ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Application of Parametric, Semi-Parametric and Non-Parametric Survival Models for Myocardial Infarction (Mi) Patients

Dr. T. Leo Alexander [(1)], Tamanna.B[(2)]

Associate Professor, Department of Statistics, Loyola College, Chennai, India[(1)]

Research Scholar, Department of Statistics, Loyola College, Chennai, India[(2)]

ABSTRACT: In the statistical area of survival analysis, an **Accelerated Failure Time Model** (**AFT model**) is a parametric model that is used to analyse the 'disease' which is a result of some mechanical process with a known sequence of intermediary stages. An AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. The Kaplan Meier Estimator is a Non-Parametric statistic used toestimate the survival function from lifetime data. The Cox Regression Model is a Semi-Parametric model, no assumption about the shape of hazard function, but make assumption about how covariates affect the hazard function. This Paper deals with Kaplan Meier curves, Cox Regression Model and AFT Models namely, exponential, log-normal, weibull and log-logistic distributions to evaluate the survival times of 500 MI patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The analysis is done on R.

Section 1 gives a brief introduction about survival analysis and we have discussed in short aboutMyocardial Infarction. Section 2 presents in detail about the technique of Kaplan Meier Estimator, Cox Regression Model and the application of AFT Models to survival data. An analysis based on the various techniques is done and the results are presented in Section 3 followed by the conclusion of the study in Section 4.

KEYWORDS: Censoring, Cox Regression, Hazard Ratio, Kaplan Meier Estimator, Log-likelihood, Survival Rates.

I. INTRODUCTION

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems.Examples of survival data are the lifetime of electronic devices, components, or systems (reliability engineering); felons' time to parole (criminology); duration of first marriage (sociology); length of newspaper or magazine subscription (marketing); and worker's compensation claims (insurance) and their various influencing risk or prognostic factors. Survival data can include survival time, response to a given treatment, and patient characteristics related to response, survival, and the development of a disease.

For the basic concepts of regression and the tests of significance we have referred to two books namely, Introduction to linear regression analysis, 5th edition (2012), Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey vining and Statistics for business and economics 11th edition (2012), David R.Anderson, Thomas A.Williams, Dennis J.Sweeney. To learn in detail about survival analysis, references were taken from the books, Cox, D.R and Oakes, D. (1984), Analysis of Survival Data. London Chapman and Hall, Klein, J.P. and Moeschberger, M.L. (1997): Survival Analysis Techniques for Censored and Truncated data, New York: Springer-Verlag and Lee ET and Wang JW. (2003) Statistical Methods for Survival Data Analysis.A case study was referred from Kay R and Kinnersley N (2002) On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: A case study in influenza. Drug Inform.J. 36, 571-579.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Myocardial Infarction (MI), commonly known as a heart attack, occurs when blood flow stops to a part of the heart causing damage to the heart muscle. The most common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck, or jaw. Often it is in the center or left side of the chest and lasts for more than a few minutes. The discomfort may occasionally feel like heartburn. A blockage can develop due to a buildup of plaque, a substance mostly made of fat, cholesterol, and cellular waste products.

The other diseases related to MI are Atrial fibrillation, Cardiogenic Shock, Congestive Heart Failure, Third-Degree Atrioventricular Block. Depending on the presence or absence of these diseases, we analyse the survival times of the MI patients.

II. TECHNIQUES INVOLVED IN KAPLAN MEIER ESTIMATOR AND AFT MODELS

II.1 APLAN MEIER ESTIMATOR

The Kaplan Meier estimator, also known as the product limit estimator, is a non parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. In other fields, Kaplan–Meier estimators may be used to measure the length of time people remain unemployed after a job loss, the time-to-failure of machine parts, or how long fleshy fruits remain on plants before they are removed by frugivores.

