USING MULTIPLE LINEAR REGRESSION TO FORECAST THE NUMBER OF ASTHMATICS

Darmesah Gabda[1], Noraini Abdullah[1], Kamsia Budin[2] & C.K. Lim[1]
[1]Programme of Mathematics with Economics
[2] Environmental Science Programme
Universiti Malaysia Sabah
Locked bag 2073, 88999 Kota Kinabalu, sabah
MALAYSIA

Abstract: - The objective of this study was to determine the association between the number of asthmatic patients in Kota Kinabalu, Sabah with the air quality and meteorological factors using multiple linear regression. The main eight independent variables with the fourth order interactions were included in the model. There were 80 possible models considered and the best model was obtained using the eight selection criteria (8SC). The result showed that the best model would represent the cause of the rise in the number of asthmatics modeled by M80.23.

Key-Words: - multiple regression, eight selection criteria, fourth-order interaction, best model, asthma.

1 Introduction

Asthmatic individuals had been identified as a population that is especially sensitive to the effects of ambient air pollutants [1]. In this study, five criteria pollutants were considered for the assessment of their associations with the number of asthmatics, namely carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and particular matter (PM_{10}). Meteorological factors such as temperature, relative humidity and rainfall also were considered as contributing causes on the increasing number of asthmatics. Many studies had showed that the different results of asthmatic association with air quality factors and meteorological factors. Hence, the impact of each of these air quality and meteorological factors need to be studied.

In this study, multiple regression was used to relate the number of asthmatics with the air quality and meteorological factors. The interaction variables were also included in the model, besides the main variables. In some problems of multiple regression, some independent variables may not be related to the dependent variable. Hence, a procedure to select an appropriate subset independent variable is required to relate with the dependent variable. Criteria on the selection of the best model played an important role in choosing the best model since the total number of variables involved were large. In this study with the help of the eight selection criteria and the level of significance, α equals 0.05, the best model was obtained. Several tests were carried out to the best model such as the individual test, global test, Wald test and randomness test.

2 Literature review

Air pollutants had an effect on the respiratory and cardiovascular health although it is still at low level below the national guideline. A study in Klang Valley showed that nitrogen dioxide had the greatest impact on the respiratory and cardiovascular morbidity. It also showed that PM_{10} and sulphur dioxide were significantly associated with the relative risks for respiratory and cardiovascular mortality[2]. Stronger associations of coughs among children with PM_{10}, $PM_{2.5}$, PM_1 and $PM_{10\text{-}2.5}$ had been reported [3]. Children were more sensitive to the effects of increased levels of PM air pollution than adults. It was also reported that asthmatic presentations had increased with each $10-\mu g/m3$ increase in PM_{10} concentration [4]. However, in some studies, there was no significant association between particulate matter (PM) with the total respiratory admissions and number of children without hyperactivity [5, 6].

The effects of interactions between the air pollutants and meteorological factors towards asthma were also studied. Besides considering the single independent variable as an explanatory to the dependent variable, interaction effects between the independent variable was suggested by [7] to be taken into the model. These

interaction effects represented the combined effects of the variables on the criterion or dependent measure. As [8] had noted, the interaction factor should be studied rather than the isolated effect of a single variable. The interpretation of the individual variables may be incomplete or misleading when interaction effects are present [8].

3 Methodology

In this study, the effects of the air quality and meteorological factors as the causes of increasing number of asthmatics admitted to the Hospital Queen Elizabeth, Kota Kinabalu Sabah (2003 – 2005) was determined by using the best selection multiple linear regression. The dependent variable was Y : the number of asthmatics and the eight independent variables were X_1: carbon monoxide (CO), X_2: ozone (O_3), X_3 sulfur dioxide (SO_2), X_4: nitrogen dioxide (NO_2), X_5: particulate matter (PM10), X_6: temperature, X_7: relative humidity and X_8: rainfall. A random response Y relating to a set of independent variables $x_1, x_2, ..., x_k$ based on the multiple linear regression model is as shown in equation (1) below [9]:

$$Y = \gamma + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad (1)$$

where; $\gamma, \beta_1, ..., \beta_k$ are unknown parameters and ε is an error term factors.

