Air pollution at a road is related to traffic volume and meteorological variables

Xiaoyu Zhu

Beijing Normal University - Hong Kong Baptist University United International College, Guangdong Province, 519087, China

Keywords: Air pollution, NO_2, meteorological variables, traffic

Abstract: Nitrogen dioxide (NO_2) is a brownish red, highly active gaseous substance. It plays an important role in the formation of ozone. NO_2 can be produced by natural lightning, and N_2 and O_2 in the air can be combined by a large amount of heat generated when lightning current passes through. Anthropogenic nitrogen dioxide mainly comes from high-temperature combustion processes such as motor vehicles and factory emissions. The purpose of this paper is to study the relationship between NO_2 concentration and traffic flow and meteorological variables.

1. Introduction

Nitrogen dioxide is nitrogen oxide, the impact of atmospheric environment and human body can not be ignored. NO_2 is one of the main culprits of atmospheric photochemical pollution and also causes acid rain. When it enters the soil with rainwater, it directly pollutes water bodies and soil. If it enters the human body, it will cause harm to human life. It can also directly stimulate respiratory organs, causing toxic reactions and harm human health.

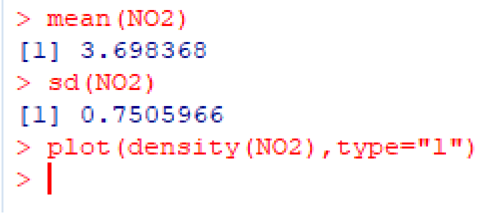
The data are a subsample of 500 observations from a data set that originate collected by the Norwegian Public Roads Administration. The response variable consist of hourly values of the logarithm of the concentration of NO_2 (particles), measured at Alnabru in Oslo, Norway, between October 2001 and August 2003.

The independent variables are the logarithm of the number of cars per hour, temperature 2 meter above ground (degree C), wind speed (meters/second), the temperature difference between 25 and 2 meters above ground (degree C), wind direction (degrees between 0 and 360), hour of day and day number from October 1.

Table1 Statistics table

logarithm of the number of cars /hour	temperature 2 meter above ground	wind speed	temperature difference between 25 and 2 meters above ground	wind direction	hour of day	day number from October 1
number	℃	(meters/second)	℃	Degrees	Hours	Number

First work out the mean and variance of NO_2 by R program. The mean of NO_2 concentration is 3.70 and the variance of it is 0.75. Then we plot the data distribution of NO_2. We can see that NO_2 is very close to a normal distribution. QQ-scatter plot was used to draw the distribution of dependent variables, which showed a linear distribution. So it was determined by linear programming with independent variables and dependent variables.



```
> mean(NO2)
[1] 3.698368
> sd(NO2)
[1] 0.7505966
> plot(density(NO2),type="l")
> 
```

