Advances and Applications in Mathematical Sciences
Volume 21, Issue 10, August 2022, Pages 6081-6102

MULTIVARIATE STATISTICS TO SURVIVAL DATA

M. MANIKANDAN, M. RAMAKRISHNAN and R. RAVANAN

[1]Research Scholar
[2]Assistant Professor
Department of Mathematics
Ramakrishna Mission Vivekananda
College, Chennai, India
E-mail: manijee211993@gmail.com
mramkey@rkmvc.ac.in

[3]Joint Director of Collegiate Education
Chennai Region, Chennai, India
E-mail: ravananstat@gmail.com

Abstract

Survival analysis is absorbed to analysis of time to event data. To handle these outcomes, as well as censored observations, where the event was not observed during follow-up, survival analysis methods should be used. Kaplan-Meier estimation can be used to create graphs of the observed survival curves, while the log-rank test can be used to compare curves from different groups. Regression and Discriminant analysis are part of Multivariate statistics. Regression analysis is one such concept which explores the relationship between two or more quantifiable variables so that one variable can be predicted from other. Discriminant analysis uses discrete response and continuous predictors to classify observations into different groups, particularly using linear (or) quadratic classification function. The aim of the study is to analyse the survival data with survival techniques, multiple regression and discriminant analysis and interpret the outcomes in all ways. Such discriminant classifiers are used to identify customers with repaying capacity of loans in banking industry it is also used in clinical diagnosis for identifying specific diseases, based on clinical and social democratic parameters obtained from the patients. The study has made use of Advanced Statistical free and open software *R* (4.1.1) and its packages "survival", "survminer", "Surv Reg CensCov" and "survreg" are used to analyze the Data.

2020 Mathematics Subject Classification: 62H10.
Keywords: Survival analysis, Kaplan-Meier estimator, Regression Analysis and Discrimination Analysis.
Received January 25, 2022; Accepted March 5, 2022

1. Introduction

Survival Analysis is the study about survival data. Survival data include survival time, event and characteristics related to event. Survival curves are generated by Kaplan-Meier method. Traditional Kaplan-Meier method used for finding survival probabilities for censored and non-censored observations. Survival to any time point is calculated as the product of the conditional probabilities of surviving each time interval. The calculations are simplified by ignoring censored times. Cox proportional Hazard (CPH) model is well known for analyzing survival data because of its simplicity as it has no assumption regrading survival distribution. CPH helps to find out hazard ratio based on coefficients. These coefficients are ease to interpret and clinically meaningful (D. Hosmer S. Lemeshow 1989). In Parametric survival models, it is considered that survival time follows known distributions as Weibull, Exponential, Log-normal, and Log-logistic distributions. Parametric models may be acceleration failure time (AFT) and PH models. The AFT models are useful for comparison of survival times whereas the PH is applicable for comparison of hazards (DG. Kleinbaum, M. Klien [8]). Parametric models are better over PH with respect to sample size and relative efficiencies (A. Nardi, M. Schemper, [2]).

Logistic Regression and Linear Discriminant Analysis are multivariate statistical methods which can be used for the evaluation of the association between various covariates and categorical outcomes. Both methodologies have been extensively applied in research, especially in Medical and Sociological Sciences. Logistic Regression is a form of regression which is used when the dependent variable is dichotomous, discrete, or categorical and the explanatory variables are of any kind. Discriminant analysis is a similar classification method that is used to determine which set of variables discriminant between two or more naturally occurring groups and to classify an observation into these known groups. In both Discriminant analysis and Logistic regression can be used to predict the probability of a specified outcome using all or a subset of available variables.

Regression analysis is performed so as to determine the correlations between two or more variables having cause-effect relations and to make predictions for the topic by using the relation. The regression using one single

independent variable is called univariate regression analysis while the analysis using more than one independent variable is called multivariate regression analysis (Tabachnick, [5]). Through univariate regression analysis, the relations between a dependent variable and independent variable are analyzed, and the equation representing the linear relations between the dependent and independent variables is formulated. The regression models with one dependent variable and more than one independent variable, however, is known as multivariate regression analysis.

Logistic regression is used to estimate the association of one or more independent (predictor) variables with a binary dependent (outcome) variable (Patrick, 2021). Linear regression not only tests for relationships but also quantifies their direction and strength (Schober and Vetter, [21]). Ping Jin et al., [22] found that Multi-task Logistic regression (MTLR) and Cox did extremely well-better than the other survival models, Kaplan-Meier estimator and accelerated failure time model on maximizing profits. This suggests that MTLR and Cox are likely good choices for predicting Reservation Price distribution predictions, in general.

