Journal of Mathematics and Computing Science, 2023,
Volume 9, No 1, 29-38

Determining the Prognostic Factors of Lung Cancer Data using Multiple Linear Regression Analysis

Siti Afiqah Muhamad Jamil[1]*, Nurain Ibrahim [2]

[1, 2] School of Mathematical Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2] Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Authors' email: afiqahjamil@uitm.edu.my*, nurainibrahim@uitm.edu.my

*Corresponding author

Received 1 May 2023; Received in revised 25 May 2023; Accepted 1 June 2023
Available online 19 June 2023

Abstract: Due to the discovery of numerous cancer types, the risk of mortality among people has significantly increased. In Malaysia, lung cancer is one of the top five cancers that affect both men and women. The purpose of this study is to examine the association between the type of lung cancer and treatment of lung cancer and its associated factors that influence the duration of survival among lung cancer patients. A retrospective cohort study was carried out among lung cancer patients from one of the general hospitals in Johor. The lung cancer data will apply the Chi-Square Test for Independent and Multiple Linear Regression analysis. Several assumptions of model diagnostic prior to data analysis also have been conducted. The results indicated that type and treatment of lung cancer significantly associated towards the duration of survival. Based on descriptive statistics, there is high frequency on male compared to female, Chinese compared to other races, Adenocarcinoma type of Non-Small Cell Lung Cancer (NSCLC) and Chemoradiotherapy (CCRT) compared to other treatments. In short, based on the analysis, it was found that gender and treatment significantly affecting the duration of survival of lung cancer patients. Therefore, this will provide insights into which factors are most strongly associated with the prognosis besides, helps in prioritize and rank the importance of different factors.

Keywords: Lung Cancer Disease, Model Diagnostic Checking, Multiple Linear Regression

1 Introduction

A statistical method known as multiple linear regression uses several explanatory variables to estimate the outcome of a response variable [1]. Besides, in biomedical science, the survival distribution of human and animal frequency are being used in predicting the probability of survival especially in fields of engineering, sociology, criminology and insurance [2]. Most of the past research applied the deep learning supervision on the lung cancer metastatic lesions, machine learning in early-detection of lung cancer in sputum, relationship of socio-economic status with lung cancer, neural network model for lung cancer detection and other approaches of data analysis [3-6].

The most prevalent cancers among males and females, according to the Malaysian Cancer Registry Report 2012-2016, are colorectal, lung, prostate, breast, and cervix cancer. There was an 11% increase in the number of newly diagnosed cancer cases, and cancer-related deaths increased by nearly 30%. As the rising cancer burden continues to place enormous physical, emotional, and financial strain on cancer patients, communities, and the nation's health care system, the rising incidence of cancer will become a major public health concern. The leading cause of morbidity and mortality in Malaysia is still lung cancer, which is a serious health concern. Smoking, including both active and passive exposure to tobacco smoke, is one of the main risk factors for lung cancer [7]. Lung cancer disease-related issues include the need for extensive tobacco control measures, late-stage diagnosis, and restricted access to specialised facilities [8]. Besides, a study has assessed the correlation of radiotherapy with the second

primary malignancy (SPMs) among the resectable lung cancer patients [9]. Patients with resectable lung cancer showed an increased risk of getting second primary solid and gastrointestinal cancers after undergoing radiotherapy. More attention must be paid to SPM prevention related to radiotherapy.

The aims of this study are examining the association between type of lung cancer and the treatment of lung cancer since, both variables highly relatable and significant towards the duration of survival of lung cancer based on the past research study. By considering several factors related to the lung cancer, this study also aims to examine the significant prognostic factors associated with the duration of survival of lung cancer disease.

2 Materials and Methods

This research has applied the retrospective cohort study where collection of data has been made years back from February 2008 until February 2017 and the data has been collected in one of the general hospitals in Johor Bahru, Malaysia. Besides, 35 patients who have been diagnosed with lung cancer has been selected for this study. Increasing the sample size using bootstrapping procedure or simulation is recommended to improve the model accuracy [10]. Additionally, this study used part of the data which represent the right censored data. The data consists of only five independent variables which are the ages, gender, races, types of lung cancer and treatments of lung cancer and one dependent variable that is the duration of survival. Besides, due to time constraints, data restrictions and inconsistency data with missing values, only selected variables have been chosen to be applied in this study.

