QUEEN'S UNIVERSITY BELFAST

ELE8096 WIRELESS SENSOR SYSTEMS

Coursework 2
Linear Regression

Author:	*Student Number:*
Zichi Zhang	40299571
Muzixiang Xiao	40344034
Jiyu Zou	40344452
Yuhang Zhang	40319754
Chuao Zheng	40336028
Yujie Yang	40348355

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

February 2, 2022

Team member roles:

Zichi Zhang

Muzixiang Xiao

Jiyu Zou

Yuhang Zhang

Chuao Zheng

Yujie Yang

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 1.

In section 2, the correlation between each pollutant will be discussed.

After that, prediction of trend of data using linear regression with R is presented in section 3 and finally the conclusion will be made at the end of the report.

1 Introduction

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 1. In section 2, the correlation between each pollutant will be discussed. After that, prediction of trend of data using linear regression with R is presented in section 3 and finally the conclusion will be made at the end of the report. [1].

2 basic statistics

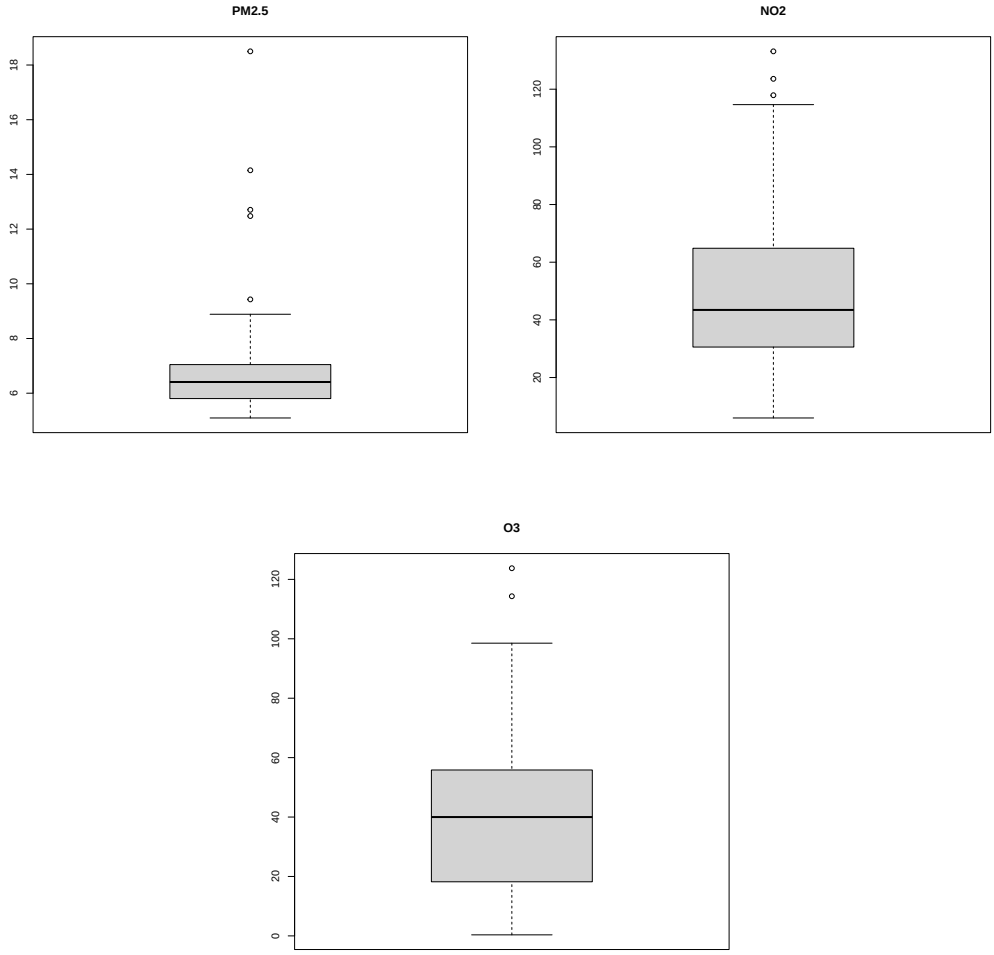


Figure 1: PM2.5 box

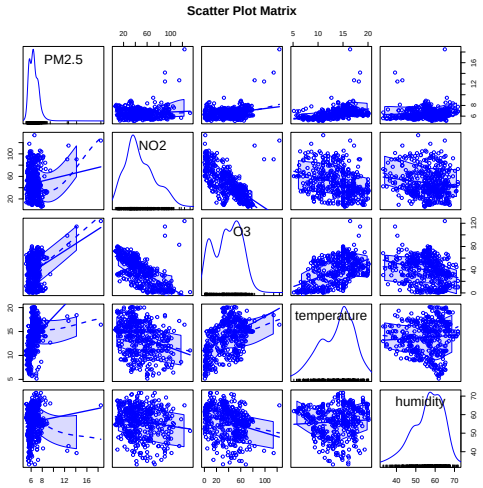
Observing the boxplot of PM2.5, you will find that the median corresponding to the middle line of the box is 6.3, which means that the concentration of 6.3 reflects the average level of PM2.5 pollutants, and the upper and lower limits of the box represent PM2. 5's top four and bottom four. This shows 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

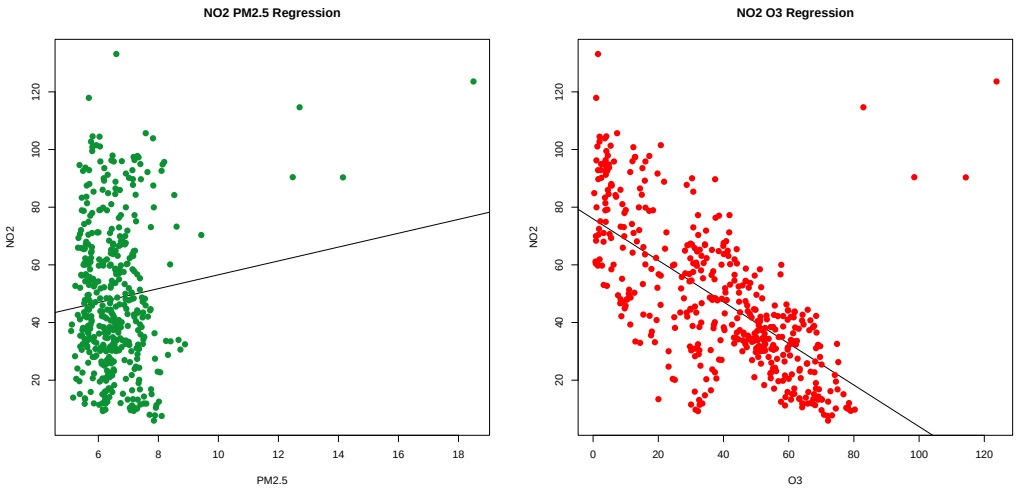
representing the maximum value of 114.65 and the minimum value of 5.97, and there are three points above the maximum line that are out of range. These three points are outliers.

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



3.1 NO2



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 build a predict model 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.

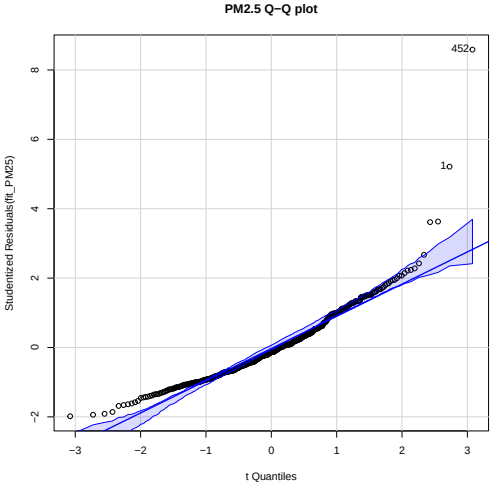
4 multi variable regression forecasting models, visualisation of results

