0.000 0.842 0.877 0.014 0.000 0.062
```

When computing variance inflation factor (VIF), a strong multicollinearity was detected. In particular, NOx has VIF of 10.55. This is also supported by the results of conditional indices. Therefore, NOx was removed from the list of predictors.

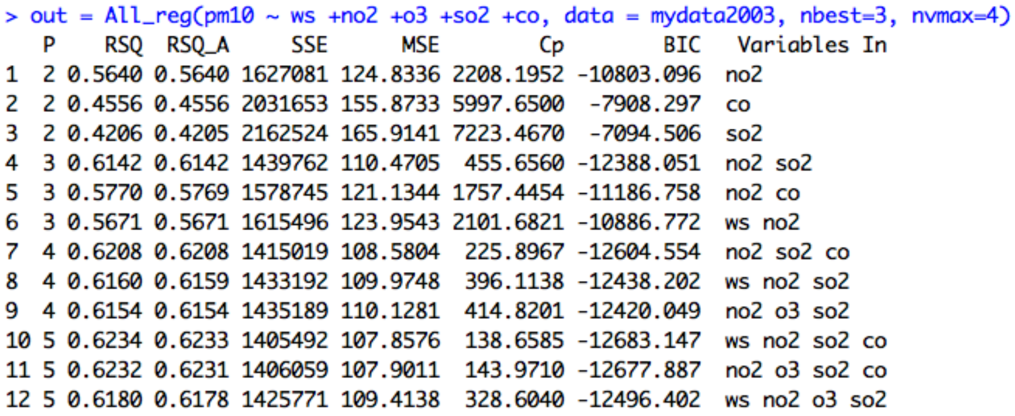
Since NO2 is a component of NOx, NOx is not very meaningful given that NO2 is in the model. So the removal of NOx should not reduce the interpretability of the model drastically. However, the VIF were reduced substantially, which is shown below.



```
> vif(lm.fit_full_noNox)
      ws      no2       o3      so2       co 
1.070032 3.360990 1.383922 1.714793 3.115272 
```

3.3 Variable selection

Here are the summary results for several variable selection criteria, including R-squared, adjusted R-squared, SSE, MSE, Cp and BIC (Also see Appendix B for visualization of these results).



```
> out = All_reg(pm10 ~ ws +no2 +o3 +so2 +co, data = mydata2003, nbest=3, nvmax=4)
   P    RSQ  RSQ_A     SSE      MSE        Cp        BIC  Variables In
1  2 0.5640 0.5640 1627081 124.8336 2208.1952 -10803.096  no2
2  2 0.4556 0.4556 2031653 155.8733 5997.6500  -7908.297  co
3  2 0.4206 0.4205 2162524 165.9141 7223.4670  -7094.506  so2
4  3 0.6142 0.6142 1439762 110.4705  455.6560 -12388.051  no2 so2
5  3 0.5770 0.5769 1578745 121.1344 1757.4454 -11186.758  no2 co
6  3 0.5671 0.5671 1615496 123.9543 2101.6821 -10886.772  ws no2
7  4 0.6208 0.6208 1415019 108.5804  225.8967 -12604.554  no2 so2 co
8  4 0.6160 0.6159 1433192 109.9748  396.1138 -12438.202  ws no2 so2
9  4 0.6154 0.6154 1435189 110.1281  414.8201 -12420.049  no2 o3 so2
10 5 0.6234 0.6233 1405492 107.8576  138.6585 -12683.147  ws no2 so2 co
11 5 0.6232 0.6231 1406059 107.9011  143.9710 -12677.887  no2 o3 so2 co
12 5 0.6180 0.6178 1425771 109.4138  328.6040 -12496.402  ws no2 o3 so2
```

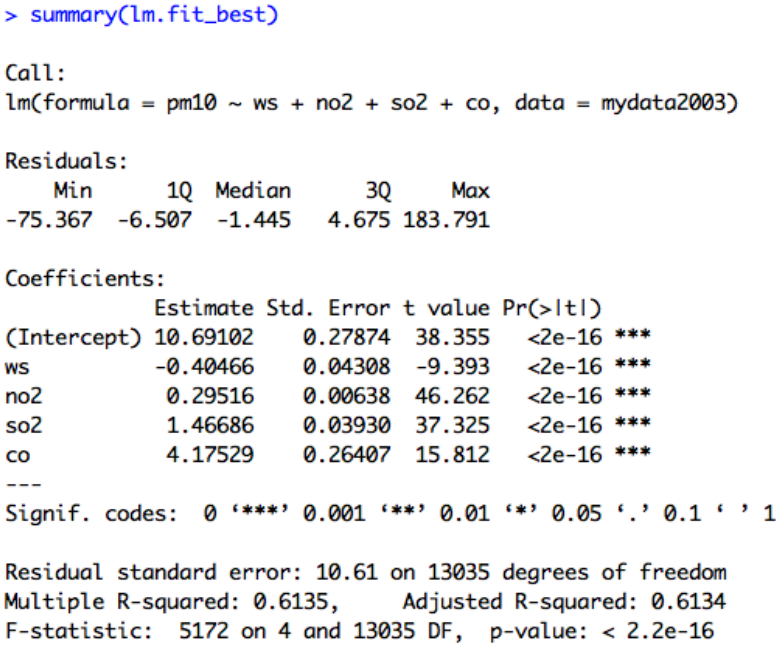
Comparatively speaking, the 10th model seems to be the most desirable one. It has the lowest Cp and BIC value and the highest value on adjusted R-squared. We therefore selected ws, no2, so2 and co as our predictors.

However, stepwise procedure with both forward and backward direction selected the full model (See appendix D). In general, we still prefer the more parsimonious model selected from all regression procedure, its performance on adjusted R-squared is comparable with the full model even without *o3*. To summarize, we chose *ws, no2, so2* and *co* as predictors.

4. Final Model

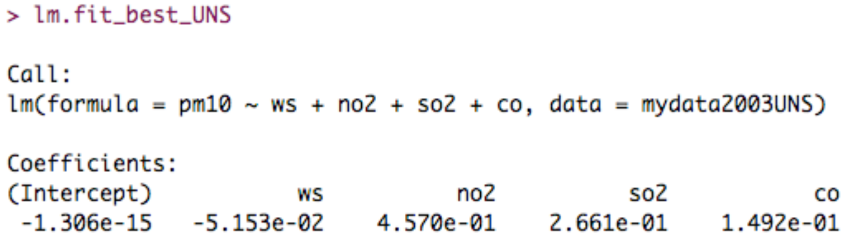
4.1 Basic Results

Here is the summary statistics of the final model we selected. All partial T tests as well as the overall F test were highly significant. Among all variables, NO2 and SO2 have the largest t values.



