Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 10, Number 6 (2014), pp. 835-839
© Research India Publications
http://www.ripublication.com

Using Multiple Linear Regression To Predict PM_{10} Concentration In Chonburi, Thailand

K. Saithanu[1] and J. Mekparyup[2]

[1,2]Department of Mathematics, Faculty of Science, Burapha University
169 Muang, Chonburi, Thailand
[1]ksaithan@buu.ac.th, [2]u.pranot@hotmail.com

Abstract

To train and validate the model of prediction PM_{10} concentration in Chonburi, the multiple linear regression equation would be simply modeled from the training data set which was measured concentration of PM_{10} and other various variables in the atmosphere for the period 2006-2008. The stepwise technique obviously illustrated a relationship between the dependent PM_{10} concentration and the nine independent variables composing of the four air pollutant and the five meteorological variables as $PM'_{10}=-44.5+0.451CO+0.023NO_2+0.0175O_3+0.619HC+0.0649Pressure-0.0148RH-0.0772Temp-0.00156SR+0.193WS$ with 0.6222 for standard error of estimation. The performance of regression model was evaluated with the mean bias error considering from the unseen data set observed in 2009. It indicated the validation data set showed its mean bias error nearly closed to zero and the standard error of estimation (0.6951) was comparable to the training data set.

Keywords: Multiple linear regression, best subset, stepwise**K. Saithanu[1] and J. Mekparyup**

Mathematics Subject Classification: 62J05

INTRODUCTION

Chonburi is one of eastern provinces in Thailand. It is located in the urban industrial area so the air pollution problem especially the pollution of PM_{10} has been extensively increasing at present corresponding to the annual concentration report in Chonburi [1]. The standard of 24 hours PM_{10} level in the atmosphere is 120 $\mu g/m^3$ in accordance with the Air Quality Index specified by the Thai Environmental Protection Department [2]. Many researchers have continually studied the characteristics of PM_{10} by means of change in temperature, humidity, precipitation and ventilation [3], [4], [5] because the PM_{10} concentration basing on both of air pollutant and meteorological variables may affect the future change in regional weather. Some researches used the advanced model like neural network to analyze and predict the PM_{10} concentration such as [6], [7], [8], [9], [10]. However, this study purposed the simple statistical tool like multiple linear regression with best subset and stepwise methods to predict PM_{10} concentration in Chonburi also then rectified the obtaining equation to satisfy the assumption of regression.

MATERIALS AND METHODS

The delegate of monitoring stations in Chonburi was the General Education Centre, Mueang District. The observational data measured during 2006-2009 here (826 cases) was depended on the dependent PM_{10} concentration and the sixteen independent variables in the atmosphere, both of nine air pollutant variables (CO, NO, NO_2, NO_X, SO_2, HC, CH_4, NMHC and O_3) and seven meteorological variables (Pressure, Temperature: Temp, Relative Humidity: RH, Wind Speed: WS, Wind Direction: WD, Sun Radiation: SR and Rain). The step of data analysis for this study was as follows.

1. Separate the whole data set into two parts. The training data set was the first part consisting 620 observations measured in the period 2006-2008. The validation data set was the remaining 206 observations in 2009.
2. Build the estimated regression equation by using the training data set. Two regression methods were combined for variable selection. The best subset technique was first employed to roughly consider the number of suitable variables for fitting regression model because it can identify models with as few predictors as possible. Once the number of proper variables was obtained, the stepwise technique was secondly used to fit the estimated regression equation because it can remove and add variables to the regression model for the objective of identifying a useful subset of the predictors. The appropriateness of model was then later diagnosed and rectified in accordance with the three assumptions of regression analysis.

2.1 Normality of the error distribution was examined by the Anderson-Darling statistic as defined in [11] with Equation 1.

$$AD = \sum_{i=1}^{n} \frac{1-2i}{n}\left[\ln\left(F_0\left(y_{(i)}\right)\right) + \ln\left(1-F_0\left(y_{(n+1-i)}\right)\right)\right] - n \quad (1)$$

where F_0 be the assumed (normal) distribution with the assumed or sample estimated parameters (μ,σ), $y_{(i)}$ be the *i*th sorted, standardized, sample value and *n* be the sample size.

2.2 Independent and constant variance of the error term was tested by the Breusch-Pagan statistic as calculated in [12] with Equation (2).

$$\chi^2_{BP} = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \tag{2}$$

where SSR^* be the regression sum of squares by regressing e^2 on the independent variables and SSE be the error sum of squares by regressing the dependent variable on the independent variables.

2.3 Multicollinearity among predictor variables was investigated by the variance inflation factor as computed in [12] with Equation 3.

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

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).