A Inclusion and Exclusion Criteria

Based on Table 1 below, the list of variables consists of five variables which are the gender, races, types of lung cancer and treatments of lung cancer which represent the categorical data and age represented the continuous data, with the dependent variable used in this study which is the duration of survival.

Table 1: List of variables

Independent variables	
Categorical data	**Continuous data**
Gender Races Types of Lung Cancer Treatments of Lung Cancer	Age (years)
Dependent variables	
Duration of survival	

B Statistical Analysis

Figure 1 represent the flowchart of analysis in this study. Statistical Package for Social Sciences (SPSS) version 26 has been used through the whole process of analysis procedure. Based on Figure 1, the analysis began with exploratory data, where descriptive statistics for each variable were explained in numerical and graphical form using values for frequency, percentage, bar chart, histogram, and other data presentation.

Next, the parametric approach assumption and model diagnostic testing on the study's variables also being observed. Chi-square test for independence will be applied to investigate the relationship between the two categorical variables of type and treatment of lung cancer. Additionally, Multiple linear regression will be applied to examine the significant factors affecting the durations of survival among lung cancer patients.

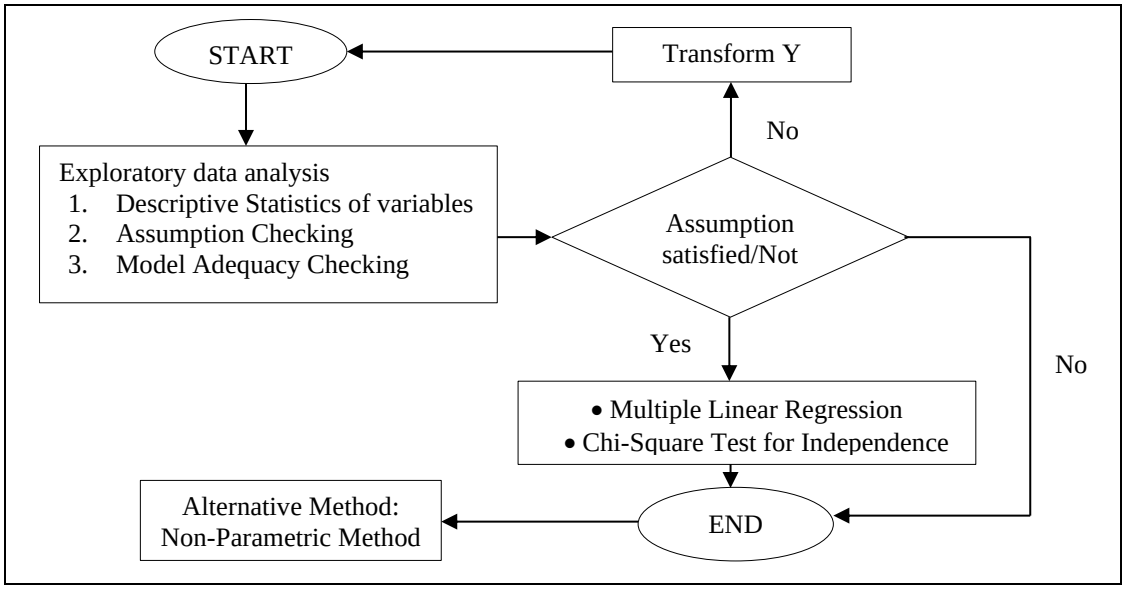

Figure 1: The flowchart of statistical analysis

C Exploratory Data Analysis

Multiple linear regression is a common statistical method to examine the relationship between independent towards the dependent variable. There are several assumptions need to be checked before analysing the multiple linear regression which are [11]:

i. The dependent variable and independent variable must be linear.
ii. There is no multicollinearity of the variables.
iii. There are no influential observations/cases.

For the linearity assumptions, since, there is only one continuous independent variable, this study applied the lack of fit test using the *p*-value approach. Accepting the null hypothesis will indicates that there is no lack of fit. Besides, for multicollinearity, VIF need to be less than 10 and the tolerance value should be greater than 0.2. Consequently, for influential observations, the value of Cook's Distance needs to be smaller than 1 for the non-influential observations.

Additionally, after model fitting, model adequacy checking need to be applied. There are several assumptions required to be checked which involved [11]:

i. Autocorrelation (Error term must be independent)
ii. Homoscedasticity (Error variance must be constant)
iii. Residuals are normally distributed.

