# Coursework 2
# Linear Regression


| *Author:* | *Student Number:* | *Team member roles:* |
|---|---|---|
| Zichi Zhang | 40299571 | Working on building multi variable regression models. |
| Muzixiang Xiao | 40344034 | Looking at data pre-processing and basic statistics: boxplot in R. |
| Jiyu Zou | 40344452 | Looking at further statistics (correlation among variables, presenting correlation matrixes, distributions and histograms). |
| Yuhang Zhang | 40319754 | Working on building multi variable regression models. |
| Chuao Zheng | 40336028 | Looking at data pre-processing and basic statistics: boxplot in R. |
| Yujie Yang | 40348355 | Looking at further statistics (correlation among variables, presenting correlation matrixes, distributions and histograms). |

| *Supervisor:* | *Module Number:* |
|---|---|
| Dr. E. Garcia-Palacios | ELE8096 |


February 6, 2022

**Abstract**

Data analysing and prediction has become one of the most useful tools recently, trend of data could be deduced by applying linear regression. For group project, it aims to analyse data of concentration of pollutants (NO3, PM2.5, O3) in Belfast for each hour, besides, the relationship between each pollutant and other parameters like humidity and temperature should be revealed after analysing data. R is used for data processing and trend prediction which will be discussed on following paragraphs. The distribution of each pollutant will be analysed firstly in section 2. In section 3, the correlation between each pollutant will be discussed. After that, prediction of trend of data using linear regression with R is presented in section 4, we build three models to compare the performance of models. In section 5, the model performance and the results are discussed. And find that multi variables regression as the model is better. Finally the conclusion will be made at the end of the report.

# 1 Introduction

As one of the most developed countries, UK spends a lot of energy per year, such as coal and natural gas. These energy guarantee the daily life of the city which also lead to air pollution. To ensure that people have a healthy environment, the detection of air quality is becoming more and more important, and the analysis of air quality data is the most critical part of it.

Linear regression is a popular analytical method in the field of data analysis. It is an attempt to model the relationship between two variables by fitting a linear equation to observed data, where one variable is an explanatory variable and the other is a dependent variable. The advantage of regression analysis is that it can allow you to essentially crunch the data to help to make better decisions for currently and future.

The regression method of forecasting means investigating on the relationships between data points, which help to predict data in the short and long term. Also, it allows people to understand how different variables impact on the trend of data predicted.

In this paper, a data set containing air quality is analyzed using linear regression. This data was collected in Belfast and it contains concentration of PM2.5, NO2, 03, and related temperature and humidity which collected via a multi-sensor platform (sensor system) each hour. .

