Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 10, Number 5 (2014), pp. 769-774
© Research India Publications
http://www.ripublication.com

Multiple Linear Regression Analysis for Estimation of Nitrogen Oxides in Rayong

J. Mekparyup[1], K. Saithanu[2] and M. Buaphan[3]

[1,2,3]Department of Mathematics, Faculty of Science, Burapha University
169 Muang, Chonburi, Thailand
[1]jatupat@buu.ac.th, [2]ksaithan@buu.ac.th, [3]mussaya58@gmail.com

Abstract

The purpose of the present study was to estimate monthly average Nitrogen Oxides in Rayong, Thailand, using multiple linear regression analysis to build the multiple regression equation with dependent variable, Nitrogen Oxides (NO_X), and independent variables, wind speed (WS), wind direction (WD), air temperature (AT) and relative humidity (RH). The results of this study found that the multiple linear regression equation for estimation monthly average Nitrogen Oxides in Rayong was $NO_x = 12.5+13.61WS'-0.000057WD'-0.103RH'$ with standard error of estimation 1.81049 and adjusted coefficient of determination 0.559.
Mathematics Subject Classification: 62J05

Keywords: multiple linear regression analysis, best subset method

Introduction

Nitrogen Oxides (NO_X), one of air pollutants, consists of Nitrogen Dioxide (NO_2), Nitric Oxide (NO) and Nitrous Oxide (N_2O). NO_X is produced from combustion at high temperature, such as industrial waste fuel combustion, burning wood, burning of natural gas, the exhaust of motor vehicles, combustion plant, and then released to the air. Breathing high level of NO_X may cause respiratory disease, bronchitis and emphysema, cardiovascular diseases, swelling of tissues in the throat and skin irritation. High level of NO_X entering environment may burn skin or eyes when contact and lead to smog conditions when react with sunlight [1].

High level of NO_X is one of the problems in Rayong, a delightful seaside province in urban areas on Thailand's eastern Gulf coast, because there are a lot of industrial estates, such as Mabthabut Industrial Estate, Eastern Hemmarat Industrial Estate, Eastern Seaboard Industrial Estate, Industrial Estate Rayong, etc. These industrial estates release air pollutants to the environment leading to air pollution. Although

NO_X is one of factors causes occurring air pollution, there are various meteorological factors [2][3]. The present study purposes to estimate NO_X in Rayong to reduce the one of factors effect to the air pollution using the multiple linear regression (MLR) analysis.

Materials And Methods

NO_X and meteorological factors, air temperature, wind speed, wind direction and relative humidity [3][4][5], were collected from Air Quality and Noise Management Bureau, Pollution Control Department, Thailand since June 2009 to May 2014.

Checking Relationship Among Factors

For checking relationship between NO_X and meteorological variables, the correlation coefficient (R) is computed following Equation 1 [6].

$$R = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}} \tag{1}$$

Generating The Multiple Linear Regression Model

MLR is one of many techniques widely use to analysis multivariate variables. In this study, there are 5 variables which are one dependent variable, monthly average Nitrogen Oxides (NO_X), and 4 independent variables, monthly average wind speed (WS), monthly average wind direction (WD), monthly average air temperature (AT) and monthly average relative humidity (RH), were used for generating the MLR model in Rayong following Equation 2;

$$NO_X = \beta_0 + \beta_1 WS + \beta_2 WD + \beta_3 AT + \beta_4 RH + \varepsilon \tag{2}$$

where ε = error of the regression model.

Considering The Best Multiple Linear Regression Equation

The best subset method is used for generating the appropriate MLR equations then considering the best equation by Mallows' C_p [7], standard error of estimation (S) and adjusted coefficient of determination (R^2_{adj}). Then the fitted equation is tested by F statistic.

Checking Multiple Regression Analysis Assumptions

After selected the fitted MLR equation, checking assumptions of multiple regression analysis is consequently tested; (I) normality of the error distribution with Anderson-Darling statistic (AD) [8]; (II) independence of the errors with Durbin-Watson statistic (DW) [9]; (III) homoscedasticity of the errors with Breusch-Pagan statistic

(BP) [10]; (IV) multicollinearity among predictor variables with Variance Inflation Factor (VIF) following Equation 3;

$$\left(VIF\right)_j = \frac{1}{1-R_j^2} \quad ; j = 1, 2, \ldots, p-1 \qquad (3)$$

where R_j^2 be the coefficient of multiple determination with independent x_j regressing on the $p-2$ other independent x variables in the model (*p* be the number of predictor variables). If checking all assumptions is not satisfied, another MLR equation will be considered by suitable Mallows' C_p or the equation will be adjusted by Box-Cox transformation method [11] until all assumptions will be met.