Survival model an estimate of $S(t)$ denoted by $\hat{S}(t)$ is estimated using Nonparametric and COX Models. The author plans to estimate, log survival time using multiple linear regression model with variables Dose and Clinic as independent variables. Discriminant analysis is used to group the observation into 'event' or 'censor' observation.

2. Survival Function and Methods

Survival function is a key term in survival analysis, along with censoring and event. The concept of a survival function is essential for the understanding of survival analysis. The survival function is defined as the probability of the outcome event not occurring up to a specific point in time, including the point of observation (t) and is denoted by

$$S(t) = P(T > t) = 1 - F(t) \tag{2.1}$$

T - Random variable denoting the time to event

$P(T > t)$ -Probability of not experience the event up to and including time t

$F(t)$-Cumulative distribution function.

The ratio of the number of events occurring during the entire study period to the total number of observations is termed the incidence rate. The hazard function is a function for calculating the instantaneous incidence rate at any given point in time and is denoted by $h(t)$.

$$h(t) = \lim_{\Delta \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \tag{2.2}$$

2.1 Parametric survival using accelerated failure time (AFT) model

Let T is a random variable of survival time and X is a column vector of the covariates $X_1, X_2, \ldots, X_p$ the AFT model defines the relationship of survival function for every time $t \epsilon T$, $S(t \mid X)$, and the covariates as follows

$$S(t \mid X) = S_0[te^{\beta^t X}] \tag{2.1.1}$$

where S_0 is the baseline survival function and $\beta^t = \beta_1, \beta_2, \ldots, \beta_p$ is a vector of regression coefficients. The factors $(e^{\beta^t X})$ in the equation is known as the accelerated factor which accelerates the survival function with covariate $X = 0$. The AFT model assumes that the effects of the covariate are fixed and multiplicative by the accelerative factor on the time scale of t. However, it does not assume that the model holds the constant hazards assumptions as in PH model.

The relationship between covariates and the survival time can be also illustrated as a linear relation between the natural logarithm of survival time and covariate X, that is

$$Y = \log T = \mu + \theta^t X + \sigma W \tag{2.1.2}$$

Where μ is the slope, $\sigma > 0$ is an unknown scale parameter, $\theta^t = (\theta_1, \theta_2, \ldots, \theta_p)$ is the vector of regression coefficients, $\theta = -\beta$, σ is a scale parameter and W is a distribution error which is a random variable and assumed to follow a certain parametric distribution. For every distribution of W, there is a related parametric for T. The name for the AFT model come

from the distribution of T rather than the parametric distribution of $\log T$. The commonly parametric distributions, which correspond to the AFT model are Weibull, Exponential, Log-logistic and Log-normal. However, the AFT models that are considered in this section are Weibull AFT model, Exponential AFT model, Loglogistic AFT model and Lognormal AFT model. The survival function of T_i, $i = 1, 2, \ldots, n$ is given by

$$S_i(t) = \Pr(T_i \geq t) = \Pr(\log T_i \geq \log t) = \Pr(Y_i \geq \log t)$$

$$= \Pr(\mu + \theta^t X + \sigma W \geq \log t)$$

$$= \Pr\left(W_i \geq \frac{\log t - (\mu + \theta^t X)}{\sigma}\right) \tag{2.1.3}$$

Table 1. The survival function and hazard functions.

Distribution	PDF	CDF	Survival	Hazard
Exponential	$\lambda e^{-\lambda t}$	$1 - e^{-\lambda t}$	$e^{-\lambda t}$	Λ
Weibull	$\lambda\gamma(\lambda t)^{\gamma-1}e^{(-\gamma t)^{\gamma}}$	$1 - e^{(-\lambda t)^{\gamma}}$	$e^{(-\lambda t)^{\gamma}}$	$\lambda\gamma(\lambda t)^{\gamma-1}$
Lognormal	$\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\log t-\mu)^2}{2\sigma^2}}$		$\frac{1}{t\sigma\sqrt{2\pi}}$ $\int_0^{\infty} \frac{1}{x} e^{-\frac{(\log t-\mu)^2}{2\sigma^2}} dx$	
Loglogistic	$\frac{\lambda k(\lambda t)^{[k-1]}}{\{[1 + (\lambda k)^k]\}^2}$	$1 - \left(\frac{1}{1 + (\lambda k)^k}\right)$	$\frac{1}{1 + (\lambda k)^k}$	$\frac{\lambda k(\lambda t)^{[k-1]}}{1 + (\lambda k)^k}$

2.2 Non-Parametric Estimator for Survival Function

The most common non-parametric approach in the literature is the Kaplan-Meier (or product limit) estimator. The Kaplan -Meier estimator works by breaking up the estimation of $S(t)$ into a series of steps or intervals based on observed event times.

