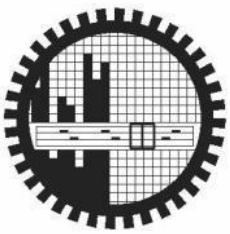
EVALUATION OF EMPLOYEE ABSENTEEISM RATE USING SURVIVAL ANALYSIS

SAIF MUHAMMAD MUSFIR RAHMAN



DEPARTMENT OF INDUSTRIAL & PRODUCTION ENGINEERING (IPE)
BANGLADESH UNIVERSITY OF ENGINEERING & TECHNOLOGY
DHAKA, BANGLADESH

September, 2015

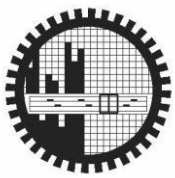
EVALUATION OF EMPLOYEE ABSENTEEISM RATE USING SURVIVAL ANALYSIS

BY

SAIF MUHAMMAD MUSFIR RAHMAN

A Thesis Submitted to the
Department of Industrial & Production Engineering,
Bangladesh University of Science and Technology
in Partial Fulfilment of the requirements for the Degree of

BACHELOR OF SCIENCE IN INDUSTRIAL & PRODUCTION ENGINEERING



DEPARTMENT OF INDUSTRIAL & PRODUCTION ENGINEERING
BANGLADESH UNIVERSITY OF ENGINEERING & TECHNOLOGY
DHAKA, BANGLADESH

September, 2015

CERTIFICATE OF APPROVAL

The thesis titled "**EVALUATION OF EMPLOYEE ABSENTEEISM RATE USING SURVIVAL ANALYSIS**" submitted by **Saif Muhammad Musfir Rahman**, Student No. 0908002, Session 2012-2013, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Bachelor of Science in Industrial & Production Engineering on 10 September, 2015.

Dr. M. Ahsan Akhter Hasin

(Thesis Supervisor)

Professor

Department of Industrial & Production Engineering

Bangladesh University of Engineering and Technology

Dhaka-1000, Bangladesh.

CANDIDATE'S DECLARATION

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

_____

Saif Muhammad Musfir Rahman

ACKNOWLEDGEMENT

At first, the author wants to convey his deepest to the almighty ALLAH, the beneficial, the merciful for granting me to bring this research work into light.

The author would like to express his sincere respect and gratitude Dr. M. Ahsan Akhtar Hasin, Professor, Department of Industrial and Production Engineering (IPE), Bangladesh University of Engineering and Technology (BUET), Dhaka, for his thoughtful suggestions, constant guidance and encouragement throughout the progress of this research work.

The author is also thankful to Mrs.Rahima Begum, HR Director, Ananta Limited for her cordial encouragement and sincere help during the data collection phase.

The author is especially grateful to Golam Kabir, Assistant Professor, Department of Industrial and Production Engineering, BUET, for his valuable suggestion and support to continue the research.

The author also expresses his heartiest thanks to all of his colleagues of the Department of Industrial and Production Engineering (IPE), Bangladesh University of Engineering and BUET), Dhaka for their cooperation and motivation during the study.

The author is grateful to all the writers and publishers of the books and journals that have taken as references while conducting this research.

With a very special recognition, the author would like to thanks his parents as well as all the members of his families, who provided their continuous inspiration, sacrifice and support encouraged me to complete the research work successfully.

ABSTRACT

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc. For example, if the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack.

In this paper, employee records of two different sized firms were recorded and then survival analysis was performed to measure the probability or rate of absenteeism. Absenteeism usually is dependent on a variety of factors. Several such factors were determined and the rate of absenteeism against them was measured using a refined algorithm. The covariate responsible for the greatest absenteeism rate was thus identified.

Table of Contents

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

pdf	: Probability Density Function
cdf	: Cumulative Distributive Function
KM estimator	: Kaplan-Meier Estimator
PHM	: Proportional Hazards Model

CHAPTER 1

INTRODUCTION

1.1 GENERAL INTRODUCTION

It is of no doubt that the work force is one of the strongest assets of any company. The success of a company or an organization depends heavily on how well its workforce is managed. One such sector of the human resource management is managing employee absenteeism. But unfortunately, no substantial mathematical model exists that an organization or a company could easily use to forecast its employee absenteeism rate. Industry studies have shown that fewer than 20% of large companies calculate the costs associated with absence and only one-third track leave utilization. Without a firm grasp on scheduled as well as unscheduled employee absences, organizations are unable to track the associated direct and indirect costs. Several industry studies have shown that the direct and indirect costs of absenteeism can amount to 15% of payroll. Yet, lost workdays- approximately 2.8 million each year for injuries and illness- are not the only expense associated worth employee absence. Survival analysis is one of way of keeping track of the lost workdays.

The term "survival analysis" pertains to a statistical approach designed to take into account the amount of time an experimental unit contributes to a study. That is, it is the study of time between entry into observation and a subsequent event. Originally, the event of interest was death hence the term, "survival analysis." The analysis consisted of following the subject until death. The uses in the survival analysis of today vary quite a bit. Applications now include time until onset of disease, time until stockmarket crash, time until equipment failure, time until earthquake, and so on. The best way to define such events is simply to realize these events are a transition from one discrete state to another at an instantaneous moment in time.

