International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Impact of Influential Observations in Survival Analysis

Pavithra V[1], Kannan R[2]

[1, 2]Department of Statistics, Annamalai University
E-Mail Id: *paviarthy[at]gmail.com* and
statkannan[at]gmail.com

Abstract: ***Whenever survival analysis is used, the most important thing to look at is the data and the reliability of each model based on the data. Especially, using survival analysis, the way of handling the data is very important because using the data for analysis there are countless flaws and especially when influence observations are present and when they are not, the reliability of the model is completely different. The purpose of this research article is to clearly see how the survival models like Kaplan-Meier, Cox Proportion Hazard Model, Minimum Survival Probability and Maximum Survival Probability results are vary with and without of influential observations in the breast cancer data.***

Keywords: Breast Cancer, Influence Observation, Kaplan-Meier Test, Log Rank Test, Cox Proportional Hazard Model, Minimum Survival probability, Maximum Survival Probability.

1. Introduction

Survival analysis is a statistical technique widely used in many fields of science, especially in the medical field, studying the time until an event of interest occurs. Events can be death, tumor recurrence, or disease development. The response variable is the time this event occurred, called survival or eventtime, that can be censored. The breast cancer data set was used exclusively for this survival study.

Breast cancer (BC) is a non-communicable disease that begins in the cells of the breast. BC is one of the leading cancers in Indian women with over 1.5 lakh new BC patients registered in India in 2018. It accounts for 14% of all cancers in women. The BC is uncommon in men, 1 in 400 men has BC. This is the most common disease among Indian women, 1 in 28 people may develop BC at some point in their lives. Unfortunately, the number of BC cases reported each year is increasing faster than ever. The BC accounts for more than 27% of all new cancer cases. There is an increasing trend in the number of new cancer patients and comparably the risk is higher in urban areas as 1 in 22 women and lower in rural areas as 1 in 60 women. In India, the average age of the high-risk group is between 40-55 years are more prone to BC (American Cancer Society (ACS) 2019-2020). The overall numbers in India are better compared to the number for developed countries like US/UK is less where in 1 in 8 women are diagnosed annually. However, due to the relatively high level of awareness of the disease in the developed nations and the many government funding that promotes early detection, most cases are detected and treated at an early stage.

On the other hand, in India, the survival rate is very low due to the large population and low awareness. 1 in 2 women diagnosed with BC will die within the next five year. One of the main causes of high mortality is lack of awareness, late diagnosis and absence of appropriate BC screening program. Most of the BCs are diagnosed at advanced stage. Many patients in the urban area are diagnosed at stage-2 and most of the cases from rural areas, these lesions are diagnosed only after they transform to metastatic tumors. The exact cause of BC is still unknown, but years of medical research have identified several risk factors. It is not yet clear why some women at very high risk do not develop BC, while some women with no risk factors may develop BC. The risk factors for BC include genetics and inheritance, late pregnancy, oral contraceptive use, early onset of menstruation, late menopause, excessive alcohol consumption, smoking, obesity in girls adolescence, increased stress and poor eating habits-these factors are due to an increase in the incidence of BC.

Through cancer, especially BC is a very dangerous disease that is prevalent worldwide (Torre et al., 2015). Cancer is a group of diseases caused by the uncontrolled growth and spread of abnormal cells throughout the body (Diabate et al., 2018). BC is expensive and has received a lot of attention from doctors and statisticians. Mortality with unstable mortality with many different prognostic (Pg) factors (Parkin et al., 2014). American Joint Committee on Cancer (AJCC) BC staging is associated with survival prognosis (American Cancer Society, 2017). This situation is indicated by reduced survival from stage-1 90%, stage-2 65%, stage-3 20% and stage-4 5% (Sinaga et al., 2017). The majority of BC cases are classified as invasive or noninvasive. Invasive BC has spread throughout the body, but noninvasive did not spread throughout the body (Abay et al., 2018). Age has a significant effect on whether women get BC. The mortality rate of BC increases with age (Rezaianzadeh et al., 2009). A study conducted by Addis Ababa University on the impact of several risk factors on BC and survival showed that stage and type of disease have a significant effect on BC survival (Kantelhardt et al., 2014).