```
> summary(lm.fit_best)

Call:
lm(formula = pm10 ~ ws + no2 + so2 + co, data = mydata2003)

Residuals:
    Min      1Q  Median      3Q     Max 
-75.367  -6.507  -1.445   4.675 183.791 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.69102    0.27874  38.355   <2e-16 ***
ws          -0.40466    0.04308  -9.393   <2e-16 ***
no2          0.29516    0.00638  46.262   <2e-16 ***
so2          1.46686    0.03930  37.325   <2e-16 ***
co           4.17529    0.26407  15.812   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.61 on 13035 degrees of freedom
Multiple R-squared: 0.6135,     Adjusted R-squared: 0.6134 
F-statistic:  5172 on 4 and 13035 DF,  p-value: < 2.2e-16
```

Here are the model coefficients of our ``best´´ model after the unit normal scaling.



```
> lm.fit_best_UNS

Call:
lm(formula = pm10 ~ ws + no2 + so2 + co, data = mydata2003UNS)

Coefficients:
(Intercept)           ws          no2          so2           co  
 -1.306e-15   -5.153e-02    4.570e-01    2.661e-01    1.492e-01  
```

4.2 Residual Analysis

Normal probability plot (See Appendix D) and residual plots for all predictors and all interaction terms (See Appendix F, G, H) were constructed to aid the analysis of the best model. Based on the normal probability plot, the residuals deviate from the expected residuals under normality substantially, which is illustrated by large deviation from the normal line at the upper right. Outlier removal was not performed because we do not have sufficient information to explain the reason for these high residuals.

