Journal of Science and Technology, Vol. 8 No. 2 (2016) p. 1-5

The use of Linear Statistical Model to Predict Tumour Size of Colorectal Cancer

M. A. Shafi[1], M. S. Rusiman[1*], N. S. A. Hamzah[1], E. N. Nasibov[2], N. A. H. M. Azmi[1]

[1]Faculty of Science, Technology and Human Development, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia.
[2]Department of Computer Science, Faculty of Science, Dokuz Eylul University 35160 Buca, Izmir, Turkey

Abstract: Colorectal cancer (CRC) is a type of cancer in the large intestine (colon), the lower part of our digestive system. Most cases of colon cancer begin as small non-cancerous clumps of cells called adenomatous polyps. The aim of this quantitative study is to identify the determinants of patient who have colorectal cancer symptoms in general hospital. The sample study included 180 patients who have colorectal cancer aged above 21 years old and received treatment at general hospital in Kuala Lumpur, Malaysia. Secondary data were obtained through doctors and nurses using cluster sampling. Based on the results of multiple linear regressions (MLR), 11 predictor variables were significant to predict tumour size of colorectal cancer. The statistical measurement error used were mean square error (MSE), root mean square error values (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

Keywords: Multiple Linear Regression, mean square error (MSE), root mean square error values (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

1. Introduction

The types of cancer in medical record reach more than 100 different diseases. One of the top three of cancer diseases faced by patient includes colorectal cancer (CRC). CRC is the most common cancer in Malaysia by 13.2% compared to the Western Countries [1,2,3]. It is noticeable that this cancer has been spreading in Malaysia during recent decades. In Malaysia, generally among citizens above 50 years’ old diagnosed with colorectal cancer, there are more male cases than female cases [4,5,6]. It contributes to higher demand of bed’s patient due to neoplasm. Neoplasms may have begun with no cancer then malignant (cancer). They are also called as a tumour [3].

At present, the cause of colorectal cancer is not completely understood and nonspecific. It may involve numerous risk factors which contribute to developing colorectal cancer. The common factors identified amongst patients are age, polyps, family history, diet, tobacco and body weight [7]. Moreover, causes of CRC typically include gastric, ovarian, small bowel and others.

The stages of CRC are based on the size of tumour developed by patient. Physical examination, biopsy and imaging test are the process of finding out the size of tumour. Colorectal cancer has four stages which are stage I and II that are called as earlier stage, while stage III and IV are called final stage. Publicly known, stage I means the condition where the cancer is kept to the inner wall of the colon, stage II means the cancer has blow-out through the wall of the colon, stage III means cancer has blow-out the blood vessel until lymph node and stage IV means the cancer has blow-out to other organs. The stages of CRC play an important factor in determining treatment options [7,8].

The guidelines to reduce and prevent development of colorectal cancer among individuals were introduced in 2017. It discussed about the screening to detection of cancer, radiotherapy to kill cancer cells and treatment or surgery of colorectal cancer. Unfortunately, less than 40% of individuals in their early stages take immediate action and keep in mind about the risk of colorectal cancer [2,5]. However, the awareness of colorectal cancer screening in Malaysia is still lower than other cancer, especially cervical cancer. This probably gives higher impact to patients in an advance stage [9].

1.1 Background of regression

Regression analysis is statistical methodology that utilizes the relationship between two or more quantitative variables. The outcome of the variables can be predicted based on data in regression analysis. This methodology can be applied in

*Corresponding author: saifulah@uthm.edu.my

penerbit.uthm.edu.my/ojs/index.php/jst

business analysis, social analysis, biological sciences and other disciplines [10].

A few examples of applications are; (i) sales of product can be predicted through relationship between sales and amount of profit, (ii) the performance of employee can be predicted through relationship between worker performance and attitude test and (iii) the length of time stay in shopping mall can be predict by relationship between the money spending and buy items. All these applications show that the main objective in regression analysis is predicting.

Recently, there are many models used by researchers in regression analysis such as multiple linear regression (MLR), logic model, fuzzy regression and others. There are many researchers studied about predicting tumour size of colorectal cancer of patients using fuzzy linear regression [1,11,12,13].