## 2 basic statistics

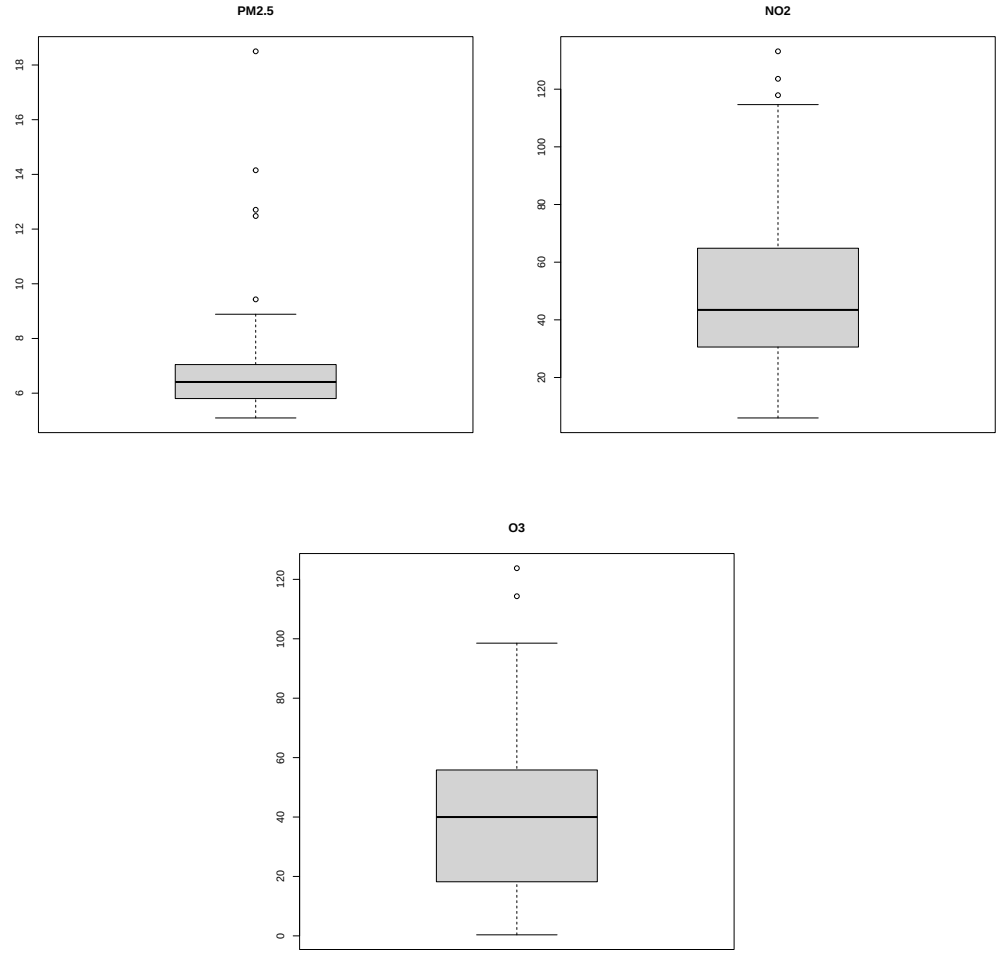



Figure 1: boxplot

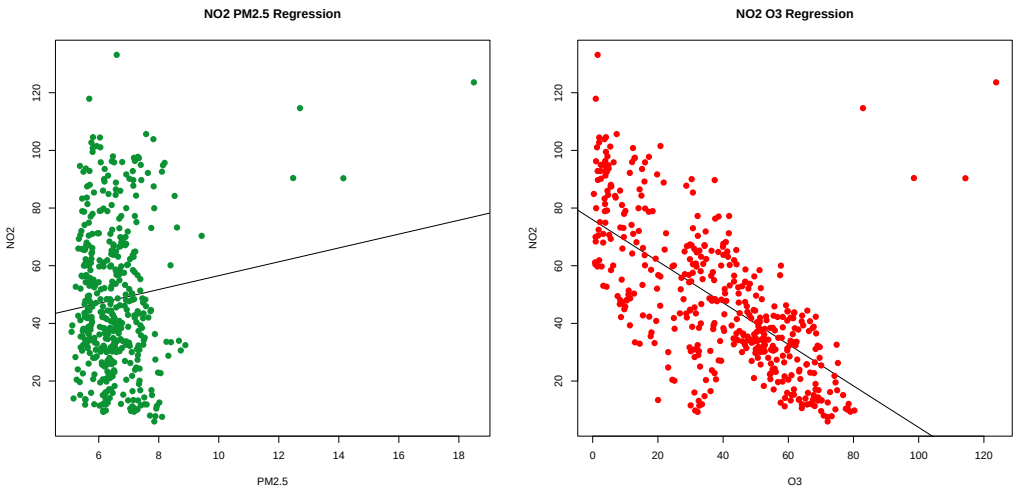
Observing the boxplot of PM2.5, it is obvious that the median corresponding to the middle line of the box is 6.3, which means that the concentration of 6.3 is the average level of PM2.5 pollutants, and the upper and lower limits of the box represent one fourth and three fourth of the value of concentration of PM2.5 in the data set. This shows that the concentration of PM2.5 pollutants mostly fluctuates from 7.1 to 5.9. There are two lines above and below the box representing the maximum value of 9.43 and the minimum value of 5.09, and there are five points above the maximum line that are out of range. These five points are outliers.

Observing the boxplot of NO2, you will find that the median corresponding to the middle line of the box is 44, which means that the concentration of 44 reflects the average level of NO2 pollutants, and the upper and lower limits of the box represent the upper four-digit fraction of NO2 respectively. and the next four digits. This shows that the concentration of NO2 pollutants mostly fluctuates from 31 to 66. There are two lines above and below the box representing the maximum value of 114.65

and the minimum value of 5.97, and there are three points above the maximum line that are out of range. These three points are outliers.

