VOL. 3, NO. 2, February 2012 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

Road Accident Models with Two Threshold Levels of Fuzzy Linear Regression

[1] Lazim Abdullah, [2] Nurnadiah Zamri

Department of Mathematics, Faculty of Science and Technology
University Malaysia Terengganu, Malaysia

[1]lazim_m@umt.edu.my, [2]nadzlina@yahoo.co.uk

ABSTRACT

It has been hypothesized that number of road accidents and road casualties are increased in line with the raise in the variables of registered vehicles, population and road length. However, the effects of these variables toward road accidents are still inconclusive. Therefore, this paper develops models based on the variables which can be used to determine road accidents in Malaysia. In order to explain the effects of these variables to number of road accidents, fuzzy linear regression models with threshold level 0.5 and 0.9 are tested. Historical data of the variables from the year 1974 to 2007 were collected to test the model. The results show that by applying a multi-variable approach of fuzzy linear regression, the models provide not only crisp output but also output range for number of road accidents in Malaysia. The model with threshold level 0.5 outperformed the latter model. The variables of registered vehicles and population were notable predictors to number of road accidents in Malaysia.

Keywords: *Road accident; fuzzy linear regression; threshold level; coefficient of determination.*

1. INTRODUCTION

One of the frightening issues when discussing the traffic in Malaysian highways is road accidents. The past thirty four years have seen rapid increase of road accidents (RA) in Malaysia. In the midst of large volume of traffics on the road, accidents which involving various types of vehicles inevitably happen. Worldwide, the number of people killed in road traffic crashes each year is estimated at almost 1.2 million, while the number of injury could be as high as 50 million – the combined population of five of the world's large cities [1]. In United States, National Highway Traffic Safety Administration (NHTSA) reported over 43,200 people died while on the road in 2005 compared to 42,636 in 2004 [2]. The world body, World Health Organization also reported that the total number of road traffic deaths worldwide and injuries is forecast to rise by some 65% between 2000 and 2020 [3], [4]. In low income and middle-income countries deaths are expected to increase by as much as 80% [5]. As many other developing countries in Asia, Malaysia is no exception in facing the tragedy. Report from Royal Malaysian Police shows that traffic accident in Malaysia have been increasing at the average rate of 9.7% per annum over the last three decades [6]. Compared to the earlier days, total number of traffic road accidents had increased from 24,581 cases in 1974 to 363,319 cases in 2007. Number of fatalities (death within 30 days after accident) also increased but at slower rate compared to total road accident from 2,303 in 1974 to 6,282 in 2007. Thus, RA has become a hot topic of discussions among public and definitely a concern for all countries.

Contributing causes of RA are diverse indeed. Most accidents are said to be attributed by the fault of the driver. Mechanical causes such as brake failure, tyre burst etc. are also there in lesser numbers. Therefore, the majority of accident causes are linked with road conditions and drivers' behaviour. These statements confirm the statistics released by Department of Transport (Dft) [7]. It can be convincingly said that a single factor is not present in an accident. Accident happens as a result of numbers of complex factors jointly responsible for accident. Notwithstanding the human factors, the other factors such as road defects and vehicle defects are also occupied in discussing factors to road accidents. Moreover, a report by Highway Planning Unit, Road Safety Section, Malaysian Ministry of Works, pointed out that road condition, population and number of registered vehicles have been associated with road accident. The increase of road accidents is said to be linked with the rapid growth in population, economic development, industrialization and motorization encountered by the country. Since 1970's, Malaysia had experienced a remarkable growth in these sectors. As a result, number of registered vehicles rose to more 16 million.

The alarming figures of accidents rate involving multiple vehicles compounded with multi unexplainable factors linked to accidents motivate the need to further explore these issues. One of the most typical solutions in explaining the effect of multi factor to a single subject is mathematical modelling. A considerable amount of literature has been published on modelling of traffic road accidents. One of the most popular approaches in accidents modelling is linear prediction model. For example, Fajaruddin [8]

developed an accident prediction model for Federal Route 50 by using multiple linear regression analysis. Bener and Crundall [9] evaluated the trend of road traffic accidents problems in the United Arab Emirates and to compare these trends with other western countries like USA and UK and also a neighbouring country like Qatar. Multiple linear regression analysis was performed to determine predictor for fatalities per 10,000 vehicles.