1.2 Aim of study

The aims of this study are to identify which determinants give the higher impact to develop colorectal cancer stages and to determine whether MLR is a good model to predict tumour size of colorectal cancer by statistical measurement error such as mean square error (MSE), root mean square error values (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

2. Material and Method

2.1 Material

The data of this study is defined as the patients aged from 21 until 90 years old who show symptoms and suffer from colon cancer in any of the four stages and are receiving treatment at general hospital, Kuala Lumpur. The patients include male and female from various ethnics.

2.2 Method

In this study, MLR model are used toward CRC data. There are some conditions and assumptions that need to be determined first. Before using the MLR model, the model should fulfill the early assumptions such as normality test and multicollinearity.

Multiple linear regression model is stated as follow,

$$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik} + \varepsilon_i(\beta) \qquad (1)$$

where:
Y_i is the value of dependent variable
B_0, β_1, β_2 and β_j are parameters
X_i is a known as independent variables
i = 1,……, n

The function for least squares method is,

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

From (1), $\varepsilon\,(\boldsymbol{\beta}) = \mathbf{Y\text{-}X\boldsymbol{\beta}}$
Then, $S(\beta) = \mathbf{(Y\text{-}\ X\boldsymbol{\beta})^T\,(Y\text{-}\ X\boldsymbol{\beta})}$

$$= \mathbf{Y^TY\text{-}\ 2\boldsymbol{\beta}^TX^TY + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta}} \qquad (2)$$

In order to minimize $S(\beta)$, we have to differentiate $S(\beta)$ with respect to β where $\left.\frac{\delta S}{\delta \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}}$ is equal to 0,

$$\left.\frac{\delta S}{\delta \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} \;=\; -2\mathbf{X}^{T}\mathbf{Y} + 2\mathbf{X}^{T}\mathbf{X}\boldsymbol{\beta} = 0$$

Hence, the least squares estimator is,

$$\hat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^{T}\mathbf{X})^{-1}\,\mathbf{X}^{T}\mathbf{Y} \qquad (3)$$

The values fit by the equation $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_i x_i$ are denoted as i, and the residuals e_i are equal to y_i-$\hat{y}$, the difference between the observed and fitted values [10].

3. Results

3.1 Assumption of MLR

The data of this study tested for three assumptions and all the assumptions were satisfied and trustworthy. The results are shown in Table 1, Figure 1 and Figure 2.

a) The constant variance of residual

Scatter plot was used to identify a condition variance of residual using SPSS software. The points on the Figure 1 appeared to be randomly scattered and shows no pattern. It can be assumed that the variance in the error terms is constant. The first assumption is satisfied.

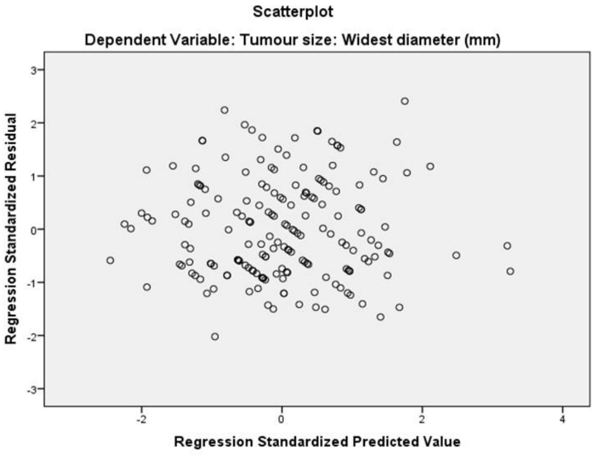

Figure 1 Scatter plot of constant variance

b) The residual of normality distributed

To test the normality, this study used Q-Q plot by tumour size variable (dependent variable) to determine the shape/trend that follows the normal distribution. The Q-Q plot in Figure 2 shows that the data are in straight line. This indicates that the data are normal and the assumption is satisfied.

