Chapter: 2

EXISTING METHODOLOGY

2.1 Survival Analysis
2.2 Need of Censoring
2.3 Logistic Regression Model
2.4 Probit Model
2.5 Accelerated Failure Time Model
2.6 Cox Proportional Hazard Model
  2.6.1 Exponential model
  2.6.2 Weibull model
The rapid advent of communication technology, and increasing requirement of some popular branded mobile phone manufacturers to design, develop and market the sophisticated handsets with various features, require a modeling technique to predict the failure rate and determine the life of the handset, pertaining to various criterions. This chapter will navigate us to accord the preference of the Cox Proportional Hazard Model over some of the other existing family models available for survival analysis of lifetime of mobile phone based on some specific parameters.

2.1 Survival Analysis: It encompasses a variety of methods for analyzing time to event data and has emerged as major area in statistical research and applied statistics during the last decade Cox and Oakes (1984) Kalbfleisch and Prentice (1980). It is a modeling approach to estimate the outcome of variable of interest, namely the time until an event occurs. Historically survival analysis comes from the study of deaths in biological organisms and it gives the length of time that an organism “survives”. But with the rapid influx of technological knowledge, it has grown it applications in the industrial and business world as well. Its relevance is seen in diverse disciplines ranging from biostatistics to engineering for reliability analysis Nelson (1990), in sociology for event-history analysis Allison (1984). Heckman and Singer (1985) and Lancasler (1979) had studied from an econometric perspective and Steinberg and Carson (1990) analyzed willingness to pay data. From the use in mechanical systems, we refer to the event as failure.
Some examples of life data analysis in non-medical fields include the amount of time until the malfunctioning of mechanical equipment, time until an ailment is reported or causes a death to a human body, time until a mobile phone shall start malfunctioning after its operation or the employment time of employee for a certain company etc. Survival time represents a time \( t \) from a particular time (i.e. the time when the event occurs) till the end of the time for the entire period. The analysis of these survival times can be successfully implemented when all the survival times are known. However such ideal situation may not always exists and there are many instances like this in research. Observations in such a case are said to be censored data.

2.2 Need of Censoring: A terminally ill patient may survive until end of the study, or a mechanical component may not show any malfunctioning during the times it is being observed. In these cases, the survival times of the observations are not known, but it is known to be at least as long as the time of the study. This is called Type I censoring when all censored data are having the same length (Lee, 1992). Censoring techniques are of three types, i.e. type I censoring, type II censoring and interval censoring.

In Type I censoring technique all the units are observed for say \( T \) hours where \( r \) failures (where \( r \) can be any number from 0 to \( n \)) are noted during the period. The (exact) failure times are say \( t_1, t_2, ..., t_r \) and there are \( (n - r) \) units that survived the entire \( T \)-hour test without failing. Considering that \( T \) is fixed in advance and \( r \) is random, we don't know how many failures will occur until the test is run. In these cases we assume the exact times of failure are recorded when the failures occur. This is also known as right
censoring technique. We in our studies have followed right censoring technique for collecting data.

Another important technique available in survival analysis is Type II censoring. This test decides in advance about the exact number of failure the researcher would like to observe say $r$ failure times and then test until they occur. In these cases $T$ is unknown until the $r^{th}$ number of unit fails. This is called Type II Censoring data. Mathematically, if we want to observe failure times $t_1, t_2, \ldots, t_r$, where $r$ is specified in advance then, test ends at time $T = t_r$, where $(n-r)$ units have survived the test. Again we assume it is possible to observe the exact time of failure for failed units only.

Type II censoring has the significant advantage as we know in advance in how many failure times our test will yield results - this helps enormously when planning adequate tests. However, an open-ended random test time is generally impractical from a management point of view and this type of testing is rarely seen.

The third type of censoring is interval censoring. Sometimes exact times of failure are not known; only an interval of time in which the failure occurred is recorded. To study the survival analysis on such data interval censoring is applied. Interval censored data generally present tied observations. If the number of ties are less then the analysis of such
data can be done by using the Cox proportional hazards model (Cox, 1972) by means of the exact partial likelihood, or by approximations method proposed by Breslow (1972, 1974), Peto (1972), Efron (1977) and Farewell & Prentice (1980). But when the number of tied observations is large, in such cases time can be considered as discrete observation and a model can be fitted to the probability of occurrence of an event since it did not occur in the previous interval. Such fits can be made using the Cox proportional hazard model for grouped data (Prentice & Gloeckler, 1978) or logistic model (Lawless, 1982).

