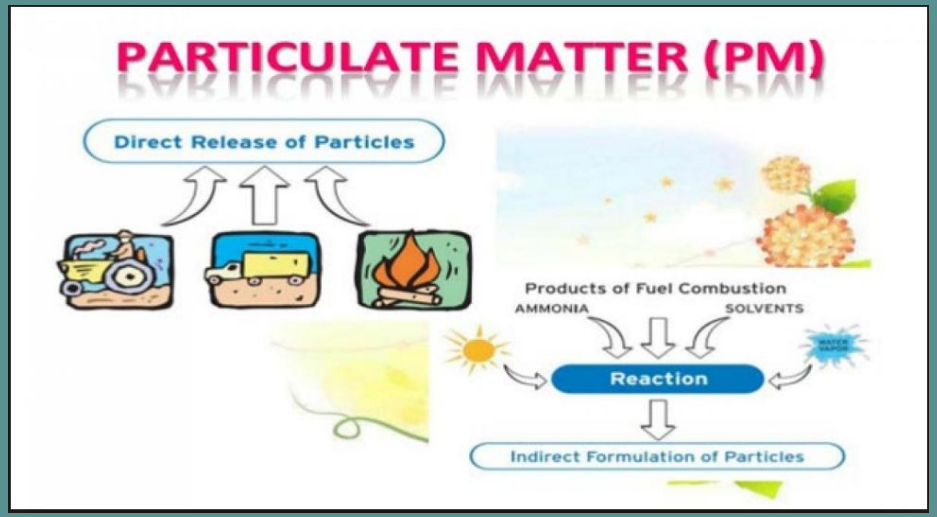
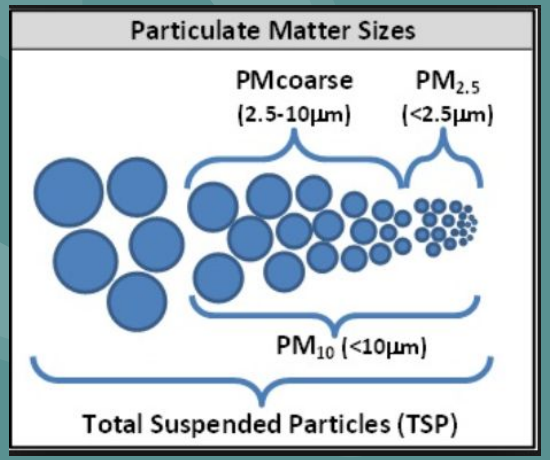
Use of Regression Analysis to Predict PM2.5 Levels in Beijing using Weather and Time Attributes

Dahee Kim, Aditya Desai, Wendy Wang &
John Riddle

What is PM2.5?

- Particulate Matter (PM)
  - Solids / Liquid Droplets
  - Diameter less than or equal to 2.5 micrometers
  - Ash, Soot, Sea Spray, Industrial Emissions, Vehicle Emissions, etc