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

$$= \prod_{t_i<t}\left(1 - \frac{d_i}{n_i}\right) \tag{2.2.1}$$

This estimator holds for all $t > 0$ and it depends only on two variables, n_i and d_i which are n_i-number in risk at time t_i, d_i - number of events at time t_i.

2.3 Semi-Parametric model. Cox PH model is one type of regression model which is commonly used in medical research for investigating the association between the survival time of patients and one or more predictors variables. This method is used to evaluate the effect of many factors on survival time, and it allow to examine the specified factors influence the rate of particular event that occurs at a particular time.

The general form of hazard function is written as

$$h(t, x, \beta) = h_0 \cdot r(x, \beta) \tag{2.3.1}$$

where h_0 reflects how hazard function changes with survival time, and $r(x, \beta)$ characterizes how hazard function changes with covariates. Cox (1972) has proposed exponential function for $r(x, \beta)$, and the hazard function is written as

$$h(t, x, \beta) = h_0 \cdot e^{x\beta} \tag{2.3.2}$$

when x changed from x_0 to x_1, the hazard ratio is

$$HR(t, x_0, x_1) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} = \frac{h_0(t) \cdot e^{x_1\beta}}{h_0(t) \cdot e^{x_0\beta}} = e^{\beta(x_1 - x_0)} \tag{2.3.3}$$

The model is termed Cox proportional hazard model, researchers are interested in the parameter β, which is interpreted as changing rate of hazard when the covariate changed by $(x_1 - x_0)$ unit, the baseline hazard function $h_0(t)$ remains unknown, so the model is called semi-parametric model.

3. Multivariate Analysis

3.1 Regression Analysis. The regression analysis can be used to identify the explanatory variables that are related to a response variable, to describe the form of the relationships involved, and to provide an equation for predicting the response variable from the explanatory variables.

On Regression methods that fall under the rubric of ordinary least squares (OLS) regression, including simple linear regression, polynomial regression, and multiple linear regression. OLS regression is the most common variety of statistical analysis today. Other type of regression models also available, including logistic regression and Poisson regression.

For most of this will be predicting the response variable from a set of predictor variables using OLS. Regression fit models of the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}; i = 1, 2, \ldots, \quad n \tag{3.1.1}$$

where n is the number of observations and k is the number of predictors variables. $\hat{Y}_i$ is the predicted value of the dependent variable for observation i. X_{ji} is the j^{th} predictors value for i^{th} observation. $\hat{\beta}_0$ is the intercept. $\hat{\beta}_j$ is the regression coefficient for the j^{th} predictor.

Our goal is to select parameter (intercept and slopes) that minimize the difference between actual response values and those predicted by the model. Specifically, model parameters are selected to minimize the sum of squared residuals:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (\hat{Y}_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}))^2$$

$$= \sum_{i=1}^{n} \varepsilon_i^2 \tag{3.1.2}$$

To properly interpret the coefficient of the OLS model satisfy a number of statistical assumptions:

- NORMALITY- for fixed values of the independent variables, the dependent variable is normally distributed.

- INDEPENDENCE- the Y_i values are independent of each other.
- LINEARITY- the dependent variable is linearly related to the independent variables.
- HOMOSCEDASTICITY- the variance of the dependent variable does not vary with the levels of the independent variables. (Constant variance).

If violate these assumptions, our statistical significance tests and confidence intervals may not be accurate.

3.2 Discriminant Analysis. Discriminant analysis is used to predict the probability of belonging to a given Class or Category based on one (or) multiple predictor variables. It works with continuous and (or) categorical predictors variables.

Linear discriminant analysis and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

3.3 Logistic Regression. Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary/categorical outcome, we use dummy variables. Researcher can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

The logistic regression method assumes that:

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative.
- There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is $\log it(p)$

$= \log\left(\frac{p}{1-p}\right)$, where p is the probabilities of the outcome.

- There is no influential values (extreme values or outliers) in the continuous predictors
- There is no high inter correlations (i.e. multicollinearity) among the predictors.

To improve the accuracy of model, make sure that these assumptions hold true for the data. Both Logistic regression and Discriminant analysis can be used for binary classification tasks.

4. Application to the two methadone treatment clinics for heroin Addicts (ADDICTS)

Retention of patients in methadone treatment was studied in a cohort of 238 heroin addicts who entered maintenance programs between February 1986 and August 1987. This Australian study by Caplehorn et al. (1991), in which two methadone treatment clinics for heroin addicts were compared and is used to assess patient time remaining under Methodone treatment. A patient survival time was determined as the time, in days, until the person dropped out of the clinic or was censored. The two clinics differed according to their live-in policies for patients. This data is downloaded from http://web1.sph.emory.edu/dkleinb/surv3.htm.

