**Simon Fraser University**

# Regression Analysis of PM2.5 Data in Shanghai

---

**Identify the factors affecting the change of PM2.5 in Shanghai**

Lao Chin Wang (301297087)

Xiaoliang Zhang (301297782)

Ao Tang (301297684)

STAT 350

Dec 2nd, 2019

# Abstract

In this report, we are going to investigate the factors that may affect the changes of PM2.5 in Shanghai. PM 2.5 refers to the atmospheric particulate matter that have a diameter of less than 2.5 micrometers. Base on our dataset, we are going to built a multiple linear regression model with 13 explanatory variables and PM2.5 will be our response variable. In order to obtain the best model, 3 explanatory variables will be removed by variable selection using AIC. After that, we perform transformation and diagnose the model. Then we test the MSE and cross validation before obtaining the best model. Finally, we compare all the MSE value and conclude that the stepwise model is the best model.

# Introduction

### 1. Dataset

This dataset was taken from UCL Machine Learning Repository which is maintained by Guanghua School of Management at Peking University (Chen, S. X. 2017, July). This dataset includes 31880 observations on 13 variables from 1st of January 2010 to 31st of December 2015.

### 2. Variables

- Reponse Variable:
    - PM 2.5: PM2.5 concentration (ug/m^3)

- Explanatory variables (After variable selection) :
    - month: month of data in this row
    - day: day of data in this row
    - hour: hour of data in this row
    - season: season of data in this row
    - DEWP: Dew Point (Celsius Degree)
    - TEMP: Temperature (Celsius Degree)
    - HUMI: Humidity (%)
    - PRES: Pressure (hPa)
    - cbwd: Combined wind direction
    - Iws: Cumulated wind speed (m/s)
    - Iprec: Cumulated precipitation (mm)

# Variable selection

First we set up a full model with all regressor variables and evaluate it with R^2, adjusted R^2, and AIC value. The best model should has largest R^2 value, adjusted R^2 value and smallest AIC value. We use "Stepwise" method to do the variable selection. The *coefficients* output shows the variables' name, then we can check the number of variables which meet each requirement that should be included in different models and provides an obviously result.

```
mfit_null = lm(PM2.5~1, data = SH)
mfit_full = lm(PM2.5~., data=SH)

step.model_suc2=step(mfit_null, data=SH, scope = list(lower=mfit_null, upper=mfit_full), direction = "forward")

step.model_suc3=step(mfit_full, data=SH, direction = "backward")

step.model_suc1=step(mfit_null, data=SH, scope = list(upper=mfit_full), direction = "both")
```

```
Coefficients:
(Intercept)       cbwdNE       cbwdNW       cbwdSE       cbwdSW         TEMP          Iws       season        month        Iprec
  914.14937    -23.66917      8.90553    -19.12138     -3.31403     -5.04544     -0.09411      6.87640     -0.98948     -0.57671
       hour         PRES         HUMI         DEWP          day
    0.28896     -0.73203     -1.03798      3.62633      0.24795
```

By Stepwise method above, we got our best model which have 14 variables. But the variable "cbwd" is a non-numerical variable, it only contain five non-numerical values and 3 of 5 are be selected. Therefore, we only keep 11 variables in the final model:

PM2.5 ~ cbwd + TEMP + Iws + season + month + Iprec + hour + PRES + HUMI + DEWP + day

Then we do a model selection before transformation to construct a linear regression model with proper variables and interaction terms. Through the *forward*, *backward* and *stepwise*, it give us three same model and with three same AIC value.

| Information Criteria | Selection Direction | Final Selected Model ( Before transformation) | AIC Value |
|---|---|---|---|
| AIC | Forward | PM2.5 ~ cbwd + TEMP + Iws + season + month + Iprec + hour + PRES + HUMI + DEWP + day | 48059.01 |
| AIC | Backward | PM2.5 ~ cbwd + TEMP + Iws + season + month + Iprec + hour + PRES + HUMI + DEWP + day | 48059.01 |
| AIC | Stepwise | PM2.5 ~ cbwd + TEMP + Iws + season + month + Iprec + hour + PRES + HUMI + DEWP + day | 48059.01 |