Since the value of dependent variable was in a discrete form, so it needs to be transformed into an interval form. The value of Y was transformed by using the logistic regression below [10];

$$\ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_8 x_8 \quad (2)$$

where p_i : proportion of y_i

The model in equation (2) above can be expressed as follows :

$$g_i = \beta_0 + \sum_{i=1}^{8} \beta_i x_i + \varepsilon \quad (3)$$

where $g_i = \ln\left[\frac{p_i}{1-p_i}\right]$

Transformation to equation (3) needs to be done to solve the heteroscedasticity problem. When n is large, it can be shown that :

$$\hat{c}_i = \frac{1}{\sqrt{n\hat{p}_i(1-\hat{p}_i)}} \quad (4)$$

Transformation of equation (3) can therefore be expressed as in equation (5) below:

$$\frac{g_i}{\hat{c}_i} = \beta_0\left(\frac{1}{\hat{c}_i}\right) + \sum_{i=1}^{8} \beta_i\left(\frac{x_i}{\hat{c}_i}\right) + \frac{\varepsilon_i}{\hat{c}_i} \quad (5)$$

The relationship between the number of asthmatics with the air pollution and meteorological factors was determined using equation (5) above. The interaction effects were also considered as explanatory or independent variables in the model. The best model to determine the causes of increasing number of asthmatics was chosen from a selection of all possible models, based on the eight selection criteria. Each of all the possible models were run using the SPSS (Statistical Software for Social Sciences) to test the significant of the model. For each coefficient $(i = 1, 2, ..., k)$ the following test was carried out. The hypothesis to test the model was:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \textit{At least one } \beta \textit{ is nonzero}$$

Hypothesis null was rejected when F-statistic (F_c) as shown in equation (4) was greater than $F^*_{k-1,n-k,\alpha}$ [11]:

$$F_c = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 \Big/ (k-1)}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \Big/ (n-k)} \quad (6)$$

where Y_i : independent observations, $\bar{Y}_i$: mean value and $\hat{Y}_i$: estimation value of Y_i .

Then, hypothesis testing on a single independent variable was carried out to determine the significant variable in the model. The hypothesis to test the single variable was;

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Hypothesis null was rejected when t-statistic (t_c) as shown in equation (7) was greater than $t^*_{n-k,\alpha/2}$ [11]:

$$t_c = \frac{\hat{\beta}_j - \beta(H_0)}{s(\hat{\beta}_j)} \quad (7)$$

where $\beta(H_0)$ is a value of β_j under H_0 and $s(\hat{\beta}_j)$ is a standard deviation of β_j.

The corresponding independent variable was eliminated from the model when the regression coefficient β_j was not significant (p-value > $\alpha/2$). The regression equation was then rerun again with the remaining variables. When there were more than one regression coefficients not significant, the independent variable with the highest p-value was eliminated from the model. The hypothesis on single independent variable was carried out until all the independent variable was significant (p-value < $\alpha/2$). The final model with all significant variables was then called the *selected model* [12].

Based on equation (5), the estimator was obtained using the least square method where the criteria was to minimize the sum of square of error (SSE), $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. In this study, the eight criteria model selection were used to select the best model. These criteria were based on minimizing the SSE multiplied by a penalty factor. The selection criteria to select the best model are as follows [11]:

i) SGMASQ : $\left(\frac{SSE}{n}\right)\left[1-\left(\frac{m}{n}\right)\right]^{-1}$

ii) AIC : $\left(\frac{SSE}{n}\right)e^{(2m/n)}$

iii) FPE : $\left(\frac{SSE}{n}\right)\frac{n+m}{n-m}$

iv) GCV : $\left(\frac{SSE}{n}\right)\left[1-\left(\frac{m}{n}\right)\right]^{-2}$

v) HQ : $\left(\frac{SSE}{n}\right)(\ln n)^{2m/n}$

vi) RICE : $\left(\frac{SSE}{n}\right)\left[1-\left(\frac{2m}{n}\right)\right]^{-1}$

vii) SCHWARZ : $\left(\frac{SSE}{n}\right)n^{m/n}$

viii) SHIBATA : $\left(\frac{SSE}{n}\right)\frac{n+2m}{n}$

where $m = k+1$ is the number of parameter in the model and n is the number of the observations.