Fig.1 Mean and variance

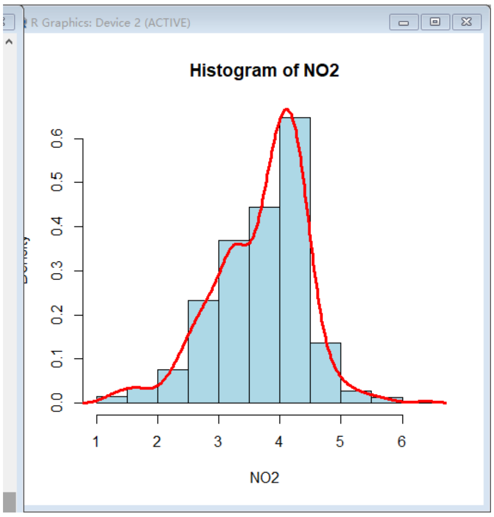


Fig.2 NO_2 distribution

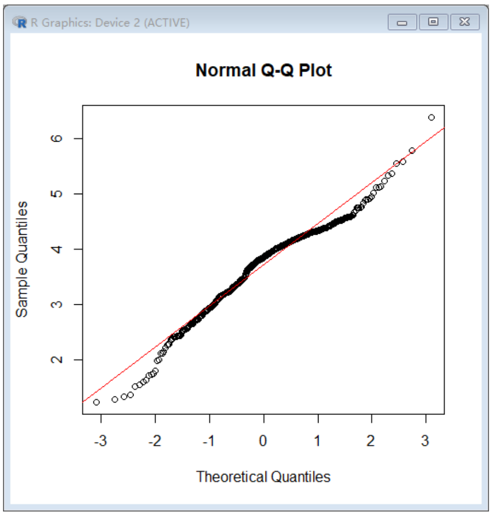


Fig.3 NO_2 qq plot

Draw the performance graph of NO_2 increasing with time. NO_2 concentration has no relation with day number from October 1. In this picture, at 200 to 400 days, there are no NO_2 points. At other times, NO_2 is present and the distribution is very similar.

Draw the performance graph of NO_2 increasing with hour. NO_2 concentration has no relation with hour of day. The NO_2 concentration distribution is similar from hour to hour. Removing these variables from the model is what we want.

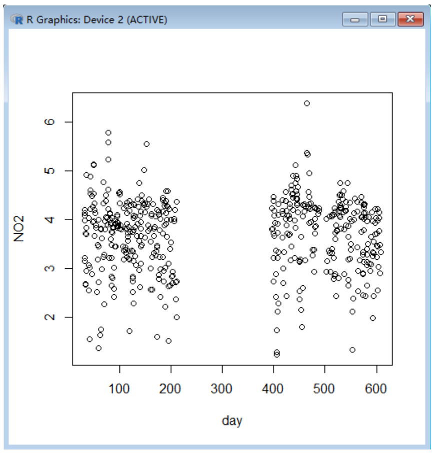


Fig.4 Plot(NO_2,day)

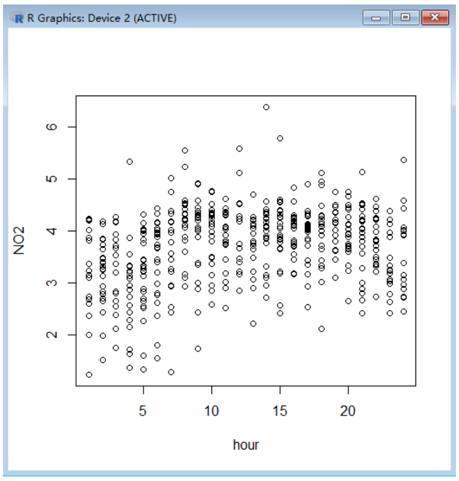
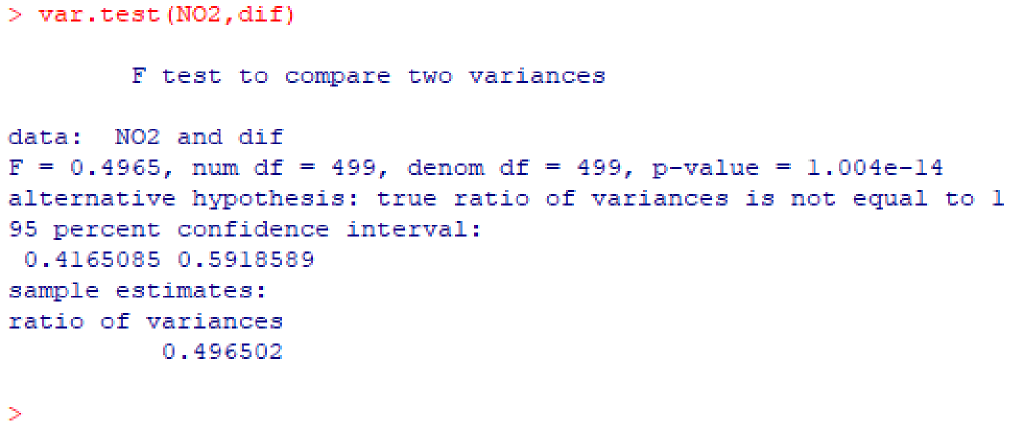


Fig.5 Plot(NO_2, hours)

We assume that temperature difference has no significant relationship with NO_2 concentration. Suppose H0: the variance of temperature difference is the same as the variance of NO_2 concentration. Ha: the variance of temperature difference is different from the variance of NO_2 concentration. Comparative analysis of variance was used.