Table 2. Summary of data set.

Variable	Description	Codes (Values, Percentage)
ID	Study ID	1-238
Days. Survival	The Time until the patient dropped out of the clinic or was censored	Days
Status	Indicates whether the patient dropped out of the clinic or was censored	1 = Dropped out (150, 63%) 0 = Censored (88, 37%)

Prison	Whether the patient has a prison record or not	1= prison record (81, 127) 0= otherwise (69, 111)
Clinic	Whether the patient taken treatment in Clinic 1 or Clinic 2	0=clinic 1 (122, 163) 1=clinic 2 (28, 75)
Dose	Maximum methadone dose	mg/day

The Addicts data sets contains censored observations. The researcher aims to analyze the Addicts data in three different ways.

- Model 1 contains all observations with censoring.
- Model 2 contain observation with event i.e., those patients depart from Clinic and
- Model 3 represents the data contains drop out from the study.

Models 2 and 3 does not contain any censoring observations.

Model 1 contains 238 Patients' observations that also includes censoring. At the time, the study is terminated, 88 are continue in the clinic after the study or dead or drop out. Researcher consider the 88 drop out as censored.

Table 3. KM survival probability for Addicts data.

Time	Risk set	Survival probability	Std. error	LCI 95%	UCI 95%
7	236	0.9960	0.0042	0.9875	1.0000
13	235	0.9920	0.0060	0.9799	1.0000
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
489	95	0.5090	0.0345	0.4452	0.5810
496	94	0.5030	0.0346	0.4398	0.5760

.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
878	13	0.1700	0.0367	0.1116	0.2600
892	10	0.1530	0.0368	0.0958	0.2450
899	9	0.1360	0.0364	0.0807	0.2300

At the end of the Seventh day, risk set in the Addict data contains 236 patients, because 2 were censored. The survival probability at 7th day after taken treatment is 0.9960. In 878th day of treatment, 13 patients were at risk set and the survival probability is 0.17.

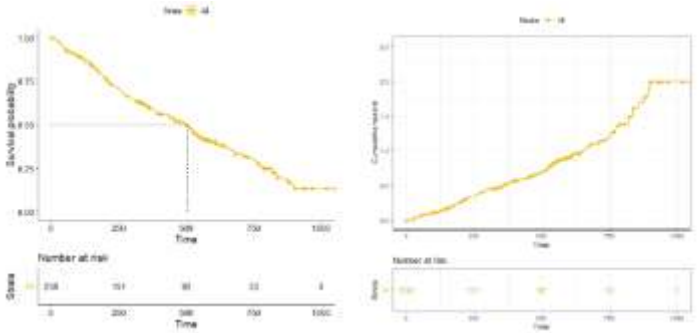

Figure 1. Non-Parametric Survival and Hazard Plots: Model 1.

The Figure 1 shows that KM curve of point estimation for addict data (Model 1), the estimate obtained are invariably expressed in graphical form. The graph plotted between estimated survival probabilities (on Y axis) and time past after entry into the study in days (on X axis) consists of vertical and horizontal lines. The survival curve is drawn as a step function. The estimated median survival time for Model 1 is 504 days.

In Table 4 shows that KM survival probability for Model 2 (patients depart from Clinic). The median survival time is 280 days.

Table 4. KM survival probability for Model 2 in Addicts data.

Time	Risk set	Survival probability	Std. error	LCI 95%	UCI 95%

7	150	0.9933	0.0066	0.9804	1.0000
13	149	0.9867	0.0094	0.9685	1.0000
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
280	76	0.5000	0.0408	0.4261	0.5868
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
892	2	0.0067	0.0066	0.0009	0.0470
899	1	0.0000	-	-	-

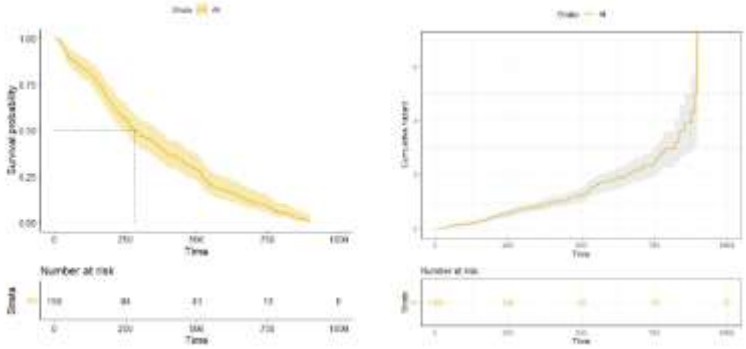

Figure 2. Non-Parametric Survival and Hazard Plots: Model 2 (patients depart from Clinic).