Besides, for the autocorrelation, as Durbin-Watson is between 1.5 and 2.5, the assumptions of error term is satisfied. Meanwhile, based on scatterplot of residual versus predicted variable, as it is randomly scattered, the assumption of homoscedasticity is satisfied. Continuously, normality assumption sis based on the Kolmogorov Smirnov Test of *p*-value approach. Accepting the null hypothesis will leads to a normal data, thus assumption normality is satisfied.

D Chi-Square Test for Independence

In order to examine the relationship between the type of lung cancer and the treatment of lung cancer, Chi-Square Test for Independence has been applied. The following hypothesis statement will be used:

H_0: There is no association between the type and treatment of lung cancer.
H_1: There is an association between the type and treatment of lung cancer.

If the $p-value < \alpha = 0.05$, we reject H_0. Hence, there is an association between the type of lung cancer and the treatment of lung cancer. Besides, this study used 95% confidence level throughout whole procedure of hypothesis test for the analysis [13].

E Multiple Linear Regression

Based on [14-15], a multiple linear regression model was used in this study. The regression model of Eq. (1) employed in the study is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{1}$$

Where y is the dependent variable, x_1 through x_5 are the 5 independent variables with the regression coefficient, β_1 until β_5. To determine the model's significance, overall F-test has been used. The F-statistic value can be calculated using the following formula in Eq. (2) below:

$$F = \frac{MSR}{MSE} \tag{2}$$

Where Mean Square Regression (MSR) is the value of the sum of squares of the predictor variables divided by the degree of freedom and Mean Square Error (MSE) is the mean square of error. The p-value given in the ANOVA can also be used to gauge the model's significance. When the p-value is less than 0.05, the model is considered significant.

These examples illustrate the hypothesis.

H_0 **:** $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1 **:** At least one $\mu_j \neq 0$ where $j = 1, 2, 3, 4, 5$

3 Results and Discussions

There are three different sections of the results which consists of the exploratory data analysis, assumptions checking, model adequacy checking, Chi-square test for independence and Multiple Linear Regression analysis.

i. Exploratory Data Analysis

Five independent variables and one dependent variable has been applied to the multiple linear regression analysis and comprised of two continuous data and four categorical data. The descriptive statistics among all the variables is summarized as in Table 2 and Table 3.

Table 2: Descriptive statistics for continuous lung cancer data

Variables (Continuous data)		**No. (%)**
Age (year)	Mean, Standard deviation (SD)	61.03, 13.033
	Median, range	63, (55, 69)
Duration of survival (months)	Mean, Standard deviation (SD)	27.10, 25.153
	Median, range	14.30, (7.8, 40.0)

Based on Table 2 above, on average, the age of lung cancer patients who were diagnosed in this study was 61.03 years old while the duration of survival is about 27.10 months. 50% patients of lung cancer aged 63 years old and above and another 50% aged below 61 years old.

Table 3: Descriptive statistics for categorical lung cancer data

Variables (Categorical data)		No. (%)
Gender	Male	21 (60.0)
	Female	14 (40.0)
Races	Malay	11 (31.4)
	Chinese	22 (62.9)
	Indian	2 (5.7)
Types of lung cancer	Adenocarcinoma, NSCLC	13 (37.1)
	Large cell carcinoma, NSCLC	7 (20.0)
	Small cell lung cancer, SCLC	8 (22.9)
	Squamous cell carcinoma, NSCLC	7 (20.0)
Treatment of lung cancer	Chemoradiotherapy, CCRT	20 (14.3)
	Chemotherapy, Chemo	5 (14.3)
	Chemotherapy and Surgery, Chemosurgery	7 (20.0)
	Chemotherapy and Targeted therapy, ChemoTarget	3 (8.6)

Based on Table 3 above, male has the highest frequency with 60% compared to female. For races, the highest frequency of lung cancer patients is Chinese with 62.9% followed with Malay and Indian. Meanwhile, for the type of lung cancer, adenocarcinoma Non-Small Cell Lung Cancer (NSCLC) with 13% followed with small cell lung cancer, large and squamous cell carcinoma. For the treatment of lung cancer, chemoradiotherapy (CCRT) is the highest preferable treatment provided for lung cancer patients in this research study.

ii. Assumption Checking

For the linearity assumption, the only continuous variable which is age is found to be linearly related towards the duration of survival of lung cancer patients based on the lack of fit test.