There are many definitions of outliers in the literature, both mathematical and more informal, as explained in more detail in (Ben-Gal, 2005). For example, (Hawkins, 1980) defines an outlier as an observation that deviates sufficiently from other observations to raise suspicion that it is produced by some other mechanism. or (Johnson et al., 1992) define outliers in a dataset that appears to be inconsistent with the

rest of the dataset. In statistics, the first attempt to analyze and identify influential observations was based on the residual (Therneau et al., (1990). In particular, in regression analysis an influential observation is one whose deletion has a large effect on the parameter estimates (Everitt and Brian, 1998). Detecting influential observations in survival data is of great importance because identifying individuals with too high or too short survival times can lead to the discovery of new prognostic factors in medical (Nardi and Schemper, 1999). The influential observation is an observation for a statistical calculation whose deletion from the dataset would noticeably change the result of the calculation (Burt et al., 2009).

The aim of this study is to examine survival and risk of death from the Adyar Cancer Institute in 2013. We investigated the influential observations in the data in our point of view and analyzing datasets containing influential observations yields inaccurate results and the accuracy of the model is greatly reduced during the analysis and the researchers are confused while describing the model. Therefore, in this paper clearly see how model accuracy varies data with and without of influential observations. We will continue to analyze the survival of the BC study using these two data approaches to overcome the deficiencies caused by BC. This BC survival analysis implemented using, which includes various models, was employed. The Kaplan-Meier (K-M) with log rank test and Cox Proportion Hazard (PH) models are most commonly utilized models (Lee and Wang (2003)). In addition, we looked at the Minimum survival probability and Maximum survival probability (Felix and Kannan, (2007) and Pavithra and Kannan (2022)).

2. Statistical Methods

2.1 Hazard Functions

The hazard function of the hold time X is denoted by h(x) and defined as individual probability fails in the time interval $(x, x+\Delta x)$ that the individual has lived for time x, the hazard function is expressed as:

$$h(x) = \left[\frac{P(x < X < x+\Delta x | X > x)}{\Delta x}\right] \quad \longrightarrow (1)$$

2.2. Cox Proportional Hazard Model

The relationship between the hazard rate and the covariate set can be expressed using the model:

$$ln\ ln[h(t)] = ln\ ln[h_0(t)] + \sum_{i=1}^{n} x_i\beta_i \quad \longrightarrow (2)$$

Where $x_1, x_2, x_3, \ldots, x_n$ are covariates.$\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ are the regression coefficients to be estimated. t is time and $h_0(t)$ is the baseline hazard rate when all covariates are zero.

2.3 The Survival Function:

Individual opportunities to survive for time x are expressed by S(x) = P(X > x). Let X be the continuous random variables, then the survival function is the complement of the Cumulative Distribution function S(x) =1-F(X) where F(X) = P(X ≤ x). The survival function is the integral of the probability density function f(x):

$$\hat{S}(x) = P(X > x) = \int_x^{\infty} f(t)dt \quad \longrightarrow (3)$$

$$f(x) = -\frac{dS(x)}{dx} \quad \longrightarrow (4)$$

Then if X is the discrete random variables, and can be obtained x_j with the probability mass function (p. m. f) $p(x_j)$ = P(X= x_j), j=1,2,3,… where $x_1, x_2, x_3, \ldots$ then the survival function for the discrete variables X is given by:

$$\hat{S}(x) = P(X > x) = \sum_{x_j > x} p(x_j) \quad \longrightarrow (5)$$

2.4. Kaplan-Meier with Log Rank Test

Estimated survival function for K-M Expressed as:

$$\hat{S}(x_{(j)}) = \hat{S}(x_{(j-i)})\hat{P}(X \geq x_j) \quad \longrightarrow (6)$$

