PM2.5 Air Pollution in Beijing

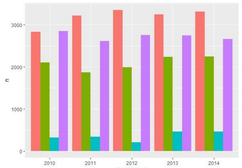
Department of Mathematics and Statistics
Nan Mi

what is PM2.5?

PM2.5 refers to dangerous atmospheric particulate mater, that have a diameter less than 2.5 micrometers.

Natural sources: soil dust, sea salt, plant pollen, spores, bacteria, volcanic eruptions, forest fires, etc.

Man-made source: power plant, metallurgy and petroleum industrial processes, motor vehicles, coal burning, wood burning, etc.



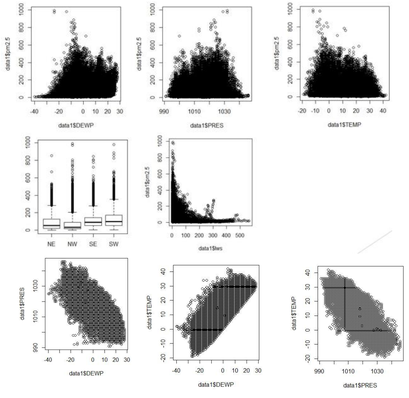
Good pm2.5<=50

moderate 50<PM2.5<=100

unhealthy 100<PM2.5<=300

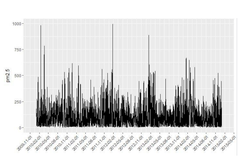
poisonous 300<PM2.5

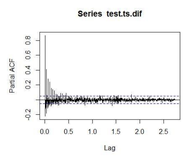
From the figures, we can see that the sorts of air quality doesn't change too much during 2010 to 2014. About 1/3 of a year, the air quality is unhealthy.

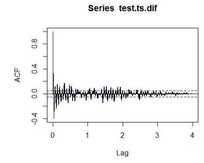


From the plot, we can see that there is somewhat relationship about the factors and pm2.5 values, and interaction terms shows some relationship.

Time series analysis








```
Augmented Dickey-Fuller Test

data:  pm
Dickey-Fuller = -23.463, Lag order = 34,
p-value = 0.01
alternative hypothesis: stationary
```

From the figures of AIC and PAIC, we can see the data are stationary. The Dickey-Fuller test also shows that the data are stationary. Therefore, time series analysis can be applied.

ARIMA (p, d, q) (Autoregressive Integrated Moving Average Model)

- **p** is the number of autoregressive terms,
- **d** is the number of non-seasonal differences needed for stationarity
- **q** is the number of lagged forecast errors in the prediction equation

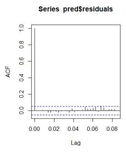
```
Series: test.ts
ARIMA(1,0,4) with non-zero mean

Coefficients:
          ar1     ma1      ma2      ma3     ma4      mean
      0.9212  0.1895  -0.1392  -0.0343  0.0429  94.0197
s.e.  0.0145  0.0298   0.0317   0.0315  0.0286  11.1435

sigma^2 estimated as 1033:  log likelihood=-7194.96
AIC=14403.91   AICc=14403.99   BIC=14440.98
```

$$y_t = 94.0197 + 0.9212y_{t-1} + 0.1892\varepsilon_{t-1} - 0.1392\varepsilon_{t-2} - 0.0343\varepsilon_{t-3} + 0.0429\varepsilon_{t-4} + \varepsilon_t$$

$$\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \varepsilon_{t-4} \sim \textit{Normal distribution}$$

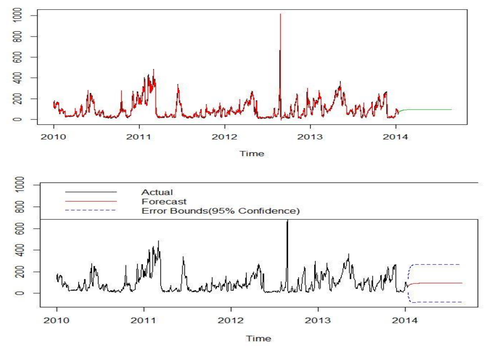
```
        Box-Ljung test

data:  pred$residuals
X-squared = 7.5389, df = 20, p-value = 0.9945
```

$\Rightarrow$ *no high lags*



The Box-Ljung test is used to test whether there exist high lags.