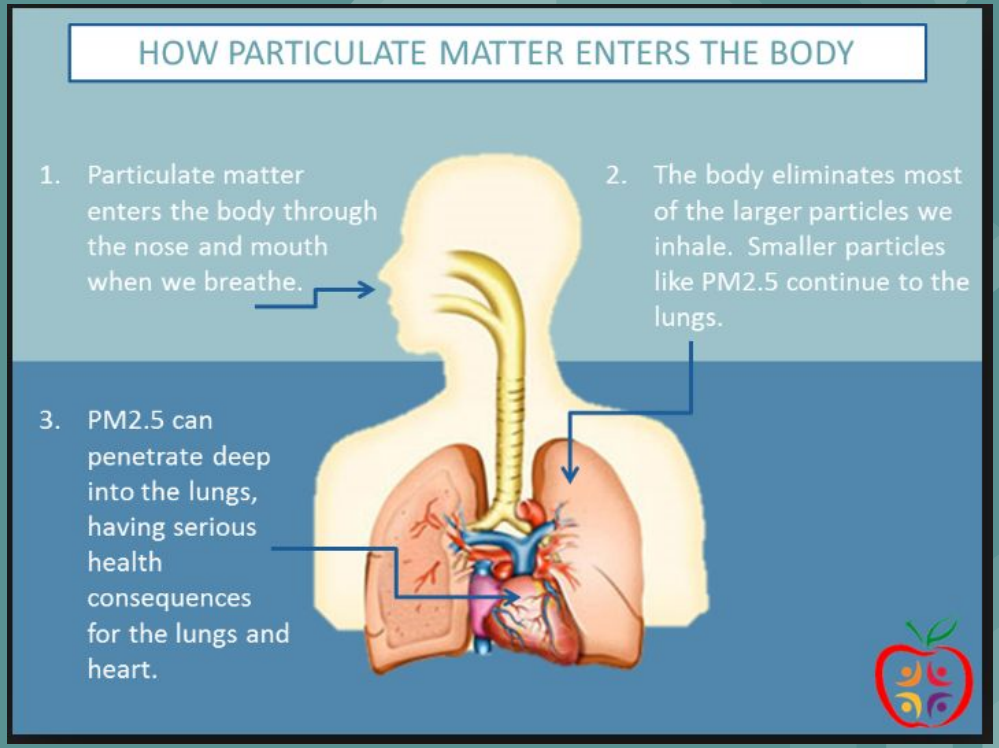




Why is it important?

- Health Problems
  - Cardiovascular
  - Age specific mortality risk
  - Respiratory

- Causes
  - Human industrial activity
  - Emissions
  - Weather?



Data Description

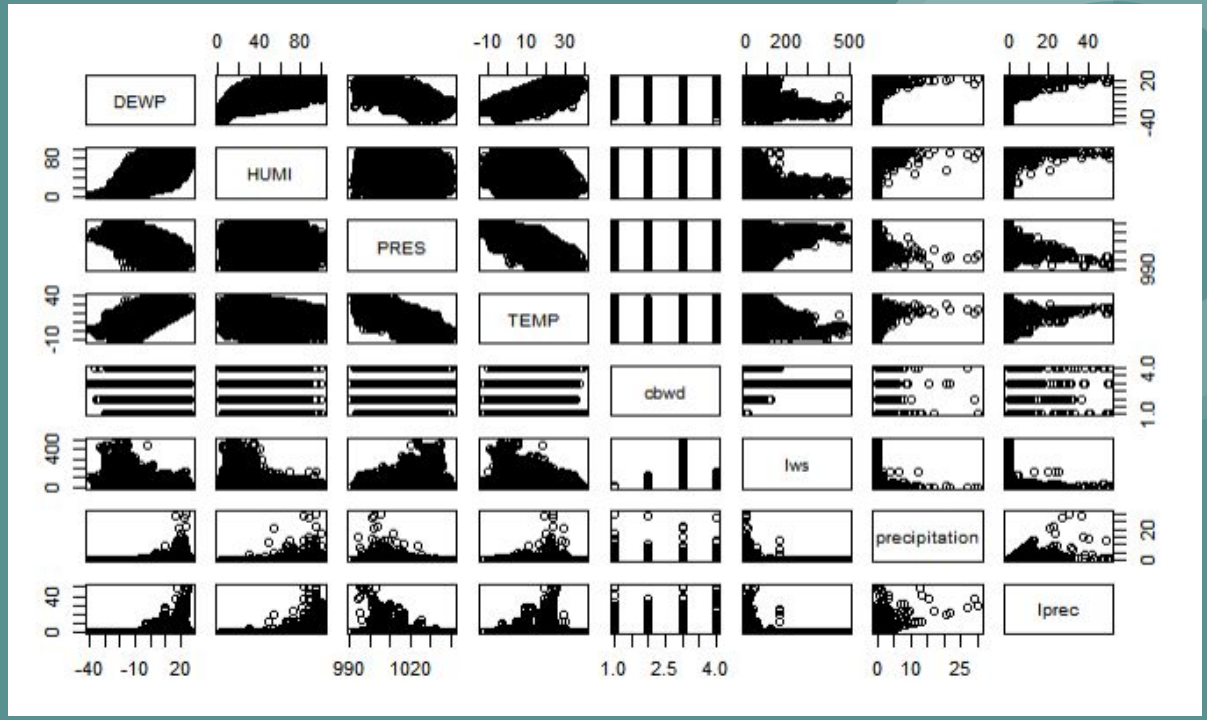
- Beijing, China
  - N = 5000
- Response Variable: PM2.5 (US Post)
- Predictors:
  - Date (Year, Month, Day, Hour)
  - Season
  - Dew Point
  - Temp
  - Humidity
  - Pressure
- Can we use real-time weather data to predict PM2.5 exposure risk?