In general, log rank is used to compare k-M survival curves formed by the following hypothesis:
H_0: There is no difference between the survival curves:
H_1: At least one difference between the survival curves:

$$\textbf{Log Rank Test} = \frac{(O_i - E_i)^2}{Var\,(O_i - E_i)} \quad \longrightarrow (7)$$

$$O_i - E_i = \sum_{j=1}^{n} m_{ij} - e_{ij} \quad \longrightarrow (8)$$

m_{ij} denotes the number of individuals who experience the event at timex_j, and e_{ij} is the value of hope. The null hypothesis will be rejected if log rank statistics $\geq \chi_{\alpha}^2$ with n-1 degrees of freedom (df) or p-value <α.

2.5 Minimum Survival Probability (MISP):

Survival probabilities are calculated on the assumption that all those that are censored, the result of interest occurred. Then, for any interval i, D_i denotes the number of deaths during i, W_i denotes the number of censored observation during i and N_i denotes the number of subjects at the beginning of i. Then MISP for time interval $i$is expressed by

$$\textbf{MISP}=1 - \frac{(D_i - W_i)}{N_i} \quad \longrightarrow (9)$$

2.6 Maximum Survival Probability (MASP)

The survival probabilities are calculated by assuming that all those who are censored at time i are alive till the end of time interval i. Hence the notations of MASP is,

$$\textbf{MASP}=1 - \left(\frac{D_i}{N_i}\right) \quad \longrightarrow (10)$$

3. Source of Breast Cancer Data

The data we used in our research were obtained from the Adyar Cancer Institute in Chennai. These data are the newly diagnosed breast cancer for 2013 and where we used the number 257 patients for our study. The data provided by the cancer center for this research: Gender, Age, Medical

History, Date of Diagnosis, laterality of the BC, Grade, Stages, Treatments (Surgery, Chemo Therapy, Radiation Therapy, Hormonal Therapy) with dates, follow-up details with dates and Alive Status.

4. Result and Discussion

4.1 Cox Proportional Hazard Model:

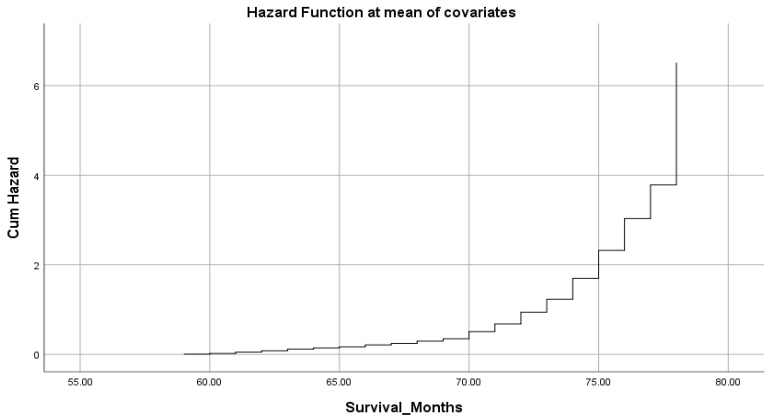

Figure 1: Hazard Function at mean of Covariate

The estimated variables by Cox regression are given in following: Age Group (X_1), Rural Urban (X_2), Medical History (X_3), Laterality (X_4), Stage (X_5), Recurrence and Metastatic (X_6), Surgery (X_7), Chemo Therapy (X_8), Radio Therapy (X_9), Hormonal Therapy (X_{10}). In step-1, the partial test shows that only Pg variables are statistically significant (P-value <5%). The backward stepwise method is used to extract the least influencing factors so that the final model is obtained in step-4 and the same method of Cox PH analysis applied all the three data set.