These methods are more suitable when the time intervals are disjoint. In the case of an overlapping interval, they cannot be directly applied (Finkelstein, 1986). According to Nelder & Wedderburn, (1972) the fitting of models for discrete times in the presence of interval censoring becomes easier when generalized linear models fit. After models are fitted, there is an interest in knowing which of them better explain the experimental data. For this, Colosimo et al. (2000) proposed two score tests from a more general distribution: the Aranda-Ordaz asymmetric transformations family, which has the logit transformation (logistic model), and complementary log-log (Cox proportional hazards model), as particular cases.

There are other well known methods for estimating unconditional survival distributions apart from Cox proportional hazard model and Logistic regression model are Probit
regression model, to estimate failure time models based on actual failure times we have accelerated failure time model. For our studies Logistic regression model was an alternative to Cox proportional hazard model as the failure of the event such as Integrated circuit failure battery problem etc, which are some of our parameter in the model, can be collected in binary form. Presence of these problems can be coded as 1 and absence as 0.

2.3 Logistic Regression Model: This model is used for prediction of the probability of occurrence of an event by fitting data to a logit function or a logistic curve. It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. A logit model always assumes the univariate or multivariate technique. It assumes the proper estimation of the probability in case of occurrence of an event, by predicting a dichotomous dependent outcome i.e. life of the handset from a set of independent variable such as battery problem, integrated circuit problem, speaker problem etc as in our case. Logistic regression is used extensively in the medical and social sciences fields, as well as marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription. The model is expressed as

\[ P_i = E(Y = 1 | X_i) = \beta_1 + \beta_2 X_i \] ........................................ (a)

where \( X \) is considered as the independent variable and \( Y = 1 \) means that a user owns a specific product under study.

\[ P_i = E(Y = 1 | X_i) = \frac{1}{1 + \exp[-(\beta_1 + \beta_2 X_i)]} = \frac{1}{1 + \exp(-Z_i)} \] ........................................ (b)
where \( Z_i = \beta_1 + \beta_2 X_i \).

This equation is known as CDF of Logit model. When \( Z_i \) varies between the - \( \infty \) to + \( \infty \), \( P_i \) ranges between 0 and 1. There exists certain problems as the equation is nonlinear with not only on \( X \) but also on the independent variables, which gives an estimation problem because we cannot estimate the unknown parameters required for a linear regression model (often known as *ordinary least squares* (OLS)). Taking the logarithm on both sides of equation \( b \) we get

\[
L_i = \ln \left[ \frac{P_i}{1 - P_i} \right] = z_i = \beta_1 + \beta_2 X_i
\]

Here \( L_i \) is the Logit function of the \( i \)th variable.

(A) For a continuous outcome variable \( Y \), the numerical value of \( Y \) at each value of \( X \)

![Figure - 20 (A)](image-url)
For a binary outcome variable, the proportion of individuals cases at each value of $X$.

The Logit model has certain advantages. It provides prominent and good results by enabling us to transform the binary variable to a continuous range variable between $-\infty$ to $+\infty$, not only that it is easily to interpret and simpler to analyze. Logistic regression follows parametric estimation that is asymptotically consistent and quite normal so that the analogue of regression t-test can be applied. However every model has certain demerits. In case of Logit model it is difficult to identify drivers of consumer acceptance evaluated on a three-category scale.

This model is widely accepted for several factors. Some of them are discussed below:
Logistic model can be used successfully to numerous investigations that examine the relationship between risk factors and various disease events. In such circumstances the event times are grouped into intervals and the logistic regression can be adapted to the analysis of such data by modeling the interval when an event occurs. Furthermore, it is shown that results from such an adaptation often leads to parameter estimates that are close to those obtained by the proportional hazards model in the grouped event time setting, *Epidemiol* (1985). Its applicability is also seen in credit scoring modeling of customer open-end accounts depending on application data and transaction behavior data. Not only that in the field of the marketing, it can be used successfully for brand presence and brand loyalty for any consumer product. It can be further used to identify the factors which effect the adoption of a particular technology such as use of new variates, fertilizers, pesticides on a farm. In case of gender studies also logit analysis has been used for affecting the decision to make status of male and female in a family.

2.4 Probit Regression Model: In order to understand the probability behavior for a binary dependent variable there should be a proper choice of Cumulative distributive function. Logit model uses a cumulative logistic function which may not be the only choice for the applications usage. In fact, in some cases, a normal Cumulative distributive function often becomes a better choice, which emerges the use of the Probit model.

In the context of the cell phone survival analysis above, the decision to own a cell phone or not, depends on the unobservable index $I_i$ in such a way that when the value of $I_i$
increases, there is a higher probability of a user to own a cell phone. This index can be express as

\[ I_i = \beta_1 + \beta_2 X_i \]

A probit model is a specification for ordinal or a binary response model which employs a probit link function. This model is most often estimated using standard maximum likelihood procedure, such an estimation being called a probit regression. The model can be represented as

\[ \Pr(Y = 1 \mid X) = \Phi(X' \beta) \]

where \( \Pr \) denotes probability, and \( \Phi \) is the Cumulative Distribution Function (CDF) of the standard normal distribution. The parameters \( \beta \) are typically estimated by maximum likelihood here we assume that response variable \( Y \) is dichotomous with only Boolean or binary results to be denoted as 0 or 1. For example \( Y \) may represent presence/absence of a certain condition, success/failure of some device, answer yes/no on a survey, etc given the condition \( X \).
The above figure gives the transformation function in the probit and logit model.

