Applied Mathematical Sciences, Vol. 9, 2015, no. 56, 2749 - 2759
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.5175

The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malaysia

Muhammad Ammar Shafi

Faculty of Science, Technology and Human Development
University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor
Malaysia

Mohd Saifullah Rusiman

Faculty of Science, Technology and Human Development
University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor
Malaysia

Abstract

Regression analysis has become popular among several fields of research and standard tools in analysing data. This structure was represented by four commonly statistical models such as multiple linear regression, fuzzy linear regression (Tanaka, 1982), fuzzy linear regression (Ni, 2005) and extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012). Colorectal cancer (CRC) was applied on CRC cases in Malaysia. The CRC patients' quality of life in order to detect the CRC at an early stage is still very poor, the programmes are mainly ad-hoc and not implemented as a national wide programme. This study aims to determine the best model to measure the tumor size at hospitals using mean square error and root mean square error. Secondary data was used where 180 patients having colorectal cancer and receiving treatment in hospitals was recorded by nurses and doctors. Based on the results, fuzzy linear regression (Ni, 2005) is the best model to predict the tumor size developed by patients after receiving treatment in hospital.

Keywords: Fuzzy Linear Regression, extended fuzzy linear regression by benchmarking models under fuzziness, multiple linear regression, mean square error, root mean square error

1. Introduction

Regression analysis has become one of the standard tools in analysing data. Its popularity is due to several reasons. The mathematical equation gained from its analysis could explain relationship between the dependent and independent variables. It provides much explanatory power, especially due to its multivariate nature. It is widely available in computer packages and easy to interpret. It has been widely used in applied sciences, economic, engineering, computer, social sciences and other fields (Agresti, 1996).

Nonlinear modelling is of interest to many researchers in modelling statistics, rather than linear modelling. The functional form obtained should be approximately near to the real data. If the functional form is far away from the real data, its mean the estimation is inconsistent, bias and so on. However, other difficulties may arise with a non-linear approach (Rousseeuw et. al, 2004). Nowadays, there are many models resulting from the regression analysis such as multiple regression, quadratic regression, cubic regression, logit model, probit model, exponential model, growth model, neural network regression and fuzzy regression. Statistic methods are suitable for use in the medical area. Other researchers in Malaysia conducted research in medical statistics such as mortality of patients in Intensive Care Unit (ICU) (Muhammad Ammar et. al, 2014; Mohd Saifullah et. al, 2012).

Colorectal cancer is cancer of the colon and rectum. The colon and rectum are two parts in the human body that play important roles in digesting food and producing waste. Colorectal cancer is one of the most common diseases malignancies in the world (Malaysian Oncological Society, 2007). According to World Health Organization in 2012, colorectal cancer is the fourth leading cause of death among cancer sufferers. This cancer is a rising threat in many countries especially in the Asian Region including Malaysia.

Colorectal cancer is the third leading cause of cancer deaths in Malaysia. World Health Organization in 2012 stated lung cancer has the highest death rate caused by cancer at 17.93%. It is followed by breast cancers at 15.83% and colorectal cancer is third in the list at 13.10% rates of cancer deaths. Data from the Ministry of Health of Malaysia in 1995 indicate an increase in colorectal cancer admission rates from 8.1% to 11.9%. Recent studies have shown increasing colorectal cancer (CRC) cases in Asian population (Wendy, 2008; Radzi, 2008). According to the Second Paper of the National Cancer Registry (2003), 14.2% of male in Malaysia suffer from colorectal cancer while for female, it is 10.1%. It can be concluded that cancer is more common among men than women. CRC comprises of four stages which ranging from an early stage, second stage, third stage and final stage. Stage I refers to when the cancer is confined to the inner lining of the colon or rectum, stage II indicates that the cancer has spread

through the wall of the colon or rectum, stage III is when the cancer spreads to nearby lymph nodes and lastly stage IV whereby the cancer spreads to distant parts of the body, such as the liver or lungs (Malaysian Oncological Society, 2007).

There are twenty five variables for independent variable and one variable for dependent variable in this study. Twenty five variables includes gender, ethnic, age, icd10, TNM (Tumor, Nodes, Metastasis) staging, family history of colon cancer, diabetes mellitus, crohn's disease, ulcerative colitis, polyp, history of cancer(s), endometrial, gastric, small bowel, hepatobiliary, urinary tract, ovarian, other cancer, intestinal obstruction, colorectal, weight loss, diarrhoea, anaemia, blood stool and abdominal pain. Finally, the dependent variable is tumour size (mm). This study aims to determine the best model to measure the tumor size at general hospital at Kuala Lumpur. The twenty five independent variables also were analysed to look the relationship toward the tumor size.

