MULTIPLE LINEAR REGRESSION IN FORECASTING THE NUMBER OF ASTHMATICS

DARMESAH GABDA, ZAINODIN HJ JUBOK, KAMSIA BUDIN & SURIANI HASSAN
School of Science and Technology
University Malaysia Sabah
Locked Bag 2073, 88999 Kota Kinabalu, Sabah
MALAYSIA
darmesah@gmail.com, zainodin@gmail.com, bkamsia@ums.edu.my, suriani@ums.edu.my
http://www.ums.edu.my

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. Four significant correlation coefficient variables were considered in the multiple linear regression. There were 32 possible models considered together with the related interaction variables and the best model was obtained using the eight selection criteria (8SC). The result showed that the best model obtained could represent the cause of the rise in the number of asthmatics.

Key-Words: - multiple regression, model selection, eight selection criteria, 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 [10]. 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 ([2], [6], [7], [9]). 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 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 [5] to be taken into the model. These interaction effects represented the combined effects of the variables on the criterion or dependent measure. As [11] 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 [11].

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 variables 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 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 $x_1, x_2, ..., x_s$ based on the multiple linear regression model is as shown in equation (1) below [13]:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_s x_s + \varepsilon \quad (1)$$

where; $\beta_0, \beta_1, ..., \beta_s$ 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 defined in equation (2) [1];

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

where p_i: proportion of y_i, $p_i = y_i / N$, (N: sample size).

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

$$h_i = \beta_0 + \sum_{r=1}^{s} \beta_r x_{ri} + \varepsilon_i \quad (3)$$

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

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):

$$w_i = \beta_0\left(\frac{1}{\hat{c}_i}\right) + \sum_{r=1}^{s} \beta_r\left(\frac{x_{ri}}{\hat{c}_i}\right) + \frac{\varepsilon_i}{\hat{c}_i} \quad (5)$$

where $w_i = \frac{h_i}{\hat{c}_i}$ for $i = 1, 2, ..., n$.

The relationship between the number of asthmatics with the air pollution and meteorological factors was determined using equation (5). 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 set of all possible models, based on the eight selection criteria (8SC). 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 $(r = 1, 2, ..., s)$ in the model (5) the following test was carried out. The hypothesis to test the model was:

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

Hypothesis null was rejected when F-statistic (F_c) as shown in equation (6) was greater than $F^*_{s-1,n-s,\alpha}$ [12]:

$$F_c = \frac{\sum_{i=1}^{n}(\hat{w}_i - \bar{w})^2 \Big/ (s-1)}{\sum_{i=1}^{n}(w_i - \hat{w}_i)^2 \Big/ (n-s)} \quad (6)$$

where w_i were independent observations with $\bar{w}$: mean value and $\hat{w}_i$: estimation value of w_i for $i = 1, 2, ..., n$.

Then, hypothesis testing on a single independent variable was carried out to determine the significant variable in the model for each r $(r = 1, 2, ..., s)$. The hypothesis to test the single variable was;

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

Hypothesis null was rejected when t-statistic (t_c) as shown in equation (7) was greater than $t^*_{n-s,\alpha/2}$ [12]:

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

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

The corresponding independent variable was eliminated from the model when a regression coefficient β_r was not significant (highest p-value > $\alpha/2$). The regression equation was then rerun 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 variables were significant (p-value < $\alpha/2$). The final model with all significant variables was then called as *selected model* [8].

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), $\sum_{i=1}^{n}(w_i - \hat{w}_i)^2$. In this work, the eight criteria model selection were used to select the best model. These criteria were based on minimizing the SSE multiplied by a corresponding penalty factor. The selection criteria to select the best model are as defined in [12]:

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

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

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 $m = s+1$ is the number of parameters in the model and n is the number of observations. The best model was chosen based on the model having most number of the eight selection 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)[12]. The best model and the initial possible model were expressed as in equations (8) and (9) with d and e are error terms:

The complete model (initial possible model);

$$w = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r + \beta_{r+1} x_{r+1} + ... + \beta_s x_s + d \quad (8)$$

The reduced model (best model);

$$w = \gamma + \beta_1 x_1 + ... + \beta_r x_r + e \quad (9)$$

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

$$H_0 : \beta_{r+1} = \beta_{r+2} = ... = \beta_s = 0$$
$$H_1 : \textit{At least one of the } \beta \textit{ is nonzero}$$

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

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

Equation (11) was used to check the residual $(z_i = w_i - \hat{w}_i)$ randomness [4]. If $z_i\,(i = 1,2,...,n)$ were independent, then the random variable;

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

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

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

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

The assumption of residual randomness was met since $|T_n| < T_{\left(v,\alpha/2\right)}$. Thus, the reduced model is the best model.