The p-value is 0.9945 which is very large; therefore, we reject the null hypothesis. We conclude that there are not high lags. Thus, the model works.



In the above figure, the red curve is the pm2.5 values based on the time series model, and the black curve is the value of the real pm2.5.

In the below figure, the red line is the forecasting for the pm2.5 value for year 2015 in Beijing based on the data. The blue dash lines are the confidence interval. All the predict values are within the confidence interval.

Therefore, the time series model ARIMA (1, 0, 4) does explain the data well.

Comparing ARIMA and Regression models

```
Residual standard error: 78.32 on 41747 degrees of freedom
Multiple R-squared:  0.2762,    Adjusted R-squared:  0.276
F-statistic:  1770 on 9 and 41747 DF,  p-value: < 2.2e-16

                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set -0.03421951 32.07985 15.56282 -14.86163 26.40248 0.1787067 -2.515054e-06
```

ARIMA (1, 0, 4) model shows much smaller MSE than the regression model (32.07985 and 78.32)

Therefore, ARIMA (1, 0, 4) model is a better choice.

A research group at Tsinghua University collected hourly PM2.5 data in Beijing from 2010 to 2014 and obtained total 48,324 hourly samples. They used the following features in their data collection.

Year	Year of data in this row
No	Row number
Month	Month of data in this row
PM2.5	PM2.5 concentration (ug/m^3)
DEWP	Dew point (â, f)
TEMP	Temperature (â, f)
PRES	Pressure (hPa)
cbwd	Combined wind direction
Iws	Cumulated wind speed (m/s)
Is	Cumulated hours of snow
Ir	Cumulated hours of rain

Linear Regression Model

```
Call:
rfit.default(formula = pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws +
    DEWP:PRES + TEMP:PRES, data = data1)

Coefficients:
                Estimate  Std. Error  t.value   p.value
(Intercept)  1.1764e+03  7.0893e+01  16.5933 < 2.2e-16 ***
DEWP        -8.8664e+01  3.6493e+00 -24.2959 < 2.2e-16 ***
TEMP         1.8426e+01  4.2073e+00   4.3796 1.192e-05 ***
PRES        -1.0323e+00  6.9431e-02 -14.8682 < 2.2e-16 ***
cbwdNW      -5.6383e+00  9.9521e-01  -5.6655 1.476e-08 ***
cbwdSE       2.9302e+01  9.7091e-01  30.1798 < 2.2e-16 ***
cbwdSW       2.4117e+01  1.0260e+00  23.5071 < 2.2e-16 ***
Iws         -1.0304e-01  6.3479e-03 -16.2324 < 2.2e-16 ***
DEWP:PRES    9.0175e-02  3.5935e-03  25.0938 < 2.2e-16 ***
TEMP:PRES   -2.2585e-02  4.1393e-03  -5.4562 4.893e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Overall Wald Test: 18170.87 p-value: 0
```

$$pm2.5 = 1176.4 - 88.664 * DEWP + 18.426 * TEMP - 1.0323 * PRES - 5.6383 * cbwdNW + 29.302 * cbwdSE + 24.117 * cbwdSW - 0.10304 Iws + 0.090175 * (DEWP:PRES) - 0.022585 * (TEMP:PRES)$$

Conclusion

- Pm2.5 data is time series data
  - Stationary
  - Trend
  - No seasonality
- Linear regression is able to predict the future pm2.5 value.
- ARIMA (1, 0, 4) is able to predict the future pm2.5 value.
- ARIMA (1, 0, 4) is more accurate than the linear regression model.