Validaiting The Multiple Linear Regression Equation

The observed data (OBS) and the estimated data (EST) are compared by time series plot and scatter plot. Then the percentage of error (PE) is computed following Equation 4.

$$\text{PE} = \frac{\left|\text{OBS} - \text{EST}\right|}{\text{OBS}} \text{X}100\% \qquad (4)$$

Results And Discussion

The highest negative correlation coefficient between NO_X and WS was found with R= –0.648 (P-value=0.000). Moreover, all correlation coefficient between NO_X and all independent variables was negative.

According to the best subset method, WS, WD and RH were selected to generate the MLR equation with Mallows' C_p= 3.1, S=1.8098 and R^2_{adj}=0.559 which was the same previous studies [1][2][3][4][5]. And the regression equation was as Equation 5 with the test statistic F=25.95 (P-value=0.000).

$$\hat{\text{NO}}_X = 34.413 - 6.777\text{WS} - 0.016\text{WD} - 0.12\text{RH} \qquad (5)$$

After obtained the appropriate equation, the assumptions was consequently determined; (I) normality of the error distribution was tested with AD=0.922 (P-value=0.018) so the distribution of error was a significant normality, (II) independence of the errors was calculated with DW=1.202 compared with critical value (D_L=1.317) so the errors was not significantly independent then the Box-Cox transformation was used and the MLR equation was regenerated as Equation 6 with S=1.81049 and R^2_{adj}=0.559;

$$\hat{\text{NO}}_X = 20.248 + 13.584\text{WS}' - 0.00005654\text{WD}' - 1.781\text{RH}' \qquad (6)$$

where $\text{WS}' = 1/\text{WS}$, $\text{WDS} = 1/\text{WD}$ and $\text{RH}' = \sqrt{\text{RH}}$ with the test statistic F=25.91 (P-value=0.000). Then the assumptions were reconsidered following; (I) the distribution of error was significantly normal (AD=0.980, P-value=0.013), (II) the errors was significantly independent (DW=1.345, D_L=1.317), (III) homoscedasticity of the errors was tested with BP=2.477 (P-value=0.1155) so the variance of error was significantly constant, (IV) VIFs were illustrated by Equation 3 for checking the multicollinearity and the results showed that there was no correlation among independent variables in the MLR equation

$(VIF_{WS'} = 1.3,\ VIF_{WD'} = 1.5 \text{ and } VIF_{RH'} = 1.3)$ [12].

After validated all assumptions, then graph of time series between the OBS and the EST were plotted to compare as Figure 1(a) and scatter plot was created as Figure 1(b) with R=0.762. Moreover, the PE was calculated by Equation 4 and the result showed in Table 1. The PE values ranged from 1 to 10 with 17.97%, 11 to 20 with 10.54%, 21 to 30 with 18.89%, 31 to 40 with 32.91%, 41 to 50 with 8.78% and 51 to 60 with 10.92%.

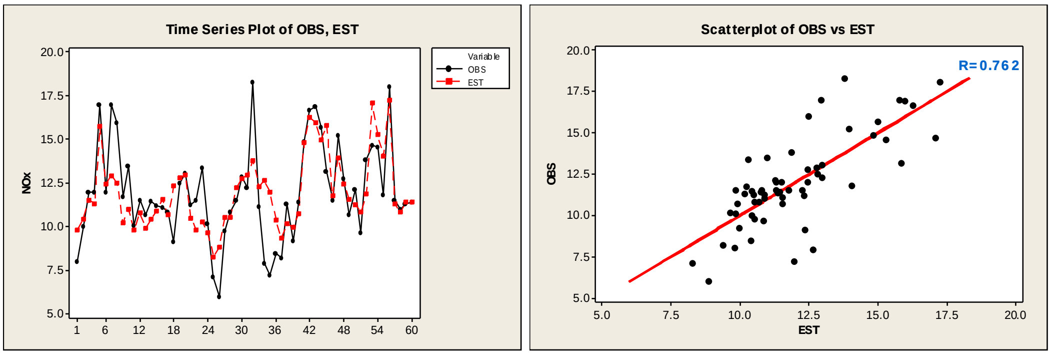

Figure 1: Comparison between OBS and EST in Rayong; (a) Time series plot, (b) Scatter plot