3 Results

Result from Pearson correlation analysis showed that X_1: carbon monoxide (CO), X_3: sulfur dioxide (SO_2), X_4: nitrogen dioxide (NO_2), X_5: particulate matter (PM_{10}), X_6: temperature, X_7: relative humidity and X_8: rainfall had a negative relationship with the number of asthmatics while X_2: ozone (O_3) had a positive relationship with the number of asthmatics. Table 1 showed the results of the Pearson correlation analysis.

Table 1. Correlation between variables

	w	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
w	1	-0.016	0.251	-0.391[a]	-0.999[b]	-0.067	-0.361[a]	-0.459[b]	-0.104
X_1	-0.016	1	0.748[b]	0.136[a]	0.028	0.424[b]	0.127	0.185	0.040
X_2	0.251	0.748[b]	1	-0.010	-0.236	0.326	0.167	0.072	-0.076
X_3	-0.391[a]	0.136[a]	-0.10	1	0.368[a]	0.565[b]	0.730[b]	0.642[b]	-0.257
X_4	-0.999[b]	0.028	-0.236	0.368[a]	1	0.050	0.330[a]	0.431[b]	0.108
X_5	-0.067	0.424[b]	0.326	0.565[b]	0.050	1	0.591[b]	0.389[a]	-0.319
X_6	-0.361[a]	0.127	0.167	0.730[b]	0.330[a]	0.591[b]	1	0.851[b]	-0.190
X_7	-0.459[b]	0.185	0.072	0.642[b]	0.431[b]	0.389[a]	0.851[b]	1	-0.023
X_8	-0.104	0.040	-0.076	-0.257	0.108	-0.319	-0.190	-0.023	1

a correlation is significant at the 0.05 level.
b correlation is significant at the 0.01 level.

Based on Table 1, the independent variables with significant correlation coefficients were chosen to be included in the multiple linear regression model. X_3: sulfur dioxide (SO_2), X_4: nitrogen dioxide X_6: temperature and X_7: relative humidity were considered to be independent variables. Since we had four independent variables, interaction variables up to the third order were included in the model. Using the four identified independent variables, Table 2 showed the steps to determine all possible models in this work.

Table 2. The number of all possible models

Number of variables	Single	1st order interaction	2nd order interaction	3rd order interaction	Total
1	4	-	-	-	4
2	6	6	-	-	12
3	4	4	4	-	12
4	1	1	1	1	4
Total	15	11	5	1	32

There were 32 models considered in this work 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 showed all the selected models with the value of 8SC.

Table 3. Values of 8SC for all selected models

Selected Model	SGMASQ	AIC	FPE	GCV	HQ	RICE	SHIBATA	SCHWARZ
M1	0.9061	0.9564	0.9565	0.9594	0.9862	0.9628	0.9509	1.0443
M2	0.0021	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023	0.0025
M3	0.9300	0.9815	0.9816	0.9847	1.0121	0.9881	0.9759	1.0718
M4	0.8443	0.8911	0.8912	0.8940	0.9189	0.8971	0.8860	0.9731
M5	0.0015	0.0016	0.0016	0.0017	0.0017	0.0017	0.0016	0.0019
M8	0.0010	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0012
M9	0.0012	0.0012	0.0012	0.0013	0.0013	0.0013	0.0012	0.0014
M11.1	0.0008	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0010
M12	0.7389	0.8203	0.8210	0.8313	0.8722	0.8445	0.8028	0.9780
M13.2	0.8464	0.8933	0.8934	0.8961	0.9211	0.8993	0.8881	0.9754
M14	0.0004	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004	0.0005
M15	0.0004	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0006
M16.2	0.8666	0.9146	0.9147	0.9176	0.9431	0.9208	0.9094	0.9987
M21.2	0.0003	0.0003	0.0003	0.0003	0.0004	0.0003	0.0003	0.0004
M22.2	0.0004	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0006
M23.2	0.7531	0.8562	0.8577	0.8746	0.9245	0.8980	0.8287	1.0668
M24.2	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0001	0.0002
M25.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005
M26.3	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0004	0.0005
M27.4	0.7449	0.8269	0.8277	0.8380	0.8793	0.8513	0.8093	0.9860
M28.3	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0003
M30.6	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0002**	**0.0002**	**0.0001**	**0.0002**
M31.9	0.0001	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0002
M32.1	0.0001	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0002

The best model was chosen based on the model having a majority minimum value of the 8SC. Results showed that model M30.6 was chosen as the best model. Table 4 showed the estimated parameters for model M30.6 where all the corresponding p-values are less than 5%.