Figure 2 shows that the survival and hazard curve for patients depart from Clinic Addicts data without censoring and also horizontal vertical line for figure shows that median survival time 280 days.

In Model 3 (Patient drop out from the study), KM survival probability shows in Table 5. 50% of patient surviving longer than at 532 days. In Figure 3 also confirms about the median survival time.

Table 5. KM survival probability for Model 3 in Addicts data.

Time	Risk set	Survival probability	Std. error	LCI 95%	UCI 95%
2	88	0.9773	0.0159	0.9466	1.0000
28	86	0.9545	0.0222	0.9120	0.9991
53	84	0.9318	0.0269	0.8806	0.9860
.	.	.	.	.	.
.	.	.	.	.	.
.		.	.	.	.
532	45	0.5000	0.0533	0.4057	0.6162
.	.	.	.	.	.
.	.	.	.	.	.
.		.	.	.	.
1052	2	0.0114	0.0113	0.0016	0.0798
1076	1	0.0000	0.0000	0.0000	0.0000

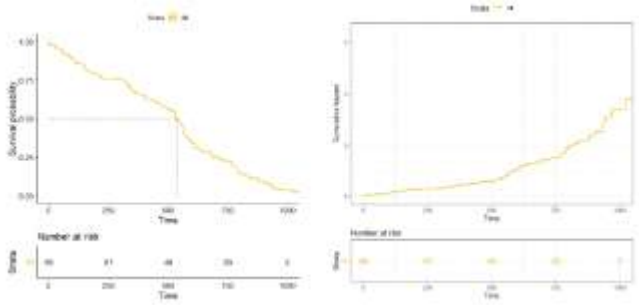

Figure 3. Non-Parametric Survival and Hazard Plots: Model 3 (Patient drop out from the study).

Table 6. Comparison of the survival probability for the three models to ADDICT data.

Survival Probability			
Time	Model 1	Model 2	Model 3
7	0.9960	0.9933	0.9773
13	0.9920	0.9867	0.9773
.	.	.	.
.	.	.	.
.	.	.	.
275	06680	0.5067	0.7614
280	0.6630	0.5000	0.7614
.	.	.	.
.	.	.	.
.	.	.	
892	0.1530	0.0067	0.0909
899	0.1360	0.000	0.0909

From the Table 6, it is observed that the survival probability is 0.66 for overall observation (Model 1), 0.50 in Model 2 and 0.76 in Model 3 at 280 days. In the same time, Model 2 patient only survives 50%.

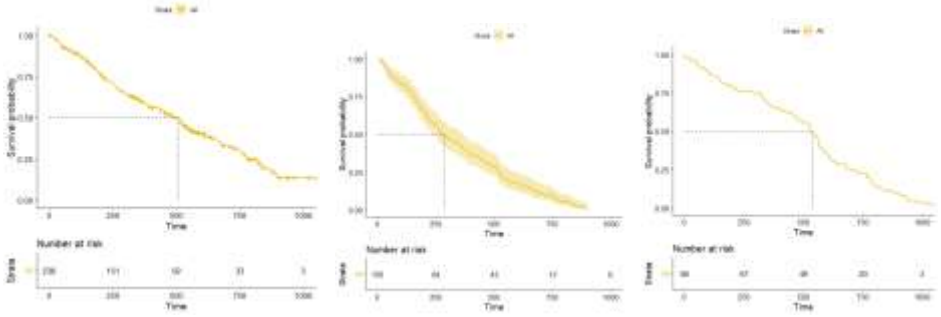

Figure 4. KM survival curves for Model 1, 2 and 3.

The above Figure 4 shows that KM curve of point estimation for Addict data (Model 1, 2 and 3), the estimate obtained are invariably expressed in graphical form. The estimated median survival time for Model 1, Model 2 and Model 3 are 504 days, 280 days and 532 days respectively. The Patient depart from the study is earlier when compared to dropout.

Semiparametric Cox model Estimation of hazard were analysed for all the three models and observed the significant covariates for survival time. The following Table shows the covariates, Hazard ratio and its corresponding p-value for all models.

Table 7. Estimation of Hazard ratio and p-value for three models.

		Dose	clinic 2
Model 1	HR	0.9663	0.3864
	P-value	< 0.01**	< 0.01**
Model 2	HR	0.97907	1.31576
	P-value	<0.01**	0.1958
Model 3	HR	0.9689	0.4368
	P-value	< 0.01**	< 0.01**

Note: ** has denote significant at 1% level.