Table 1: Percentage error of Equation 6

Year	Month	OBS	EST	PE
2009	6	8.00	9.81	22.68
	7	10.00	10.43	4.33
	8	12.00	11.52	4.04
	9	12.00	11.32	5.63
	10	17.00	15.77	7.25
	11	12.00	12.45	3.74
	12	17.00	12.94	23.90
2010	1	15.98	12.49	21.84
	2	11.74	10.24	12.80
	3	13.47	10.99	18.43
	4	10.08	9.83	2.50
	5	11.51	10.78	6.32
	6	10.70	9.90	7.52
2011	6	7.12	8.28	16.30
	7	5.98	8.86	48.04
	8	9.75	10.52	7.95
	9	10.83	10.54	2.67
	10	11.52	12.24	6.24
	11	12.88	12.78	0.77
	12	12.27	12.96	5.68
2012	1	18.31	13.78	24.72
	2	11.16	12.31	10.22
	3	7.92	12.64	59.61
	4	7.23	11.97	65.45
	5	8.45	10.39	22.91
	6	8.21	9.37	14.16
2013	6	10.69	11.55	8.04
	7	12.12	11.26	7.09
	8	9.66	10.85	12.29
	9	13.82	11.88	14.07
	10	14.68	17.07	16.32
	11	14.57	15.28	4.85
	12	11.81	14.05	18.97
2014	1	18.05	17.23	4.53
	2	11.54	11.30	2.08
	3	11.01	10.87	1.24
	4	11.33	11.39	0.55
	5	11.40	11.44	0.30

	7	11.47	10.41	9.23		7	11.30	10.17	9.99
	8	11.23	10.88	3.09		8	9.21	9.98	8.43
	9	11.08	11.55	4.25		9	11.41	10.75	5.79
	10	10.83	10.70	1.23		10	14.86	14.83	0.18
	11	9.13	12.34	35.17		11	16.67	16.26	2.44
	12	12.52	12.81	2.32		12	16.90	15.97	5.50
2011	1	13.06	12.97	0.71	2013	1	15.68	14.99	4.38
	2	11.26	10.51	6.72		2	13.16	15.83	20.25
	3	11.52	9.84	14.56		3	11.54	11.76	1.89
	4	13.38	10.29	23.09		4	15.22	13.95	8.36
	5	10.17	9.65	5.10		5	12.76	12.44	2.50

Acknowledgement

The authors wish to thank to the Air Quality and Noise Management Bureau, Pollution Control Department, Thailand for kind support collecting data.

REFERENCES

[1] Gurjar, B.R., Butler, T.M., Lawrence, M. G., & Lelieveld, J., 2008, "Evaluation of emissions and air quality in megacities," Atmospheric Environment, 42(7), 1593-1606.

[2] Banerjee, T., Singh, S.B., & Srivastava, R. K., 2011, "Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India," Atmospheric Research, 99(3), 505-517.

[3] Kavuri, N.C., Paul, K.K., & Roy, N., 2012, "Regression modeling of gaseous air pollutants and meteorological parameters in a steel city, Rourkela. In 2nd International Science Congress (ISC-2012), 8th-9th December 2012, Mathura, UP, India.

[4] Shi, J.P., & Harrison, R.M., 1997, "Regression modelling of hourly NO_X and NO_2 concentrations in urban air in London," Atmospheric Environment, 31(24), 4081- 4094.

[5] Banerjee, T., Singh, S.B., & Srivastava, R.K., 2011, "Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India," Atmospheric Research, 99(3), 505-517.

[6] Dowdy, S., & Wearden, S., 1983, "Statistics for Research," Wiley, New York.

[7] Hocking, R.R., & Leslie, R.N., 1967, "Selection of the best subset in regression analysis," Technometrics, 9(4), 531-540.

[8] Lewis, P.A.W., 1961, "Distribution of the Anderson-Darling Statistic," The Annals of Mathematical Statistics, 32(4), 1118-1124.

[9] Durbin, J., & Watson, G.S., 1951, "Testing for Serial Correlation in Least Squares Regression II," Biometrika, 38(2), 159-177.

[10] Breusch T.S., & Pagan, A.R., “A Simple Test for heteroscedasticity and Random Coefficient Variation,” Econometrica, 47(5), 1287-1294.

[11] Sakia, R.M., 1992, “The Box-Cox transformation technique: a review,” The statistician, 169-178.

[12] O’brien, R.M., 2007, “A caution regarding rules of thumb for variance inflation factors,” Quality & Quantity, 41(5), 673-690.