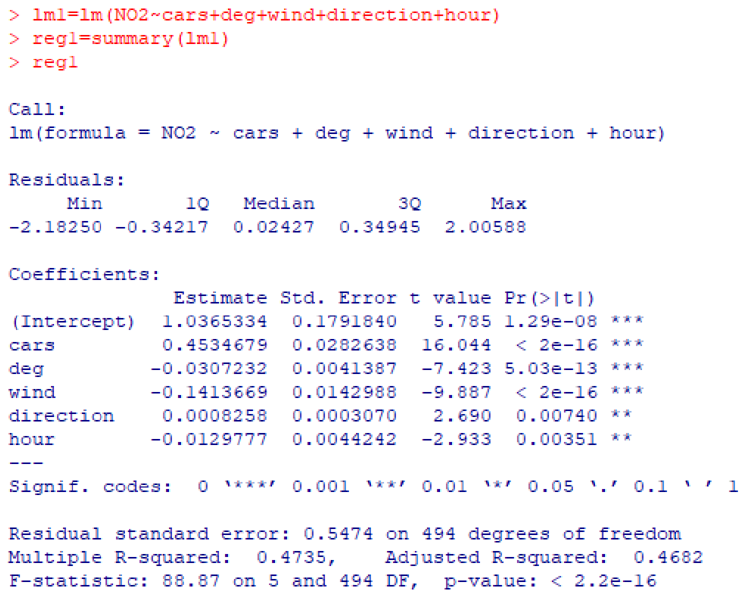


```
> var.test(NO2,dif)

        F test to compare two variances

data:  NO2 and dif
F = 0.4965, num df = 499, denom df = 499, p-value = 1.004e-14
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4165085 0.5918589
sample estimates:
ratio of variances
          0.496502

>
```

2. T-test for dif

H0 is rejected because the p-value is less than 0.025. There is a sufficient evidence to show a difference in variance of NO_2 and difference of temperature.

Build the simple linear model for the NO_2 concentration.



```
> lm1=lm(NO2~cars+deg+wind+direction+hour)
> reg1=summary(lm1)
> reg1

Call:
lm(formula = NO2 ~ cars + deg + wind + direction + hour)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.18250 -0.34217  0.02427  0.34945  2.00588 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.0365334  0.1791840   5.785 1.29e-08 ***
cars         0.4534679  0.0282638  16.044  < 2e-16 ***
deg         -0.0307232  0.0041387  -7.423 5.03e-13 ***
wind        -0.1413669  0.0142988  -9.887  < 2e-16 ***
direction    0.0008258  0.0003070   2.690  0.00740 ** 
hour        -0.0129777  0.0044242  -2.933  0.00351 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5474 on 494 degrees of freedom
Multiple R-squared:  0.4735,    Adjusted R-squared:  0.4682 
F-statistic: 88.87 on 5 and 494 DF,  p-value: < 2.2e-16
```

3. Linear regression

The qq-plot is drawn which include the independent variables. The picture show that their linear relationship is not a straight line. Try to remove a direction variable and draw the qq-plot again. As can be seen from the comparison of the following two figures, the relationship between independent variables and dependent variables is more linear after removing the variable direction.So we choose the three variables Car + Deg + Wind.

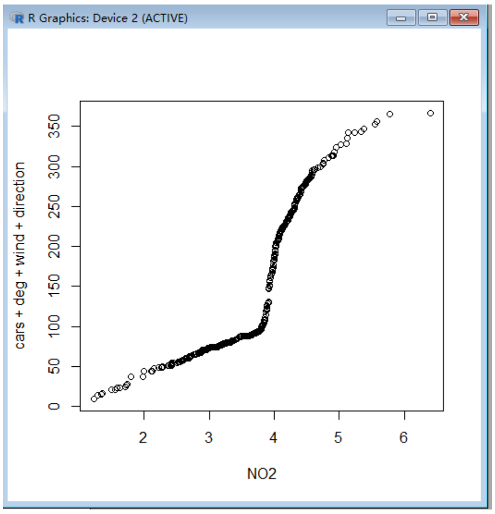


Fig.6 Qqplot(NO2,cars+deg+wind+direction)