Method of Analysis

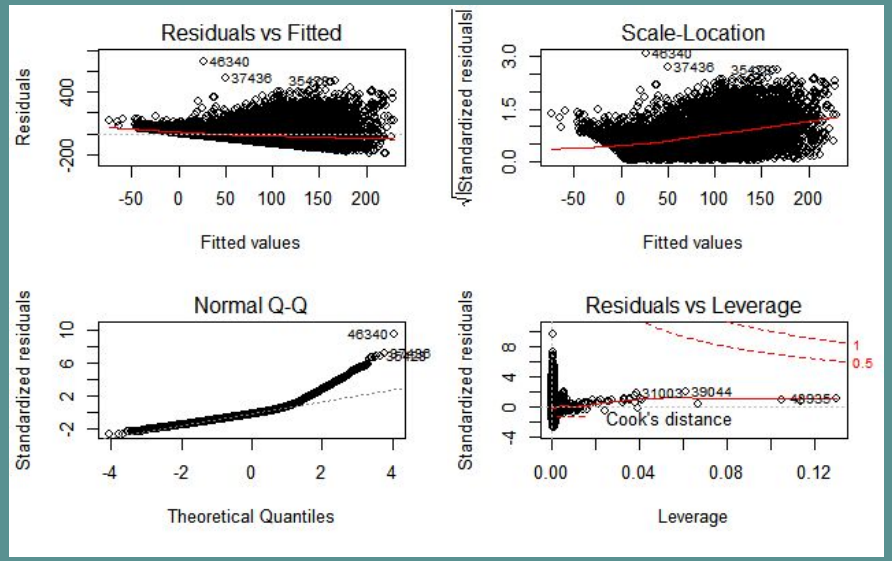
- Build Model (Multi-Linear Regression)
  - Scatter Plots
  - Diagnostic Plots
  - ANOVA
- Analysis

```
Call:
lm(formula = beijing$PM_US.Post ~ beijing$DEWP + beijing$HUMI +
    beijing$PRES + beijing$TEMP + beijing$cbwd + beijing$Iws +
    beijing$precipitation + beijing$Iprec, data = beijing)

Residuals:
    Min      1Q  Median      3Q     Max
-190.15  -44.48  -12.07   26.10  695.83

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          949.68259  100.16638   9.481  < 2e-16 ***
beijing$DEWP          -1.00593    0.22909  -4.391 1.13e-05 ***
beijing$HUMI           1.48678    0.07799  19.064  < 2e-16 ***
beijing$PRES          -0.87427    0.09709  -9.005  < 2e-16 ***
beijing$TEMP          -2.12359    0.22379  -9.489  < 2e-16 ***
beijing$cbwdNE       -29.55163    1.89969 -15.556  < 2e-16 ***
beijing$cbwdNW       -37.99321    1.58671 -23.945  < 2e-16 ***
beijing$cbwdSE         9.43092    1.45940   6.462 1.06e-10 ***
beijing$Iws           -0.18884    0.01269 -14.882  < 2e-16 ***
beijing$precipitation -1.24591    0.98251  -1.268    0.205
beijing$Iprec         -3.19450    0.29246 -10.923  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72.84 on 19051 degrees of freedom
Multiple R-squared:  0.2856,    Adjusted R-squared:  0.2852
F-statistic: 761.4 on 10 and 19051 DF,  p-value: < 2.2e-16
```