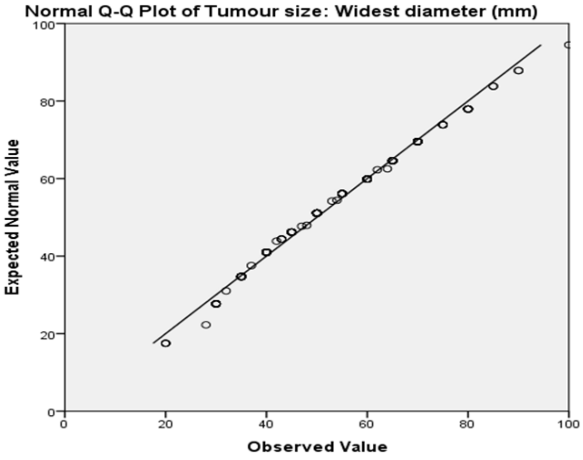

Figure 2 Scatter plot of constant variance

c) Multicollinearity checking

Table 1 shows that the tolerance values are less than 0.99 and all VIF values are less than 10. This indicates that the multicollinearity among independent variables do not exist.

Table 1 Coefficients of tolerance values, VIF and eigen values for actual data

Variable	Tolerance	VIF	Eigen value
(Constant)			18.145
Gender	0.864	1.158	0.731
Age (years)	0.860	1.163	0.672
Ethnic Group	0.876	1.141	0.578
ICD 10 Site	0.845	1.183	0.519
TNM Staging	0.856	1.168	0.474
Family History	0.840	1.190	0.468
Crohn's Disease	0.869	1.151	0.418
Polyp	0.863	1.158	0.354
History of cancer	0.787	1.271	0.334
Endometrial	0.877	1.141	0.313
Gastric	0.858	1.166	0.287
Small bowel	0.843	1.186	0.281
Hepatobiliary	0.891	1.123	0.249
Urinary tract	0.844	1.185	0.245
Ovarian	0.83	1.205	0.221
Other cancer	0.839	1.192	0.189
Intestinal Obstruction	0.847	1.181	0.171
Colorectal	0.841	1.19	0.154
weight loss	0.854	1.171	0.119
diarrhea	0.843	1.186	0.095
blood stool	0.886	1.129	0.043
anemia	0.867	1.154	0.076
abdominal	0.787	1.27	0.012

3.2 Analysis of multiple linear regression

Majority of researchers in medical health will use MLR model to predict the tumour size of colorectal cancer (CRC). This study also applied MLR model toward CRC data. There are 25 predictor variables and tumour size of colorectal cancer as a response or dependent variables. The results in Table 2 showed that 11 predictor variables were significant toward tumour size. All the significant variables are age at diagnosis, icd10 site, TNM staging, family history, Crohn’s disease, history of cancer, gastric, ovarian, intestinal obstruction, anemia and abdominal. The MLR model is shown as below:

$\hat{Y}$ = 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 anemia – 3.994 abdominal.

Table 2 Summary of parameter estimation multiple linear regression model

Independent variables	Beta (β)	Sig. Value
(Constant)	76.056	*0.000
Gender	1.977	0.286
Age at diagnosis (years)	-0.421	*0.000
Ethnic Group	2.231	0.114
ICD 10 Site	3.459	*0.027
TNM Staging	0.961	*0.040
Family History	-16.738	*0.000
Diabetes Mellitus	0.832	0.679
Crohn's Disease	5.035	*0.012
Ulcerative colitis	1.508	0.432
Polyp	-1.577	0.396
History of cancer	5.557	*0.007
Endometrial	-0.869	0.648
Gastric	-6.517	*0.001
Small bowel	-3.33	0.087
Hepatobiliary	-1.047	0.589
Urinary tract	2.597	0.176
Ovarian	12.865	*0.000
Other cancer	3.248	0.084
Intestinal Obstruction	-4.35	*0.022
Colorectal	1.227	0.511
Weight loss	3.063	0.1
Diarrhea	0.75	0.695
Anemia	-7.943	*0.003
Blood stool	-2.158	0.233
Abdominal	-3.994	*0.038

*Significant at 0.05

The summarization of measurement errors value is shown in Table 3.