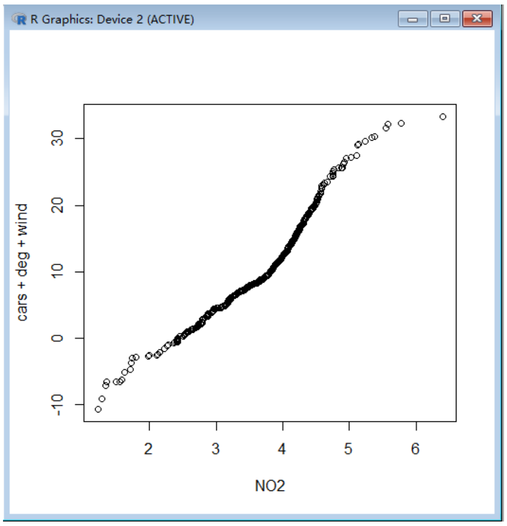
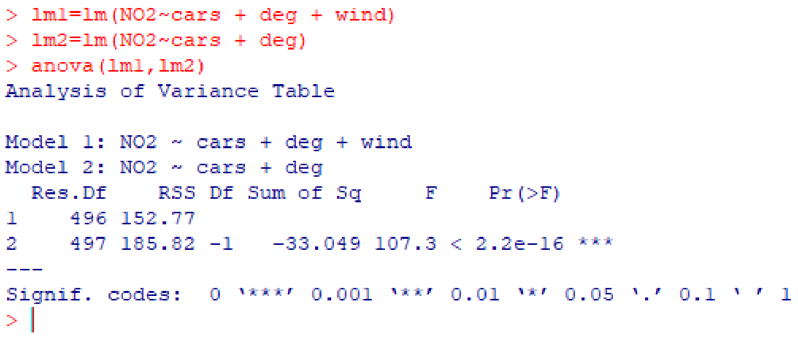


Fig.7 qqplot(NO2,cars+deg+wind)

Compare the two models: NO_2~Car + Deg + Wind and NO_2~Car + Deg.Compare the two models: NO_2~Car + Deg + Wind and NO_2~Car + Deg. ANOVA was used to compare the two models and select a more appropriate model to express the relationship between variables.



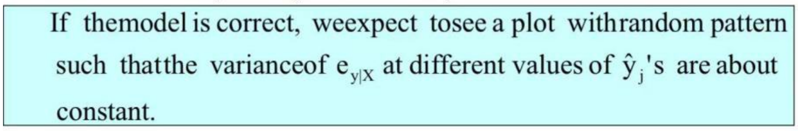
```
> lm1=lm(NO2~cars + deg + wind)
> lm2=lm(NO2~cars + deg)
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: NO2 ~ cars + deg + wind
Model 2: NO2 ~ cars + deg
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1    496 152.77                                 
2    497 185.82 -1   -33.049 107.3 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Fig.8 ANOVA table

From the P-value <0.05 in the ANOVA table, it can be concluded that the first model is better.We plot the residual against estimated response to check the correctness of the model.

a) Residual against estimated response plots:

$$e_j \text{ or } r_j \text{ against } \hat{y}_j$$



If themodel is correct, weexpect tosee a plot withrandom pattern such thatthe varianceof $e_{y|X}$ at different values of $\hat{y}_j$'s are about constant.

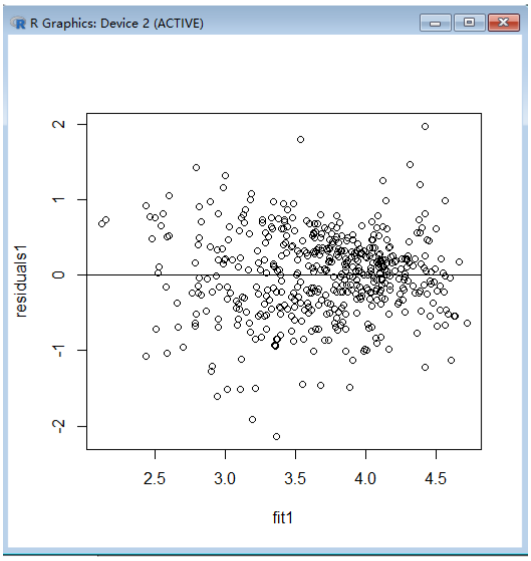


Fig.9 diagnostic plot

- **An outliers test:**

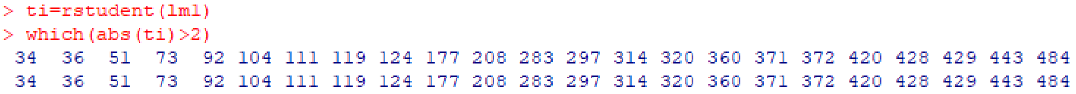
$$t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}}$$

where

$$s_{-i}^2 = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}$$

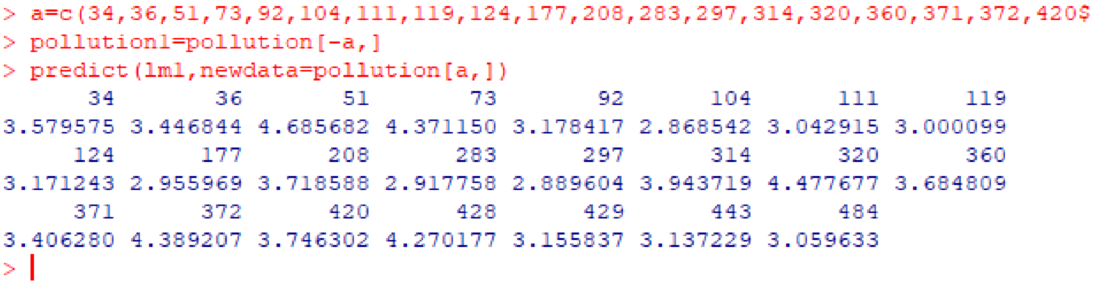
4. Outlier Equation

For a regression model with p parameters: Any observation with hii>2p/n has potential for exerting strong influence on the results. This does not apply for data set with 2p/n>1.Pay attention on observation (s) that have hii>2n/p or |ti|>2. Using the outliertest by R to check the outliers.