H_0: There is no lack of fit data
H_1: There is lack of fit data

Table 4: ANOVA Lack of Fit

	F-value	**Df**	***P*-value**
Deviation from linearity	1.184	23	0.407

Based on Table 4, since the p-value $= 0.407 > \alpha = 0.05$, we accept H_0. Thus, there is no lack of fit data and linearity assumption is satisfied. For the second assumption which is the multicollinearity assumption. The results of VIF and tolerance as in Table 5 below:

Table 5: Tolerance and VIF values

Variables	**Tolerance**	**VIF**
Age	0.906	1.104
Races	0.925	1.081
Gender	0.837	1.195
Type	0.776	1.288
Treatment	0.723	1.383

Multicollinearity happens when the VIF is greater than 10 and the tolerance value is less than 0.2. Based on Table 5, this assumption is satisfied because the tolerance values are all greater than 0.2 and the VIFs are less than 10.

Table 6: *Cook's Distance* values

	Minimum	**Maximum**
Cook's Distance	0.000	0.397

Additionally, for the third assumption, checking the influential cases was examined based on the value of Cook's Distance. If the value of Cook's Distance is higher than 1, observation will be considered as influential. As in Table 6, the Cook's Distance value are under 1, thus, there is no influential cases in this study and this third assumption is satisfied.

iii. Model Adequacy Checking

In model adequacy checking, this study examines the autocorrelation assumption or in other words it is the error of independent term. Durbin-Watson test was applied as in Table 7 below:

Table 7: Durbin-Watson Test

Durbin-Watson Value
1.894

Based on the value of Durbin-Watson, since, $1.5 < DW = 1.894 < 2.5$ or close to 2, thus, the assumption of autocorrelation has met. The next diagnostic checking is the homoscedasticity assumption or checking for the constant error variance:

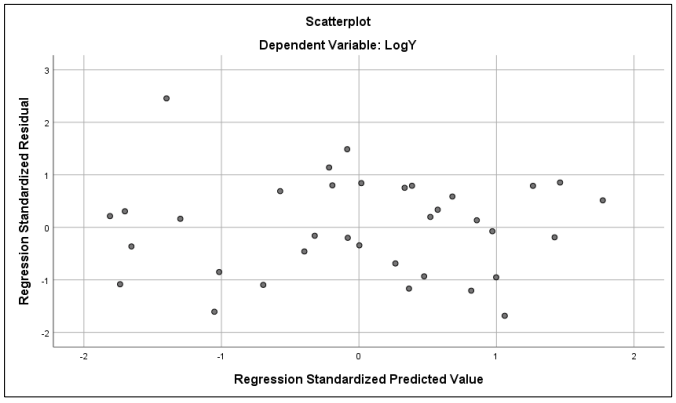

Figure 2: The scatterplot of residual versus predicted variable

By referring to Figure 2, since the scatterplot has shown randomly scattered and has no pattern, hence error variance is constant and the assumption of homoscedasticity has met. Besides, for the normality assumption, Kolmogorov test has been applied and the summary result can be seen as in Table 8 below:

Table 8: Kolmogorov-Smirnov Test

Kolmogorov-Smirnov (p-value)
0.053

H_0: Data comes from normal distribution.
H_1: Data comes from not normal distribution

Using the p-value approach, since p-value $= 0.053 > \alpha = 0.05$, we accept H_0. Thus, normality assumption has met and this study concluded that data comes from normal distribution.

iv. Chi-Square Test for Independence

In order to examine the relationship between type of lung cancer and the treatment of lung cancer, Table 9 below, showed the output of the chi-square test.