The β regression coefficient of the obtained models are all positive (β>0) with the value of exp(β)>0, meaning that all factors included in the model influence the event speed (death). That is, the risk of failure of depending on advanced stage of BC is 1.608 times greater than those lower stages. The risk of death of BC patients with recurrent and metastatic is 0.613 times greater than those that do not have recurrent and metastatic.

Table 1: Stepwise Method for with and without influential observation data

Stepwise Method	Independent Variables	With Influence Observation Data				Without Influence Observation Data			
		β	Wald	Sig.	exp(β)	β	Wald	Sig.	exp(β)
Step-1	X1	-.147	2.876	.090	.864	-.138	2.505	.114	.871
	X2	-.108	.404	.525	.898	-.103	.360	.549	.902
	X3	-.067	.145	.703	.935	-.062	.122	.727	.940
	X4	-.073	.220	.639	.930	-.022	.020	.888	.978
	X5	.535	11.386	.001	1.708	.533	11.058	.001	1.704
	X6	-.422	2.717	.099	.656	-.478	3.529	.060	.620
	X7	.210	.250	.617	1.234	.157	.123	.726	1.170
	X8	-.137	.228	.633	.872	-.157	.286	.593	.854
	X9	.243	1.446	.229	1.276	.190	.856	.355	1.209
	X10	-.051	.075	.784	.950	-.107	.312	.576	.898
Step-2	X1	-.109	1.937	.164	.897	-.109	1.921	.166	.896
	X3	-.067	.158	.691	.935	-.053	.099	.754	.948
	X5	.464	11.632	.001	1.590	.475	11.909	.001	1.608
	X6	-.396	2.481	.115	.673	-.462	3.405	.065	.630
Step-3	X1	-.095	1.844	.175	.910	-.098	1.941	.164	.907
	X5	.467	11.831	.001	1.595	.478	12.091	.001	1.613
	X6	-.396	2.476	.116	.673	-.461	3.398	.065	.630
Step-4	X5	.464	11.710	.001	1.591	.475	11.971	.001	1.608
	X6	-.422	2.829	.093	.656	-.489	3.831	.050	.613

Table 2: Overall Score in Stepwise Method for with and without influential observation data

Types Of Data	Stepwise Method	-2 Log Likelihood	Overall (score)		
			Chi-square	df	Sig.
With Influence Observation data	Step-1	1570.734	22.716	10	.012
	Step-2	1573.573	19.361	4	.001
	Step-3	1573.730	19.112	3	.000
	Step-4	1575.568	17.311	2	.000
Without Influence Observation data	Step-1	1536.930	24.914	10	.000
	Step-2	1539.205	21.125	4	.000
	Step-3	1539.304	20.939	3	.000
	Step-4	1541.238	19.055	2	.000

In Table-1: The result of cox proportional hazard analysis showed that the most significant pg variable to the probability of death was the presence of advanced stage and tumor recurrence with metastatic. Table-2 indicated the overall score for the data set of without influential observation methods are more appropriate and highly significant for all the backward stepwise methods in the cox proportion analysis. Meanwhile, the With Influence Observation data leads slightly reduced the efficiency and lower the precision of the model estimates when compared to the results of without influential observation.

4.2. Kaplan-Meier Analysis:

The K-M estimated the probability of survival curve for with and without influential observations method. The following figures are according to the two impact variables of stages and recurrence with metastasis.