1.2 RATIONALE OF THE STUDY

In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Why not use linear regression to model the survival time as a function of a set of predictor variables? First, survival times are typically positive numbers; ordinary linear regression may not be the best choice unless these times are first transformed in a way that removes this restriction. Second, and more importantly, ordinary linear regression cannot effectively handle the censoring of observations.

Unlike ordinary regression models, survival methods correctly incorporate information from both censored and uncensored observations in estimating important model parameters. The dependent variable in survival analysis is composed of two parts: one is the time to event and the other is the event status, which records if the event of interest occurred or not. One can then estimate two functions that are dependent on time, the survival and hazard functions. The survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. The survival function gives, for every time, the probability of surviving (or not experiencing the event) up to that time. The hazard function gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time. While these are often of direct interest, many other quantities of interest (e.g., median survival) may subsequently be estimated from knowing either the hazard or survival function. It is generally of interest in survival studies to describe the relationship of a factor of interest (e.g. treatment) to the time to event, in the presence of several covariates, such as age, gender, race, etc. A number of models are available to analyze the relationship of a set of predictor variables with the survival time. Methods include parametric, nonparametric and semi-parametric approaches.

1.3 OBJECTIVE OF THE STUDY

The objective of the study is as follows-

1. To develop a predictive model so as to determine the rate or probability of absenteeism in a given time.
2. To determine the factor which is most responsible for employee absenteeism.

3. To calculate the survival probability, hazard rate and log likelihood functions of data sets
4. To verify the models by data splitting

1.4 OUTLINE OF METHODOLOGY

In order to perform survival analysis on the data, the first step involved is collecting the data from two different sized firms. The data is then tabulated with the absence/presence status with 1 signifying absence (event) and 0 is not absent corresponding to the time taken. Kaplan-Meier estimator is then used to calculate the probability of an event (absence) to occur corresponding to a particular time. Non-parametric model (Weibull distribution) is then applied to the data and the results obtained are compare to that of Kaplan-Meier. But practically, absenteeism is dependent on other factor and these factors (covariates) are then injected into the model to obtain survival rate. Cox proportional hazards model is applied in order to take the covariates in consideration and the effectiveness of the results are compared with the other two.

1.5 ORGANIZATION OF REPORT

The Report is divided into six sections or chapters. Chapter 1 is entitle “Introduction” and presents the introductory idea to survival analysis and what the report constitutes. The objective and methodology od the thesis have been outlined here.

Chapter 2, entitled “Literature Review” comprises of the recent research works in employee absenteeism and survival analysis. Importance of the previous works on these topics are also discussed briefly.

The theoretical background of survival analysis and the solving techniques including Kaplan-Meier, parametric distribution and the Cox Proportional Hazard Model have been discussed briefly in Chapter 3 entitle “Theoretical Background”.

Chapter 4 is entitle “Problem Formulation”. It encompasses how the problem has been tackeld with and how it has been modeled in order to perform survival analysis calculations.

To establish the robustness of the proposed model, it has been implemented to evaluate the employee absenteeism rate of two industries: Ananata Apparel Ltd. and ABC Air Ltd. The formulation of the data sets in order to solve with a statistical software package are discussed here. The data set is then solved using R and the necessary steps are discussed in Chapter 5 entitled "Application of the Model".

Finally, a summary of the outcomes of the thesis and strengths and weaknesses of the proposed approach and recommendations for future research is given in Chapter 6 entitled: "Conclusions and Recommendations" and references are provided at the end.

CHAPTER 2

LITERATURE REVIEW

2.1 LITERATURE REVIEW ON EMPLOYEE ABSENTEEISM

Most of the studies dealing with employee absenteeism have been carried out by psychologists and social scientists involved with management and administration. If one were to single out the two or three most often cited journals in which these studies appear they would be the Journal of Applied Psychology, the Academy of Management Journal, and the Journal of Organizational Behavior and Management. Absenteeism has been the focus of much research by economists and other social scientists in recent years. However, as mentioned above, few economists have included work environmental issues in their studies on absenteeism. An important exception is the work by Johansson and Palme (1996). They explicitly model the everyday economic choice of being absent from work in relation to occupation and personnel characteristics that may influence this choice. Cross-sectional data from the 1981 Swedish Level of Living Survey are used in combination with information on individual work absence from the National Social Insurance Board register. The data set consists of 1967 individuals for the year 1981. Two types of working condition variables are used: occupation-specific risk indices and self-reported information from the survey. Johansson and Palme find that risk, i.e., accidents at work and work-related diseases, will increase work absence. In addition, they find significant effects related to different self-reported job characteristics, which indicate that individuals with a work profile involving low stress and outdoor work have, on average, a lower absence rate. Drago and Wooden (1992) use a cross-national data set to test different theoretical frameworks for predicting self-reported employee absence rates. They regard shift-work as a negative working condition and find that workers on shift-work are more prone to be absent. In addition, they include indices for job satisfaction and work group cohesion and find that work group cohesion leads to low absence where job satisfaction is high and to high absence when job satisfaction is low. Another contribution that is relevant for this study is by Brown et al. (1999), which investigate the effects on absenteeism of two types of employer-sharing plans – profit sharing

and employee share ownership – in 127 French firms over the years 1981–1991. They find that both plans were associated with reductions in absenteeism. Absence is defined as the total number of absence events in the firms divided by the total number of employees employed by the firms.

