International Journal of Science, Environment
and Technology, Vol. 3, No 3, 2014, 931 – 943

ISSN 2278-3687 (O)

SEMI-PARAMETRIC SURVIVAL MODEL

H. Alexis Selvaraj[1] and Dr. G. Stephen Vincent[2]
[1]Assistant Professor of Statistics,
Periyar E.V.R College, Tiruchirappalli-23
[2]Associate Professor of Statistics (Rtd.),
St. Joseph's College (Autonomous), Tiruchirappalli-2
E-mail: alexis.selvaraj@gmail.com

Abstract: Several statistical methods have been proposed for modelling survival analysis data. The survival methods may be divided into two broad categories such as proportional hazard approaches and accelerated failure time models (Bradburn et al. 2003(1)). The Cox model is the most frequently used regression model in survival analysis (Bender et al.2005). Here the Cox regression model which is the semi-parametric survival model has been used to identify the relationship between the selected variables and the survival time after the completion of treatment. As a semi-parametric survival model it should be satisfied the assumption of randomness condition. It has been verified using chi-square test and logistic regression model. The main objective of the study is to identify the covariates which are prognostic factors of breast cancer after completion of treatment.

Keywords: Covariates, Semi-parametric survival model, Cox PH model.

INTRODUCTION

Statistical modelling approach is used to explore the relationship between the survival experience of a patient and the explanatory variables (Collett, 1994). It is usual to work with the survivor function for descriptive analyses and the hazard function for assessing the relationship between explanatory variables and survival time. Since hazard function does not involve the cumulative history of events, it is considered as the main vehicle of statistical modelling. The Cox-Proportional hazard model (Cox,1972) is the most commonly used multivariable approach for analysing survival time data in medical research (Bradburn et al.2003(1)). The first and foremost assumption of Cox proportional hazard mode is that the censorings are at random. It has been verified using chi-square test and logistic regression model. There are two approaches to this censored data regression model, the approach originally proposed by Cox and the counting process approach. The objective of model building in survival analysis is to identify a set of potential explanatory variables that contribute towards the hazard function. Any statistical model contains more than one covariate to predict the outcome and is called a multivariable model. Collett (1994)

*Received Apr 14, 2014 * Published June 2, 2014 * www.ijset.net*

recommends that the “general strategy” for model building. In this paper an attempt has been made to find out the significant prognostic factors of five years survival of breast cancer patients.

MATERIALS AND METHODS

The relevant lifetime data on the patients of breast cancer is obtained from one of the reputed hospital in Tiruchirappalli, Tamil Nadu from 1st January2009 to 31st December 2009. Among the 523 diagnosed during this period, 478 have completed the treatment and fulfilled the inclusion criteria. The remaining 45 patients have been excluded either because of they moved to other hospitals or who haven’t completed the treatment were excluded from the present study.

Model Building and Strategy for Model Selection:

The objective of model building in survival analysis is to identify a set of potential explanatory variables that contribute towards the hazard function. For identifying the contribution or association of a variable with the survival time, a lot of procedures are available. The choice of the statistical model covariates mainly depends upon the objective of the study. Generally any statistical model contains more than one covariate to predict the outcome and is called a multivariable model. Bradburn et at. (2003(1)) explains the three possible scenarios as to why a study may use a multivariate model. They are

i. A single factor is under investigation for its association with survival, but several other factors exist.

ii. A collection of factors of known relevance is under investigation for their ability to predict survival and

iii. Where a collection of factors are under investigation for their potential association with survival, possibly with known additional factors.

The breast cancer study is a combination of (ii) and (iii) scenarios.

Sample size considerations

The power of survival analysis is related to the number of events rather than the number of participants (Bradburn, 2003). Simulation works have suggested that at least 10 events need to be observed for each covariate considered, and anything less will lead to problems, for example, the regression co-efficient becomes bised (Peduzzi, et al.,1995). The breast cancer data used in the present study consists of 151 deaths and 10 covariates implying approximately 15 events per covariate.

Conversion of continuous variables

When the dependence of the hazard function on a variate, which takes a wide range of value is to be modelled, it is ideal to convert the continuous variable as a categorical variable (Collect, 1994). For this purpose, the following procedure is adopted:

i. The values of the variate are first grouped into four or five categories containing approximately equal number of observations.

ii. A factor is then defined whose level corresponds to this grouping.

In this study, the continuous variables, namely, the age of the woman, tumour size and treatment duration have been converted into categorical variables based on the clinical importance and number of observations.