Table 4. The best model (M30.6)

	Unstandardized			
Variable	Coefficients	Std. Error	t	p-value
Constant	-3.615	0.359	10.067	$2.730\text{x}10^{-11}$
X_4	-1138.535	2.553	445.911	$1.384\text{x}10^{-60}$
X_6	0.067	0.005	13.984	$6.223\text{x}10^{-15}$
X_7	0.022	0.002	14.173	$4.330\text{x}10^{-15}$
X_{67}	$2.980\text{x}10^{-4}$	$1.978\text{x}10^{-5}$	-15.063	$8.267\text{x}10^{-16}$

Thus, the equation of the model M30.6 can be expressed as follows:

$$\hat{w} = -3.615 - 1138.535X_4 + 0.067X_6 + 0.022X_7 - 2.98x10^{-4}X_{67} \quad (12)$$

Using the Wald Test on the best model, the completed model (M30) was taken as the initial possible model and M30.6 as the reduced model. This was done to justify the action of removing the insignificant variables. Thus a best model is obtained.

The complete model (M30):

$$w = \beta_0 + \beta_3X_3 + \beta_4X_4 + \beta_6X_6 + \beta_7X_7 + \beta_{34}X_{34} + \beta_{36}X_{36} + \beta_{37}X_{37} + \beta_{46}X_{46} + \beta_{47}X_{47} + \beta_{67}X_{67} + \varepsilon$$

The reduced model (M30.6):

$$w = \beta_0 + \beta_4X_4 + \beta_6X_6 + \beta_7X_7 + \beta_{67}X_{67} + \varepsilon$$

The hypothesis is:

$$H_0 : \beta_3 = \beta_{34} = \beta_{36} = \beta_{37} = \beta_{46} = \beta_{47} = 0$$
$$H_1 : \textit{At least one of the } \beta \textit{ is nonzero}$$

The F_C value as in equation (10) was 1.0202 and the F critical value was 2.21, hence null hypothesis (H_0) was accepted. Thus, the reduced model was justified to be the best model. Based on equation (11), T_n equals 0.1235 and $T_{v,\alpha/2} = 2.045$. Since $|T_n| < T_{v,\alpha/2}$, the assumption of randomness residual from the best model was met.

4 Conclusion

There were 32 possible models considered in this work to forecast the number of asthmatics. The 8SC was used to determine the best model [12]. From all the selected models evaluated, it was determined that M30.6 was the best model. This model was further justified as the best model using the Wald Test and residual randomness test. This work had found out that besides several main variables, the first order interaction were significant in the best model. It showed that the effects of the interaction variables should always be considered, as stated and recommended by [11]. However, taking a large number of independent variables in a model can also cause problem of multicolinearity [3]. Hence, further work along this topic is also required.

References:

[1] Abdullah, M.: *Analisis regresi*. Dewan Bahasa Dan Pustaka, Kuala Lumpur,1994.

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

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

[4] Ismail, B.M. Siska, C.N., Yosza, D.: Unimodality tests for global optimization of single variable function using statistical method. *Malaysian Journal of Mathematical Sciences*, Vol 1, No.2, 2007, pp. 1-11.

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

[6] 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., Ruzita, M.S.: A study of health impact and risk assessment of urban air pollution in the Klang Valley, Malaysia. *Buletin Kesihatan Masyarakat*, Vol 1, No.2, 2004, pp. 1-11.

[7] Johnston, F.H., Kavanagh, A.M., Bowman, D.M.J.S., Scott, R.K.: Exposure to bushfire smoke and asthma: An ecological study. *The Medical Journal of Australia*, Vol 176, No. 11, pp. 535-538.

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

[9] Mar, T.F., Larson, T.V., Stier, R.A., Claoborn, C., Koenig, J.Q.: An analysis of the association between respiratory symptoms in

subjects with asthma and daily air pollution in Spokane, Washington. *Medical Journal Watch*, Vol 16, 2004. pp. 809-815

[10]Peden, D.B.: Pollutants and asthma: Role of air toxics. *Environmental Health Perspevtives*, Vol 110, No. 4, 2002, pp. 565-567.

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

[12] Ramanathan, R.*: Introductory Econometrics with Applications 5th Edition.* Thomson Learning, South-Western Ohio,.2002.

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