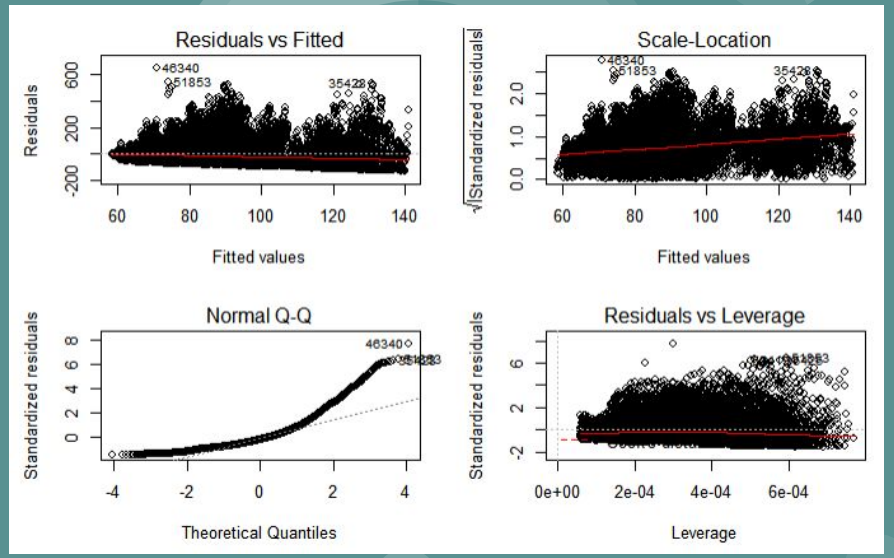
Scatter Plot Matrix of Predictors



Diagnostic Plot



Diagnostics of Weather Descriptors



Diagnostics of Time Attributes

ANOVA

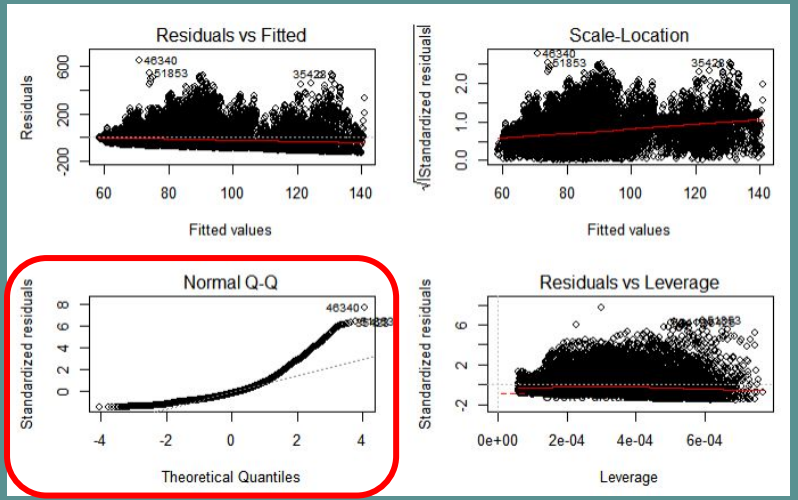
```
Analysis of Variance Table

Response: beijing$PM_US.Post
                          Df    Sum Sq  Mean Sq   F value    Pr(>F)
beijing$DEWP               1    663018   663018  124.9689 < 2.2e-16 ***
beijing$HUMI               1  27702569 27702569 5221.5123 < 2.2e-16 ***
beijing$PRES               1     22896    22896    4.3156   0.03778 *
beijing$TEMP               1   1002122  1002122  188.8847 < 2.2e-16 ***
beijing$cbwd               3   8797111  2932370  552.7071 < 2.2e-16 ***
beijing$Iws                1   1224860  1224860  230.8674 < 2.2e-16 ***
beijing$precipitation      1    352680   352680   66.4748 3.761e-16 ***
beijing$Iprec              1    632996   632996  119.3102 < 2.2e-16 ***
Residuals              19051 101074481     5305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Method of Analysis (cont)

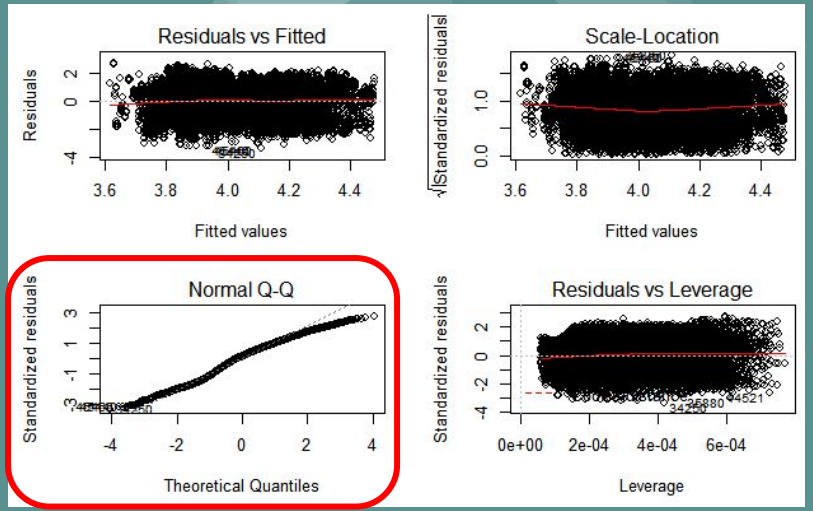
- Log transformation on Y
- Stepwise Regression
- Anova
- Confidence Interval and Prediction

Log transformation on Y

Example) Time attributes



Before log transformation on PM_US.Post



After log transformation on PM_US.Post

Stepwise Regression