Application of Cox PH Model to Breast Cancer Survival Data

For fitting Cox PH model to this data, the survival time after completion of the treatment has been considered as the dependent variable and the following variables have been considered as prognostic variables, namely, age of the woman at diagnosis, place of residence at diagnosis, associated diseases, if any, stage of the disease at diagnosis, size of the tumour at diagnosis, nucleus status at diagnosis, type of the cell status, treatment provided, duration of radiotheraphic treatment and recurrence of the disease curing the follow up period. A null model has been fitted without any explanatory variable. The statistics -2 $\log\widehat{L}$ has been noted. The variables have been entered as explanatory variables separately. Their -2 $\log\widehat{L}$ statistical value and its difference have been noted and they are shown in the following table number 1. The regression co-efficient and its corresponding Hazard ratio value with 95% confidence intervals are shown in table number 2.

Among the 10 variables fitted separately, the following five variables -2 $\log\widehat{L}$ statistics has been found be significant compared to the null model -2 $\log\widehat{L}$ statistics. These include age of the woman, nucleus status of cell, stage of the disease at diagnosis, duration of radiotheraphic treatment and any recurrence during the follow up period. All the five variables have been fitted as explanatory variables simultaneously. Then one variable has been omitted and the corresponding the -2 $\log\widehat{L}$ statistics has been noted. The results of comparison between each model with the full model of all the five significant variables in step i are shown in table number 3. Among them the variable nucleus change in the cell has been found to be non-significant indicating that it could be omitted from the selected five variables.

Table Number: 1

Values of -2 log$\hat{L}$ for Univariate Model (Analysis –I)

Variable	-2 log$\hat{L}$ value	Difference	Degrees of Freedom	'P' Value
Null Model	1779.251	-	-	-
Age	1764.450	14.8	3	0.002
Place of Residence	1777.941	1.309	1	0.253
Associated Disease	1776.243	3.008	1	0.083
Tumour Size	1779.088	0.163	1	0.687
Histology	1776.247	2.824	2	0.244
Change in Nucleus	1775.660	3.590	1	0.058
Stage	1738.652	40.599	3	<0.001
Type of Treatment	1778.604	0.646	1	0.421
Duration	1764.395	14.856	3	0.002
Recurrence	1691.06	88.191	1	<0.001

The next step is keeping the four variables recurrence of the disease, the stage of the disease, duration of radiotherapy treatment and age of the woman, -2 log$\hat{L}$ has been calculated. Again this -2 log$\hat{L}$ has been compared with the models of one variable omitted at a time. The results are shown in table number 3. All the variables are found to be significant indicating that these four variables are important and they influence on the survival of the breast cancer patients.

Table Number: 2
Hazard Ratios from the Cox PH Model
(Univariate Analysis-I)

Variable		Parameter Estimate	Standard Error	P<Chi .Sq.	Hazard Ratio	95% Hazard Ratio	
						Upper Limit	Lower Limit
Age (in Years)	< 40	-	-	-	-	-	-
	40-49	0.409	0.217	0.059	1.506	0.984	2.305

	50-59	0.294	0.234	0.209	1.342	0.848	2.123
	>60	0.505	0.233	0.031	1.658	1.048	2.622
Place	Urban	-	-	-	-	-	-
	Rural	0.195	0.168	0.248	1.215	0.873	1.689
Associated Disease	Yes	0.505	0.313	0.107	0.603	0.327	1.115
Tumour Size	<4.0	-	-	-	-	-	-
	>4.0	-0.066	0.165	0.687	0.936	0.677	1.293
Histology	SCC	-	-	-	-	-	-
	Poorly Diff. SCC	0.387	0.275	0.159	1.473	0.859	2.524
	Others	-0.186	0.229	0.417	0.831	0.530	1.301
Nucleus	No	-	-	-	-	-	-
	Yes	0.417	0.231	0.071	1.517	0.965	2.384
Stage	I	-	-	-	-	-	-
	II	0.527	0.270	0.051	1.694	0.998	2.876
	III	1.251	0.276	<0.001	3.495	2.034	6.006
	IV	2.928	0.517	<0.001	18.699	6.795	51.460
Type of Treatment	Surgery +Radiotherapy	-	-	-	-	-	-
	Radiotherapy	0.202	0.258	0.433	1.224	0.739	2.027
Duration of Radiotheraphic Treatment	>75 days	0.815	0.243	0.001	2.259	1.403	3.636
	61-75	0.593	0.196	0.003	1.810	1.232	2.659
	<45	0.361	0.250	0.150	1.435	0.878	2.343
	46-60 days	-	-	-	-	-	-
Recurrence	Yes	1.677	0.167	<0.001	5.349	3.858	7.415
	No	-	-	-	-	-	-

The final model with the above four variables has been fitted. The results of the regression co.efficient and the corresponding hazard ratio with 95% confidence limits are shown in table number 4. The hazard ratio of the variable recurrence has been found to be 4.3, which indicates that the chance of death within five years after completion of the treatment is 4.3 times higher for a woman with the recurrence of the disease within five years compared to the woman who has no recurrence during the study period.