With new development of fuzzy theory, Tanaka et al. [10] introduced a new linear regression model by integrating with fuzzy numbers. Despite the long existence of fuzzy linear regression, this method has never been tested to RA data. Linear prediction models of fuzzy linear regressions have been vastly investigated by many researchers. The first model proposed by Tanaka et al [10] with the purpose to minimize fuzziness as an optimal criterion. The method was further developed by minimizing the total spread of the output [11],[12], [13]. In 2005, Chang and Ayyub [14] proposed a new method as an extension of Tanaka et al method's by introducing linear programming model to reduce spread of the model. In their model, Chang and Ayyub [14] provide an open interpretation of threshold level *h* in solving the linear programming by passing the arbitrary of $h \in [0,1]$ to prerogative of decision makers. It looks like the choices of *h* and its effect to the forecasting efficiency was unspoken. The decision in choosing the right threshold values has heightened the need for exploring its effect to performance of the forecasting model. It is hypothesized that variations of *h* may contribute to the model efficiency.

The value of '*h*' is referred to as the threshold value between the estimated fuzzy regression model and the collected data, which thus determines the range of the possibility distributions of the fuzzy parameters [15]. Since *h* is subjectivity selected by a decision maker as an input to the model, the selection of a proper value is important in fuzzy regression [15]. The *h* values used in previous research vary widely, ranging from 0 to 0.9. Tanaka and Watada [16] suggested that the selection of the *h* value be based upon the sufficiency (sample size) of the data set available. Set *h*=0 when the data set becomes smaller compared to some ideal size. However, Savic and Pedrycz [17] recommended that the *h* value not exceed 0.9. Bardossy et al. [18] suggested that selection of an *h* value be dependent upon the decision maker's belief in the model, generally recommending an *h* value between 0.5 to 0.7. In some literature, for example, Heshmathy & Kandel, [12]; and Tanaka et al., [10] is consistently used *h*=0.5. What makes this technique attractive is the model performance is heavily depending on *h* values. The *h* values are selected arbitrarily by decision makers as an input to the model. Based on these premises, this paper intends to model Malaysian road accidents data using a computational tool of fuzzy linear regression with threshold *h*=0.5, and 0.9. Subsequently, the performance of the two models is also examined. This paper has been organized in the following way. Section II describes the fuzzy linear regression model of Chang and Ayyub [14]. Section III presents the computation for the modelling along with error analyses. Conclusions are finally drawn in Section IV.

2. FUZZY LINEAR REGRESSION

The algorithm of fuzzy regression analysis [14] is partly reproduced as to make this paper self contained.

Linearity of variables can be written as

$$\tilde{y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \tilde{A}_2 x_2 + \ldots + \tilde{A}_j x_j + \ldots + \tilde{A}_N x_N = \tilde{A}_x \tag{1}$$

where $\tilde{y}$ is the fuzzy output, $x = [x_1, x_2, \ldots, x_N]^T$ is the real-valued input vector of independent variables and $\tilde{A} = [\tilde{A}_0, \tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_N]$ is a vector of the model's fuzzy parameters. The fuzzy parameters $\tilde{A}_j = (\alpha_j, c_j)$, $j = 0,1,\ldots,N$ with its membership function as shown below:

$$\mu_{\tilde{A}_j}(a_j) = \begin{cases} 1 - \dfrac{|\alpha_j - a_j|}{c_j} & , \ \alpha_j - \alpha_j \le a_j \le \alpha_j + c_j \\ 0, & otherwise \end{cases} \tag{2}$$

where a_j is the centre value of the fuzzy number and c_j the spread.

Hence, the fuzzy linear regression model can be rewritten as follows:

$$\tilde{y} = (\alpha_0, c_0) + (\alpha_1, c_1) x_1 + (\alpha_2, c_2) x_2 + \ldots + (\alpha_N, c_N) x_N \tag{3}$$

The estimated output $\tilde{y}$ can be obtained by using the extension principle. The derived membership function of the fuzzy number $\tilde{y}$ is

$$\mu_{\tilde{y}}(y) = \begin{cases} 1 - \dfrac{|y - \alpha^T x|}{c^T |x|}, \ x \neq 0, \\ 1, \quad otherwise \\ 0, \quad otherwise \end{cases} \tag{4}$$

where $|x| = (|x_1|, |x_2|, \ldots, |x_N|)^T$, the central value of $\tilde{y}$ is $\alpha^T x$, and the spread (range) of $\tilde{y}$ is $c^T |x|$.