From the above Table 7, it is observed that the covariates dose and Clinic are highly significant in Models 1 and 3 but the covariate Clinic is not significant in Model 2. Hazard for a patient taking treatment in clinic 2 is higher when compared to clinic 1.

Table 8. AIC values for different parametric models.

Parametric Models	Model 2	Model 3
AIC (Exponential)	2051.881	1265.884
AIC (Weibull)	2029.729	1238.502
AIC (Log-logistic)	2056.763	1267.656
AIC (Lognormal)	2060.190	1294.293

Weibull model is the best parametric model for both Model 2 and 3 because its AIC value is less compared to other models.

Table 9. Hazard Ratio and Event Time Ratio using Weibull distribution.

	Model 2		Model 3	
	Dose	Clinic2	Dose	Clinic 2
HR	0.9792	1.3133	0.9784	0.5457
ETR	1.0149	0.8253	1.0127	1.4169

In Model 2, hazard for a patient taking treatment in clinic 2 is higher when compared to clinic 1 but it is less in model 3, the same way of interpretation also observed in Weibull AFT model.

When using survival methods for addicts data, the researcher observed Dose and clinic are significant variables for survival time. Using Multivariate analysis, Discriminant and Logistic Analysis, the researcher taken censored indicator (binary 1/0) as dependent variable and Dose and Clinic are predictors. Confusion Matrix for Discriminant and Logistic Analysis given below Table 10. In Both Multivariate Analysis, 71% of the observations are correctly classified, only 29% are misclassified.

Table 10. Confusion Matrix for Discriminant and Logistic Analysis.

Discriminant			Logistic	
Predicted	0	1	0	1
0	47	28	39	20
1	41	122	49	130
Total	88	150	88	150

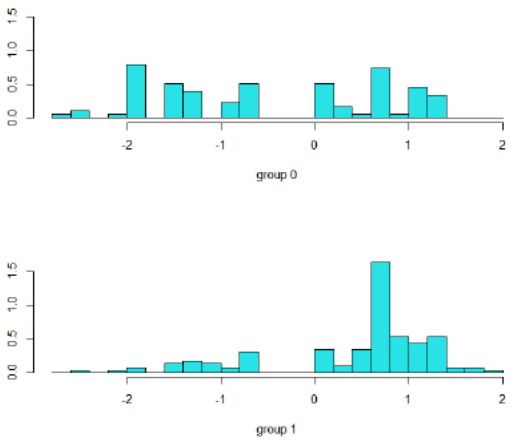

Figure 5. Classification plot for discriminant function.

The Addicts dataset has a binary response (outcome, dependent) variable called status. There are Three predictor variables in the data, but taking only two variable Dose and Clinics in the Heroin addict data to estimates the logistic regression model. Below Table shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. Identified the variables Dose and Clinic 2 are statistically significant because their P-values are less than 0.05.

Table 11. Summary of the logistic regression model.

	Estimate	Std. error	Z	P-value
Intercept	2.6673	0.6638	4.018	< 0.01**
Dose	-0.0263	0.0105	-2.510	<0.05*
Factor (clinic2)	-1.5424	0.3049	-5.058	< 0.01**

Note: 1)** has denote significant at 1% level. 2) * has denote significant at 5% level.

In the classification analysis also identified the covariates Dose and Clinic are significantly related with censor variable.

Finally, researcher analyse Addicts data as two different data i.e., patient depart (Model 2) and drop out (Model 3) with the help of Multiple Linear

Regression. It is observed that, In Model 2, the variable Dose is significant but Clinic is not statistically significant. In Model 3, both variables are significant and then contribution to log(time) is positive for both predictors.

Table 12. Summary estimation of Model 2 and 3.

	Model 2		Model 3	
	Estimate	P-value	Estimate	P-value
Intercept	4.147	< 0.01**	3.427	< 0.01**
Dose	0.024	< 0.001**	0.027	< 0.01**
Clinic 2	-0.277	0.164	0.465	<0.05*

Note: 1) ** has denote significant at 1% level. 2) * has denote significant at 5% level.

From the above table, it is observed that increase in Dose results in increase in the log(time). This increase is comparatively less in Clinic1 in Model 2 compare to Model 3.