The chance of death is 7.1 times higher for a woman with stage IV of the disease compared to the woman with stage I. Similarly, the chance of death is 2.6 times higher for a woman in disease status of stage III level compared to the woman at stage I disease level.

The other two variables, age of the woman and duration of radiotheraphic treatment, are found to be non-significant at 5% level of significance. However, the chance of death is higher for older women compared to the younger group and for the woman with longer or shorter duration of radiotheraphic treatment compared to the normal/ideal duration of radiotheraphic treatment days.

The stage i.e. severity of the disease has been measured at four level as described by FIGO. In this data, only 5 patients have been in stage IV level i.e., disease spread to other organs of the body. Clinically, it is very severe and difficult to treat. For the above analysis, stage IV has been included, because of its clinical importance though it consists of only 5 patients. However, statistically it is insignificant. Hence, another analysis has been performed with the above selected four variables, omitting the patients with severity of disease at level IV. This analysis is called analysis II. The total number of observations in analysis II is 473.

Table Number: 3

Values of -2 log$\hat{L}$ for selected significant models in first Step (Analysis I)

Variable	-2 log$\hat{L}$ value	Difference	df	P value
A+N+S+D+R	1647.787	131.463	11	<0.001
A+N+S+D	1741.542	94	-	-
A+S+D+R	1649.898	2.11	1	NS
S+R+A+N	1653.381	5.59	1	<0.05
S+R+D+N	1656.819	9.032	1	<0.01
A+N+D+R	1667.939	20.152	1	<0.01

S+R+D+A	1649.898	-	-	-
S+R+D	1659.705	98.07	1	<0.001
S+R+A	1655.264	5.366	1	<0.05
S+D+A	1714.577	64.679	1	<0.001
R+D+A	1671.179	21.281	1	<0.001
S+R+D+A	1649.898	-	-	-
S+R+D+Place	1648.948	0.95	1	NS
S+R+D+Associated disease	1649.331	0.567	1	NS
S+R+D+Tumour size	1646.903	2.995	1	NS
S+R+D+Histology	1648.299	1.599	1	NS
S+R+D+Type of Treatment	1649.885	0.013	1	NS

Where

A-Age of the woman, N-Change in nucleus, S-Stage of the disease, D-Duration of the radiotheraphic treatment, R-Recurrence during the follow-up period

Table Number: 4
Hazard Ratios from the Cox PH Model
(Multivariable Analysis-I)

Variable		Parameter Estimate	Standard Error	P<Chi. Sq.	Hazard Ratio	95% Confidence Interval	
						Lower Limit	Upper Limit
Recurrence		1.479	0.175	<0.001	4.3387	3.111	6.185
Duration of Radiotheraphic Treatment	46-60 days	-	-	-	-	-	-
	≤45	0.373	0.259	0.149	1.452	0.875	2.410
	61-75	0.365	0.203	0.072	1.441	0.967	2.147
	≥75	0.468	0.251	0.062	1.591	0.977	2.611
Age(in	<40	-	-	-	-	-	-

Years)	40-49	0.115	0.222	0.606	1.122	0.725	1.735
	50-59	0.394	0.240	0.101	1.484	0.926	2.375
	≥ 60	0.351	2.440	0.151	1.420	0.880	2.293
Stage of Diseases	I	-	-	-	-	-	-
	II	0.385	0.277	0.164	1.469	0.855	2.526
	III	0.970	0.288	0.001	2.639	1.501	4.638
	IV	1.967	0.536	<0.001	7.148	2.50	20.439

Analysis II

As in analysis I, all the 10 variables have been first entered as explanatory variables separately. Their corresponding -2 log$\hat{L}$ statistics has been calculated and compared with the null model -2 log$\hat{L}$. The results are shown in table number 5. There has been no wide change in their significance level of the variables observed, compared to above fitted model. The results are shown in table number 6.

In analysis II also, all the five variables that are significant in analysis I have been found to be significant at 5% level. As in analysis I, again all the five variables recurrence of the disease, duration of radiotherapy treatment, stage of the disease at diagnosis, age of the woman and change in nucleus of the cell have been fitted as explanatory variables. The process of omitting each variable and assessing the significance level of each variable has been done. In Analysis II also, the changes in the nucleus of the cell has been found to be insignificant. Hence, among the remaining four variables, the significance level of each variable has been tested by omitting each variable at a time. The selected four variables are confirmed for their significance. The results are shown in table number 7.