Observing the boxplot of O3, you will find that the median corresponding to a line in the middle of the box is 40, then it can be shown that the concentration of 40 reflects the average level of O3 pollutants, and the upper and lower limits of the box represent the O3. Upper four-digit score and lower four-digit score. This shows that the concentration of O3 pollutants mostly fluctuates from 18 to 56. There are two lines above and below the box representing the maximum value of 98.52 and the minimum value of 0.35. There are two points above the maximum line that are out of range, which are outliers.

## 3 more advanced statistics




The plots show a negative trend between NO2 and O3. While the trend is less clear between NO2 and PM2.5, there is a slight positive association. That is, more O3 is associated with a lower NO2, and higher PM2.5 is associated with a higher NO2.

We can use “cor” command in R to calculate the correlation with two variables below:

```
> cor(data$NO2, data$PM2.5)
[1] 0.1036955
> cor(data$NO2, data$O3)
[1] -0.6475529
```

The calculation results show the same conclusion as the image, the correlation between NO2 and PM2.5 is only 0.1036955 which means a slight positive association, and the correlation between NO2 and O3 is -0.6475529 which shows there is a negative trend.

We have the formula

$$R^2 = 1 - \frac{SSE}{SST} \tag{1}$$

By using “summary” command in R, we can get the value of $R^2$ to show the predict model good or not. We denote the modoel NO2 $\sim$ O3 with model1 and the modoel NO2 $\sim$ PM2.5 with model2, and a predict model is built by “lm” command in R:

```
> model1 = lm(NO2 ~ O3, data=data)
> summary(model1)

Call:
lm(formula = NO2 ~ O3, data = data)
... ...
Multiple R-squared:  0.4193,    Adjusted R-squared:  0.418

> model2 = lm(NO2 ~ PM2.5, data=data)
> summary(model2)

Call:
lm(formula = NO2 ~ PM2.5, data = data)
... ...
Multiple R-squared:  0.01075,   Adjusted R-squared:  0.008554
```

These show that the $R^2$ of model1 is 0.4193, and the $R^2$ of model2 is 0.01075 which means model1 is much better than model2.

# 4 multi variable regression forecasting models, visualisation of results

We denote the model NO2 $\sim$ O3 + PM2.5 with model3, and build a predict model by “lm” command in R and show the results by “summary” command:

```
> model3 = lm(NO2 ~ O3 + PM2.5, data=data)
> summary(model3)

Call:
lm(formula = NO2 ~ O3 + PM2.5, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-54.701 -11.939  -0.712  12.328  56.725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 30.72543    5.04722   6.088 2.46e-09 ***
O3          -0.82857    0.03827 -21.649  < 2e-16 ***
PM2.5        7.54685    0.79694   9.470  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.7 on 449 degrees of freedom
Multiple R-squared:  0.516,     Adjusted R-squared:  0.5138
F-statistic: 239.3 on 2 and 449 DF,  p-value: < 2.2e-16
```

## 5 model performance and discussion of results

The results show that the $R^2$ of model3 is 0.516 which is larger than the model1 and model2. Also, the adjusted R-squared shows the same results. From code in section 4, line 17, the number of stars represent the significance of the group of data to the model. And the stars of O3 and PM2.5 both are three stars which means they both are very significant to this model. The prediction performance could be improved by using multi variable regression forecasting model. For the current model, several ways could be chosen to improve accuracy of the prediction. For example adding temperature and humidity data to linear regression could be a possible option to optimize the model. Moreover, adding different weights to various variables is also a good way to optimize the model.

## 6 Conclusions

In conclusion, the analysis and comparison of the NO2, O3 and PM2.5 data by using R show that NO2 is positively associated with PM2.5 and negatively associated with O3. The $R^2$ of the multivariate regression model is closer to 1 than that of the single-variable regression model when both PM2.5 and O3 are used as independent variables of NO2. Hence, the regression equation fitted by the multivariate regression model is better than the single-variable regression equation, and its performance is also better, and the final forecasting results are more accurate by using multivariate regression. Several options that could improve better results are mentioned above and the combination of them with multivariate regression could produce more accurate forecasting trend of data.