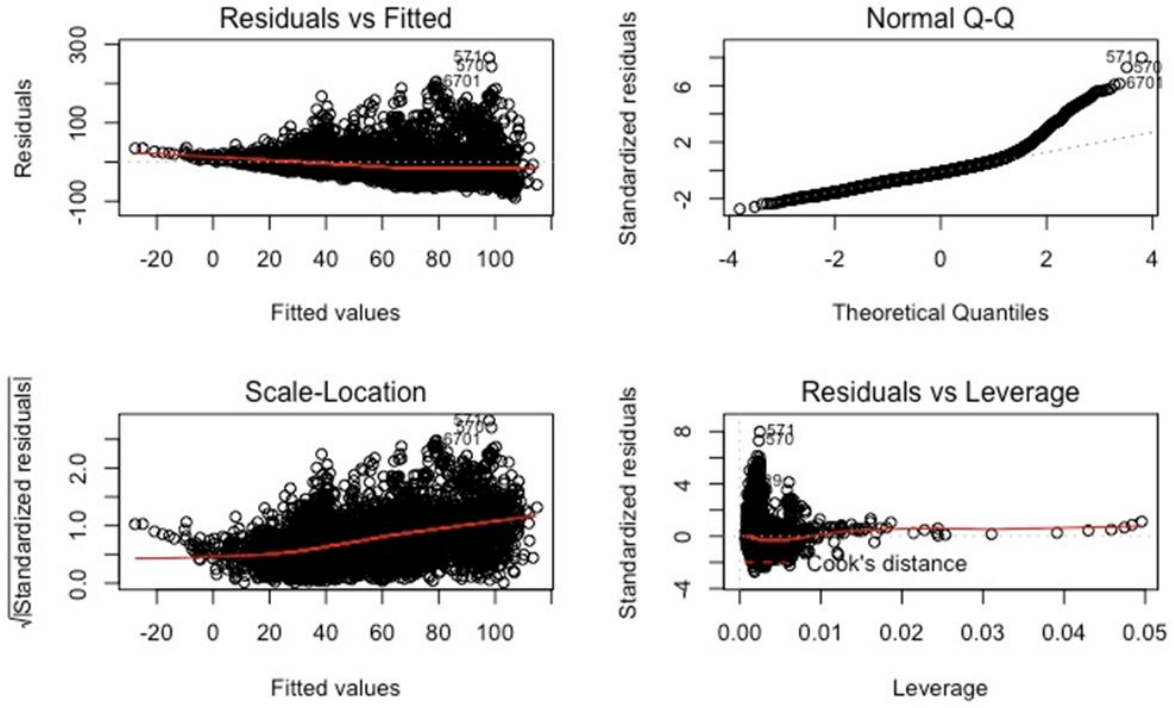
(Model selection before transformation)

Note: We use AIC value rather than BIC value to do the Model selection, because we have a very big sample size and it will lead to a big variance. Use AIC value select model can effective reduce the big variance, also can make the model be better.

# Assumptions

In this part, we assume that if the data are appropriate to use the linear regression models, then these conditions are satisfied:

- Linearity
- Normality
- Constant Variance (Equal Variance)
- Reduce Influential point



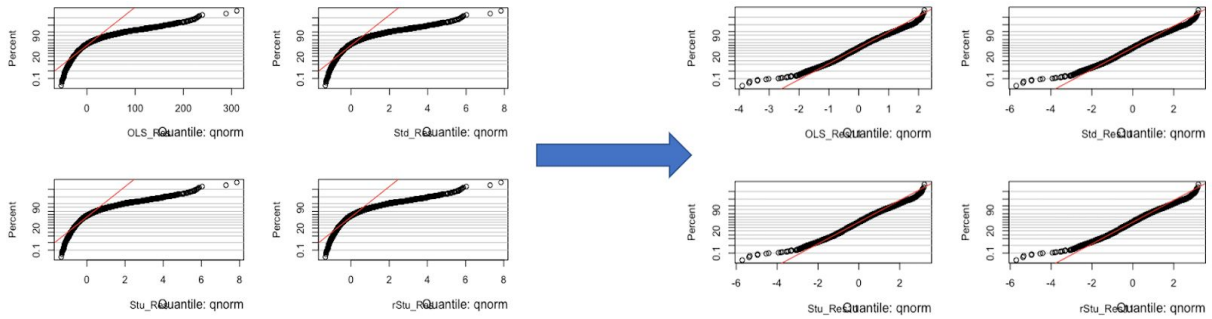

(Final model plot of regression model before transformation)

However, through the plots above, it is obvious that the conditions are not satisfied. Then we need to transform the regression model and try to get some improvements in final model.