To determine the fuzzy coefficients $\tilde{A}_j = (\alpha_j, c_j)$, the following linear programming (LP) problem is formulated:

$$\text{Minimize} \quad J = \sum_{j=O}^{N}\left(c_j \sum_{i=1}^{M} \left| x_{ij} \right| \right) \tag{5}$$

$$\text{subject to} \quad \sum_{j=0}^{N} \alpha_j x_{ij} + (1-h) \sum_{j=0}^{N} c_j \left| x_{ij} \right| \geq y_{i,} \tag{6}$$

$$\sum_{j=0}^{N} \alpha_j x_{ij} - (1-h) \sum_{j=0}^{N} c_j \left| x_{ij} \right| \leq y_i \tag{7}$$

$c_j \geq 0$, $\alpha_j \in R$, $j = 0, 1, 2, \ldots, N$, $x_{i0} = 1$,

$i = 1, 2, \ldots, M$, $\quad 0 \leq h \leq 1$

where J is the total fuzziness of the fuzzy regression model. The h value, which is between 0 and 1, is a threshold level to be chosen by the decision maker. This term is referred to as a degree of fitness of the fuzzy linear model to its data. Each observation y_i has at least h degree of belonging to $\tilde{y}$ as $\mu_{\tilde{y}_i}(\tilde{y}_i) \geq h \quad (i = 1, 2, \ldots, M)$. Therefore, the objective of solving the LP problem is to determine the fuzzy parameters $\tilde{A}_j$ such that the total vagueness J is minimized subject to $\mu_{\tilde{y}_i}(\tilde{y}_i) \geq h \quad (i = 1, 2, \ldots, M)$. It is noted that the fuzzy regression contains all samples within its range. This indicates that fuzzy linear regression expresses all possibilities, which the samples embody and exist for the system under consideration.

3. COMPUTATIONS

This study set out three major variables affecting the total number of RA. The explanatory variables express in the model are registered vehicles (RV), road length (RL) and population (PO). These three major variables are assigned as independent variables and RA as dependent variables. The historical data of RA, RV, RL and PO from the year 1974 to 2007 were employed as input data.

The variables of the model are positioned in such way that it fulfils requirement of linearity. Correlation coefficients of variables are calculated for each explanatory variable separately in order to determine the best fit variable among these variables. Strength of relationship among the variables is presented in Table I.

Table I: Pearson Correlation Coefficients Of RA Model Variables

	RA	PO	RV	RL
RA	1.00			
PO	0.947	1.00		
RV	0.993	0.959	1.00	
RL	0.849	0.966	0.862	1.00

With the correlation coefficients more than 0.8, it shows that the relationship between RA and the variables are very strong. Thus this paper hypothesized that the variables are feasible in describing RA in Malaysia.

A fuzzy regression model with threshold value, h=0.5, 0.9 are fed into the computation. It is posited that the threshold level, h= 0.5, 0.9 may offer a different RA forecasting model.

The correlation coefficients presented in Table I indicates that the variables are fit for fuzzy linear regression model. The fuzzy regression model with these variables can be written as follows.

$$RA = (\tilde{A}_0) + (\tilde{A}_1)RV + (\tilde{A}_2)PO + (\tilde{A}_3)RL \tag{8}$$

where RA is the response variable, RV, PO and RL are the real-valued input vector of explanatory variables and $\tilde{A} = [\tilde{A}_0, \tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_N]$ is a vector of the model's fuzzy parameters. Then, all the vector of the model's fuzzy parameters can be changed as $\tilde{A}_j = (\alpha_j, c_j)$, $j = 0, 1, \ldots, N$ that has been shown as below:

$$RA = (c_0, p_0) + (c_1, p_1)RV + (c_2, p_2)PO + (c_3, p_3)RL \tag{9}$$

Fuzzy numbers $c_0 \ldots c_3$ are the centres and $p_0 \ldots p_3$ are the spreads of the model. By solving the linear programming problem (equations 5, 6 and 7), the fuzzy coefficients are obtained. Table II shows the centres and spreads of the model when h= 0.5.

Table II: Coefficients Of Three Variables Regression Model (H=0.5)

Type	Name	Value
Center	c_0	35436.695312
	c_1	0.000000
	c_2	0.000220
	c_3	0.000000
Spread	p_0	458.141052
	p_1	0.022814
	p_2	0.000000
	p_3	0.000000

Therefore, the fuzzy regression model with threshold level, h=0.5 for three independent variables is written as

$$RA = (35436.695312,458.141052) + (0,0.022814)RV + (0.000220,0)PO + (0,0)RL \quad (10)$$

The value of $(35436.695312,458.141052)$ is a fuzzy intercept and (0.022814, 0), (0, 0.000220) and (0, 0) are the fuzzy slopes. Surprisingly, the coefficients tell that the variable of RL was not determined to have a notable effect on RA.