Absenteeism is viewed as costly and disruptive for organisations. For instance, Barmby et al. (2002) demonstrate for nine industrialised countries that a significant proportion of work hours are lost through absence, ranging from 1.8% to 6% of average weekly hours for Switzerland and Sweden, respectively. Psychologists have suggested that absenteeism in stressful situations might be implicitly condoned by management because it is seen as a form of temporary relief for the employee, and could therefore be regarded as an efficient response (Steers and Rhodes, 1978). Devising an appropriate policy response to absenteeism is therefore complex, even more so when it is realised that worker absence can occur involuntarily because of physical or mental ill-health, or because of voluntary shirking behaviour by workers (Barmby et al., 1994; Johansson and Palme, 2002). Disentangling shirking behaviour (i.e. voluntary absence) from involuntary absence is extremely difficult to do in practice. Previous research has attempted to do so using absence spell length. However, as Driver and Watson (1989) argue this is unreliable because long spells of absence could be voluntary and a sequence of short spells could be indicative of recurring sickness. Consequently, this paper does not explicitly try to distinguish between voluntary and involuntary absence. Instead, we take the approach of modeling the employee absenteeism as functions of general covariates. Biorn et al. (2010) utilized a statistical model to explore and empirically separate cohort, time and age effects in worker absence behavior.

2.2 LITERATURE REVIEW ON SURVIVAL ANALYSIS

Survival analysis has been used in many different researches, but never before as a tool in evaluating employee absenteeism rate. Seval Kul (2010) showed the use of survival analysis for clinical pathways. He found that clinical pathways also produce data suitable for application to survival analysis when the primary variable of the research is time until the predefined event occurs. Bradburn, Clark, Love and Altman performed multivariate data analysis using survival analysis- choosing a model and assessing its adequacy and fit. The most notable work in survival

analysis involves the non-parametric estimation from incomplete observations done by Kaplan and Meier (2008). This paper broke free from the traditional application of Kaplan-Meier estimator. In their 1958 paper introducing the product-limit estimator, Kaplan and Meier noted that the assumption of independence between censoring and survival times "deserves special scrutiny". Many authors since then have showed that informatively censored observations can bias estimates of the survival function. Peterson gave sharp bounds on the marginal survival function without making further assumptions about the association between survival and censoring. However, in most situations these bounds are considered too wide for practical use. We can conceptualize informative censoring as being composed of a part that can be explained by measured factors that are prognostic for both censoring and survival and the remaining part that is not explained by measured factors. When all the factors prognostic for both censoring and survival are available, Satten, Datta and Robins (2001) proved that the marginal survival function is identifiable and Robins and colleagues proposed methods for estimating the function.

The general approach taken when there are no measured prognostic factors is to model the association between censoring and survival. The parameters of the models, which correspond to the degree of association, are varied over a plausible range and the resulting estimates are used to place bounds on the survival function. Fisher and Kanarek considered such a model. They assumed that being lost to follow-up occurs simultaneously with an event that alters survival by an amount associated with a scale parameter α. In their model, when $\alpha = 1$, censoring has no effect on survival, while $\alpha < 1$ contracts survival and $\alpha > 1$ stretches it by $\alpha(t-c)$, where t-cis the survival time following the censoring event. Lagakos and Williams considered a model with an exponential survival function, an unspecified function c(y)that measures the relative odds of observing a failure at $y = \min(t,c)$, and a parameter θthat corresponds with the degree of association between censoring and survival with 0 indicating death immediately following censoring and 1 indicating non-informative censoring. Slud and Rubinstein introduced a known function $\rho(t)$specifying the hazard ratio between censored and uncensored subjects over time and showed how to calculate boundson the survival function given bounds on $\rho(t)$. Klein and Moeschberger took a similar approach with a fixed parameter (θ) representing the hazard ratio comparing the censored to uncensored. They demonstrated the relationship between this parameter and Kendall's coefficient of concordance (τ) and discussed the use of that measure to specify the plausible range of association between censoring and survival. Zheng and Klein

showed that a known copula defining the dependence between censoring and survival is sufficient to identify the marginal survival function. The copula is chosen to be monotone in a given parameter and bounds for the survival function are estimated by varying the parameter over a range. A relatively new idea in survival analysis is the Cox proportional hazards model. Abeysekara and Sooriyarachchi (2008) tested the proportional hazards assumption in the Cox proportional hazards model by the use of Schoenfeld's global test. Smith, AK Ryan (2006) employed Cox regression to model time until event while simultaneously adjusting for influential covariates and accounting for problems such as attrition, delayed entry and temporal biases.

CHAPTER 3

THEORETICAL BACKGROUND

3.1 BASICS OF SURVIVAL ANALYSIS

The term "survival analysis" pertains to a statistical approach designed to take into account the amount of time an experimental unit contributes to a study. That is, it is the study of time between entry into observation and a subsequent event. Originally, the event of interest was death hence the term, "survival analysis." The analysis consisted of following the subject until death. The uses in the survival analysis of today vary quite a bit. Applications now include time until onset of disease, time until stock market crash, time until equipment failure, time until earthquake, and so on. The best way to define such events is simply to realize that these events are a transition from one discrete state to another at an instantaneous moment in time. Of course, the term "instantaneous", which may be years, months, days, minutes, or seconds, is relative and has only the boundaries set by the researcher.