A plot of the Kaplan–Meier estimator is a series of declining horizontal steps which, with a large enough sample size, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant. An important advantage of the Kaplan–Meier curve is that the method can take into account some types of censored data, particularly right-censoring, which occurs if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up. On the plot, small vertical tick-marks indicate individual patients whose survival times have been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is the complement of the empirical distribution function.

II.2 COX REGRESSION MODEL

The Cox Regression procedure is useful for modelling the time to a specified event, based upon the values of given covariates. One or more covariates are used to predict a status (event).

The central statistical output is the hazard ratio. Data contain censored and uncensored cases. Similar to logistic regression, but Cox regression assesses relationship between survival time and covariates.

II.3 EXPONENTIAL DISTRIBUTION

Under the exponential distribution, the hazard function is constant. Fitting data to the exponential thus requires estimation of only the single parameter λ which must be > 0.The exponential distribution is probably of greatest use as a starting point for understanding failure-time distributions. The constant hazard means that the exponential distribution plays a role in failure-time models that is somewhat similar to that of the normal distribution in linear statistical modelling.

The exponential distribution is related to the extreme-value distribution. Specifically, T has an exponential distribution with parameter λ, denoted as

$$T \sim E(\lambda), \text{ iff } Y = \log T = \alpha + W ,$$

where $\alpha = -\log\lambda$ and W has a standard extreme value (min) distribution.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

II.4 WEIBULL DISTRIBUTION

An important generalization of the exponential distribution allows for a power dependence of the hazard function on time. The Weibull distribution (including the exponential distribution as a special case) can be parameterised as either a proportional hazards model or an AFT model, and is the only family of distributions to have this property.

A simple empirical check for the applicability of the Weibull distribution to a data set can be derived from the Weibull survival function: using life table estimates for the survival function, a plot ofln(-ln(S(t))) against ln(t) should be approximately a straight line, with the slope being an estimate for p and the ln(t) intercept an estimate for –logλ.
The Weibull is also related to the extreme-value distribution:

$$T \sim W(\lambda,p) \text{ iff } Y = \log T = \alpha + \sigma W,$$

where W has the extreme value distribution, $\alpha = -\log\lambda$ and $p = 1/\sigma$

II.5 LOG-NORMAL DISTRIBUTION

Analytical expressions for the lognormal involve indefinite integral, but the salient feature of the lognormal is that the hazard function is 0 at time 0, increases to a maximum, and then decreases, asymptotically approaching 0 as t goes to infinity. The latter property means that the lognormal is probably not useful for studies involving long lifetimes (although these are quite rare in ecology) because a hazard of 0 is implausible. The lognormal is a two-parameter distribution in which both p and λ are assumed > 0.
T has a lognormal distribution iff

$$Y = \log T = \alpha + \sigma W,$$

where W has a standard normal distribution.

As one might guess from the name, under the lognormal distribution the log failure times are normally distributed. This suggests that simple empirical checks for the lognormal distribution can be done by log-transforming the data and using any of the standard methods of testing for normality. If T has a lognormal distribution then this means that Y=ln(T) has a normal distribution, described by expectation μ and a variance σ^2. In general the distribution of a failure or survival time is skew. Skew distributions may be modelled by means of a lognormal distribution or a gamma distribution as well.

II.6 LOG-LOGISTIC DISTRIBUTION

T has a log-logistic distribution iff

$$Y = \log T = \alpha + \sigma W,$$

where W has a standard logistic distribution.

We find that the log-logistic survivor function is $\frac{1}{1+(\lambda t)^p}$.

we find the hazard function as $\lambda(t) = \frac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p}$.

The logit of the survival function S(t) is linear in logt. This fact provides a diagnostic plot: if you have a non-parametric estimate of the survivor function you can plot its logit against log-time; if the graph looks like a straight line then the survivor function is log-logistic.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

III. ANALYSIS IN ASSESSING THE SURVIVAL TIMES OF MI PATIENTS

III.1 ANALYSIS ON GENDER OF THE PATIENTS BASED ON KAPLAN MEIER ESTIMATOR

We present the survival times of the patients as a graphical representation. Based on whichever curve covers more of the Area Under the Curve (AUC), we draw conclusions. The following Kaplan Meier curve is plotted for the variable gender.