In general, the residual plots for all variables and all interaction terms have no clear patterns. Therefore, we did not consider adding interaction term to the model and transforming any variables.

However, many of them displayed substantial heteroscedasticity, especially for wind speed and CO. More concretely, when wind speed or CO concentration become larger, the variance tend to become smaller. To resolve this problem, the structure of the variance should be modeled in a more flexible way. For instance, weighted least square method can be adopted in future research.

5 Discussion

Interpretation of the relationships between PM10 and predictor variables:

1. Wind Speed

Our model suggests that larger wind speed tends to be associated with smaller OM10. Indeed, research has found a significant inverse relationship between wind speed and PM10 concentration, as wind can take away all the ambient particles in the air (Alshitawi, 2009).

2. Nitrogen Dioxide and Sulfur dioxide (NOx & SO2)

Our model suggests that NOx and SO2 are positively correlated with PM10, which can be explained by their chemical origination. Nitrogen oxides and sulfur dioxides are major by-products emitted from coal combustion from industrial facilities and power plants, and these two gases also serve as precursors for some aerosols and

PM10 (Int Panis, 2008). That is, some particulate matters are derived from the oxidation of gases like NOx and SO2. Thus, the more fossil fuel consumed by the city, the higher the concentrations of NOx, SO2 and PM10 will detected in the atmosphere.

3. Ground-level Ozone (O3)