The origin of survival analysis goes back to mortality tables from centuries ago. However, it was not until World War II that a new era of survival analysis emerged. This new era was stimulated by interest in reliability (or failure time) of military equipment. At the end of the war these newly developed statistical methods emerging from strict mortality data research to failure time research, quickly spread through private industry as customers became more demanding of safer, more reliable products. As the uses of survival analysis grew, parametric models gave way to nonparametric and semi-parametric approaches for their appeal in dealing with the ever-growing field of clinical trials in medical research. Survival analysis was well suited for such work because medical intervention follow-up studies could start without all experimental units enrolled at start of observation time and could end before all experimental units had experienced an event. This is extremely important because even in the best-developed studies, there will be subjects who choose to quit participating, who move too far away to follow, or who will die from some unrelated event. The researcher was no longer forced to withdraw the experimental unit and all associating data from the study, instead techniques called censoring enabled researchers to analyze incomplete data due to delayed entry or withdrawal from the study. This was

important in allowing each experimental unit to contribute all of the information possible to the model for the amount of time the researcher was able to observe the unit.

The last great strides in the application of survival analysis techniques has been a direct result of the availability of software packages and high performance computers which are now able to run these difficult and computationally intensive algorithms relatively efficiently.

3.2 TOOLS USED IN SURVIVAL ANALYSIS

Time is continuous, which results in the probability of an event at a single point of a continuous distribution being zero. We are challenged to define the probability of these events over distribution. This is best described by graphing the distribution of event times. To ensure the readers will start with the same fundamental tools of survival analysis, a brief descriptive section of these important concepts will follow.

3.2.1 The Cumulative Distribution Function

The cumulative distribution function (cdf) is very useful in describing the continuous probability distribution of a random variable, such as time, in a survival analysis. The cdf of a random variable T, denoted $F_T(t)$, is defined by $F_T(t) = P_T(T \leq t)$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties of a distribution function F(t) can be listed as a consequence of the knowledge of probabilities. Because F(t) has the probability $0 \leq F(t) \leq 1$, then F(t) is a nondecreasing function of t, and as t approaches ∞, F(t) approaches 1.

3.2.2 The Probability Density Function

The probability density function (pdf) is also very useful in describing the continuous probability distribution of a random variable. The pdf of a random variable T, denoted $f_T(t)$, is defined by $f_T(t) = d F_T(t) / dt$. That is, the pdf is the derivative or slope of the cdf. Every continuous random variable has its own density function, the probability $P(a \leq T \leq b)$ is the area under the curve between times a and b.

3.2.3 The Survival Function

Let $T \geq 0$ have a pdf f(t) and cdf F(t). Then the survival function takes on the following form:

$S(t) = P\{T > t\}$

$= 1 - F(t)$

That is, the survival function gives the probability of surviving or being event-free beyond time t. Because S(t) is a probability, it is positive and ranges from 0 to 1. It is defined as $S(0) = 1$ and as t approaches ∞, S(t) approaches 0. The Kaplan-Meier estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step idea. The Kaplan-Meier estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. The survival curve describes the relationship between the probability of survival and time.

3.2.4 The Hazard Function

The hazard function h(t) is given by the following:

$h(t) = P\{ t < T < (t + \Delta) \mid T > t\}$

$= f(t) / (1 - F(t))$

$= f(t) / S(t)$

The hazard function describes the concept of the risk of an outcome (e.g., death, failure, hospitalization) in an interval after time t, conditional on the subject having survived to time t. It is the probability that an individual dies somewhere between t and $t + \Delta$, divided by the probability that the individual survived beyond time t. The hazard function seems to be more intuitive to use in survival analysis than the pdf because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time t.

3.2.5 Left and Right Censoring

The most common form of incomplete data is right censoring. This occurs when there is a defined time (t=0) where the observation of time is started for all subjects involved in the study. A right censored subject's time terminates before the outcome of interest is observed. For example, a subject could move out of town, die of an unexpected cause, or could simply choose not to participate in the study any longer. Right censoring techniques allow subjects to contribute to the model until they are no longer able to contribute (end of the study, or withdrawal), or they have an event. Conversely, an observation is left censored if the event of interest has already occurred when observation of time begins. For the purposes of this study we focused on right censoring.

The following graph shows a simple study design where the observation times start at a consistent point in time (t=0). The X's represent events and the O's represent censored observations. Notice that all observations are classified with an event, or they are censored at time of separation or at the end of the study period. Some subjects have events early in the study period and others have events at the end of the study period. Likewise some subjects leave early, but most do not have an event during the entire study and are simply right censored at the end. There is no need for left censoring or truncation techniques in this simple example.

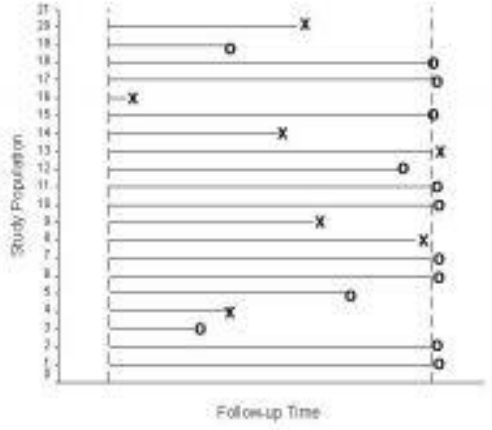


Figure 3.1 : Right Censoring of Data

3.3 THE KAPLAN-MEIER ESTIMATE OF THE SURVIVAL FUNCTION

The life table [6] is the earliest statistical method to study human mortality rigorously, but its importance has been reduced by the modern methods, like the Kaplan-Meier (K-M) method. In clinical studies, individual data is usually available on time to death or time to last seen alive. The K-M estimator for the survival curves is usually used to analyze individual data, whereas the life table method applies to grouped data. Since the life table method is a grouped data statistic, it is not as precise as the K-M estimate, which uses the individual values. We only describe the K-M estimate here.