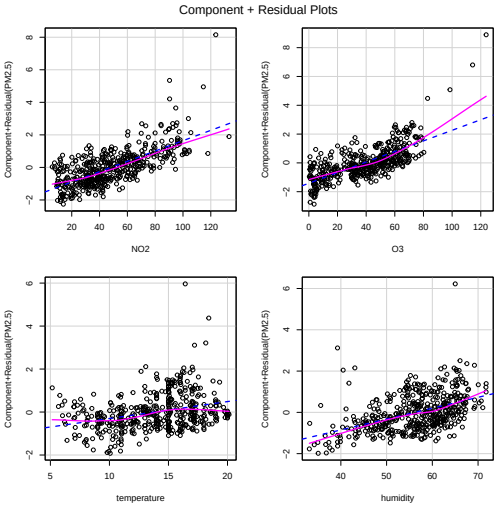
We denote the model NO2 $\sim$ O3 + PM2.5 with model3, and build a predict model by "lm" command in R and show the results by "summary" command:

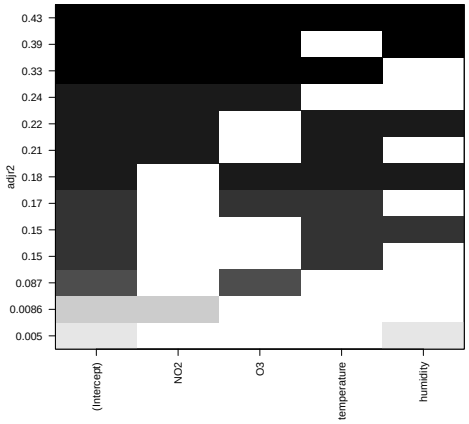
```
> model3 = lm(NO2 ~ O3 + PM2.5, data=data)
> summary(model3)

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
```

The results show that the R^2 of model3 is 0.516 which is larger than the model1 and model2. 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 is improved by using multi variable regression forecasting model.

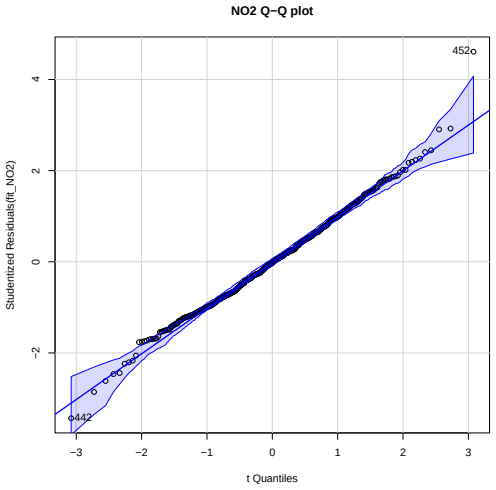
4.1 PM2.5

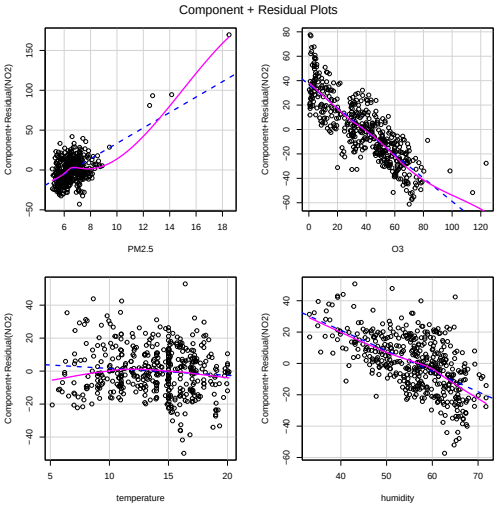

PM2.5 Q−Q plot
452
1
Studentized Residuals(fit_PM25)
t Quantiles

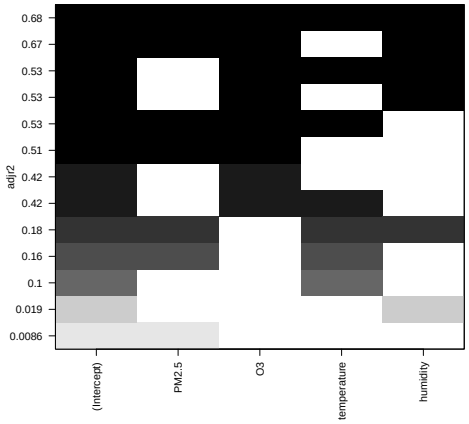

Component + Residual Plots
Component+Residual(PM2.5)
NO2
O3
temperature
humidity


adjr2
0.43
0.39
0.33
0.24
0.22
0.21
0.18
0.17
0.15
0.15
0.087
0.0086
0.005
(Intercept)
NO2
O3
temperature
humidity