Although O3 was not included in our best model, the full model suggests a positive correlation between O3 and PM10, which is supported by the fact that O3 and PM10 both have NOx as their common precursor.

4. Carbon monoxide (CO)

Our model suggests a positive correlation between CO and PM10. CO is formed when carbon in the fuel is not burned completely, so its level increases when conditions are poor for engine combustion. Transportation is the main source of particulate matters in cities, especially along busy roads. Hence the concentrations of both PM10 and CO will increase as the volume of traffic increases

6 Conclusion

From the seven initial variables, we determined five predictors for modeling PM10 concentration, including wind speed, NO2, SO2, and CO, which are also justified by theories and previous findings.

There are several limitations of the present model. In future research, heteroscedasticity should be accounted by changing the structure of the variance. Moreover, the autocorrelation of the data should be considered by using time series models, since there is reason to believe that PM concentration has time dependencies. We did not remove any outlier in the present study, but this should be considered when more information is available so that these observations can be explained.

References

Alshitawi, Mohammed, Hazim Awbi, and Norhayati Mahyuddin. (2009) The Effect of Outdoor Conditions and Air Change Rate on Particulate Matter (PM10) Concentration in a Classroom. *The Twelfth International Conference on Air Distribution in Rooms.*

Int Panis, L.L.R. (2008). The Effect of Changing Background Emissions on External Cost Estimates for Secondary Particulates. *Open Environmental Sciences, 2*, 47–53.

Kan, H., Chen, B., & Hong, C. (2009). Health impact of outdoor air pollution in China: Current knowledge and future research needs. *Environmental Health Perspectives, 117*(5), 187. doi:10.1289/ehp.12737

Nevada Division of Environmental Protection (n.d.). Particulate Matter Pollution Fact Sheet. Retrieved from: https://ndep.nv.gov/baqp/monitoring/docs/particulate_matter.pdf

United States Environmental Protection Agency (2013) *Particulate Matter*. Retrieved from: http://www.epa.gov/pm/

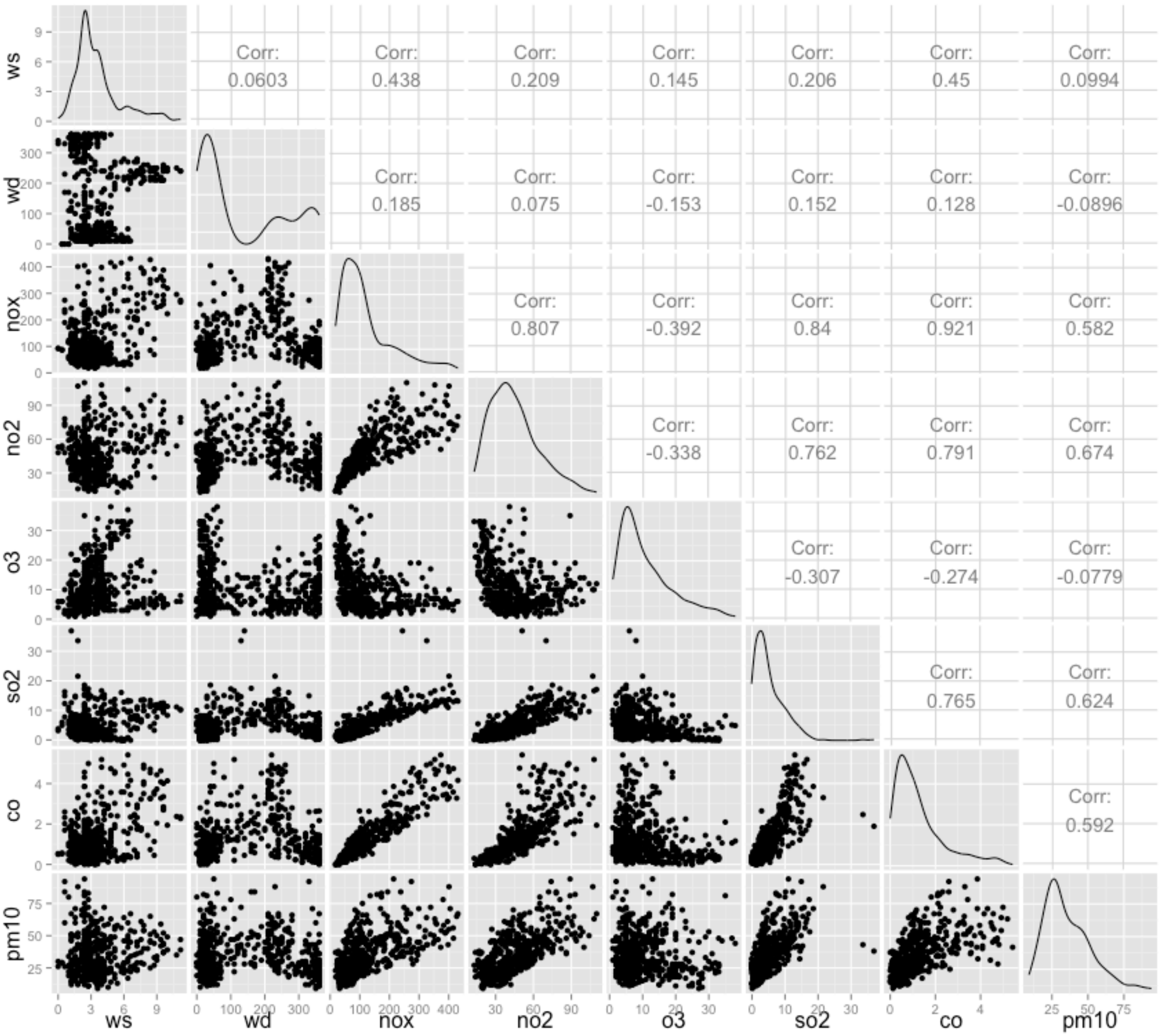
World Health Organization (2014). *Ambient (outdoor) air quality and health*. Retrieved from: http://www.who.int/mediacentre/factsheets/fs313/en/

Smyth, S., Jiang, W., Yin, D., Roth, H., &Giroux, E. (2006) Evaluation of CMAQ O3 and PM2.5 performance using Pacific 2001 measurement data, *Atmospheric Environment, 40*(15), 2735-2749

Appendix A

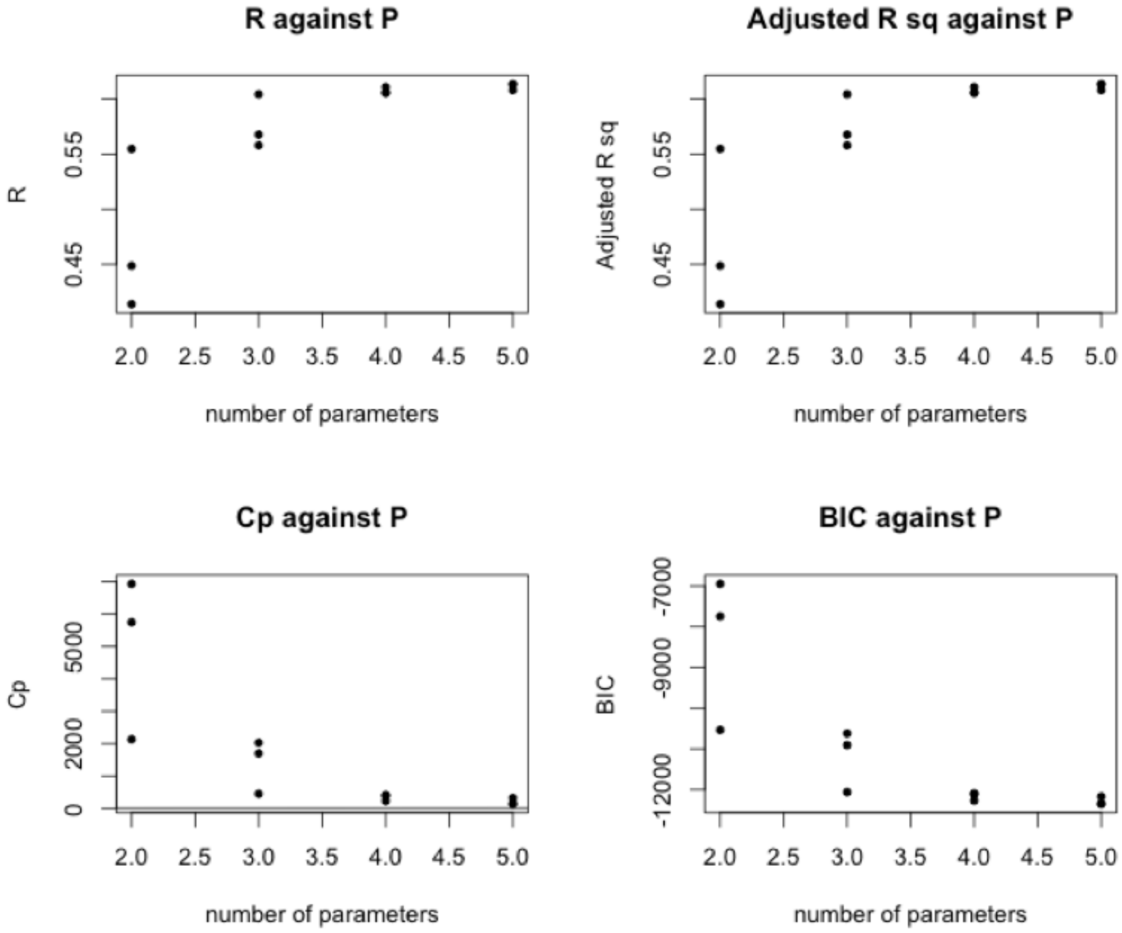
The scatter plots, correlation matrices and frequency plots.

(Only the first 500 observations were visualized)



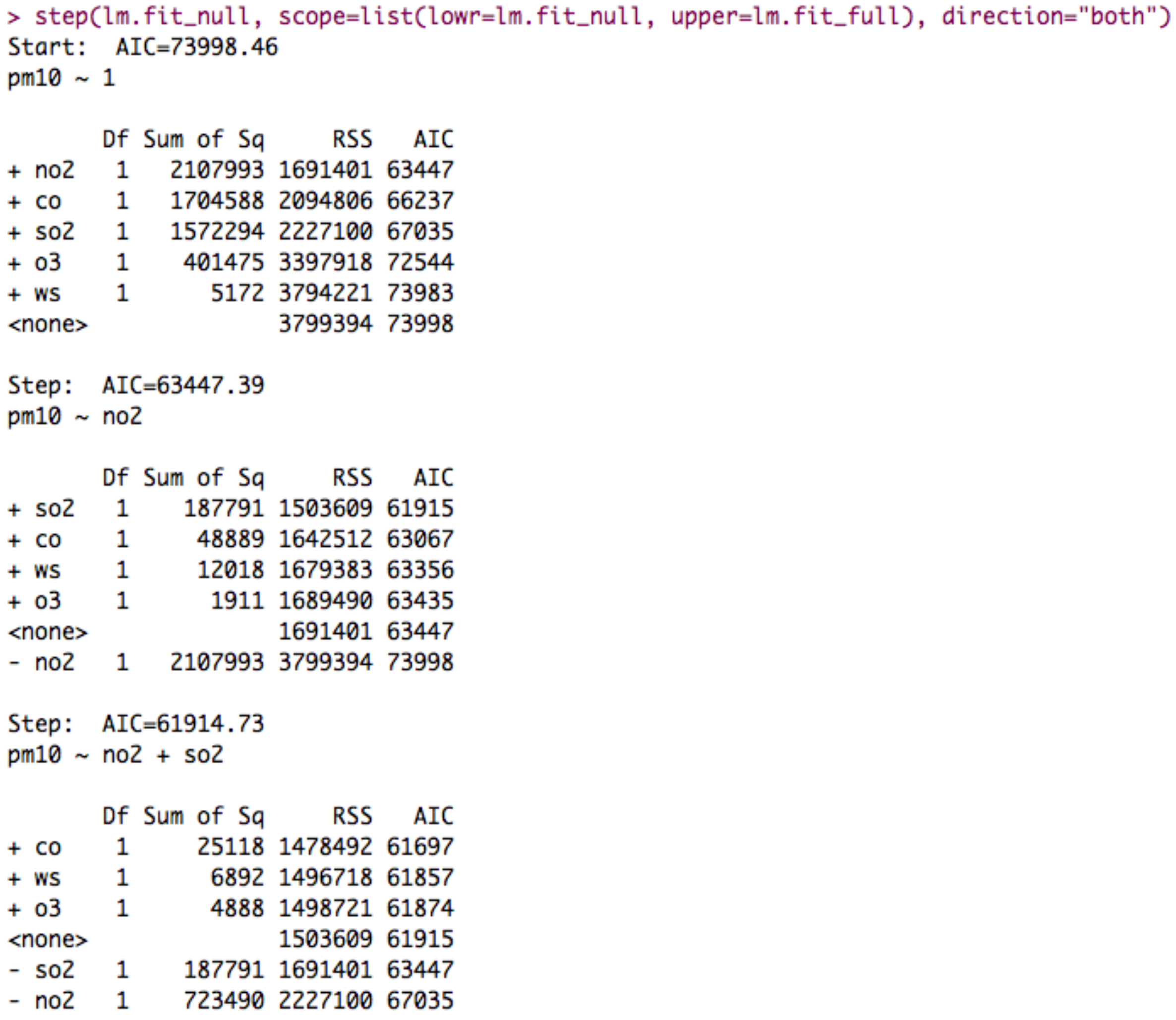
Appendix B

The computed variable selection criteria on the best model.

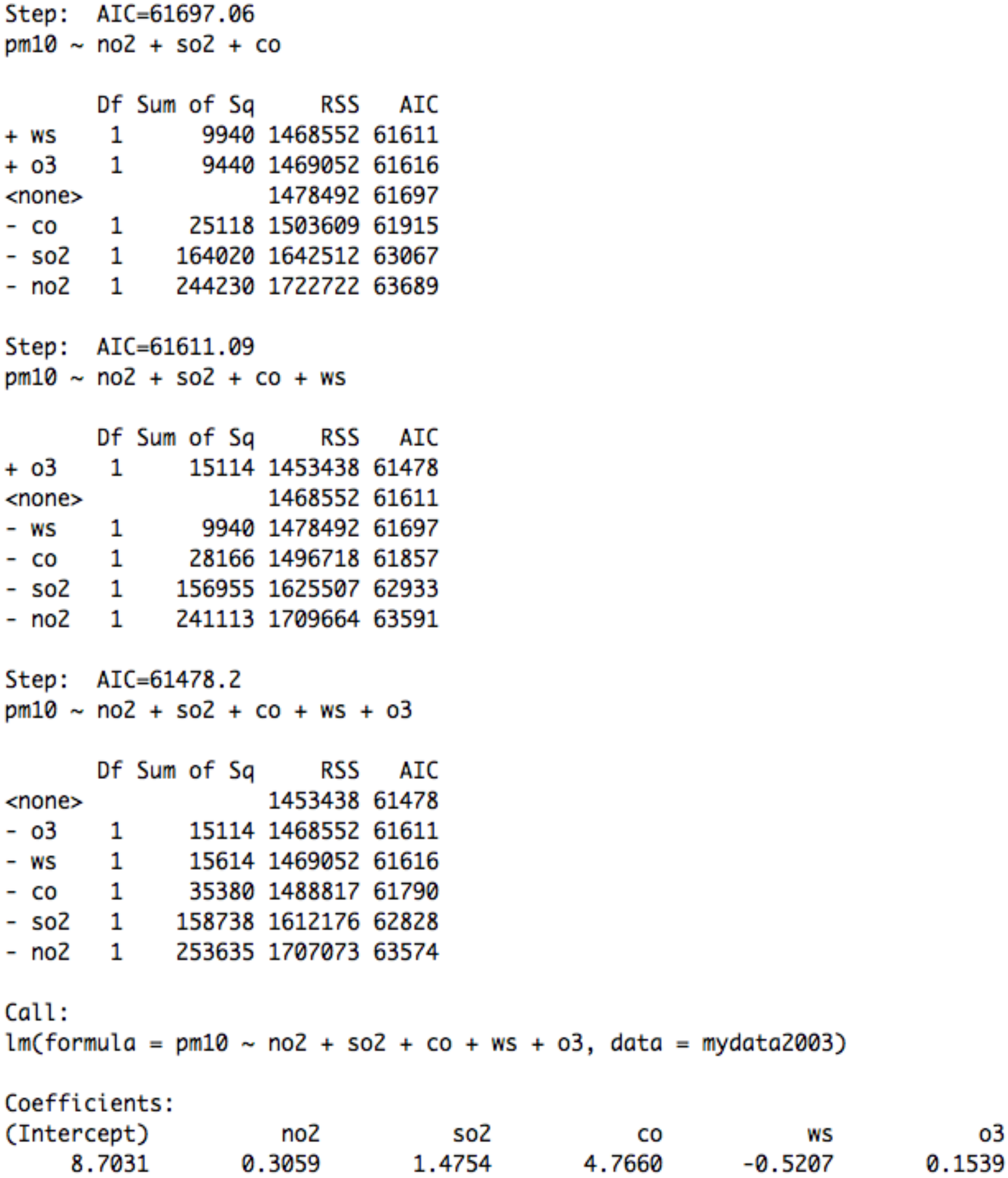


Appendix C

The output for stepwise regression procedure



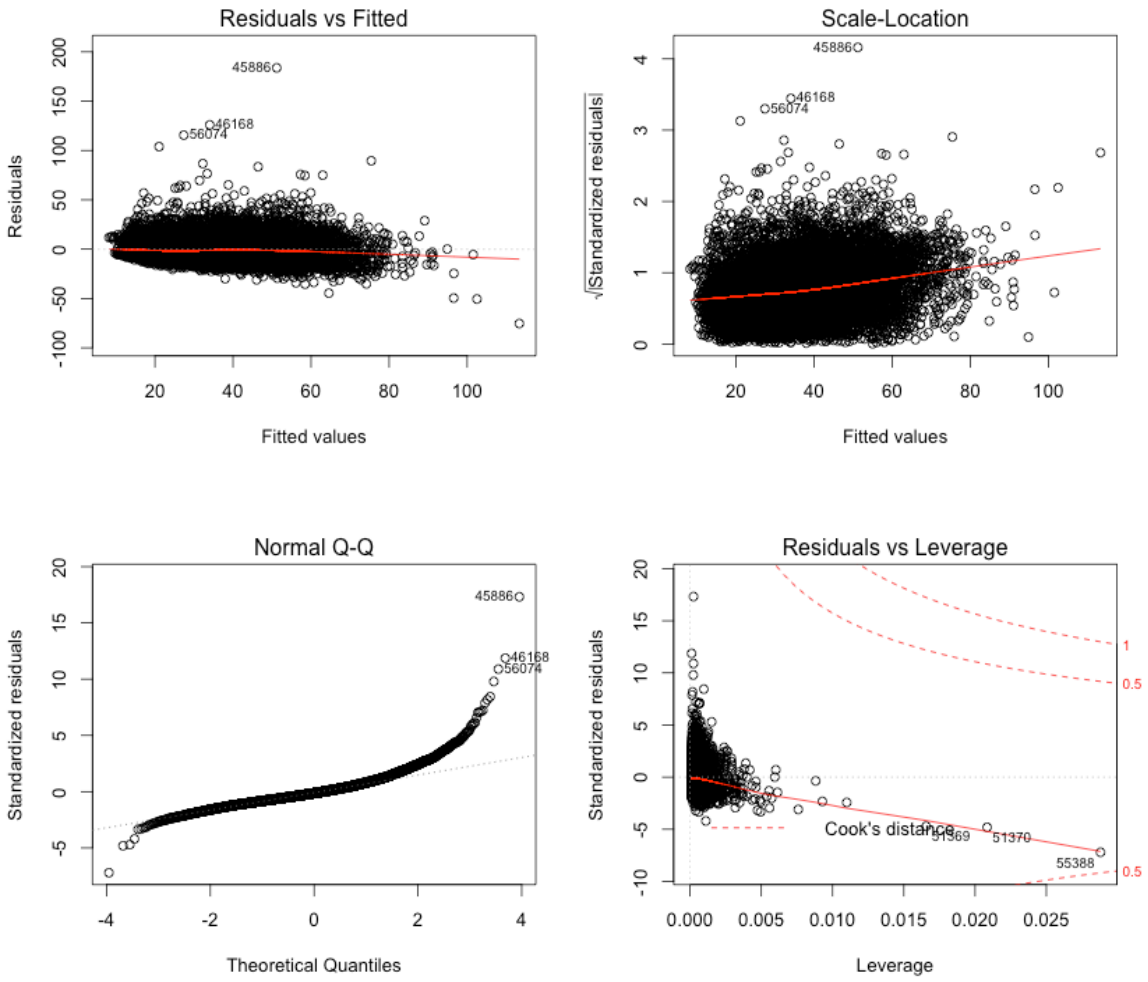
```
> step(lm.fit_null, scope=list(lowr=lm.fit_null, upper=lm.fit_full), direction="both")
Start:  AIC=73998.46
pm10 ~ 1