Table 9: Chi-Square Test for Independence

Pearson Chi-Square Value	Df	p-value
35.736	9	0.000

Based on Table 9, since $p - value = 0.000 < \alpha = 0.05$, we reject H_0. Thus, we can conclude that, there is an association between type of lung cancer and the treatment of lung cancer.

v. Multiple Linear Regression

The F-value and p-value of overall significance of model could be seen as in Table 10 below:

Table 10: Overall Model Multiple Linear Regression

Model	F	P-value
Regression	3.874	0.007

Since p- value$= 0.007 < \alpha = 0.05$, we reject H_0. Therefore, indicating that at least one of the predictor variables affecting the duration of survival of lung cancer. In order to check for the model performances, R-squared and adjusted R-squared are as follow:

Table 11: R-Squared and Adjusted R-Squared

Adjusted R-Squared Value
0.304

Based on the value of adjusted R-Squared, 30.4% of the total variation in duration of survival is explained by the age, races, gender, type, treatment and status of lung cancer patients. Another 60.6% is explained by other factors.

vi. Individual t-test

The independent variables included in the study has been examine thoroughly by using individual t-test. This test would help in identifying the predictor variables which are not significant to be removed from the regression model.

Table 12: Individual t-test

Independent Variables	t-value	Sig.	Results
Age, x_1	-1.925	0.064	Not Significant
Races, x_2	-0.829	0.414	Not Significant
Gender, x_3	**2.318**	**0.028**	**Significant**
Type, x_4	0.387	0.701	Not Significant
Treatment, x_5	**3.361**	**0.002**	**Significant**

Based on Table 12, two variables that were significant are gender and treatment with both p-value of the individual t-test is less than alpha, 0.05. Meanwhile, the other variables which are the age, races, and type of lung cancer were not significant. This means, gender and treatment statistically affecting the duration of survival. The final model has removed variables age, races and type of lung cancer.

Since this study wanted to elaborate each group related to the significant categorical variable, the final model has been re-run using the significant variables of dummy. The suggested model as in Table 13 meanwhile, the final model can be seen in Table 14 below:

Table 13: Coefficients of Suggested Model

Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std. Error**	Beta		
(Constant)	.971	.404		2.402	.023
Age, x_1	-.010	.005	-.289	-1.925	.064
Races, x_2	-.101	.122	-.123	-.829	.414
Gender, x_3	.336	.145	.363	2.318	**.028**
Type, x_4	.025	.064	.063	.387	.701
Treatment, x_5	.308	.092	.565	3.361	**.002**

Table 14: Coefficients of Overall Final Model

Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std. Error**	**Beta**		
(Constant)	1.481	.241		6.139	.000
Gender, x_3	.247	.152	.266	1.621	.115
Treatment=Chemotherapy, $x_{5(Chemotherapy)}$	-.919	.315	-.708	-2.917	.007
Treatment=Chemo and Surgery, $x_{5(Chemosurgery)}$	-.553	.282	-.487	-1.960	.059
Treatment=Chemoradiotherapy, $x_{5(Chemoradiotherapy)}$	-.278	.256	-.303	-1.087	.286

The final model:

$$\log y = 1.481 + 0.247x_3 - 0.919x_{5(Chemotherapy)} - 0.553x_{5(Chemosurgery)} - 0.278x_{5(Chemoradiotherapy)}$$

where,
Y = duration of survival
x_3 = gender
x_5 = treatment_chemo
x_5 = treatment_chemosurgery
x_5 = treatment_CCRT

Interpretation of coefficient:

β_1 = 0.247. This indicates that the mean of duration of survival will be 0.247 (1.766 months) times higher in male compared to female.
β_2 = −0.919. This indicates that the mean of duration of survival will be 0.919 (8.299 months) times lower in treatment of Chemotherapy compared to Chemotherapy and Targeted Therapy.
β_3 = −0.553. This indicates that the mean of duration of survival will be 0.553 (3.573 months) times lower in treatment of Chemotherapy and Surgery compared to Chemotherapy and Targeted Therapy.
β_4 = −0.278. This indicates that the mean of duration of survival will be 0.278 (1.897 months) times lower in treatment of Chemoradiotherapy (CCRT) compared to Chemotherapy and Targeted Therapy.

Therefore, male has higher duration of survival compare to female while for the treatment of lung cancer, chemotherapy and targeted therapy having longer duration of survival followed with CCRT, chemotherapy and surgery, and lastly, chemotherapy only.

4 Conclusion and Recommendation

Descriptive statistics of the continuous and categorical data used in this study were presented at the beginning. The data have been presented in percentages and frequency for the categorical data while mean and standard deviation for the continuous data. Besides, this study aimed in examining the relationship between the type of lung cancer with the treatment of lung cancer. This information would clarify that different type of lung cancer will statistically associate with different types of treatment. In this study, it is proven that type of lung cancer and types of treatment have an association with each other. Since Chi-Square test for independence is one of the non-parametric methods, there was no assumption needed prior to data analysis.