```
Start:  AIC=-8940.68
log(beijing$PM_US.Post) ~ beijing$DEWP + beijing$HUMI + beijing$PRES +
    beijing$TEMP + beijing$cbwd + beijing$Iws + beijing$precipitation +
    beijing$Iprec

                         Df Sum of Sq   RSS     AIC
- beijing$precipitation  1      0.24 11912 -8942.3
<none>                              11912 -8940.7
- beijing$DEWP           1     21.31 11933 -8908.6
- beijing$Iprec          1     69.77 11981 -8831.3
- beijing$HUMI           1    104.67 12016 -8775.9
- beijing$PRES           1    112.23 12024 -8763.9
- beijing$TEMP           1    212.15 12124 -8606.2
- beijing$Iws            1    446.91 12358 -8240.6
- beijing$cbwd           3   2386.70 14298 -5465.4

Step:  AIC=-8942.29
log(beijing$PM_US.Post) ~ beijing$DEWP + beijing$HUMI + beijing$PRES +
    beijing$TEMP + beijing$cbwd + beijing$Iws + beijing$Iprec

                         Df Sum of Sq   RSS     AIC
<none>                              11912 -8942.3
+ beijing$precipitation  1      0.24 11912 -8940.7
- beijing$DEWP           1     21.32 11933 -8910.2
- beijing$Iprec          1    101.74 12014 -8782.2
- beijing$HUMI           1    104.57 12016 -8777.7
- beijing$PRES           1    112.33 12024 -8765.4
- beijing$TEMP           1    212.31 12124 -8607.5
- beijing$Iws            1    447.01 12359 -8242.1
- beijing$cbwd           3   2387.84 14300 -5465.6
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
log(beijing$PM_US.Post) ~ beijing$DEWP + beijing$HUMI + beijing$PRES +
    beijing$TEMP + beijing$cbwd + beijing$Iws + beijing$precipitation +
    beijing$Iprec

Final Model:
log(beijing$PM_US.Post) ~ beijing$DEWP + beijing$HUMI + beijing$PRES +
    beijing$TEMP + beijing$cbwd + beijing$Iws + beijing$Iprec


                        Step Df  Deviance Resid. Df Resid. Dev      AIC
1                                               19051   11911.58 -8940.677
2 - beijing$precipitation  1 0.2423645       19052   11911.82 -8942.289
```

Climate attributes after the transformation

```
Start:  AIC=930.18
log(beijing$PM_US.Post) ~ beijing$year + beijing$month + beijing$day +
    beijing$hour + beijing$season

                 Df Sum of Sq   RSS     AIC
- beijing$hour    1      0.235 20003  928.40
<none>                        20003  930.18
- beijing$season  1     50.929 20054  976.65
- beijing$day     1    108.011 20111 1030.83
- beijing$year    1    152.867 20156 1073.30
- beijing$month   1    293.892 20297 1206.21

Step:  AIC=928.4
log(beijing$PM_US.Post) ~ beijing$year + beijing$month + beijing$day +
    beijing$season