       Df Sum of Sq     RSS   AIC
+ no2   1   2107993 1691401 63447
+ co    1   1704588 2094806 66237
+ so2   1   1572294 2227100 67035
+ o3    1    401475 3397918 72544
+ ws    1      5172 3794221 73983
<none>              3799394 73998

Step:  AIC=63447.39
pm10 ~ no2

       Df Sum of Sq     RSS   AIC
+ so2   1    187791 1503609 61915
+ co    1     48889 1642512 63067
+ ws    1     12018 1679383 63356
+ o3    1      1911 1689490 63435
<none>              1691401 63447
- no2   1   2107993 3799394 73998

Step:  AIC=61914.73
pm10 ~ no2 + so2

       Df Sum of Sq     RSS   AIC
+ co    1     25118 1478492 61697
+ ws    1      6892 1496718 61857
+ o3    1      4888 1498721 61874
<none>              1503609 61915
- so2   1    187791 1691401 63447
- no2   1    723490 2227100 67035
```



```
Step:  AIC=61697.06
pm10 ~ no2 + so2 + co

       Df Sum of Sq     RSS   AIC
+ ws    1      9940 1468552 61611
+ o3    1      9440 1469052 61616
<none>              1478492 61697
- co    1     25118 1503609 61915
- so2   1    164020 1642512 63067
- no2   1    244230 1722722 63689

Step:  AIC=61611.09
pm10 ~ no2 + so2 + co + ws

       Df Sum of Sq     RSS   AIC
+ o3    1     15114 1453438 61478
<none>              1468552 61611
- ws    1      9940 1478492 61697
- co    1     28166 1496718 61857
- so2   1    156955 1625507 62933
- no2   1    241113 1709664 63591

Step:  AIC=61478.2
pm10 ~ no2 + so2 + co + ws + o3

       Df Sum of Sq     RSS   AIC
<none>              1453438 61478
- o3    1     15114 1468552 61611
- ws    1     15614 1469052 61616
- co    1     35380 1488817 61790
- so2   1    158738 1612176 62828
- no2   1    253635 1707073 63574

Call:
lm(formula = pm10 ~ no2 + so2 + co + ws + o3, data = mydata2003)