Table Number: 5

Values of -2 log$\hat{L}$ for Univariate Model (Analysis –II)

Variable	-2 log$\hat{L}$ value	Difference	Degrees of Freedom	'P' Value
Null Model	1718.123	-	-	-
Age	1703.599	14.524	3	0.002
Place of Residence	1717.271	0.852	1	0.356
Associated	1717.331	0.792	1	0.373

Disease				
Tumour Size	1714.462	3.661	1	0.056
Histology	1718.076	0.047	1	0.829
Change in Nucleus	1715.055	3.068	2	0.216
Stage	1715.133	2.990	1	0.084
Type of Treatment	1690.822	27.302	2	<0.001
Duration	1705.046	13.077	3	<0.004
Recurrence	1629.77	88.346	1	<0.001

Table Number: 6
Hazard Ratios from the Cox PH Model
(Univariate Analysis-II)

Variable		Parameter Estimate	Standard Error	P<Chi .Sq.	Hazard Ratio	95% Hazard Ratio	
						Upper Limit	Lower Limit
Age (in Years)	<40	-	-	-	-	-	-
	40-49	0.392	0.218	0.072	1.480	0.965	2.271
	50-59	0.351	0.238	0.140	1.420	0.891	2.268
	>60	0.452	0.239	0.050	1.571	0.985	2.505
Place	Urban	-	-	-	-	-	-
	Rural	0.160	0.172	0.352	1.174	0.828	1.647
Associated Disease	Yes	-0.576	0.328	0.079	0.562	0.296	1.068
Tumour Size	<4.0	-	-	-	-	-	-
	>4.0	-0.036	0.167	0.829	0.964	0.695	1.339
Histology	SCC	-	-	-	-	-	-
	Poorly Diff. SCC	0.359	0.284	0.206	1.431	0.821	2.496
	Others	-0.255	0.238	0.284	0.775	0.486	1.235
Nucleus	No	-	-	-	-	-	-
	Yes	0.383	0.231	0.098	1.467	0.932	2.308
Stage	I	-	-	-	-	-	-

	II	0.527	0.270	0.051	1.693	0.997	2.875
	III	1.250	0.276	<0.001	3.491	2.031	5.998
Type of Treatment	Surgery +Radiotherapy	-	-	-	-	-	-
	Radiotherapy	0.229	0.265	0.388	1.257	0.748	2.114
Duration of Radiotheraphic Treatment	>75 days	0.780	0.247	0.002	2.181	1.345	3.537
	61-75	0.599	0.199	0.005	1.750	1.185	2.583
	<45	0.273	0.259	0.292	1.314	0.791	2.185
	46-60 days	-	-	-	-	-	-
Recurrence	Yes	1.708	0.169	<0.001	5.516	3.958	7.688
	No	-	-	-	-	-	-

As in analysis I, all the remaining variables of step I have been entered one by one. No significant contribution has been assessed by the remaining variables. Hence, all the four variables, recurrence of the disease during follow-up, stage of the disease, duration of the radiotheraphic treatment and age of the woman, have been considered for the final model. As in analysis I, the recurrence of the disease and stage of the disease have been found as significant variables. The other two variables, namely, age of the woman and duration of the treatment have been found to be insignificant. This analysis II confirms that omitting of Stage IV patients has not altered the role of the other prognostic variables. The results are shown in table number 8.

Table Number: 7

Values of -2 log$\widehat{L}$ for selected significant models in first Step(Analysis II)

Variable	-2 log$\widehat{L}$ value	Difference	df	P value
R+S+D+A+N	1594.916	-	-	-
R+S+A+N	1600.141	5.225	1	<0.05
R+S+D+A	1596.993	2.077	1	NS
R+S+D+N	1602.837	7.921	1	<0.01
R+D+A+N	1609.740	14.824	1	<0.001
S+D+A+N	1663.400	68.484	1	<0.001
R+S+D+A	1596.993	-	-	-

S+D+A	1666.384	69.391	1	<0.001
R+D+A	1612.388	15.395	1	<0.001
R+S+A	1602.008	5.015	1	<0.05
R+S+D	1605.591	8.598	1	<0.01
R+S+D+A	1596.993	-	-	-
S+R+D+Place	1596.589	0.404	1	NS
S+R+D+Associated disease	1596.032	0.961	1	NS
S+R+D+Type of Treatment	1596.773	0.22	1	NS
S+R+D+Histology	1595.163	1.83	1	NS
S+R+D+Tumour size	1594.386	2.607	1	NS