The best model was chosen based on the model having most of the eight criteria with the least value. The Wald Test was carried out to test whether the best model from the selected model (reduced model) was acceptable than the initial selected model (complete model)[11]. The best model and the initial possible model were expressed as in equations (8) and (9):

The complete model (best model);
$$Y=\gamma+\beta_1 x_1+...+\beta_g x_g+\beta_{g+1}x_{g+1}+...+\beta_k x_k+\varepsilon \quad (8)$$

The reduced model (initial possible model);
$$Y=\gamma+\beta_1 x_1+...+\beta_g x_g+\varepsilon \quad (9)$$

The hypothesis used to carry out the Wald Test is given;

$H_0: \beta_{g+1} = \beta_{g+2} = ... = \beta_k = 0$

H_1 : *At least one* β *is nonzero*

The null hypothesis would be rejected when the F-statistic as shown in equation (10) was greater than $F_{k-g,n-(k+1),\alpha}$;

$$F_C = \frac{\left(SSE_{Reduced\ model} - SSE_{Complete\ model}\right)/(k-g)}{\left(SSE_{Complete\ model}\right)/(n-[k+1])} \quad (10)$$

Equation (11) below was used to check the residual randomness [13]. If $z_i\,(i=1,2,...,n)$ were independent on i, then the random variable;

$$T_n = R\sqrt{\frac{(n-k)}{\left(1-R^2\right)}} \tag{11}$$

followed a t-distribution with $v = n-k$ degrees of freedom,

where $R = \dfrac{\frac{1}{n}\sum_{i=1}^{n} iz_i - \overline{z}\overline{K}}{S_z S_1}$ and

$\overline{z} = \dfrac{1}{n}\sum_{i=1}^{n} z_i$, $S_z^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(z_i - \overline{z}\right)^2$, $\overline{K} = \dfrac{n+1}{2}$ and $S_1 = \dfrac{n^2-1}{12}$

The assumption of residual randomness was met since $|T_n| < T_{\alpha,n}$.

4 Results

Result from Pearson correlation analysis showed that X_1: carbon monoxide (CO) and X_2: ozone (O3) had a positive relationship with the number of asthmatics while X_3 sulfur dioxide (SO2), X_4: nitrogen dioxide (NO2), X_5: particulate matter (PM10), X_6: temperature, X_7: relative humidity and X_8: rainfall had a negative relationship with the number of asthmatics. Table 1 below showed the results of the Pearson correlation analysis.

Table 1: Correlation between variables

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	1	0.01	0.12	-0.09	-0.19	-0.03	-0.11	-0.14	-0.23
X_1	0.01	1	0.77	0.33	0.06	0.61	0.55	0.57	-0.25
X_2	0.12	0.77	1	0.10	-0.19	0.44	0.32	0.26	-0.53
X_3	-0.09	0.33	0.10	1	0.23	0.61	0.66	0.60	-0.14
X_4	-0.19	0.06	-0.19	0.23	1	-0.02	0.16	0.20	0.34
X_5	-0.03	0.61	0.44	0.61	-0.02	1	0.65	0.55	-0.40
X_6	-0.11	0.55	0.32	0.66	0.16	0.65	1	0.96	0.035
X_7	-0.14	0.57	0.26	0.60	0.20	0.55	0.96	1	0.101
X_8	-0.23	-0.25	-0.53	-0.14	0.34	-0.40	0.04	0.10	1

Based on Table 1 above, independent variables with correlation coefficients more than 0.1 were chosen to be included in the multiple linear regression model. X_2: ozone (O3), X_4: nitrogen dioxide X_6: temperature, X_7: relative humidity and X_8: rainfall were considered to be independent variables. Since we had five independent variables, interaction variables up to the fourth order were included in the model. Using five independent variables, Table 2 below showed the the steps to determine all possible models in this study.