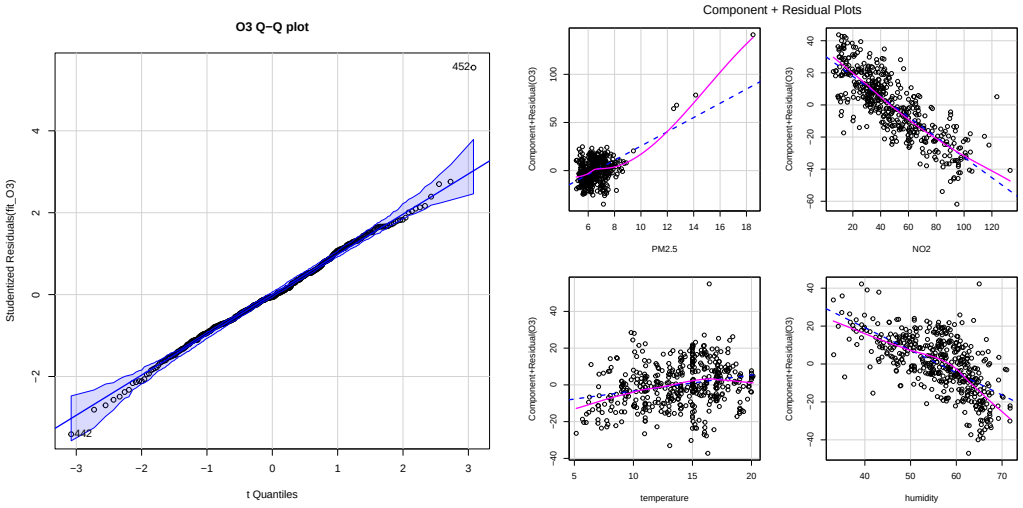
4.2 NO2

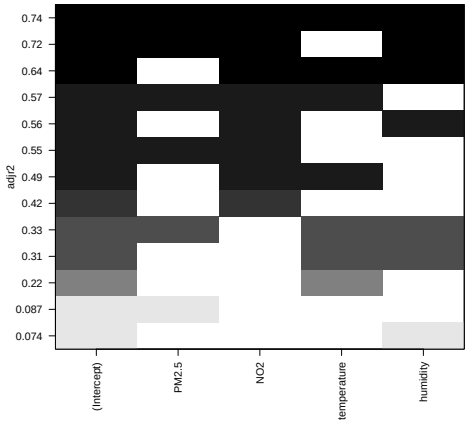

NO2 Q−Q plot
452
442
Studentized Residuals(fit_NO2)
t Quantiles


Component + Residual Plots
Component+Residual(NO2)
PM2.5
O3
temperature
humidity


adjr2
0.68
0.67
0.53
0.53
0.53
0.51
0.42
0.42
0.18
0.16
0.1
0.019
0.0086
(Intercept)
PM2.5
O3
temperature
humidity

4.3 O3





5 model performance and discussion of results

We can use the function ”outlierTest” to find the outliers in data. The output is:

```
> outlierTest(fit_PM25)
     rstudent unadjusted p-value Bonferroni p
452 8.585952         1.5164e-16   6.8539e-14
1   5.212401         2.8566e-07   1.2912e-04
> outlierTest(fit_NO2)
     rstudent unadjusted p-value Bonferroni p
452 4.611271         5.2351e-06    0.0023663
> outlierTest(fit_O3)
     rstudent unadjusted p-value Bonferroni p
452 5.543088         5.0943e-08   2.3026e-05
```

Then, we delete the outliers and make regression again, the model performance will be better.

6 Conclusions

References

[1] U. E. P. Agency, “Basic information about no2,” https://www.epa.gov/no2-pollution/basic-information-about-no2#Effects, 2021.