Where

A-Age of the woman, N-Change in nucleus, S-Stage of the disease, D-Duration of the radiotheraphic treatment, R-Recurrence during the follow-up period

Table Number: 8
Hazard Ratios from the Cox PH Model
(Multivariable Analysis-II)

Variable		Parameter Estimate	Standard Error	P<Chi. Sq.	Hazard Ratio	95% Confidence Interval	
						Lower Limit	Upper Limit
Recurrence		1.54	0.176	<0.001	4.685	3.315	6.620
Duration of Radiotheraphic Treatment	46-60 days	-	-	-	-	-	-
	≤45	0.384	0.263	0.144	1.468	0.877	2.460
	61-75	0.358	0.205	0.081	1.431	0.957	2.139
	≥75	0.435	0.255	0.088	1.544	0.938	2.543
Age(in Years)	<40	-	-	-	-	-	-
	40-49	0.131	0.222	0.555	1.140	0.73	1.763
	50-59	0.389	0.243	0.109	1.475	0.917	2.375
	≥ 60	0.310	0.249	0.213	1.363	0.837	2.219

Stage of Diseases	I	0.373	0.277	0.178	1.452	0.8414	2.497
	II	0.967	0.288	0.001	2.630	1.495	4.625
	III	1.967	0.536	<0.001	7.148	2.50	20.439

Summary and Conclusion

Modelling the survival data for prognostic factors in cancer research is on the rise recently in India (Swaminathan, 2002). The main reason is the paucity of follow up information, which is so vital in survival studies. The first and for most assumption of the Cox proportional hazard model is that the censorings are at random. This has been verified. The verification of adequacy of the number of events that are being studied at all levels of factors is desirable. This has been taken care of by suitably classifying the levels of factors in such a way that at each level, there are adequate numbers of events. The stage of the disease has few sample sizes at level IV but proportion of the event is cent percent. Hence, analysis II has been performed by omitting stage IV to ascertain the suitability of the Cox-regression model. It confirms that the inclusion of the variable stage IV level would not alter the validity of the model. The sample size considerations in concern, the over all sample size is 15 per variable. In general, the model with four variables is reliable to make a decision about the prognostic variables of the breast cancer survival. Here, the following four variables have been identified as significant prognostic factors of breast cancer survival for Cox-regression model:

i. Recurrence of the disease

ii. Stage of the disease at diagnosis

iii. Duration of radiotheraphic treatment

iv. Age of the woman at diagnosis

The findings reported by Wong, 2003 is almost similar to the findings of the present study though their purpose of applying the Cox-regression is mainly for comparison between the treatment regimens or comparison of dosage level of radiotheraphy, their ultimate aim being identification of the prognostic factors of five-year survival probability of breast cancer patients.

Based on this, two analyses have been done, i.e., with stage IV covariate and without stage IV covariate. Finally, a model with four covariates, namely, recurrence of the disease, age of the woman, duration of radiotheraphic treatment and stage of the disease, has been identified as the prognostic factors of breast cancer survival after the completion of treatment.

References

[1] Bender R, Augustin T and Blettner M (2005): Generating Survival Times to Stimulate Cox Proportional Hazards Models, Statistics in Medicine; 24:1713-1723.

[2] Bradburn MJ, Clark TG, Love SB, Altman DG(2003(1)): Survival Analysis Part II: Multivariate Data Analysis-Choosing a Model and Assessing its adequacy and fit. British Journal of Cancer. 89:605-611.

[3] Collett D. (1994): Modelling Survival Data in Medical Research, London: Chapman and Hall/CRC.

[4] Cox DR(1972): Regression Model and Life Tables, Journal of Royal Statistical Society.(13)34; 187-220.

[5] Peduzzi P, Concato J Feinstein AR and Holford TR(1995): Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis-II: Accuracy and Precision of Regression Estimates. Journal of Clinical Epidemiology; 48:1503-1510.

[6] Swaminathan R. (2002): Some Statistical Models in Cancer Survival and their Applications, Unpublished Ph.D thesis, Department of Statistics, University of Madras, Chennai.

[7] Wong, FCS, Tung SY, Leung TW, Sze WK, Wong VYW, Collin MM et al(2003): Treatment Results of High-Dose-Rate Remote after loading Brachytheraphy for Cervical Cancer and Retrospective Comparison of Two Regimans. International Journal of Radiation Oncology Biol. Phys.; 55(5):1254-1264.