# Transformation

We firstly do the transformations on the simple linear model by separating the multiple regression model variables and then we add them up to test the full model.

## 1. Log Transformation & Normality Test




Through the Normality Probability plots, after the log transformation, most of points lie on the diagonal line. Then the assumption of normality of the residuals are satisfied.

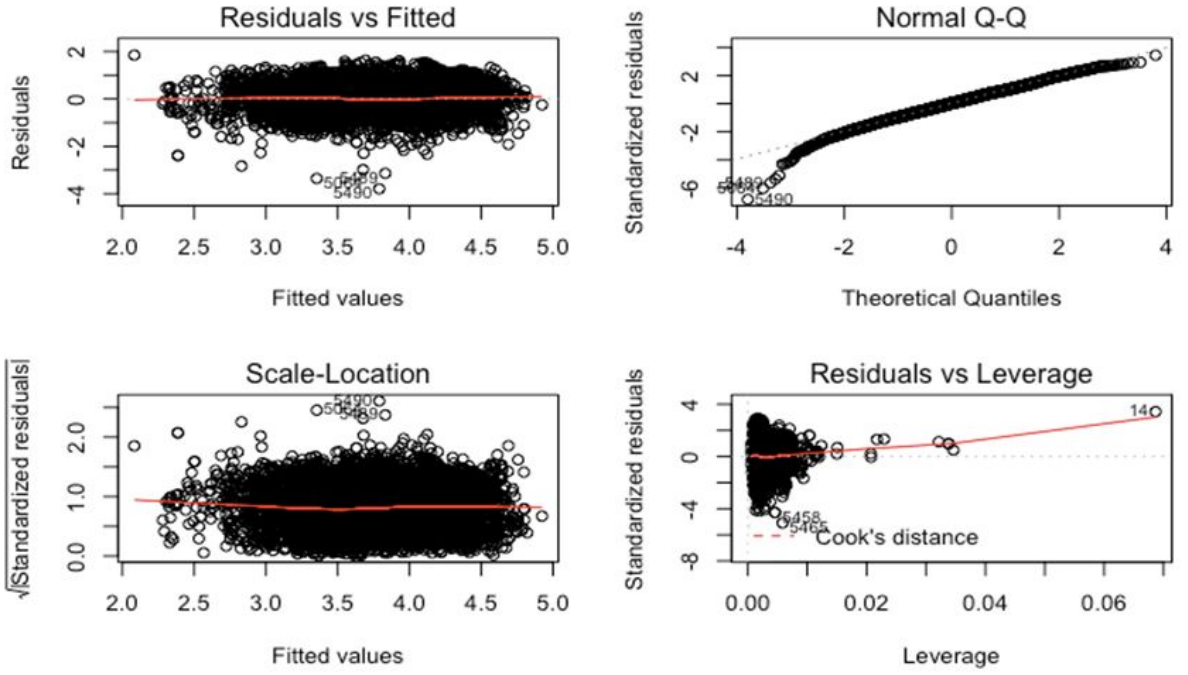
## 2. Model After Transformation

$$\log(PM2.5) \sim cbwd + TEMP + \log(Iws+1) + season + \log(month) + \log(Iprec+1) + hour + PRES + HUMI + \log(DEWP+17) + day$$

## 3. Final Model Analysis

| Transformation | Estimated σ | R^2 | Adj. R^2 | F-statistic | P-value |
|---|---|---|---|---|---|
| Before | 33.36 | 0.3075 | 0.3060 | 216.7 | < 2.2e-16 |
| After | 0.5576 | 0.3665 | 0.3652 | 282.4 | < 2.2e-16 |

Here is the summary of final model. The residual standard error has been decreased a lot. The R^2, adjusted R^2 and F-statistic all get increased. P-value keeps the same, which is smaller than 0.001. This proves that the transformation does improve our regression model.




(Final model plot of regression model after transformation)

From the final model plots after transformation above, the regression model becomes more linearity and in the Normal Q-Q plot, Scale location plot and Residuals Leverage plot of the final model plots, the normality, constant variance and Influential case also have significant improvements compared with the model before transformation. Although there are still problems left, the model after transformation is better for our regression model.