Suppose that r individuals have failures in a group of individuals. Let $0 \leq t_{(1)} < \ldots < t_{(r)} < \infty$ be the observed ordered death times. Let r_j be the size of the risk set at $t_{(j)}$, where risk set denotes the collection of individuals alive and uncensored just before $t_{(j)}$. Let $d_{(j)}$ be the number of observed deaths at $t_{(j)}$, $j = 1, \ldots, r$. Then the K-M estimator of S(t) is definened by

$$\hat{S}(t) = \prod_{j:t<t} \left(1 - \frac{d_j}{r_j}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

This estimator is a step function that changes values only at the time of each death. In fact, K-M estimator will be shown next to maximize the likelihood in the discrete case.

Suppose that the distribution is discrete, with atoms h_j at finitely many specifed points $0 \leq \tau_2 < \tau_2 < \ldots < \tau_j$. The survival function S(t) may be expressed in terms of the discrete hazard function h_j as

$$S(t) = \prod_{j|\tau_j<t} \left(1 - h_j\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

To derive the full likelihood from a sample of n observations, we first collect all the terms corresponding to the atom τ_j. Let $b_i = j$ if the ith individual dies at τ_j. The contribution to the total log likelihood is

$$\log h_{b_i} + \sum_{k<b_i} \log(1 - h_k)$$

Let $e_i = j$ if the ith individual is censored at τ_j; the log likelihood contribution to the total likelihood is

$$\sum_{k \le e_i} \log(1 - h_k)$$

Then the total log likelihood is given by

$$l = \sum_{death\ i} \log h_{b_i} + \sum_{death\ i} \left[\sum_{k<b_i} \log(1 - h_k \right] + \sum_{censor\ i} \left[\sum_{k \le e_i} \log(1 - h_k) \right] \ldots\ldots\ldots\ldots (3)$$

where d_j is the number of observed death at τ_j, c_j is the number censored at $(\tau_j; \tau_{j+1})$; and r_j is the number of living and uncensored at τ_j:

If h_j is the solution of

$$\frac{\partial l}{\partial h_j} = \frac{d_j}{h_j} - \frac{r_j - d_j}{1 - h_j} = 0, \ldots\ldots\ldots (4)$$

then

$$\widehat{h_j} = d_j / r_j \ldots\ldots\ldots\ldots\ldots (5)$$

This maximizes the likelihood since the total log likelihood function is concave down. So that the K-M estimator of the survival function is

$$\hat{S}(t) = \prod_{j|\tau_j<t} (1 - \frac{d_j}{r_j}) \ldots\ldots\ldots (6)$$

Therefore, the K-M estimator is the maximum likelihood estimator.

The K-M estimator gives a discrete distribution. If the observations are modeled to come from unknown continuous distribution, the maximum likelihood estimator does not exist.

3.4 Parametric models

The key theoretical constraint is that the domain of *f(t)* must be R^+. This rules out the Normal distribution, for instance. Suitable distribution families include:

Table 3.1 : Examples of Parametric distribution

	exponential	Gompertz	Weibull	log-logistic
$f(t)$	$\lambda \exp(-\lambda t)$	$\lambda\kappa^t \exp\{\lambda(1-\kappa^t)/\log\kappa\}$	$\lambda\kappa t^{\kappa-1} \exp(-\lambda t^\kappa)$	$\frac{\lambda\kappa t^{\kappa-1}}{(1+\lambda t^\kappa)^2}$
$F(t)$	$1-\exp(-\lambda t)$	$1-\exp\{\lambda(1-\kappa^t)/\log\kappa\}$	$1-\exp(-\lambda t^\kappa)$	$1-\frac{1}{1+\lambda t^\kappa}$
$S(t)$	$\exp(-\lambda t)$	$\exp\{\lambda(1-\kappa^t)/\log\kappa\}$	$\exp(-\lambda t^\kappa)$	$\frac{1}{1+\lambda t^\kappa}$
$h(t)$	λ	$\lambda\kappa^t$	$\lambda\kappa t^{\kappa-1}$	$\frac{\lambda\kappa t^{\kappa-1}}{1+\lambda t^\kappa}$

This paper focuses on the Weibull distribution model for the parametric form.

3.5 CPX PROPORTIONAL HAZARDS MODEL

3.5.1 Linear Regression

In linear regression, we use a predictor or "independent" variable (not independent in the statistical sense) x to explain some of the uncertainty in a "dependent" variable y. If the ith individual has independent and dependent variables x_i and y_i, respectively , the linear model is $y_i = \beta_0 + \beta_i x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$. Note that this is a model, and it depends on certain assumptions, e.g. that the relationship is linear and errors are Gaussian. Note also that as $\varepsilon_i \in (-\infty, \infty)$ and $\beta_0 + \beta_i x_i \in (-\infty, \infty)$, so must $y_i \in (-\infty, \infty)$.

3.5.2 Survival Regression

How can we do likewise for survival data? We choose to focus on models for the hazard function, as this allows statements such as "the risk to males is X times the risk to females" more readily than using the survival function as our basis.