                 Df Sum of Sq   RSS     AIC
<none>                        20003  928.40
+ beijing$hour    1      0.235 20003  930.18
- beijing$season  1     50.954 20054  974.90
- beijing$day     1    108.058 20111 1029.10
- beijing$year    1    152.866 20156 1071.52
- beijing$month   1    293.922 20297 1204.46
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
log(beijing$PM_US.Post) ~ beijing$year + beijing$month + beijing$day +
    beijing$hour + beijing$season

Final Model:
log(beijing$PM_US.Post) ~ beijing$year + beijing$month + beijing$day +
    beijing$season


           Step Df  Deviance Resid. Df Resid. Dev      AIC
1                                19056   20002.65 930.1765
2 - beijing$hour  1 0.2347251     19057   20002.88 928.4002
```

Time attributes after the transformation

ANOVA

```
Analysis of Variance Table

Response: log(beijing$PM_US.Post)
                       Df  Sum Sq Mean Sq   F value    Pr(>F)
beijing$DEWP            1  1146.7  1146.7 1834.0354 < 2.2e-16 ***
beijing$HUMI            1  3533.3  3533.3 5651.0481 < 2.2e-16 ***
beijing$PRES            1     1.3     1.3    2.0108    0.1562
beijing$TEMP            1   386.0   386.0  617.3682 < 2.2e-16 ***
beijing$cbwd            3  2917.3   972.4 1555.2721 < 2.2e-16 ***
beijing$Iws             1   456.8   456.8  730.6295 < 2.2e-16 ***
beijing$precipitation   1    32.2    32.2   51.5189 7.352e-13 ***
beijing$Iprec           1    69.8    69.8  111.5908 < 2.2e-16 ***
Residuals           19051 11911.6     0.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = log(beijing$PM_US.Post) ~ beijing$DEWP + beijing$HUMI +
    beijing$PRES + beijing$TEMP + beijing$cbwd + beijing$Iws +
    beijing$precipitation + beijing$Iprec, data = beijing)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5890 -0.5149  0.0314  0.5423  3.4124

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            18.6409672  1.0873915  17.143  < 2e-16 ***
beijing$DEWP            0.0145197  0.0024869   5.838 5.36e-09 ***
beijing$HUMI            0.0109540  0.0008466  12.938  < 2e-16 ***
beijing$PRES           -0.0141211  0.0010540 -13.398  < 2e-16 ***
beijing$TEMP           -0.0447508  0.0024294 -18.420  < 2e-16 ***
beijing$cbwdNE         -0.4708214  0.0206227 -22.830  < 2e-16 ***
beijing$cbwdNW         -0.6257801  0.0172251 -36.330  < 2e-16 ***
beijing$cbwdSE          0.2635080  0.0158430  16.632  < 2e-16 ***
beijing$Iws            -0.0036830  0.0001378 -26.735  < 2e-16 ***
beijing$precipitation  -0.0066407  0.0106660  -0.623    0.534
beijing$Iprec          -0.0333384  0.0031749 -10.504  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7907 on 19051 degrees of freedom
Multiple R-squared:  0.4177,    Adjusted R-squared:  0.4174
F-statistic:  1366 on 10 and 19051 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: log(beijing$PM_US.Post)
                Df  Sum Sq Mean Sq F value    Pr(>F)
beijing$year     1    48.7  48.735  46.429 9.786e-12 ***
beijing$month    1   245.7 245.730 234.100 < 2.2e-16 ***
beijing$day      1   106.7 106.668 101.620 < 2.2e-16 ***
beijing$hour     1     0.3   0.259   0.247    0.6192
beijing$season   1    50.9  50.929  48.519 3.378e-12 ***
Residuals    19056 20002.6   1.050
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = log(beijing$PM_US.Post) ~ beijing$year + beijing$month +
    beijing$day + beijing$hour + beijing$season, data = beijing)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4174 -0.7047  0.1436  0.7570  2.8032