```
> ti=rstudent(lm1)
> which(abs(ti)>2)
 34  36  51  73  92 104 111 119 124 177 208 283 297 314 320 360 371 372 420 428 429 443 484 
 34  36  51  73  92 104 111 119 124 177 208 283 297 314 320 360 371 372 420 428 429 443 484 
```

Removing these data and predict the new dataset.



```
> a=c(34,36,51,73,92,104,111,119,124,177,208,283,297,314,320,360,371,372,420$
> pollution1=pollution[-a,]
> predict(lm1,newdata=pollution[a,])
      34       36       51       73       92      104      111      119 
3.579575 3.446844 4.685682 4.371150 3.178417 2.868542 3.042915 3.000099 
     124      177      208      283      297      314      320      360 
3.171243 2.955969 3.718588 2.917758 2.889604 3.943719 4.477677 3.684809 
     371      372      420      428      429      443      484 
3.406280 4.389207 3.746302 4.270177 3.155837 3.137229 3.059633 
> |
```

5. Predict Data

The linear relationship between NO2 concentration and traffic volume and meteorological variables was established by using R program.

The linear relationship between NO2 concentration and traffic volume and meteorological variables was established by using R program. Select the variables that have the greatest influence on the dependent variable from the independent variables. Scatter plot and T-test were used to exclude variables with low influence. Then select the appropriate linear model by comparing the linear model. Then draw the residual diagnostic plots. Use outlierstest to find the points of high impact. Finally, the new data are used to verify the equation.

Find the simple linear relationship is NO_2=1.0365334+0.4534679cars-0.0307232deg-0.1413669wind+0.0008258direction-0.012977hour

References

[1] Aldrin, M., and I. H. Haff. "Generalised additive modelling of air pollution, traffic volume and meteorology." Atmospheric Environment 39.11(2005):2145-2155.

[2] Kolluru, Ssr, et al. "Association of air pollution and meteorological variables with COVID-19 incidence: Evidence from five megacities in India." Environmental Research 195.9(2021):110854.

[3] Dai, Z., et al. "Meteorological Variables and Synoptic Patterns Associated with Air Pollutions in Eastern China during 2013–2018." International Journal of Environmental Research and Public Health 17.7(2020):2528.