2. Methodology

The statistical software SPSS 20, Excel 2010 and Matlab 2008 were used to analyse the data. Exploratory data analysis was used to explore the behaviour of the data. Some explorations on the demographic characteristic were applied. For categorical variables, data was presented in frequency and percentage distribution while for continuous variable, descriptive statistics measured by mean or average, standard deviation and correlation of variables were applied. The parameter and model were stated as:

Multiple Linear Regression

Regression analysis was developed by Sir Francis Galton in 19th century. Galton had studied the relation between heights of parents and children and he noted that the heights of children of both tall and short parents appeared to "revert" or "regress" to the mean group. He considered this tendency to be a regression to "mediocrity". He developed a mathematical description of this regression tendency. The term regression persists to describe statistical relations between variables (Kutner, 2004).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik} + \varepsilon_i(\beta), \quad i=1,\ldots,N \tag{3.1}$$

$$\text{or } Y = X\beta + \varepsilon \tag{3.2}$$

The function for least squares method is,

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = S(\beta) = \sum_{j=1}^{d} \varepsilon_j^2 \quad \text{or } \varepsilon^T \varepsilon$$

From (3.1), $\varepsilon(\beta) = Y - X\beta$

Then, $S(\beta) = (Y - X\beta)^T (Y - X\beta)$

$= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$

To minimize $S(\beta)$, we have to differentiate $S(\beta)$ with respect to β where $\frac{\delta S}{\delta \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}}$ is equal to 0,

$$\frac{\delta S}{\delta \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{Y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = 0$$

Hence, the least squares estimator is,

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\ \mathbf{Y}$$

The values fit by the equation $\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$ are denoted *i*, and the residuals ε_i are equal to $y_i - \hat{y}$, the difference between the observed and fitted values.

Fuzzy Linear Regression (Tanaka 1982)

Hideo Tanaka (1982) is the researcher who first proposed the fuzzy linear regression. Fuzzy model existed where human estimation and some systems is influential and should deal with fuzzy structure. In order to estimate the fuzzy parameters $A_e^* = (\alpha_e, \varsigma_e)$ the following linear programming problem should be solved.

$$\min_{\alpha,\varsigma} = \varsigma_1 + ... + \varsigma_g$$

subject to $\varsigma \geq 0$ and

$$\alpha^T x_e + (1-H)\sum_f \varsigma_f \mid x_{ef} \mid \ \geq\ y_e + (1-H)\varepsilon_e$$

$$-\alpha^T x_e + (1-H)\sum_f \varsigma_f \mid x_{ef} \mid \ \geq\ -y_e + (1-H)\varepsilon_e \qquad (3.3)$$

We can obtain the best fitted model for the given data by solving the conventional linear programming problem in (3.3).

The fuzzy linear regression model (FLRM) can be stated as,

$$Y = A_0(\alpha_0, \varsigma_0) + A_1(\alpha_1, \varsigma_1)\ x_1 + \ .\ .\ .\ + A_g(\alpha_g, \varsigma_g)\ x_g \qquad (3.4)$$

Fuzzy Linear Regression (Ni, 2005)

This model has been approached by Yongshen Ni on 2005 and this model is an extension approach from fuzzy linear regression model by Hideo Tanaka on 1982. Ni (2005) proposed the fuzzy linear regression model as,

$$\min_{\alpha,\varsigma} = \varsigma_1 + ... + \varsigma_g$$

Subject to $\varsigma \geq 0$ and

$$\alpha^{\mathrm{T}} x_e + (1-H)\sum_f \varsigma_f \mid x_{ef} \mid \; \geq \; y_e$$

$$-\alpha^{\mathrm{T}} x_e + (1-H)\sum_f c_f \mid x_{ef} \mid \; \geq \; -y_e \qquad (3.5)$$

Extended FLR by Benchmarking Models under Fuzziness (Chung, 2012)
William Chung proposed the Extended FLR by benchmarking models under fuzziness as,

$$\min \frac{xjk-\bar{x}k}{sk} \leq dk \leq \max \frac{xjk-\bar{x}k}{sk} \qquad (3.6)$$

where

xjk is the sample value

$\bar{x}k$ is the sample average

sk is the standard deviation

dk is the inequality either negative values or positive values

subject to equation:

$$\hat{\mathrm{Y}}(x) = \mathrm{A0} + \mathrm{A1}\,(\frac{x1-\bar{x}1}{s1}) + \mathrm{A2}\,(\frac{x2-\bar{x}2}{s2}) + \ldots\ldots + \mathrm{An}\,(\frac{xn-\bar{x}n}{sn}) \qquad (3.7)$$

3. Results

Demographically, the majority showed that most colorectal cancer of patients in general hospital at Kuala Lumpur was Malay and male. Most of them were in the average of 61 years old and cancer colon for most patients is in the colon side.