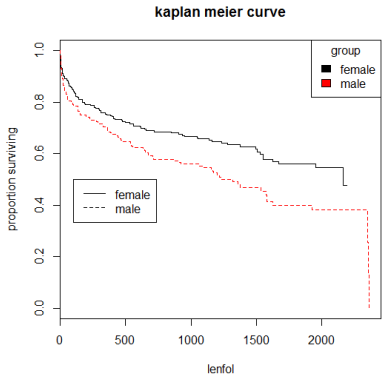

Figure III.1

From the Kaplan Meier curve we can infer that males have higher risk of death (lower survival rates) than females, in this data.

III.2 COX REGRESSION RESULTS

A particular variable of interest in this data is the gender of the patients. The following table presents the regression results for the variable gender.

Table III.1

	Coef	exp(coef)	se(coef)	Z	Pr(>|z|)	lower 95% CI	upper 95% CI
gender	0.3815	1.4645	0.1376	2.773	0.00556**	1.118	1.918

Likelihood ratio test=**7.6**on 1 df**, p=0.005843**
Wald test =**7.69** on 1 df, **p=0.005555**
Score (logrank) test =**7.78** on 1 df, **p=0.005275**

Gender is encoded as a numeric vector 1: female, 0: male. We see that the p-value for the log-rank test is less than 0.05. Therefore we reject the null hypothesis (ie: there is no significant difference between male and female). We conclude that there is significant difference between the two groups of gender. The estimated hazard ratio = 1.4645 indicates that males have higher risk of death (lower survival rates) than females, in this data.

III.3 ANALYSIS ON MI ORDER BASED ON KAPLAN MEIER ESTIMATOR

We present the survival times of the patients as a graphical representation. Based on whichever curve covers more of the Area Under the Curve (AUC), we draw conclusions. The following Kaplan Meier curve is plotted for the variable MI Order (first versus recurrent).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

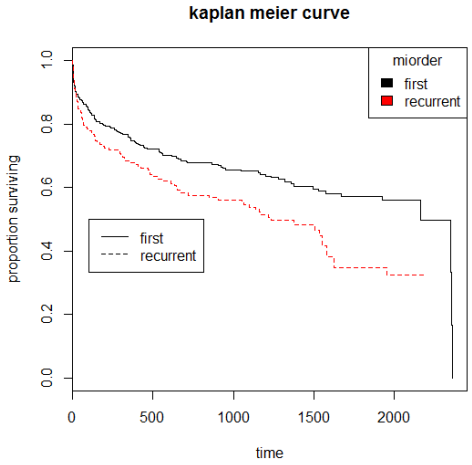

Figure III.2

From the Kaplan Meier curve we can infer that patients with recurrent MI have higher risk of death (lower survival rates) than the patients with MI occurring for first time, in this data.

III.4 COX REGRESSION RESULTS

A particular variable of interest in this data is the gender of the patients. The following table presents the regression results for the variable gender.

Table III.2

	Coef	exp(coef)	se(coef)	Z	Pr(>|z|)	lower 95% CI	upper 95% CI
Gender	0.4266	1.5320	0.1391	3.067	0.00216**	1.166	2.012

Likelihood ratio test=**9.14**on 1 df**, p=0.002497**
Wald test = **9.41**on 1 df, **p=0.002163**
Score (logrank) test =**9.55** on 1 df, **p=0.002001**

The Cox regression results are interpreted as follows:

MI Order is encoded as a numeric vector.0: first, 1: recurrent.We see that the p-value for the log-rank test is less than 0.05. Therefore we reject the null hypothesis (ie: there is no significant difference between first and recurrent MI). We conclude that there is significant difference between the two groups of MI Order. The estimated hazard ratio = 1.5320 indicates that patients with MI recurring have higher risk than patients with MI occurring for the first time, in this data.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