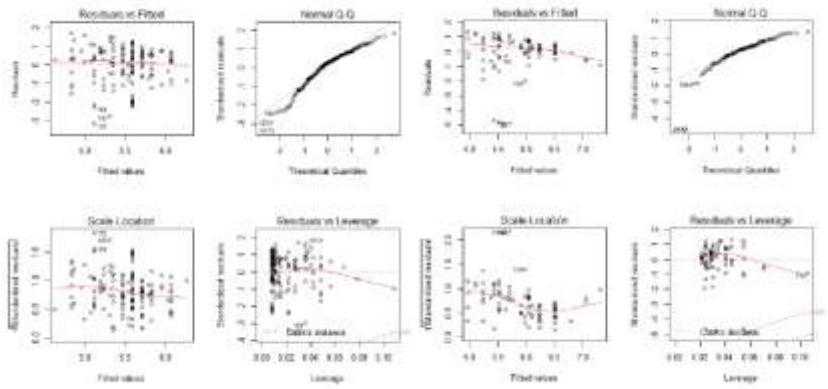

Figure 6. Diagnostic plots for Model 2 and 3.

The above graph shows that the Model 2 and 3 satisfies the assumptions for Multiple linear regression.

Durbin-Watson Test is used to test the presence of autocorrelation in a time series data.

Table 13. Durbin-Watson Test.

	Lag	Autocorrelation	D-W Statistic	P-value
Model 2	1	0.2402	1.5013	<0.01**
Model 3	1	0.0870	1.8104	0.342

From the above Table it is observed that P-value (<0.01) for Model 2 is significant but Model 3, it is not significant in Durbin-Watson Test. So, the Model 3 suggests a lack of auto correlation and conversely an independence of errors. The lag value (1 is this case) indicates that each observation is being compared with the one next to it in the data set. Although, appropriate for the time dependent data, the test is less applicable for data that is not clustered in the fashion. Homoscedasticity (Non constant variance score test) the score tests are significant for the Model 2 and 3. All the assumption for simple regression also apply for multiple regression with one addition. If two of the independent variables are highly related, this leads to a problem called Multicollinearity. Variance inflation factor (VIF), VIF scores should be close to 1 but under 5 is fine and 10+ indicates that the variable is not needed and can be removed from the model. All the values in this analysis have scores close to 1 for Model 2 and 3.

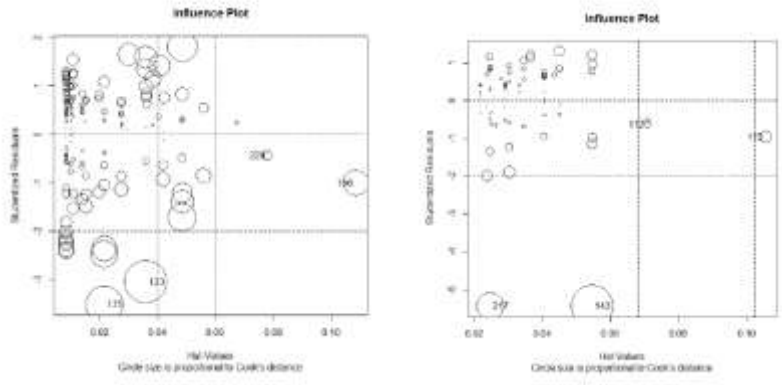

Figure 7. Influence plot for model 2 and 3.

The above plot shows that the observations 175 and 123 are outliers, 229 and 156 are influential observation and observed no high leverage point in Model 2. In model 3, it shows that 217 and 143 are outliers, 112 have high leverage and 115 is influential observation.

Table 14. Time values are listed according to covariates for the two models 2 and 3.

Dose	Model 2		Model 3	
	Clinic 1	Clinic 2	Clinic 1	Clinic 2
25mg	87	66	96	153
50mg	159	121	189	301
75mg	290	220	371	591

The Patient who is taking treatment in Clinic 1 and Dose 50mg/day will be expected to depart at 159 days but it is 121 days for patient taking treatment in Clinic 2. It is understood that the survival time for Clinic 1 is higher than Clinic 2 in Model 2 but it is contradicted in Model 3. *R* software is immensely useful to generate suitable graphical plots that ease the comparison and makes inferences more lucid.

5. Summary and Conclusion

In this study, the researcher aims to analyze the Addicts data in three different ways. Model 1 contains all observations with censoring, Model 2 contain observation with event i.e., those patients depart from Clinic and Model 3 represents the data contains drop out from the study. Models 2 and 3 does not contain any censoring observations. Comparison of survival probability in the three models and corresponding Survival Curves are obtained using Nonparametric estimation. Among all Parametric survival models, it is observed that Weibull model is fitted for ADDICTS data because its AIC value is less compared to other models. Semiparametric Cox model estimation of hazard were analysed for all the three models and observed Dose and Clinic are the significant covariates for survival time. Multiple regression used for modelling the log survival time and the discriminant analysis for identifying the censored observations. When using Multivariate analysis like Logistic regression and Discriminant Analysis were used for Model 1 in ADDICTS data, 71% of the observations are correctly classified, only 29% are misclassified in both methods. The covariates Dose and Clinic are significant to the censor indicator (1/0) in Logistic Regression. Multiple linear regression method used for Model 2 and 3 and the predictor variables

are statistically significant for Model 2 and 3. The Patient who is taking treatment in Clinic 1 and Dose 50mg/day will be expected to depart at 159 days but it is 121 days for patient taking treatment in Clinic 2. It is understood that the survival time for Clinic 1 is higher than Clinic 2 in Model 2 but it is contradicting in Model 3. From Cox and Regression Models, we arrive at the same interpretation. Using Survival Models, classification techniques and multiple linear regression for ADDICTS data results with the same significant covariates, namely the 'Dose' and 'Clinic'.