Table 3 Result of statistical measurement error in MLR

Model	MSE	RMSE	MAE	MAPE
Multiple linear regression (MLR)	129.558	11.3826	10.349	20.798

3.3 Summary of results

In order to predict the tumor size of colorectal cancer and its significant determinants, this study performed multiple linear regressions (in mm). From the model, there are 11 predictor variables were significant to predict the size of tumor .

4. Conclusion and Discussions

This study applied MLR model to analyze the data. The MSE, RMSE, MAE and MAPE had been used to measure the effectiveness of the model in predicting the tumor size of colorectal cancer, based on the factor and causes of colorectal cancer.

Only 11 independent variables are significant in this study and ovarian give the higher chances to develop CRC among patient by 12.865. The variable includes are age, icd10, TNM staging, crohn's disease, history of cancer and ovarian are directly proportional to the tumor size. The rest of variables are inversely proportional to the tumor size.

Acknowledgements

This research work is supported by FRGS (Fundamental Research Grant Scheme) grant (Vot 1498), Ministry of Higher Education, Malaysia.

References

[1] M.A.B. Shafi, M.S.B. Rusiman, N.S.H.C. Yusof, (2014). "Determinants status of patient after receiving treatment at Intensive Care Unit: A case study in Johor Bahru," *2014 International Conference on Computer Communications and Control Technology (I4CT)*, pp. 80 – 82.

[2] R.T. Greenlee, M.B. Hill-Harmon, T. Murray, S. Bolden, P.A. Wingo, (2001). "Cancer Statistic, " *CA cancer J Clin*, vol. 51, pp. 15-36.

[3] Health indicator, Ministry of health, Malaysia. (2010). "Indicators for Monitoring and Evaluation of Strategy Health for All," Ministry of Health, Malaysia.

[4] M.S. Haerian, B.S. Haerian, H. Rooki (2009). "Comparison of Survival between Patients with Hereditary Non Polyposis Colorectal Cancer (HNPCC) and Sporadic Colorectal Cancer," *Asian Pac J Cancer Prev.*, vol. 10, pp. 497-500.

[5] C.K. Kong, A.C. Roslani, C.W. Law, S.K.D. Law SKD, K. Arumugam (2010). "Impact Of Socio-Economic Class On Colorectal Cancer Patient

Outcomes in Kuala Lumpur And Kuching, Malaysia''. *Asian Pasific J Cancer Prev*, vol. 11, pp. 969-74

[6] B. Monghimi-dehkordi, A. Safaee (2012). "An Overview of Colorectal Cancer Survival Rates and Prognosis in Asia''. *World J Gastrointestinal Oncol*, vol. 4, pp. 71.

[7] Malaysian Oncological Society Novartis Corporation (Malaysia). (2007). *The Lancet Oncology*, vol. 8, pp. 773-783.

[8] Q.M. Akhtar, M. Raj and J. Menon. (2012). "Screening for Colorectal Cancer in Malaysia Consensus/ Clinical Practice Guidelines''. *Academy of Medicine, Malaysia*, pp. 2-12.

[9] K.L. Goh, F.K. Quek, G.T.S. Yeo (2005). "Colorectal Cancer In Asians: A Demographic And Anatomic Survey In Malaysian Patients Undergoing Colonoscopy''. *Aliment Pharmachol Ther*, vol. 22, pp. 859-64.

[10] M.H. Kutner, C.J. Nachtsheim, (2004). Applied Linear Statistical Models. Fifth edition, Mc Graw Hill.

[11] M.A. Shafi, M.S. Rusiman (2015). "Status of Patient After Receiving Treatment Using Multinomial Logistic Regression at Intensive Care Unit in Johor''. *Australian Journal of Basic and Applied Sciences*, *9(8) Special 2015*, pp 29-34

[12] M.A. Shafi, M.S. Rusiman (2015). "The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital Of Malaysia''. *Applied Mathematical Sciences,* vol. 56, pp. 2749-2759.

[13] M.A. Shafi, M.S. Rusiman (2015). "The Effective Model of Linear Regressions for Colorectal Cancer Stages in General Hospital: A Case Study in Kuala Lumpur''. *Australian Journal of Basic and Applied Sciences,* Vol. 9, pp 7-10