III.5 AFT MODELS

Table III.3

	exponential		log-normal		weibull		log-logistic	
	Value	p	Value	P	value	p	Value	p
(Intercept)	10.4733	4.14E-35	12.8251	1.30E-14	12.711	1.54E-16	12.957	9.50E-16
age	-0.0491	4.71E-13	-0.0843	7.43E-11	-0.0793	3.64E-10	-0.0831	1.65E-10
hr	-0.0138	1.03E-05	-0.0264	3.82E-05	-0.0216	9.01E-05	-0.0253	3.48E-05
diasbp	0.01562	2.95E-03	0.02137	1.85E-02	0.02082	1.96E-02	0.02082	2.34E-02
bmi	0.06218	3.19E-04	0.07489	1.21E-02	0.08684	3.31E-03	0.07445	1.14E-02
sho	-1.5353	9.10E-08	-2.9366	2.94E-06	-2.2958	5.26E-06	-2.8391	2.72E-06
chf	-0.8464	3.89E-08	-1.2991	3.85E-05	-1.2691	3.43E-06	-1.3551	3.35E-06

In the above table we have displayed those variables which are significant (those whose p-values are <0.05). For each of the four models we find the -2(log-likelihood) and the AIC values.

Table III.4

	Exponential	Log-normal	Weibull	Log-logistic
log-likelihood	-1697.4	-1644.7	-1638.7	-1640.1
-2(loglikelihood)	3394.8	3289.4	3277.4	3280.2
AIC	3426.769	3323.48	3311.389	3314.15

Depending on the values in the above table we draw conclusions regarding the best model for this data.

IV. CONCLUSION

Lower values of -2LogLikelihood suggest a better model. One way of selecting an appropriate parametric model is to base the decision on minimum (AIC) and also based on the -2 LL(Log-Liklihood). For the parametric models presented in the table 3.4, -2LL and the AIC values of weibull is the minimum. Thus we conclude that Weibull is the best suited model for this data. In Weibull distribution the variables namely age of the patient, sex to which the patient belongs, the heart beat rate per minute, the diastolic blood pressure, the BMI of the patients during the start of the study, cardiogenic shock(whether the person has got SHO previously), complete heart failure(whether the patient has experienced CHF) are all significant at 5% level. Thus we can conclude that all these 7 variables included in the study have significant impact on the occurrence of event.

The coefficient of the variable MI Order (whether it is recurrent MI or First MI) in Weibullmodelis -0.121636. So we now have exp(-0.121636)=0.8854. Hence we can interpret that the variable MI order has 88.54% impact on our dependent variable. In the Weibull model, the age of the patient also has a significant impact on the occurrence of the event. The coefficient of age is -0.07931 exp(-0.079315)=0.9237. Hence we can interpret that the variable age is 92.37% important. Similar interpretations can be made using the coefficients of variables in all the models.

Acknowledgement: The authors thank Mr Kannedari Siva Naga Raju, Research Scholar, Department of Statistics, Loyola College, Chennai-34 for his help and suggestions towards the article.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

REFERENCES

1. Cox, D.R and Oakes, D. (1984), Analysis of Survival Data. London Chapman and Hall.
2. Introduction to linear regression analysis 5th edition (2012), Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey vining.
3. Kay R and Kinnersley N (2002) On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: A case study in influenza. Drug Inform.J. 36, 571-579.
4. Klein, J.P. and Moeschberger, M.L. (1997): Survival Analysis Techniques for Censored and Truncated data, New York: Springer-Verlag
5. Lee ET and Wang JW. (2003) Statistical Methods for Survival Data Analysis.
6. Statistics for business and economics 11th edition (2012), David R.Anderson, Thomas A.Williams, Dennis J.Sweeney