A natural first guess for such as regression survival model would be

$$h(t, x) = \beta_0 + \beta_1 x. \quad \text{.....(7)}$$

There is no "error" term as the randomness is implicit to the survival process. Here we have used the notation $h(t, x)$ to be the hazard function for an individual whose "independent" variable has the value x, while β_0 is a baseline hazard function (for the time being assumed constant in time t) for individuals with $x = 0$.

However, this is a *bad* model. The range of $\beta_0 + \beta_1 x$ may extend below zero for certain values of β_1 or x, but the range of $h(t, x)$ must be $[0, \infty)$.

Luckily , a similar problem has arisen and been solved in generalized linear modeling. There, the predictors are incorporated into different distributions for the dependent variable. F or a Poisson model, the mean must be positive, and the exponential function is used as the *canonical link function* between covariates and mean. We thus follow suit by exponentiating the covariate terms:

$$h(t, x) = \exp(\beta_0 + \beta_1 x) = h_0 \exp(\beta_1 x) > 0. \ldots\ldots\ldots (8)$$

3.5.3 The Cox Proportional Hazards Model

We therefore consider the following generalisation:

$$h(t,x) = h_0(t, \boldsymbol{\alpha}) \exp(\beta^T x), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

where $\boldsymbol{\alpha}$ are some parameters influencing the baseline hazard function.

Note that we have decomposed the hazard into a product of two items:

• $h_0(t, \boldsymbol{\alpha})$, a term that depends on time but not the covariates; and

• $\exp(\boldsymbol{\beta}^T \boldsymbol{x})$, a term that depends on the covariates but not time.

This is the Cox PHM. The beauty of this model, as observed by Cox, is that if you use a model of this form, and you are interested in the effects of the covariates on survival, then you do not need to specify the form of $h_0(t, \boldsymbol{\alpha})$. Even without doing so you may estimate β. The Cox PHM is thus called a semi-parametric model, as some assumptions are made (on $\exp(\boldsymbol{\beta}^T \boldsymbol{x})$) but no form is pre-specified for the baseline hazard $h_0(t, \boldsymbol{\alpha})$.

To see why it is called the PHM, consider two individuals with covariates x_1 and x_2(which we can treat for simplicity as scalars). Then the ratio of their hazards at time t is

$$\frac{h(t, x_1)}{h(t, x_2)} = \frac{h_0(t, \boldsymbol{\alpha}) \exp(\boldsymbol{\beta} x_1)}{h_0(t, \boldsymbol{\alpha}) \exp(\boldsymbol{\beta} x_2)} = \exp\{\beta(x_1 - x_2)\} \ldots\ldots\ldots\ldots\ldots\ldots (10)$$

If $\beta = 0$ then the hazard ratio for that covariate is equal to $e^0 = 1$, i.e. that covariate doesn't affect survival. Thus we can use the notion of hazard ratios to test if covariates influence survival. The hazard ratio also tells us how much more likely one individual is to die than another at any particular point in time. If the hazard ratio comparing men to women were 2, say, it would mean that, at any instant in time, men are twice as likely to die as women.

Note however that this is a model, it could be wrong. There may be an interaction between covariates and time, in which case hazards are not proportional. In the next chapter we learn how to check for violations of the proportional hazards assumption and in the chapter that follows that we extend the PHM to incorporate such interactions. Similarly, there is no reason why we should expect the log of the hazard function to be linear in the covariates. For now, we consider the proportional hazards assumption to be appropriate.

3.5.4 Partial Likelihood for unique failure times

Let us use the notation $\varphi_i = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x_i})$, i.e. φ_i is proportional to the hazard rate for individual *i* (the constant of proportionality being the baseline hazard function).

The partial likelihood for β is

$$l_p(\beta, x) = \prod_{i=1}^{m}\left[\frac{\phi_i}{\sum_{j\epsilon R(t_i)}\phi_j}\right]^{\delta_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11)$$

3.6 PARTIAL LIKELIHOOD FOR REPEATED FAILURE TIMES

The case when two or more individuals are recorded as failing at the same time is more complex. The exact partial likelihood for β is considered last. First consider two approximations. The notation will be simpler (!) if we use the following notation:

- $t_{(i)}$ is the *i*th ordered unique failure time (so if four failures occur at times 1, 1, 3, 3, $t_{(1)} = 1$ and $t_{(2)} = 3$);
- I is the total number of unique failure times;
- *D(t)* is the set of individuals who fail at time *t*.

3.6.1 Breslow's Method

$$l_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \epsilon D(t_{(i)})} \phi_j}{\left(\sum_{j \epsilon R(t_{(i)})} \phi_j\right)^{|D(t_{(i)})|}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

Note that $|D(t_{(i)})|$ is the number of individuals that fail at time $t_{(i)}$.

Breslow's method is the default for many statistical pack ages. But it is not the default for R. R uses Efron's partial likelihood, as it is considered a closer approximation to the exact partial likelihood.