3.1. Multiple Linear Regression

Firstly, assumptions of multiple linear regression should be fulfilled before the data is analyzed. This assumes constant variance, normality and multicollinearity. All the assumptions were satisfied and hence, the results will be trustworthy.

Multiple linear regression is one of the common models used by statistician especially in medical health. This model was used to study and analyze twenty five predictor variables and to further detecting colorectal cancer stage. After the analysis was done; only eleven predictor variables were significance for

colorectal cancer. The significant variables were age at diagnosis, icd10 site, TNM staging, family history, crohn's disease, history of cancer, gastric, ovarian, intestinal obstruction, anaemia and abdominal. The significance of variables was measured by *p*-value which must be less than 0.05.

Other than that, ANOVA analysis had been conducted. The results show that the mean square error term is 129.558 and RMSE is 11.3826. The *p*-value for the F test statistic was less than 0.05, indicating strong evidence of alternative hypothesis against the null hypothesis. The squared multiple correlation R^2 = SSR/SST = 21396.590/41348.550= 0.517, indicate that 51.7% of the variability in the tumour size variable is explained by the twenty five independent variables.

All the significant variables influence the colorectal cancer effects. The estimated multiple linear regression models for symptoms and factors of colorectal cancer model are as follow:

$\hat{Y}$ = 76.056 + 0.421 age + 3.459 icd10 + 0.961 TNM Staging – 16.738 family history + 5.035 Crohn's disease + 5.557 history of cancer – 6.517 gastric + 12.865 ovarian – 4.350 intestinal obstruction – 7.943 anaemia – 3.994 abdominal.

3.2. Fuzzy Linear Regression (Tanaka 1982)

In this study, fuzzy linear regression by Tanaka (1982) was used to model tumor size faced by patients. The evaluation of fuzzy linear regression by Tanaka (1982) was based on mean square error and root mean square error. Mean square error for this model was 142.794 and root mean square error was 11.950.

The estimated fuzzy linear regression (Tanaka 1982) model for colorectal cancer of patients is as follow:

$\hat{Y}$ = -4.550 + (6.417, 0.426) gender – (0.008, 0) age + (5.545, -0.589) ethnic + (8.794, 0.539) icd10 + (2.675, 0.473) TNM Staging + (16.992, 2.932) family history + (0.320, 1.197) diabetes mellitus – (2.810, -0.745) Crohn's disease – (2.395, -1.611) ulcerative colitis + (3.291, -1.013) polyp – (3.980, 0.117) history of cancer + (1.127, 0.041) endometrial + (8.001, -0.465) gastric + (6.299, 0.661) small bowel + (1.972, -1.220) hepatobiliary – (2.859, 0.659) urinary tract – (11.855, -1.157) ovarian + (0.499, 0.604) other cancer + (7.482, -0.148) intestinal obstruction – (1.422, -0.121) colorectal – (1.402, 1.004) weight loss + (2.072, -2.013) diarrhoea + (10.915, 1.613) blood stool + (6.392, -0.499) anaemia + (8.790, 0.755) abdominal.

3.3. Fuzzy Linear Regression (Ni, 2005)

The model has been employed by this study to detect tumor size among patients. Twenty five of predictor variables had been used against tumour size (mm).

The evaluation of fuzzy linear regression by Ni (2005) was also mean square error and root mean square error. Mean square error for this model was 110.844 and root mean square error was 10.528.