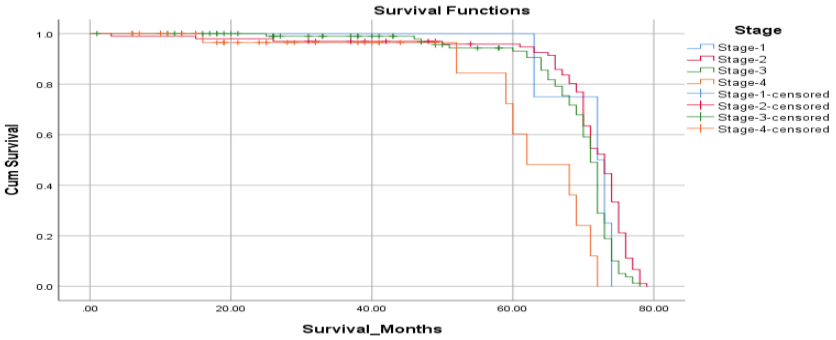

Figure 2: Tumor stage for with influential Observations

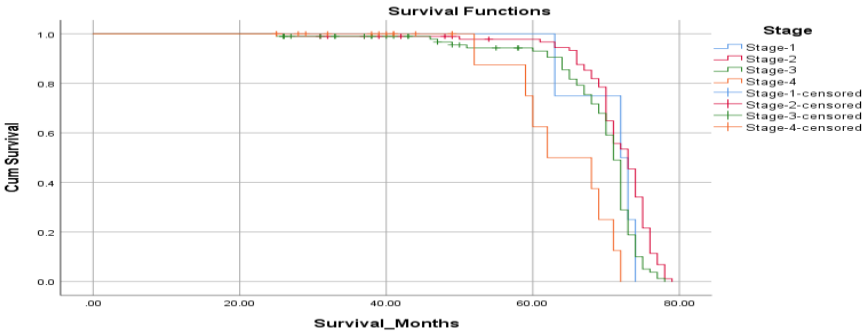

Figure 3: Tumor stage for without influential Observations

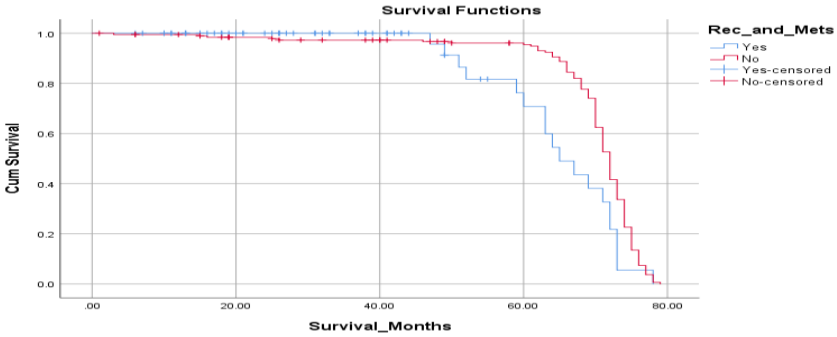

Figure 4: Recurrence and Metastatic for with influential Observations

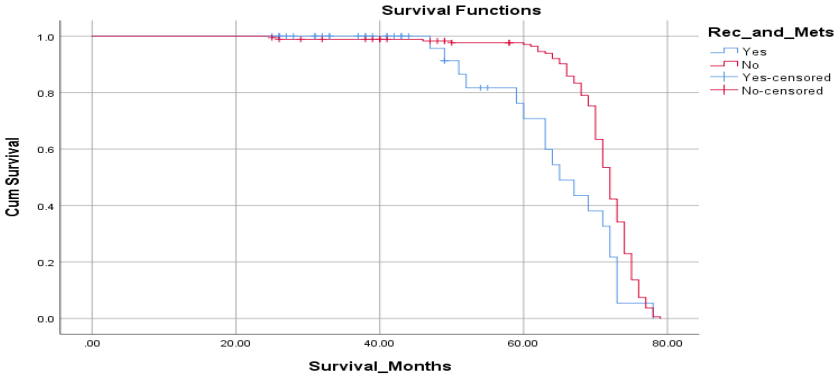

Figure 5: Recurrence and Metastatic for without influential Observations

The stages and recurrence with metastasis are the most important factor variable in the BC and these can determine the conditions of the cancer patients. In figure-2 & 3 survival probability shown that the cancer Stages of patient with BC and clearly seen which stage is mostly survived or died in these BC. The survival rate of stage-1 and stage-2 patients was very high compared to stage-3 and stage-4 and risk rate of BC patients in stage-1 and stage-2 was very low when compared to other stages. In figure-4 & 5 survival probability shown that the recurrence with metastatic cancer patient and clearly seen who have not spread the cancer they are only mostly survived in these BC.