Figure - 21

The above figure gives the transformation function in the probit and logit model.

Probit model has been applied in diverse areas. It has been applied in finance for predicting financial crisis in India in case of early warning system. To predict the severity of injury of truck passenger car rear end collision and also predicting medical malpractice claims.

2.5 An Accelerated Failure Time Model (AFT model) is one of the parametric models that provide a substitute to the some commonly used proportional hazards models. A proportional hazards model presumes that the effect of a covariate is to multiply the hazard by some constant, whereas an AFT model assumes that the effect
of a covariate is to multiply the predicted event time by some constant. AFT models can therefore be framed as linear models for the logarithm of the survival time. According to Bolshev (1976), this model is more suited to study the aging population Meeker and Esobar (1998), Viertl (1988), Bagdonavicius and Nikulin (1995, 2001). The survival and the cumulative hazard function under a covariate realization $x(\cdot)$ under AFT are given by

$$S_{n,t}(t) = G\left(\int_0^t r[x(s)]ds\right) \quad \text{and} \quad \Lambda_{n,t}(t) = \Lambda_0\left(\int_0^t r[x(s)]ds\right), \quad x(\cdot) \in E \text{ respectively.}$$

where $G = S_{x_0(t)} = S_0$, $\Lambda_0(t) = -\ln S_0(t)$, $x_0(\cdot)$ is a given stress and $x_n \in E$.

### 2.6 Cox Proportional Hazard Model:

It is the most popular model in the family of proportional hazards model. The popularity of the model exists because of the simple semi-parametric estimation procedures that can be applied even when the form of survival distribution function is not specified. It provides the primary information desired from a survival analysis, hazard ratios and adjusted survival curves, with a minimum number of assumptions. It is a robust model where the regression coefficients closely approximate the results from the correct parametric model.

Not only that, it can be used to study the impact of the covariates on the survival of the handsets. In most situations, we are more interested in the parameter estimates than the
shape of the hazard. The Cox PH model is well-suited to this goal. The detail of this model has been discussed in the coming chapters. This model consists of two factors. One is the baseline hazard rate and the other is the exponential term involving explanatory variables. The baseline hazard rate can follow any non parametric distribution like normal, exponential, log-normal or Weibull. However this model also has certain limitations like it has no estimated intercept term, neither has it provided an equation that can be used to predict survival time nor does it provide group-specific hazard rates.

2.6.1 Exponential Model: It is the simplest type of parametric model that is used for the determination of baseline hazard of Cox proportional hazard model and it assumes that the baseline hazard is constant over time. It has a constant hazard function which is given by

\[ h(t) = \lambda \]

and its survival function is

\[ S(t) = \exp \{-\lambda t\} \]

The assumption that the baseline hazard is constant over time can be evaluated in several ways. The baseline hazard is assumed to be constant within each time period, but can vary between time periods. Thus a large \( \lambda \) implies a high risk and a short survival. Conversely, a small \( \lambda \) indicates a low risk and a long survival. This distribution has the memoryless property meaning that how long an individual has survived does not affect its future survival (Lee, 1992).
The exponential distribution is not so flexible because it has only one parameter, the scale parameter $\lambda$. By adding a shape parameter the distribution becomes more flexible and can fit to more kinds of data.

### 2.6.2 Weibull Model

The more generic nature of the exponential distribution arrives at a point to include the shape parameter which is known as the Weibull distribution. We have discussed our choice of Weibull distribution in the coming chapters.

We can predict the flexibility of Weibull distribution from its hazard function. When $\gamma=1$, the Weibull distribution becomes the exponential distribution with $\theta = \lambda$ and the hazard rate remains constant as time increases and when $\gamma=2$ it is the Rayleigh distribution. For $3 \leq \gamma \leq 4$, it is close to the normal distribution and when $\gamma$ is large, say $\gamma \geq 10$ it is close to the smallest extreme value distribution (Nelson, 1982). When $\gamma > 1$ the hazard rate increases as time increases, and for $\gamma < 1$ the hazard rate decreases.

We have applied Cox proportional hazard model with Weibull as a hazard rate for baseline estimation. We have conducted empirical studies on mobile handset on five prominent handset manufacturing companies. We have collected 3100 data for our studies. The data collection methods have been discussed in details in the respective chapters.