The estimated fuzzy linear regression parameter by (Ni, 2005) for colorectal cancer of patients is as follow:

$\hat{Y}$ = -8.50e-14 + (6.417, 6.144e-14) gender – (0.008, 5.149e-16) age + (5.545, -4.767e-14) ethnic + (8.795, -6.025e-14) icd10 + (2.675, -5.659e-15) TNM Staging + (16.991, -2.468e-14) family history + (0.320, 1.732e-14) diabetes mellitus – (2.810, -1.818e-14) Crohn's disease –(2.394, 4.363e-15) ulcerative colitis + (3.291, -7.960e-15) polyp – (3.979, -2.942e-15) history of cancer + (1.127, 1.369e-14) endometrial + (8.001, 2.759e-15) gastric + (6.299, -1.873e-14) small bowel + (1.973, -3.666e-16) hepatobiliary – (2.859, 1.236e-14) urinary tract – (11.855, 9.250e-15) ovarian + (0.498, -4.120e-15) other cancer + (7.428, -2.677e-14) intestinal obstruction – (1.422, 1.850e-14) colorectal - (1.402, 1.746e-14) weight loss + (2.072, -2.706e-14) diarrhoea + (10.915, 1.990e-14) blood stool + (6.392, -1.123e-14) anaemia + (8.790, -9.496e-16) abdominal.

3.4. Extended Fuzzy Linear Regression By Benchmarking Models Under Fuzziness (Chung, 2012)

Twenty five of predictor variables had been used against tumour size (mm). Extended fuzzy linear regression benchmarking models under fuzzy environment can be developed by means of fuzzy linear regression among variables. A normalization of coefficient is applied to this model to get better results.

There were twenty five independent variables analysed and a new coefficient produced and mean square error for the benchmarking under fuzziness model colorectal cancer was 802.3071. Furthermore, root mean square error value was 28.32503.

The estimated parameter of fuzzy linear regression benchmarking under fuzziness model for colorectal cancer of patients is as follow:

$\hat{Y}$ = 76.056 + 715.266 gender - 3.806 age + 615.218 ethnic + 1029.533 icd10 + 85.430 TNM Staging - 9959.389 family history + 326.564 diabetes mellitus + 1962.528 Crohn's disease + 557.707 ulcerative colitis - 572.597 polyp + 2106.910 history of cancer - 325.925 endometrial - 2405.275 gastric - 1244.425 small bowel -402.856 hepatobiliary + 960.352 urinary tract + 5241.343 ovarian + 1172.459 other cancer - 1589.287 intestinal obstruction + 442.145 colorectal + 1106.008 weight loss + 277.794 diarrhoea - 2844.003 blood stool - 1096.827 anaemia - 1429.962 abdominal.

3.5. Summary of Results

Table 1: Summary of models

Models Of Linear Regression	MSE	RMSE
Multiple Linear Regression	129.558	11.382
Fuzzy Linear Regression (Tanaka 1982)	142.794	11.950
Fuzzy Linear Regression (Ni, 2005)	110.844	10.528
Extended FLR By Benchmarking Models Under Fuzziness (Chung, 2012)	802.3071	28.325

Based on Table 1, the present study proves that fuzzy linear regression (Ni, 2005) model is the best model to predict tumor size faced by patients in general hospital around Kuala Lumpur. Furthermore, analysis of error was based on two methods, mean square error and root mean square error. It was also used as deciding criteria between multiple linear regression, fuzzy linear regression (Tanaka 1982), fuzzy linear regression (Ni, 2005) and extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) model. In conclusion, fuzzy linear regression (Ni, 2005) is the best model with the smallest value of error which are 110.844 (MSE value) and 10.52 (RMSE value).

4. Conclusions

This study applied four models of linear regression which were multiple linear regression, fuzzy linear regression (Tanaka 1982), fuzzy linear regression (Ni, 2005) and extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) model. The best model to predict tumor size were based on the result of mean square error and root mean square error. Fuzzy linear regression (Ni, 2005) model proved to be the best model in predicting stages of colorectal cancer patients in General Hospital around Kuala Lumpur.

It can be concluded that the tomour size is directly proportional to gender, ethnic, icd10, TNM staging, family history, diabetes mellitus, polyp, endometrial, gastric, small bowel, hepatobiliary, other cancer, intestinal obstruction, diarrhoea, blood stool, anaemia and abdominal. Furthermore, the tomour size is inversely proportional to age, crohn's disease, ulcerative colitis, history of cancer, urinary tract, ovarian, colorectal and weight loss. The vagueness of tomour size can be represented as the fuzziness of the constant parameter gender, ethnic, age, icd10,

TNM staging, family history of colon cancer, diabetes mellitus, Crohn's disease, ulcerative colitis, polyp, history of cancer(s), endometrial, gastric, small bowel, hepatobiliary, urinary tract, ovarian, other cancer, intestinal obstruction, colorectal, weight loss, diarrhoea, anaemia, blood stool and abdominal pain. Mean square error value is 110.844 and root mean square error is 10.528.