Coefficients:
(Intercept)          no2          so2           co           ws           o3
     8.7031       0.3059       1.4754       4.7660      -0.5207       0.1539
```

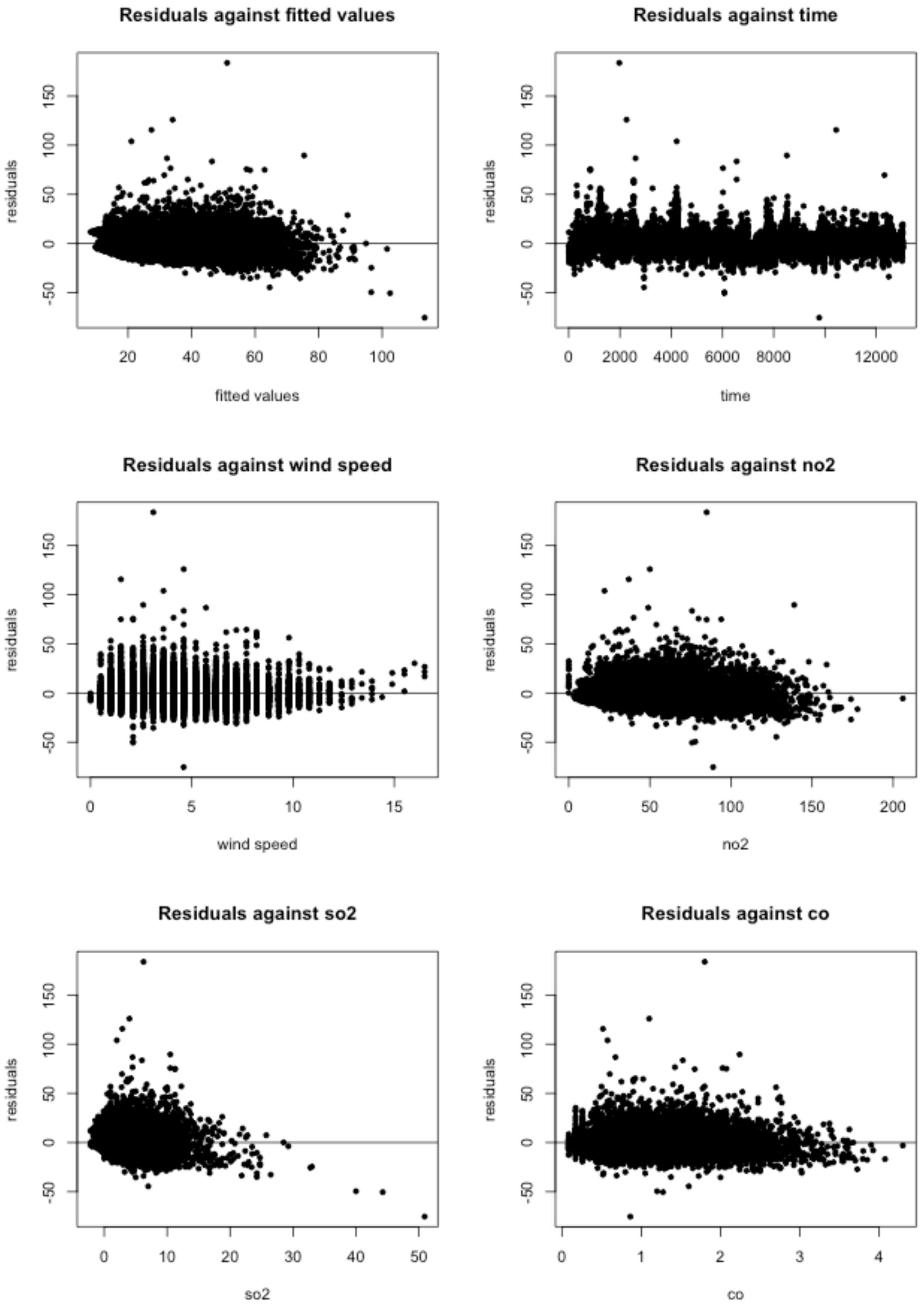
Appendix D

Some summary plots for the best model



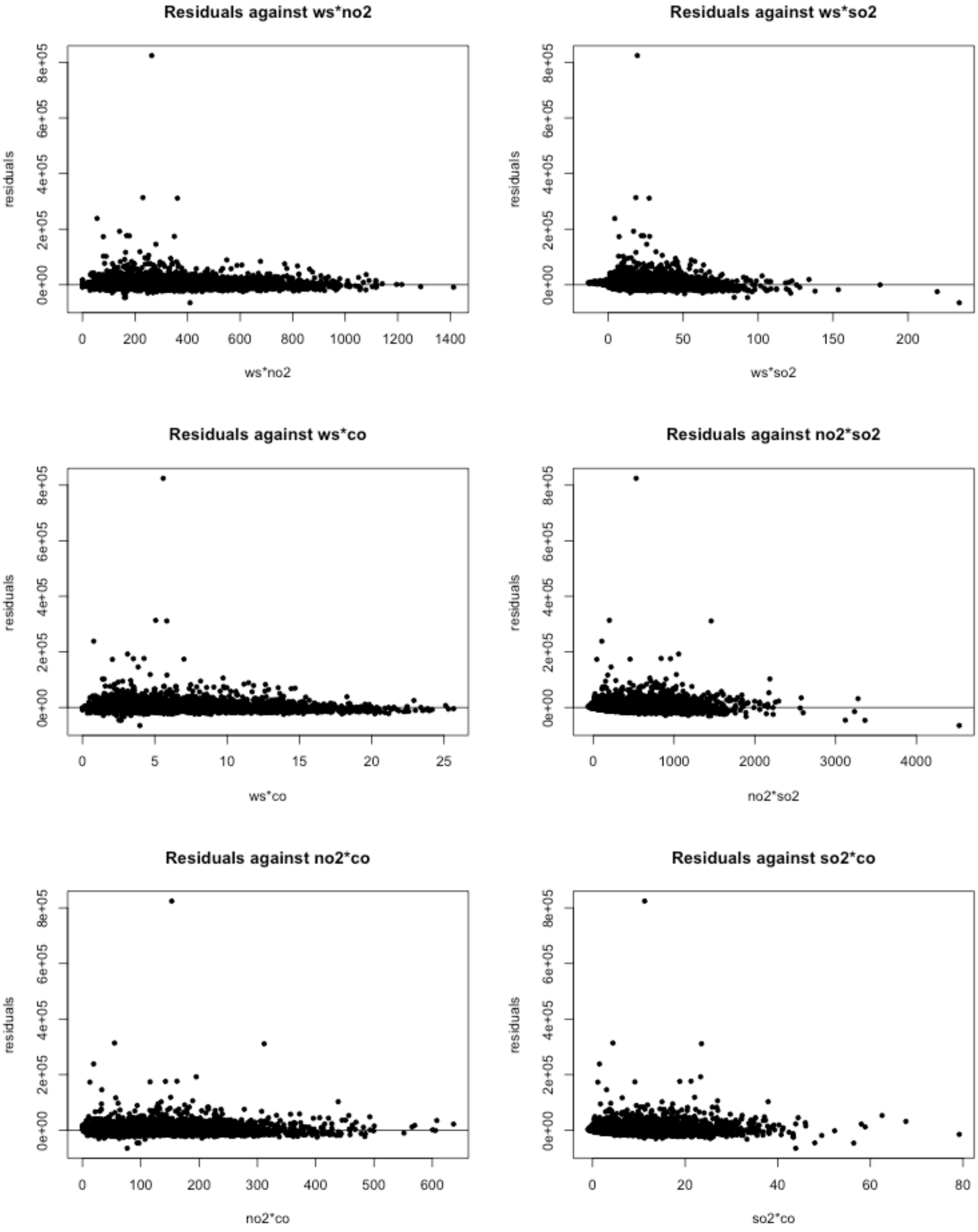
Appendix E

The residuals plots for all predictors for the best model



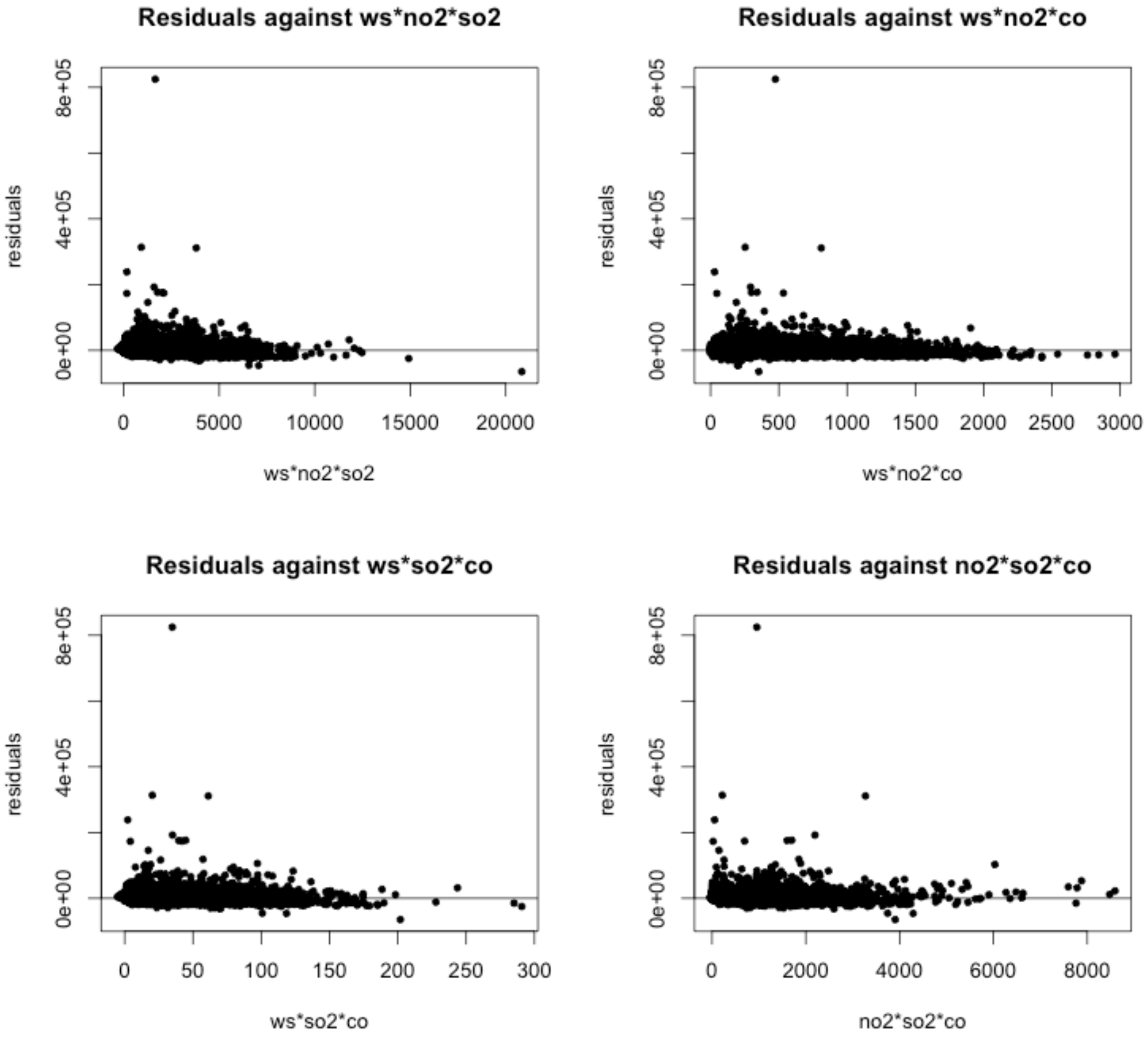
Appendix F

The residuals plots for all two-way interaction terms for the best model



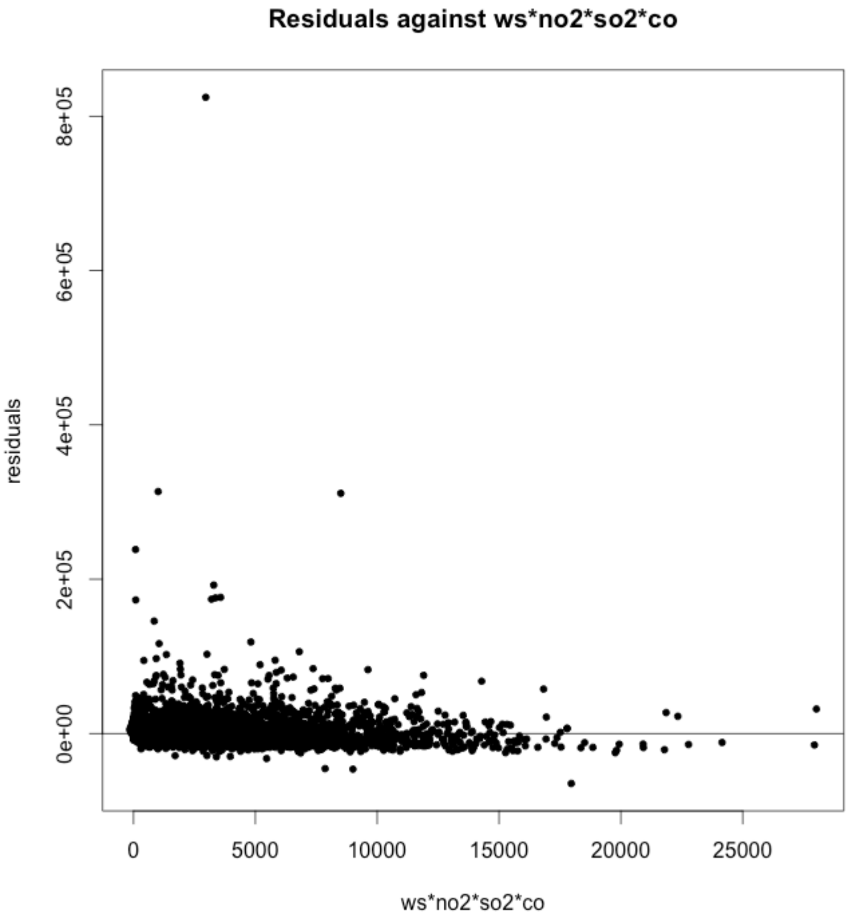
Appendix G

The residuals plots for all three-way interaction terms for the best model



Appendix H

The residuals plots the four-way interaction terms (all predictors) for the best model



Appendix I

Miscellaneous Information

Data Source

The data was retrieved from the Open Air Project at the following link.

http://www.openair-project.org/Downloads/ExampleDataSets.aspx

Code

All code that associated with this project can be found from this link:

https://github.com/QihongL/STAT333_OpenAirProject

Other Documentations

This is the link to the Google drive we created for this project. It has contains various information about the data, references. It also contains the proposal and the written report.

https://drive.google.com/open?id=0B28y5jiaX0HbfkdFcE1NWlI3ZDZNWHJ1RWlDeXM0QV9la0IxYWdONnR6OVNTRS1QN0hnMkU&authuser=0