References

[1] A. V. Peterson, Expressing the Kaplan-Meier estimation as a function of empirical sub survival functions, Journal of the American Statistical Association 72 (1977), 854-858.

[2] A. Nardi and M. Schemper, Comparing Cox and Parametric Models in Clinical Studies, Statistics in Medicine 22 (2003), 597-610.

[3] Abdul-Fatawu Majeed, Accelerated failure time models: An application in insurance attrition, The Journal of Risk Management and Insurance, Bangkok, Thailand 24(2) (2020), 12-35.

[4] Alfensi Faruk, The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data, Journal of Physics: Conference Series: 974 (2018).

[5] B. G. Tabachnick and L. S. Fidell, Using multivariate statistics (3rd ed.), New York: Harper Collins, (1996).

[6] MS. Brandon George, Samantha Seals PhD and Inmaculada Aban PhD, Survival analysis and regression models, NIH Public Asses 21(4) (2014), 686-694.

[7] D. Kalbfleisch and R. L. Prentice, Statistical Analysis of Failure Time Data, John Wiley and Sons, New York, (1980).

[8] D. G. Kleinbaum and M. Klein, Survival analysis: A self-learning text, New York, Springer Series, (1996).

[9] D. R. Cox, Regression models and life tables, Journal of the Royal Statistical Society, Series B 34 (1972), 187-220.

[10] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, John Wiley and Sons, Inc., New York, (1989).

[11] E. L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, Journal of American Statistical Association 53 (1958), 457-481.

[12] Ellen Marshall and Sofia Maria Karadimitriou, Multiple regression in R, University of Sheffield.

[13] Gulden Kaya Uyanik and Nese Guler, A study on multiple linear regression analysis, Procedia - Social and Behavioral Sciences, Elsevier 106 (2013), 234-240.

[14] J. R. Caplehorn and J. Bell, Methadone dosage and retention of patients in maintenance treatment, National Center for Biotechnology Information 154(3) (1991), 195-199.

[15] Jorge Luis Romeu, Ph.D. An example of survival analysis data applied to Covid-19, (2020).

[16] K. McGarigal et al., Multivariate Statistics for Wildlife and Ecology Research Springer-Verlag New York, (2000).

[17] M. D. Patrick Schober, Logistic regression in medical research, International Anesthesia Research Society (IARS) 132(2) (2021), 365-366.

[18] Mukesh Kumar, et al., Parametric survival analysis using R: Illustration with lung cancer data, Wiley Cancer Reports, (2019).

[19] Mushtak and A. K. Shiker, Multivariate statistical analysis, British Journal of Science 6(1) (2012), 55-66.

[20] Nikolaos Pandis, Associate Editor of Statistics and Research Design, Logistic Regression Bern., Switzerland, and Corfu., Greece (2017), 824-826.

[21] P. Schober and T. R. Vetter, Linear regression in medical research, International Anesthesia Research Society (IARS) 132 (2021), 108-109.

[22] Ping Jin et al., Using survival prediction techniques to learn consumer-specific reservation price distributions, PLOS ONE, 16(4) (2021).

[23] Aviral Gupta and Akshay Sharma, Review of Regression Analysis Models, International Journal of Engineering Research and Technology (IJERT) 6(8) (2017), 58-61.

[24] R. I. Kabacoff, R in Action Data analysis and graphics with R, second edition Manning Publication Co, Shelter Island, (2015).

[25] Shankar Khanal, Accelerated failure time models: an application in the survival of acute liver failure patients in India, International Journal of Science and Research 3(6) (2014), 161-166.

[26] Stuart W. Grant, Graeme L. Hickey and Stuart J. Head, Statistical primer: multivariable regression considerations and pitfalls, European Journal of Cardio-Thoracic Surgery 55 (2019), 179-185.

[27] T. W. Anderson, An introduction to multivariate statistical analysis, John Wiley and Sons, New York, Second Edition, (1971).

[28] W. Weibull, A statistical theory of the strength of materials, Generalstabens Litografiska Anstalts Forlag, Stockholm, (1939).