4.3. Log Rank Test

The log rank test to determine if there is a difference between the survivals curves. The log rank test of significant or not significant in Pg variables are given in following table.

Table 3: Log Rank test used for Pg variable affecting Survival of BC

Log Rank Test / Pg Variable	df	With Influence Observation data		Without Influence Observation data	
		χ^2	Sig.	χ^2	Sig.
Age Group	4	4.518	.340	7.175	.270
Area	1	.250	.617	.347	.619
Medical History	1	.067	.796	.138	.710
Laterality	1	.798	.372	.918	.518
Stage	3	25.353	.000	28.874	.000
Recurrence & Metastatic	1	6.742	.009	9.654	.001
Surgery	1	.713	.399	.981	.596
Chemo Therapy	1	2.132	.144	3.712	.100
Radiation Therapy	1	2.482	.115	4.694	.055
Hormonal Therapy	1	.137	.711	.421	.516

Based on the Log Rank Test in Table-3, the equality of survival distribution of the BC variables Cancer Stages and Recurrence with metastatic were statistically recorded a p-value (0.000 and 0.001) makes a significant difference and other variables have statistically no significant difference. Meanwhile, beauty of this survival study is in the handling the impact of with and without of influence observation, the without of influence observation data results had outperforming when compared to with influence observation.

4.4. Comparison of Survival probabilities

Table 4: Survival Probability for the dataset of with and without of influence observation

Handling Of Data Technique	Method Of Probability	BC Five Year Survival Probability by Percentage				
		1st Year	2nd Year	3rd Year	4th Year	5th Year
With Influence Observation data	MISP	95%	84%	76%	68%	64%
	MASP	96%	87%	80%	71%	66%
	K-M	96%	87%	81%	70%	65%
Without Influence Observation data	MISP	95%	85%	78%	70%	65%
	MASP	96%	87%	80%	72%	67%
	K-M	96%	86%	79%	71%	69%

Table-4: Shows the cumulative survival probabilities at the end of each year from the date of completion of treatment through different methods. These estimates are obtained by using MISP, MASP and K-M methods. In general, by all the methods estimates of the cumulative probabilities have been decreased as the survival period has increased. The higher probabilities have been estimated by MASP. i.e., the estimates of MISP and MASP provide the two extreme values of the survival band within which the true survival probability lies. The three estimates are similar but not identical. The overall five-year survival probability (%) for the BC patients has been found to be 69%, which is very much similar to other method. However, this overall survival probability may not be an appropriate one, since the stage of the disease at diagnosis is one of the significant factors associated with the number of deaths occurred.

5. Conclusion

The K-M, Cox PH, MISP and MASP survival results of the study showed that age, medical history, resident, laterality of breast, stage, recurrence, metastasis, surgery, chemo therapy, radiation therapy and hormone therapy affected the time to death of BC patients 2013 at Adyar Cancer Hospital. The K-M estimated the survival month of the BC is 69 months. The analyses Cox PH found main factor behind the poor survival time is that the treated patients is already in the advanced stage and recurrent with metastatic. The comparison between the MISP, MASP and K-M analysis the MASP and K-M showed similar together and most useful to survival analysis. The beauty of this survival analysis of with and without influential observations data studies, here we have clearly outlined how to handling influence observation in survival analysis. We report that among the two datasets used in this survival analysis, the without of influential observations dataset are more suitable for all type of analysis. The information loss is slightly high and the model accuracy is tiny low for the with influential observations data when compared to the other method of dataset. So, it's best to avoid using the influential observations in dataset when dealing with survival Analyses.