Table 2: The number of all possible models

Number of variables	Single	Interaction				
		1st order	2nd order	3rd order	4th order	Total
1	5	-	-	-	-	5
2	10	10	-	-	-	20
3	10	10	10	-	-	30
4	5	5	5	5	-	20
5	1	1	1	1	1	5
Total	31	26	16	6	1	80

There were 80 models to be considered in this study where the best model was selected to forecast the number of asthmatics.

The best model was chosen from the selected models by using 8SC. Table 3 below showed all the selected models with the value of 8SC.

Table 3: Value of 8SC for all selected models

Model	SGMASQ	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SHIBATA
M53.5	2.669	2.891	2.892	2.912	3.027	2.936	3.299	2.855
M54.5	2.690	2.913	2.914	2.935	3.051	2.936	3.324	2.877
M72.12	2.615	2.832	2.833	2.853	2.965	**2.936**	**3.231**	2.796
M73.6	2.198	2.718	2.747	2.931	3.120	4.894	4.038	2.473
M75.12	2.703	2.928	2.929	2.949	3.066	2.936	3.341	2.891
M78.15	2.163	2.767	2.824	3.114	3.276	6.292	4.490	2.420
M79.20	**1.965**	2.514	2.565	2.830	2.977	6.292	4.079	**2.198**
M80.23	2.006	**2.480**	**2.507**	**2.674**	**2.848**	4.894	3.685	2.256

The best model was chosen based on the model having a majority minimum value of the 8SC. Results showed that model M80.23 was chosen as the best model. The equation of the model M80.23 can be expressed as follows:

$$\hat{Y} = 13.716 - 0.731X_6 + 0.013X_8 + 0.003X_{67} + 9685.419X_{246} - 646.395X_{678} - 48.685X_{2467} + 0.048X_{24678} \tag{12}$$

Using the Wald Test on the best model, the completed model (M80) was taken as the initial possible model and M80.23 as the reduced model. This was done to test the significance of the omitted variables in the best model.

The complete model (M80):

$$Y = \gamma + \beta_2 X_2 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{24} X_{24} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} + \beta_{46} X_{46} + \beta_{47} X_{47} + \beta_{48} X_{48} + \beta_{67} X_{67} + \beta_{68} X_{68} + \beta_{78} X_{78} + \beta_{246} X_{246} + \beta_{247} X_{247} + \beta_{248} X_{248} + \beta_{267} X_{267} + \beta_{268} X_{268} + \beta_{278} X_{278} + \beta_{467} X_{467} + \beta_{468} X_{468} + \beta_{478} X_{478} + \beta_{678} X_{678} + \beta_{2467} X_{2467} + \beta_{2468} X_{2468} + \beta_{2478} X_{2478} + \beta_{2678} X_{2678} + \beta_{4678} X_{4678} + \beta_{24678} X_{24678} + \varepsilon$$

The reduced model (M80.23):

$$Y = \gamma + \beta_6 X_6 + \beta_8 X_8 + \beta_{67} X_{67} + \beta_{246} X_{246} + \beta_{678} X_{678} + \beta_{2467} X_{2467} + \beta_{24678} X_{24678} + \varepsilon$$

The hypothesis is:

$$H_0 : \beta_2 = \beta_4 = \beta_7 = \beta_{24} = \beta_{26} = \beta_{27} = \beta_{28} = \beta_{46} = \beta_{47} = \beta_{48} = \beta_{68} = \beta_{78} = \beta_{247} = \beta_{248} = \beta_{267} = \beta_{268} = \beta_{278} = \beta_{467} = \beta_{468} = \beta_{478} = \beta_{2468} = \beta_{2478} = \beta_{2678} = \beta_{4678}$$