Consequently, in order to determine the prognostic factors of lung cancer affecting the duration of survival, a multiple linear regression is being examined in this study. Several assumptions such as the linearity, multicollinearity and influential cases have been observed and validated. In addition, model diagnostics checking such as the autocorrelation test, homoscedasticity test and normality test have been verified and all of the assumptions were satisfied. Based on the results obtained from the analysis of Multiple Linear Regression, only gender and treatment have significant effect towards the duration of survival. All the remaining variables were not significant. It is suggested that performing the data using survival would be significant because assigning the data with censored observation will narrow down the focus of data analysis.

Acknowledgements

We would like to thank Universiti Teknologi MARA, Shah Alam for their support in this research study. The study protocol of data was approved by the NMRR-16-1815-31982 (IIR), Medical Research & Ethics Committee, MREC, MOH, Malaysia.

References

[1] A. Hayes, Multiple Linear Regression (MLR) Definition. *Investopedia Official Portal. Retrieved from*, 2022.

[2] S. Markar, C. Gronnier, A. Duhamel, J. Y. Mabrut, J. P. Bail, N. Carrere, ... & C. Mariette, The impact of severe anastomotic leak on long-term survival and cancer recurrence after surgical resection for esophageal malignancy. Annals of surgery, 262(6), 972-980, 2015.

[3] Y. Cao, L. Liu, X. Chen, Z. Man, Q. Lin, X. Zeng, & X. Huang, Segmentation of lung cancer-caused metastatic lesions in bone scan images using self-defined model with deep supervision. *Biomedical Signal Processing and Control*, *79*, 104068, 2023.

[4] S. Maurya, S. Tiwari, M.C. Mothukuri, C.M. Tangeda, R.N.S. Nandigam, & D.C. Addagiri, A review on recent developments in cancer detection using Machine Learning and Deep Learning models. *Biomedical Signal Processing and Control*, *80*, 104398, 2023.

[5] A. Gupta, C. H. Omeogu, J. Y. Islam, A.R. Joshi, & T.F. Akinyemiju, Association of area-level socioeconomic status and non–small cell lung cancer stage by race/ethnicity and health care–level factors: analysis of the National Cancer Database. *Cancer*, *128*(16), 3099-3108, 2022.

[6] M. E. Lemieux, X.T. Reveles, J. Rebeles, L.H. Bederka, P.R. Araujo, J.R. Sanchez, ... & V. I. Rebel, Detection of early-stage lung cancer in sputum using automated flow cytometry and machine learning. *Respiratory Research*, *24*(1), 1-16, 2023.

[7] S. Jusoh, N.N. Naing, N. Wan-Arfah, W.N. Hajidah, W. N. Arifin, L.S. Wong, ... & S. Selvaraj, Prevalence and Factors Influencing Smoking Behavior among Female Inmates in Malaysia. In *Healthcare* (Vol. 11, No. 2, p. 203). MDPI, 2023.

[8] D.C. Montgomery, E.A. Peck, & G.G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[9] H. Ismail, M. F. M. Hanan, H. M. Yusoff, & A. B. Pillai, FACTORS ASSOCIATED WITH FAILURE TO QUIT SMOKING AMONG SMOKING CESSATION CLINIC ATTENDEES IN MALAYSIA. *Malaysian Journal of Public Health Medicine*, *23*(1), 1-10, 2023.
[10] S. A. M. Jamil, M. A.A. Abdullah, S. L. Kek, O. R. Olaniran, & S. E. Amran, Simulation of parametric model towards the fixed covariate of right censored lung cancer data. In *Journal of Physics: Conference Series* (Vol. 890, No. 1, p. 012172), IOP Publishing, 2017.
[11] M. H. Kutner, Applied linear statistical models, 2005
[12] G. K. Uyanık, & N. Güler, A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234-240, 2013.
[13] B. Zhou, R. Zang, P. Song, M. Zhang, F. Bie, G. Bai, ... & S. Gao, Association between radiotherapy and risk of second primary malignancies in patients with resectable lung cancer: a population-based study. *Journal of Translational Medicine*, *21*(1), 1-10, 2023.
[14] D. C. Montgomery, E. A. Peck, & G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
[15] G. K. Uyanık, & N. Güler, A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234-240, 2013.