With the similar fashion, the fuzzy linear regression with threshold level, 0.9 is obtained. Table III shows the centres and spreads of the model when h= 0.9

Table III: Coefficients Of Three Variables Regression Model (H=0.9)

Type	Name	Value
Center	c_0	19687.054688
	c_1	0.000000
	c_2	0.000122
	c_3	0.000000
Spread	p_0	458.141052
	p_1	0.022814
	p_2	0.000000
	p_3	0.000000

Similarly, the fuzzy regression model with threshold level, h=0.9 for three independent variables is written as

$$RA = (19687.054688,458.141052) + (0,0.022814)V + (0.000122,0)PO + (0,0)RL$$

The value of $(19687.054688,458.141052)$ is a fuzzy intercept and (0, 0.022814), (0.000220, 0) and (0, 0) are the fuzzy slopes. Again, the coefficients tell that the variable of RL was not determined to have a notable effect on RA.

Based on the model and Equation (8), the value of Y_i^L and Y_i^U are obtained. Therefore, graphs for upper, lower, centre are presented in Figure I and Fig II.

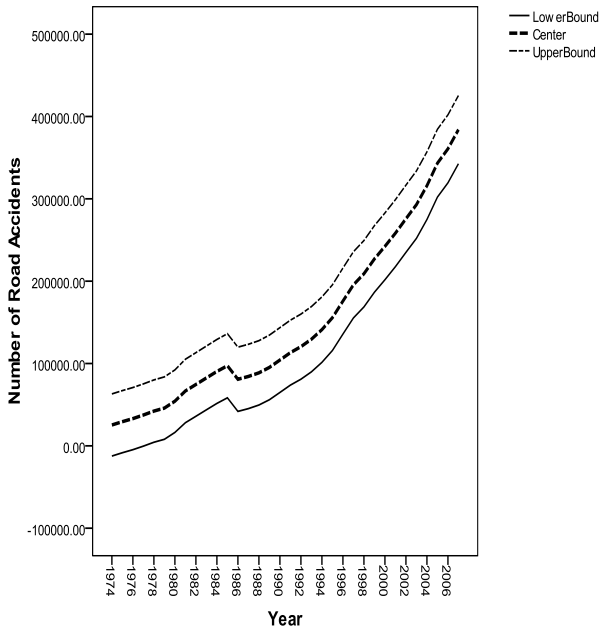

Figure I: Upper and Lower Bound of the Fuzzy Regression Model With H=0.5

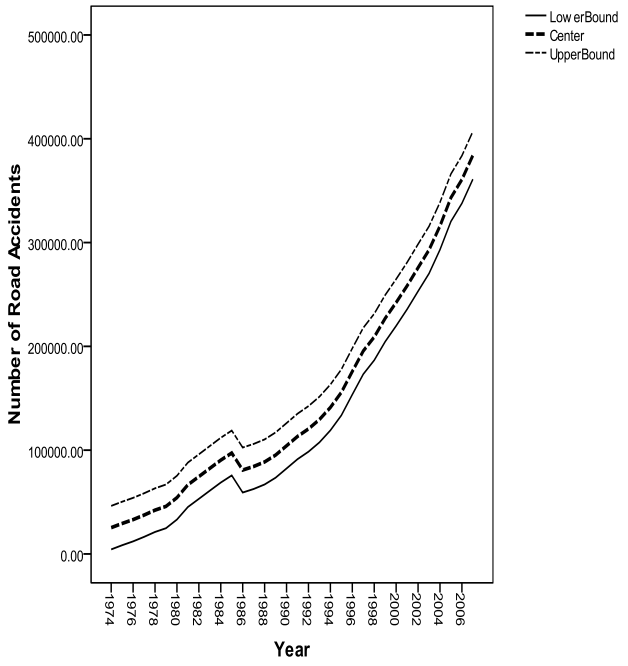

Figure II: Upper and Lower Bound of the Fuzzy Regression Model With H=0.9

The model explains the behaviour of RA in crisp output and also output range. The upper and lower bounds of each model explain the fuzziness of accidents data. However the trend lines with centres and spreads are far from conclusive in knowing the performance of the model. Therefore the models need to be examined further. One of the popular analyses in testing model performances in linear predicting model is error analysis. The models are compared with the observed data to see the magnitude of errors. Performance of the model with the two threshold levels are scrutinized. Coefficient of determination (R^2) is computed to indicate the model's performance. It is found that the two threshold levels, h=0.5, 0.9 yield different performances. The R^2 for h= 0.5 and h= 0.9 are 0.9158796 and 0.7662332 respectively. The variations in RA of the two models are approximately explained by 92 percent and 76 percent variations in RV, RL and PO. Therefore the fuzzy linear regression model with h= 0.5 is performed better than the other model.