H_1 : *Atleast one* β *is nonzero*

The F_C value as in equation (10) was 0.4251 and the F critical value was 2.544, hence null hypothesis (H_0) was accepted. Thus the best model was justified. Based on equation (11), Tn(value) equals 0.3590 and $T_{\alpha,n} = 1.697$. Since $|T_n| < T_{\alpha,n}$, the assumption of randomness residual from the best model was met.

5 Conclusion

80 possible models were considered in this study to forecast the number of asthmatics. 8SC was used to determine the best model [11]. As a result, the best model from all selected models was M80.23. This model was further justified as the best model using the Wald Test and residual randomness test. This study had found out that the first, second, third and fourth order interactions were significant in the best model. It showed that the effects of the interaction variables should be considered, as stated by [8]. However, taking a large number of independent variables also can cause the problem of multicolinearity [14]. Hence, further study on this topic is also required.

References:

[1] Peden, D.B. 2002. Pollutants and Asthma: Role of Air Toxics. *Environmental Health Perspevtives.* 110(4): 565-567.

[2] Jamal, H. H., Pillay, M. S., Zailina, H., Shamsul, B. S., Sinha, K., Zaman, H. Z., Khew, S. L., Mazrura, S., Ambu, S., Rahimah, A. dan Ruzita, M.S. 2004. A study of health impact and risk assessment of urban air pollution in the Klang Valley, Malaysia. *Buletin Kesihatan Masyarakat* 1(2): 1-11.

[3] Mar, T. F., Larson T. V., Stier, R. A., Claoborn, C. dan Koenig, J. Q. 2004. An analysis of the association between respiratory symptoms in subjects with asthma and daily air pollution in Spokane, Washington. *Medical Journal Watch* 16: 809-815.

[4] Johnston, F. H., Kavanagh, A. M., Bowman D. M. J. S. dan Scott, R. K. 2002. Exposure to bushfire smoke and asthma: An ecological study. *The Medical Journal of Australia* **176**(11): 535-538.

[5] Fusco, D., Forastiere, F., Michelozzi, Ostro, B., Arca, M. dan Perucci, C. A. 2001. Air pollution and hospital admissions for respiratory conditions in Rome, Italy. *European Respiratory Journal* 17: 1143-1150.

[6] Lewis, P.R and Corbett, S.J. Bushfires, air pollution and asthma. *The Medical Journal of Australia* **176**(11): 517.

[7] Jaccard, J., Turrisi, R., & Wan, C.K. *Interaction Effects in Multiple Regression.* Sage:Newbury Park. 1990.

[8] Pedhazur, E.J., and Schmelkin, L.P. *Measurement, design and analyisis: An integrated approach.* Erlbaum:New Jersey. 1991.

[9] Wackerly, D.D., Mendenhall III, W., and Scheaffer, R.L. *Mathematical Statistics with Applications, 6th Edition.* Thomson Learning:South Western Ohio. 2002.

[10] Abdullah, M. *Analisis Regresi*. Dewan Bahasa Dan Pustaka. Kuala Lumpur. 1994.

[11] Ramanathan, R. *Introductory Econometrics with Applications. Ed. ke-5*. Thomson Learning: South-Western Ohio. 2002.

[12] Lind, D.A., Marchal, W.G. & Wathen, S.A. *Statistical Techniques in Business & Economics.* McGraw-Hill:Boston. 2005.

[13] Ismail, B.M. 2007 Unimodality tests for Global Optimization of single variable function using statistical method. *Malaysian Journal of Mathematical Sciences.* 1(2): 1-11.

[14] Gujarati, D.N. *Basic Econometrics 3rd Edition.* McGraw-Hill: New York. 2002.