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.564e+02  2.092e+01  12.260  < 2e-16 ***
beijing$year   -1.253e-01  1.038e-02 -12.068  < 2e-16 ***
beijing$month  -4.166e-02  2.489e-03 -16.733  < 2e-16 ***
beijing$day     8.631e-03  8.509e-04  10.144  < 2e-16 ***
beijing$hour   -5.045e-04  1.067e-03  -0.473    0.636
beijing$season  4.854e-02  6.968e-03   6.966 3.38e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.025 on 19056 degrees of freedom
Multiple R-squared:  0.02211,    Adjusted R-squared:  0.02186
F-statistic: 86.18 on 5 and 19056 DF,  p-value: < 2.2e-16
```

Confidence Interval and Prediction

Examples)

```
        Welch Two Sample t-test

data:  beijing$TEMP and log(beijing$PM_US.Post)
t = 119.56, df = 19369, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.858034 10.186659
sample estimates:
mean of x mean of y
14.062900  4.040553
```

T-test of TEMP
(Climate attribute)

```
        Welch Two Sample t-test

data:  beijing$month and log(beijing$PM_US.Post)
t = 93.851, df = 22745, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.312594 2.411251
sample estimates:
mean of x mean of y
 6.402476  4.040553
```

T-test of month
(Time attribute)

Results

Parameter Estimates Table for Time Attributes

	Point estimate	Standard error	t-statistic	p-value	Confidence Intervals (95%)
year	2014	0.005516841	215814	<2.2e-16	[2009.866, 2009.903]
month	4	0.02402212	93.851	<2.2e-16	[2.312594, 2.411251]
day	6	0.06318058	178.85	<2.2e-16	[11.25447, 11.50389]
hour	10	0.05038173	145.97	<2.2e-16	[7.335631, 7.535314]
season	1	0.008131457	-152.99	<2.2e-16	[-1.714375, -1.671002]

Results

Parameter Estimates Table for Climate Attributes

	Point estimate	Standard error	t-statistic	p-value	Confidence intervals (95%)
DEWP	-10	0.1033333	-16.202	<2.2e-16	[-1.881705, -1.475554]
HUMI	20	0.184845	255.8	<2.2e-16	[46.95918, 47.68440]
PRES	1014	0.07398337	1472.5	<2.2e-16	[1011.166, 1011.457]
TEMP	10	0.08349293	119.56	<2.2e-16	[9.858034, 10.186659]
Iws	7.00	0.3360042	50.521	<2.2e-16	[16.32066, 17.63819]
precipitation	0.2	0.004592824	-453.43	<2.2e-16	[-4.006154, -3.971668]
Iprec	3.0	0.01556899	-221.31	<2.2e-16	[-3.858664, -3.790914]

Results

The final regression models are:

Predicted PM2.5 Level from climate = 18.641 + 0.0145(DewP) + 0.011(Humi) - 0.0141(Pres) - 0.0448(Temp) - 0.4708(cbwdNE) - 0.6258(cbwdNW) + 0.264(cbwdSE) - 0.0037(Iws) - 0.0335(Iprec)

Note: Precipitation predictor ignored as per AIC stepwise results

Predicted PM2.5 Level from time = $2.564*10^2$ - 0.1253(Year) - 0.04166(month) + 0.00863(day) + 0.04854(season)

Note: Hour predictor ignored as per AIC stepwise results

Discussion

Takeaways

- Usable model for predicting PM2.5 levels based on weather and time attributes
- PM2.5 is the result of a combination of weather attributes, not a single parameter
- Precipitation does not play significant role in PM2.5 level
- Model can be used in Beijing to make PM2.5 level predictions
    - Inform health recommendations and industry practices

Limitations

- Model can only be used for Beijing PM2.5 predictions
- Application in other locations may not be as accurate or reliable
- If unrealistic values are entered into the model i.e. humidity above 100%, model will provide incorrect PM2.5 prediction

References

- "Public Health: Sources and Effects of PM2.5," *LAQM support*. [Online]. Available: https://laqm.defra.gov.uk/public-health/pm25.html. [Accessed: 22-Jul-2019].

- S. X. Chen, "PM2.5 Data of Five Chinese Cities Data Set," *UCI Machine Learning Repository: PM2.5 Data of Five Chinese Cities Data Set*, 18-Jul-2017. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/PM2.5 Data of Five Chinese Cities. [Accessed: 10-Jun-2019].