Acknowledgements. The research work is supported by ERGS (Exploratory Research Grant Scheme) grant (vot E020), Ministry of Higher Education, Malaysia.

References

[1] Agresti, A. 1996. An Introduction to Categorical Data Analysis.*New York: John Wiley & Sons, Inc.*

[2] Bin Shafi, Muhammad Ammar; Bin Rusiman, Mohd Saifullah; Che Yusof, Nur Syaliza Hanim. 2014. Determinants status of patient after receiving treatment at Intensive Care Unit: A case study in Johor Bahru.
http://dx.doi.org/10.1109/i4ct.2014.6914150

[3] Center M. M., Jemal A., Ward E. 2009. International trends in colorectal cancer incidence rates. Cancer Epidemiol Biomarkers, 19, 1688-94.
http://dx.doi.org/10.1158/1055-9965.epi-09-0090

[4] Dubois, D. J. and Henry, P. 1980. Fuzzy set and systems: Theory and applications. Academic Press Inc.

[5] Dunn, J. 1974. A fuzzy relative of the ISODATA process and its use in detecting compact well separated cluster. Cybernetics 3(3):32-57.
http://dx.doi.org/10.1080/01969727308546046

[6] Hideo Tanaka et al. (1982), Linear Regression Analysis with Fuzzy Model. Transactions on systems; vol.smc-12, 903-907.
http://dx.doi.org/10.1109/tsmc.1982.4308925

[7] Kutner er al. Applied Linear Statistical Models. Fifth Edition, 2004.

[8] Jemal A, Siegel R, Ward E. 2008. Cancer Statistics. C.A: A Cancer J Clin, 58, 71-96. http://dx.doi.org/10.3322/ca.2007.0010

[9] Malaysian Oncological Society Novartis Corporation (Malaysia) Sdn. Bhd. The Lancet Oncology 2007: 8: 773-783. 6. American.

[10] Ministry Of Health, Malaysia. HEALTH FACTS 1995. Information and Documentation System Unit. Planning & Development Division.

[11] Mohd Saifullah Rusiman, Robiah Adnan, Efendi Nasibov & Kavikumar Jacob (2012). Adjustment of an Intensive Care Unit (ICU) Data in Fuzzy C-Regression Models. Journal of Science and Technology. Vol 4 (2), pp. 99-108.

[12] Natrah MS, SharifaEzat WP, Syed MA, Mohd Rizal AM, Saperi S. 2012. Quality of Life in Malaysian Colorectal Cancer Patients: A Preliminary Result. APJCP, 13, 1-6.
http://dx.doi.org/10.7314/apjcp.2012.13.3.957

[13] National Cancer Registry, Ministry of Health Malaysia. 2006. Malaysian Cancer Statistics – Data and Figure Peninsular Malaysia 2006.

[14] The First Annual Report of the National Cancer Patient Registry-Colorectal Cancer 2007-2008. 2010.

[15] Rousseeuw, P. J., S. Van Alest, K. Van Driessen, and J. Agulló (2004). Robust Multivariate Regression, Technometrics 46 293–305.
http://dx.doi.org/10.1198/004017004000000329

[16] Second Paper of the National Cancer Registry, 2003, Case Studies on Decision for Cervical Cancer Screening among Working Women, 4(2).

[17] Tanaka, H., S. Uejima and K. Asai, 1982. Linear regression analysis with fuzzy model. IEEE Trans.Man. Cybernet, 12 (6): 903-907.
http://dx.doi.org/10.1109/tsmc.1982.4308925

[18] Yongshen Ni (2005), Fuzzy Correlation and Regression Analysis. University of Oklahoma Graduate College; UMI number: 3163014.

[19] Wendy & Radzi, 2008, MED J MALAYSIA VOL 63 NO 4 OCTOBER 2008. EDITORIAL Cell Therapy Centre, Universiti Kebangsaan Malaysia Medical Centre.

[20] William Chung (2012), Using the Fuzzy Linear Regression Method to Benchmark the Energy Efficiency of Commercial Buildings. Applied Energy; 45-49. http://dx.doi.org/10.1016/j.apenergy.2012.01.061

[21] World Health Organization 2012, Publications of the World Health Organization are available on the WHO web site (www.who.int).

[22] World Health Organization Data 2010, Publications of the World Health Organization are available on the WHO web site (www.who.int).

Received: February 7, 2015; Published: April 3, 2015