3.6.2 Efron's Method

$$l_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \epsilon D(t_{(i)})} \phi_j}{\left[\sum_{j \epsilon R(t_{(i)})} \phi_j - \frac{k-1}{|D(t_{(i)})|} \sum_{j \epsilon D(t_{(i)})} \phi_j\right]} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (13)$$

3.7 PARAMETRIC PROPORTIONAL HAZARDS MODEL

The parametric proportional hazards model is the parametric versions of the Cox proportional hazards model. It is given with the similar form to the Cox PH models. The hazard function at time *t* for the particular patient with a set of *p* covariates (x_1, $x_2 \ldots x_p$) is given as follows:

$$h(t|x) = h_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = h_0(t) \exp(\beta' x) \ldots\ldots\ldots\ldots (14)$$

The key difference between the two kinds of models is that the baseline hazard function is assumed to follow a specific distribution when a fully parametric PH model is fitted to the data, whereas the Cox model has no such constraint. The coefficients are estimated by partial likelihood in Cox model but maximum likelihood in parametric PH model. Other than this, the

two types of models are equivalent. Hazard ratios have the same interpretation and proportionality of hazards is still assumed. A number of different parametric PH models may be derived by choosing different hazard functions. The commonly applied models are exponential, Weibull, or Gompertz models.

3.7.1 Weibull PH Model

Suppose that survival times are assumed to have a Weibull distribution with scale parameter λ and shape parameter γ, so the survival and hazard function of a W (λ, γ) distribution are given by

$$S(t) = \exp(-\lambda t^{\gamma}), h(t) = \lambda\gamma(t)^{\gamma-1} \dots\dots\dots\dots\dots (15)$$

with λ, γ > 0. The hazard rate increases when λ > 1 and decreases when γ < 1 as time goes on. When γ = 1, the hazard rate remains constant, which is the special exponential case.

Under the Weibull PH model, the hazard function of a particular patient with covariates (x_1, x_2…, x_p) is given by

$$h(t|x) == \lambda\gamma(t)^{\gamma-1} \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = \lambda\gamma(t)^{\gamma-1} \exp(\beta' x) \dots\dots\dots (16)$$

We can see that the survival time of this patient has the Weibull distribution with scale parameter λexp($\boldsymbol{\beta}'\boldsymbol{x}$) and shape parameter γ. Therefore the Weibull family with fixed γ possesses PH property. This shows that the effects of the explanatory variables in the model alter the scale parameter of the distribution, while the shape parameter remains constant.

The corresponding survival function is given by

$$S(t|\boldsymbol{x}) = \exp\{\exp(\beta' x)\, \lambda t^{\gamma}\} \dots\dots\dots\dots\dots\dots\dots (17)$$

After a transformation of the survival function for a Weibull distribution, we can obtain

$$\log\{-\log S(t)\} = \log\lambda + \gamma \log t \dots\dots\dots\dots\dots (18)$$

The log{- log $S(t)$} versus log(t) should give approximately a straight line if the Weibull distribution assumption is reasonable. The intercept and slope of the line will be rough estimate of log λ and γ respectively. If the two lines for two groups in this plot are essentially parallel, this means that the proportional hazards model is valid. Furthermore, if the straight line has a slope nearly one, the simpler exponential distribution is reasonable. In the other way, for a exponential distribution, there is log $S(t) = -\lambda t$. Thus we can consider the graph of log $S(t)$ versus t. This should be a line that goes through the origin if exponential distribution is appropriate.

Another approach to assess the suitability of a parametric model is to estimate the hazard function using the non-parametric method. If the hazard function were reason-ably constant over time, this would indicate that the exponential distribution might be appropriate. If the hazard function increased or decreased monotonically with increasing survival time, a Weibull distribution or Gompertz distribution might be considered.

CHAPTER 4

PROBLEM FORMULATION

4.1 DATA COLLECTION

The entire research was divided into two parts. One focused on a large sized firm and the other on a small sized firm. Ananta Apparels was considered as the firm having huge number of employees. Employee absence data was over a period of 25 working days. The same process was applied for a small sized firm (ABC Airlines travel agency) and their employee records were obtained. The data was recorded according to the following table format.

The complete data collected is provided in Appendix A.

4.2 FORMULATION

Each of the data entered is given a status of 0 or 1. 1 signifying occurrence of event (in this case absence) and 0 indicating no absence recorded in the measured time period. In the gender column, 1 is assigned to male and 2 to female. The similar assignment goes for size of the firms. Modeling the working condition was a bit tricky. In a large apparel factory working condition varies from department to department. Hence the working condition column has been modeled as representative of the departments, the numbers signifying the different departments. It is important to notice that the numbers do not represent the ranking of the departments in terms of working conditions but only the fact that someone working at the Production department will have a different working environment than someone in HR. The employees in the smaller firm, however, will have the same working environment regardless of the department they are in and hence are classified as one whole sector of the working condition.

For example, Md. Abdullahil Hadi of Ananta Apparels has been attending office for 21 consecutive days before taking leave. Hence he is given a status of 1 on day 21. The remaining four days he was present and hence has been given a status of 0 for the remaining 4. Mr. Hadi has been entered under the gender of 1 (male) and his firm size is also 1 (large). Under the

working condition, he has been assigned the number 26 which signifies that he is working at the storage department. The working conditions corresponding to each of the departments are given below:

1 – Accounts

2 – Admin

3 – Compliance

4 – House Keeping

5 – Human Resources

6 – ICT

7 – ISO

8 – Medical

9 – Operation

10 – Planning

11 – Purchase

12 – Sample

13 – Security

14 – Technical

15 – Wash

16 – Welfare

17 – Button

18 – Cutting

19 – Embroidery

20 – Eyelet

21 – Finishing

22 – Maintenance

23 – Production

24 – Quality

25 – Quality Assurance

26 – Storage

27 – Work Study

28 – Others

29 – Small Sized firm

R code is then run on the data set using Kaplan-Meier, Weibull distribution and Cox propotyional hazards model.