**4. Final Selected Model From Stepwise**

For the final model, we do the model selection again and check which one is the best. According to the table below, all of the three model selections show the same model, they also have the same as well as the minimum AIC value, then the best final model is :
log(PM2.5) ~ cbwd + TEMP + log(Iws+1) + season + log(month) + log(Iprec+1) + hour + PRES + HUMI + log(DEWP+17) + day

| Information Criteria | Selection Direction | Final Selected Model ( Before transformation) | AIC Value |
|---|---|---|---|
| AIC | Forward | log(PM2.5)~ cbwd +TEMP +log(1+Iws) +season +log(month) +log(1+Iprec) +hour+PRES +HUMI+ log(DEWP+17) +day | -7985.76 |
| AIC | Backward | log(PM2.5)~ cbwd +TEMP +log(1+Iws) +season +log(month) +log(1+Iprec) +hour+PRES +HUMI+ log(DEWP+17) +day | -7985.76 |
| AIC | Stepwise | log(PM2.5)~ cbwd +TEMP +log(1+Iws) +season +log(month) +log(1+Iprec) +hour+PRES +HUMI+ log(DEWP+17) +day | -7985.76 |

(Model selection before transformation)

Compared with the model before transformation, the AIC value has been decreased a lot, which also shows the regression model has great improvements after transformation.

# 5-Fold Cross-Validation

The Cross-Validation can be used to estimate the test error associated with the statistical learning method to evaluate its performance (James, G. 2017). We randomly dividing the data set into 5 independent groups with equal size. Our goal is to determine how well a given model can be expected to perform on independent data. The first group is treated as a validation set, and the remaining 4 groups as training set. we need find a model fit on the training set, then the mean squared error is computed on the observations in the validation set. This procedure is repeated 5 times; each time, a different group of observation is treated as validation set.

After that we sum the 5 MSE together and find mean of them, which give us the cross-validation error, and then we select the model for which the resulting estimated test error is smallest.

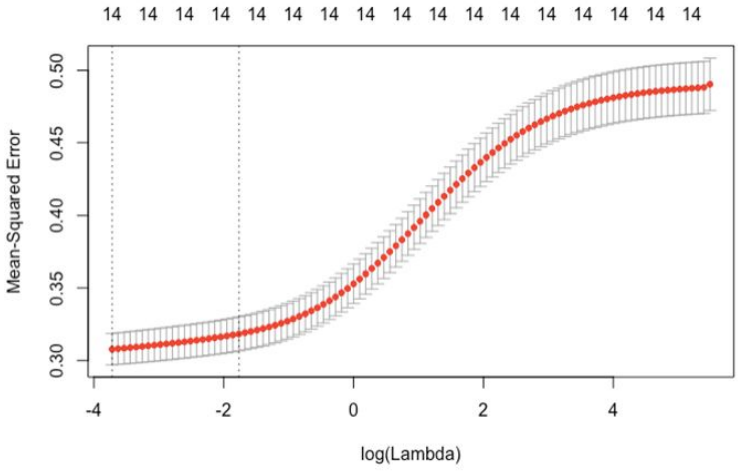
# Analysis of Ridge and Lasso Regression

Both Ridge and Lasso shrinks the coefficient estimates towards zero. That is a way to improve the fitted model, and shrinking the coefficient estimates can significantly reduce variance in our model.

**1.Ridge Regression**

Formula: $\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$

Ridge regression is similar to least squares, but least squares generate only one set of coefficient estimates. Ridge regression will produce a different set of coefficient estimates . It shrinks the value of coefficients but doesn't reach zero, which means no variable selection in this process. Parameter λ control the impact of coefficient estimates, as λ increase the shrinkage penalty increase, ridge regression coefficient estimates will approach zero (James, G. 2017).