4. CONCLUSIONS

The research was designed to determine the model of road accident. A fuzzy regression model was developed in describing road accident in Malaysia over the period of 1974 to 2007 using three predictors. The threshold level h=0.5, 0.9 were accounted in this paper. The model structure was developed after considering linearity assumption of the model's variables. The model for road accident uses registered vehicles, population and road length as variables based on one response variable which is road accident. The model shows that the variables of registered vehicles and population provide higher impact to the number of road accident. Perhaps the policy makers could consider these two variables in managing road accident in Malaysia.

ACKNOWLEDGMENT

This research was supported by Malaysian Ministry of Higher Education under the Fundamental Research Grant Scheme, No. 59172.

REFERENCES

[1] M . Peden, R.Scurfield, , D. Sleet, D. Mohan, A. A., Hyder, E.Jarawan, ,and C. Mathers, World report on road traffic injury prevention. Geneva: World Health Organization. 2004.

[2] NHTSA. http://www.nhtsa.dot.gov/poRAl/site/nhtsa/menuitem. Accessed on 08 April 2008.

[3] E. Kopits, and M. Cropper, Traffic fatalities and economic growth (Policy Research Working Paper No. 3035). Washington: The World Bank. 2003.

[4] C.J.L Murray, and A.D. Lopez, The global burden of disease: a comprehensive assessment of moRAlity and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Boston: Harvard School of Public Health. 1996.

[5] T. Toroyan and M. Peden,. Youth and road safety. Geneva: World Health Organization. 2007.

[6] Royal Malaysia Police, Statistics of road accident and death. http://www.rmp.gov.my/rmp . Accessed on April 08, 2008.

[7] Department for Transport. DfT [Online]. Available: http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dfttransstats_032188.pdf. Accessed on Jun 03, 2008 .

[8] M, Fajaruddin, Basil. Blackspot Study and Accident Prediction Model Using Multiple Liner Regression. First International Conference on Construction In Developing Countries (ICCIDC–I) Advancing and Integrating Construction Education, Research & Practice August 4-5, 2008, Karachi, Pakistan.

[9] A. Bener, and D.Crundall. Road Traffic Accidents in The United Arab Emirates Compared to Western Countries. *Journal of Advances in Transportation Studies*, Section A 6, 2005.

[10] H.Tanaka, , S.Uejima, , and K.Asai, . Linear regression analysis with fuzzy model. *IEEE Transactions on System, Man and Cybernetics,* 12, 903–907. 1982.

[11] P. T. Chang, S. A Konz,. and E. S. Lee, Applying fuzzy linear regression to VDT legibility. *Fuzzy Sets and Systems*, 80, 197–204. 1996.

[12] B. Heshmati, and A. Kandel, Fuzzy linear regression and its applications to forecasting in unceRAin environment. *Fuzzy Sets and Systems*, 15, 159–191. 1985.

[13] G. Peters, Fuzzy linear regression with fuzzy intervals. *Fuzzy Sets and Systems*, 63, 45–55, 1994.

[14] Y-H. O Chang,and B. M.Ayyub, Reliability analysis in fuzzy regression. *Proceeding of the Annual Conference of the North America Fuzzy Information Processing Society* (pp. 93–97). Allentown, PA, US. 1993.

[15] S.Ö. Kemal, Modeling car ownership in Turkey using fuzzy regression. *Journal of TranspoRAtion and Technology*, 29(3), 233-248. 2006.

[16] H. Tanaka, and J. Watada, Possibilistic linear systems and their application to the linear regression model. *Fuzzy Sets and Systems*, 27, 275–289. 1988

[17] D.Savic, , and D. Pedrycz, Evaluation of fuzzy regression models. *Fuzzy Sets and Systems*, 39, 51–63. 1991.

[18] A.Bardossy, , R.Hagaman, , L.Duckstein, , and I. Bogardi, Fuzzy least squares regression: theory and application. In J. Kacprzyk, M. Fedrizi, (Eds.), Fuzzy regression analysis (pp.181-193). Omtech Press, Warsaw and Physica-Verlag, Heidel-bergpp, 1992.