3. Validate the performance of regression model. The mean bias error was the criterion for verifying model suitability as considered with Equation 4.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right) \tag{4}$$

where $\hat{y}_i$ be the *i*th predicted value of PM_{10} concentration from the regression model, y_i be the *i*th observation of PM_{10} concentration and *n* be the number of observations in the validation data set.

RESULTS

Regarding of the best subset, the first model roughly considered were the model with the smallest Mallow C-p value. This model contained eleven predictor variables (CO, NO_2, O_3, HC, CH_4, NMHC, Pressure, RH, Temp, SR and WS). When the regression model would be fitted with stepwise method, the two variables (CH_4 and NMHC) were removed from this model with the large p-value of *t*-statistic. However, this obtaining regression equation was against to normality assumption. The Box-Cox transformation was then required to transform the PM_{10} concentration. The adjusted of estimated regression equation was then revised as

$PM'_{10} = \sqrt{PM_{10}}$ =–44.5+0.451*CO*+0.0234NO_2+0.0175O_3+0.619*HC*+0.0649*Pressure*–0.0148*RH*–0.0772*Temp*–0.00156*SR*+0.193*WS* with $S = 0.6222$. The diagnostics for regression analysis assumption resulted as follows.

1. The error distribution was normal with the P-value of *AD* test equal to 0.146.
2. Variance of the error term was independent and constant with the test statistic of Breusch-Pagan $\chi^2_{BP} = 6.5012$ which was less than the critical value (16.919).
3. The *VIF* of all these nine predictor variables were less than 5. That means no relationship among these nine variables.

Finally, the performance of regression model was measure from the validation data set with $MBE = -0.5926$ and $S = 0.6951$.

DISCUSSION

The suitable multiple linear regression model based on combining the best subset and stepwise methods could be potentially predicted the PM_{10} concentration in Chonburi. This obtaining estimated regression equation was also investigated and rectified in accordance with the assumption of regression analysis. Moreover, the MBE of the validation data set was close to zero so it presented the good model performance. That means it could well predict the PM_{10} concentration.

ACKNOWLEDGEMENT

The authors really wish to thank Air Quality and Noise Management Bureau, Pollution Control Department for helpfully supporting all data.

REFERENCES

[1] Khaenamkaew, P. Iamraksa, P., Raksawong, S., Wongsontam, K., Angwanisakul, C., & Khuntong, S., 2011, “Annual Concentration Report and Emission Sources Analysis of the Air Pollutants Measured by the AQM Station,” Research Exhibition “Research in Kasetsart University 2011” in National Agricultural Fair.

[2] Office of Natural Resources and Environmental Policy and Planning, “Notice of the National Environment Committee NO.28 (B.E.2550) on Air Quality Standard,” Retrieved October 25, 2011, from http://www.legalbase.pti.org/Law.aspx?lid=3596

[3] Mickley, L. J., Jacob, D. J., Field, B. D., & Rind, D., 2004, "Effects of future climate change on regional air pollution episodes in the United States," Geophysical Research Letters, 31(24).

[4] Liao, H., Chen, W. T., & Seinfeld, J. H., 2006, "Role of climate change in global predictions of future tropospheric ozone and aerosols," Journal of Geophysical Research: Atmospheres (1984–2012), 111(D12).

[5] Heald, C. L., Henze, D. K., Horowitz, L. W., Feddema, J., Lamarque, J. F., Guenther, A., ... & Fung, I., 2008, "Predicted change in global secondary organic aerosol concentrations in response to future climate, emissions, and land use change," Journal of Geophysical Research: Atmospheres (1984–2012), 113(D5).

[6] Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., ... & Cawley, G., 2003, "Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki," Atmospheric Environment, 37(32), 4539-4550

[7] Arampongsanuwat S., & Meesad, P., 2010, “Development of a Prediction

Model of PM_{10} in Bangkok Using Artificial Neural Networks," The 6th National Conference on Computing and Information Technology.

[8] Durao, R., & Pereira, M.J., 2012, "MLP based models to predict PM10, O3 concentrations, in Sines industrial area," Geophysical Research Abstract, 14, EGU2012-13448.

[9] Arampongsanuwat, S., & Meesad, P., 2011, "Development of a Prediction Model of PM10 in Bangkok Using Support Vector Regression and Radial Basis Function Network," The 7th National Conference on Computing and Information Technology.

[10] Mekparyup, J., & Saithanu, K., 2013, "Development of Neural Network Technique for Prediction of PM_{10} Concentration in the Industrial Area, at the East of Thailand," Applied Mathematical Sciences. 7(93), 4631-4638.

[11] Romeu, J. L., 2003, "Anderson-Darling: a goodness of fit test for small samples assumptions," Selected Topics in Assurance Related Technologies. 10(5), 1-6.

[12] Michael, H. K., John, N., Christopher, J. N., & William Li., 2005, "Applied Linear Regression Models, 5th Edition," New York: McGraw-Hill.