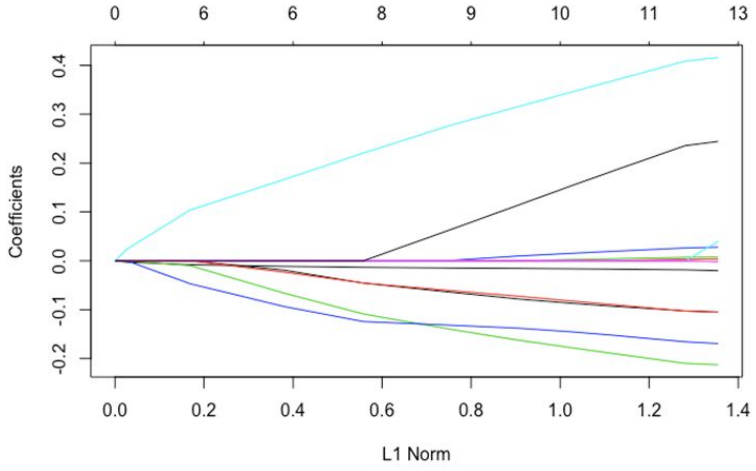



- We plot the mean squared error with log (λ ) to find selected number of variables which is 14 variables.
- Then, the second vertical dash line shows the best λ in the model, which is 0.170832.
- After that, we calculate the mean squared error is 0.3256117.

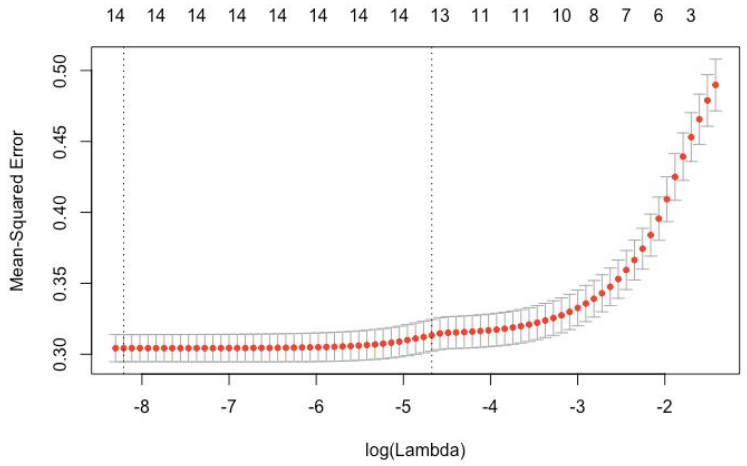
**2. Lasso Regression**

Formula: $\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$

For Lasso regression if a group of predictors are highly correlated, lasso only pick one of them and shrinks the others to zero. Since when the parameter λ is large, that forcing some of the coefficient estimates to be exactly equal to zero (James, G. 2017).




We plot the model of lasso regression, we can see that moving left to right, if λ is zero, it give us all coefficient is zero. So, depending on the value of λ , the lasso can produce a model involving any number of variables.




- We plot the mean squared error with log ( λ ) to find selected number of variables which is 13 variables.
- Then, the second vertical dash line shows the best λ in the model, which is 0.00933.
- After that, we calculate the mean squared error is 0.3217645.

### 3.Comparing the Lasso and Ridge Regression

Compare the minimum MSE from Lasso and Ridge we can found Lasso has lower minimum MSE. The reason is Lasso shrinks coefficient estimates to zero. It is kind of a trade-off between bias and variances. As λ increases, the variance will decreases and the bias will increases (James, G. 2017). Also, in this case, Lasso only selected 13 variables and Ridge selected 14 variables. So, the bias might be difference between this two methods. Hence, the minimum MSE for Lasso is slightly smaller than that of the Ridge.

# Conclusion

```
> rbind(mse.allsub, mse.ridge, mse.lasso,mse.both, mse.forward,mse.backward)
                  [,1]
mse.allsub   0.3113274
mse.ridge    0.3256117
mse.lasso    0.3217645
mse.both     0.3102568
mse.forward  0.3102568
mse.backward 0.3102568
```

We combine the MSE of allsubsets, ridge, lasso, stepwise, forward and backward method, so that we can compare all the data together.

From our result above, MSE of stepwise, forward, backward give the same result and show the smallest number among all MSE value.

On the other hand, the MSE of allsubsets also give a good result while the MSE of both ridge and lasso give a relatively large number. However, the MSE of lasso is slightly smaller than the MSE of ridge since Lasso only includes 13 variables while ridge selects 14 variables. Nevertheless, stepwise, forward and backward model only select 11 variables.

There is no doubt that stepwise, forward and backward model give the best model since more suitable variables related to our response value are chosen than ridge and lasso model, which indicates a more efficient model.

# References


1. Chen, S. X. (2017, July 8). PM2.5 Data of Five Chinese Cities Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/PM2.5 Data of Five Chinese Cities.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*. New York: Springer.