6. Recommendation

Health professionals, governments and NGO should raise awareness of early cancer screening and should also encourage women to be diagnosed at an early stage to improve mortality risk, and cancer screening facilitation and

scheduling should be planned and scheduled in rural areas of this region to elucidated mortality risk.

References

[1] Abay M., Tuke G., Zewdie E., Abraha TH., Grum T and Brhane E. (2018). Breast self-examination practice and associated factors among women aged 20-70 years attending public health institutions of Adwa town, North Ethiopia. BMC Research Notes, Vol-11, Issue-1, pp 1-7.
[2] Amin MB., Edge SB., Greene FL., Stephen B., Compton CC., Gershenwald JE., Brookland RK., Laura Meyer., Grees DM., Byrd DR., Winchester DP. (2017). American Joint Committee On Cancer Staging manual. 8th Edition, New York: Springer.
[3] American Cancer Society. Stages of Breast Cancer (Internet-2017-2018).
[4] American Cancer Society. Breast Cancer Fact & Figures (Internet-2019-2020).
[5] Ben-Gal I. (2005). Outlier detection. In Data Mining and Knowledge Discovery Handbook, pages 131-146, Springer.
[6] Burt James E., Barber Gerald M., Rigby David L. (2009), Elementary Statistics for Geographers, Guilford Press, p. 513, ISBN 9781572304840.
[7] Diabate M., Coquille L., and Samson A. (2018). Parameter estimation and treatment optimization in a stochastic model for immunotherapy of cancer. arXiv Preprint ArXiv, 1806.01915.
[8] Everitt and Brian (1998). The Cambridge Dictionary of Statistics. Cambridge, UK New York: Cambridge University Press. ISBN 0-521-59346-8.
[9] Felix AJ., and Kannan R. (2007). Statistical models in survival analysis, Chap-3, pp 50-51.
[10] Hawkins DM. (1980). Identification of outliers, volume 11. Springer
[11] Johnson RA., WichernDW., and Education P. (1992). Applied multivariate statistical analysis, volume 4. Prentice hall Englewood Cliffs, NJ.
[12] Kantelhardt E., Zerche P., Mathewos A., Trocchi P., Addissie A., Aynalem A., Wondemagegnehu T., Ersumo T., Reeler A., Yonas B., Tinsae M., Gemechu T., Jemal A., Thomssen C., Stang A., and Bogale S. (2014). Breast cancer survival in Ethiopia: A cohort study of 1,070 women. International Journal of Cancer, Vol-135, Issue-3, pp 702-709.
[13] Lee ET., and Wang J. (2003). Statistical methods for survival data analysis. Edition-3, John Wiley & Sons. New York.
[14] Nardi A. and Schemper M. (1999). New Residuals for Cox Regression and Their Application to Outlier Screening. Biometrics; Vol-55, Issue-2, pp 523-529.
[15] ParkinDM., Bray F., Ferlay J., and Jemal A. (2014). Cancer in Africa 2012. Cancer Epidemiology and Prevention Biomarkers, Vol-23, Issue-6, pp 953-966.
[16] PavithraV and Kannan R. (2022). Impact of Missing Data in Survival Analysis. High Technology Letters. Vol-28, Issue-9, pp 374-384.
[17] Rezaianzadeh A., Peacock J., Reidpath D., Talei A., HosseiniSV., and Mehrabani D. (2009). Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran. BMC Cancer, Vol-9, Issue-1, 168.
[18] Sinaga ES., Ahmad RA., and Hutajulu SH. (2017). Beritakedokteranmasyarakat. Vol-33, FakultasKedokteran, UniversitasGadjahMada.
[19] TherneauTM.,Grambsch PM., and Fleming TR. (1990). Martingale-Based Residuals for Survival Models. Biometrika, Vol-77, Issue-1, pp 147-160.
[20] Torre L. A., Bray F., Siegel RL.,Ferlay J., Lortet-Tieulent J., and Jemal A. (2015). Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians, Vol-65, Issue-2, pp 87-108.