CHAPTER 5

APPLICATION OF THE MODEL

5.1 APPLICATIONS IN R

The model is applied as mentioned above in R. Survival probability is first estimated using Kaplan-Meier. It shows the probability of survival, ie, the probability until an absence occurs in a particular time. Once an estimate is obtained from Kaplan-Meier, parametric model (Weibull) can then be run to get a more accurate value as well as a goodness fit test for the data. Finally, the data will be run against a Cox proportional Hazards Model in order to identify which covariate is causing the most absenteeism. The necessary codes for the application to run on R is given in Appendix B.

5.2 RESULTS

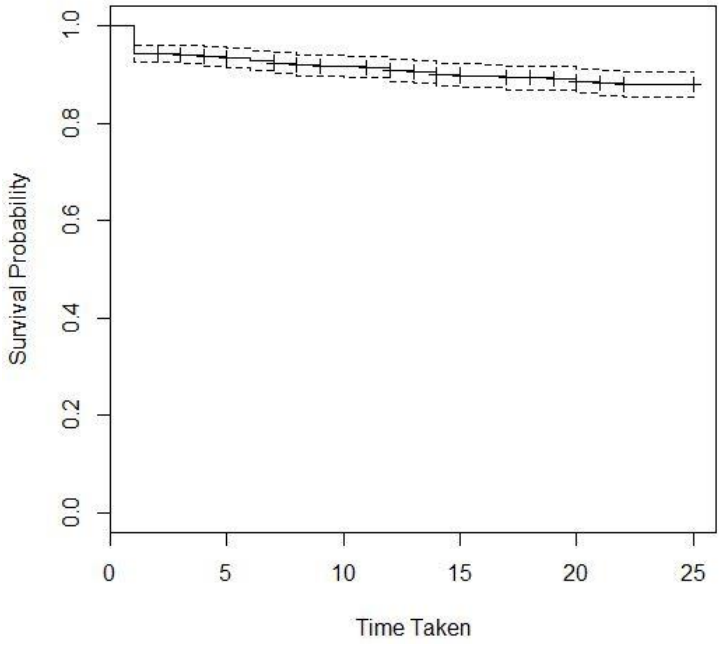


Figure 5.1: (Kaplan-Meier) Survival Curve for all the data

Table 5.1 : Survival probability of the general data

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	634	36	0.943	0.00919	0.925	0.961
3	597	2	0.940	0.00943	0.922	0.959
4	593	2	0.937	0.00966	0.918	0.956
5	587	2	0.934	0.00989	0.915	0.953
6	581	4	0.927	0.01033	0.907	0.948
7	577	2	0.924	0.01054	0.904	0.945
8	574	3	0.919	0.01085	0.898	0.941
9	569	1	0.918	0.01095	0.896	0.939
10	565	1	0.916	0.01105	0.895	0.938
11	564	1	0.914	0.01115	0.893	0.936
12	561	4	0.908	0.01153	0.886	0.931
13	556	1	0.906	0.01163	0.884	0.929
14	552	4	0.900	0.01200	0.876	0.923
15	547	1	0.898	0.01209	0.875	0.922
16	544	1	0.896	0.01218	0.873	0.921
17	543	2	0.893	0.01236	0.869	0.918
19	536	1	0.891	0.01244	0.867	0.916
20	534	3	0.886	0.01271	0.862	0.912
21	529	3	0.881	0.01296	0.856	0.907
22	525	1	0.880	0.01304	0.854	0.906

```
> 
```

```
Call:
survreg(formula = Surv(Time, Status) ~ Sex + Firm.Size + Working.Condition,
    data = data1, dist = "weibull")
                    Value Std. Error       z        p
(Intercept)        5.7832      1.271  4.5486 5.40e-06
Sex                0.0366      0.908  0.0403 9.68e-01
Firm.Size          1.7815      0.798  2.2334 2.55e-02
Working.Condition -0.0405      0.033 -1.2283 2.19e-01
Log(scale)         0.6430      0.112  5.7180 1.08e-08

Scale= 1.9

Weibull distribution
Loglik(model)= -443.1   Loglik(intercept only)= -446.2
        Chisq= 6.3 on 3 degrees of freedom, p= 0.098
Number of Newton-Raphson Iterations: 8
n=633 (1 observation deleted due to missingness)
```

Figure 5.2 : Result of Weibull distribution

```
Call:
coxph(formula = Surv(Time, Status) ~ Sex + Firm.Size + Working.Condition,
    data = data1)

  n= 633, number of events= 75
   (1 observation deleted due to missingness)

                     coef exp(coef) se(coef)      z Pr(>|z|)
Sex                -0.01769   0.98246  0.47762 -0.037   0.9704
Firm.Size          -0.92747   0.39555  0.40727 -2.277   0.0228 *
Working.Condition   0.02113   1.02136  0.01718  1.230   0.2186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
Sex                   0.9825     1.0179    0.3853    2.5053
Firm.Size             0.3956     2.5281    0.1780    0.8788
Working.Condition     1.0214     0.9791    0.9875    1.0563

Concordance= 0.588  (se = 0.034 )
Rsquare= 0.01   (max possible= 0.779 )
Likelihood ratio test= 6.17  on 3 df,   p=0.1036
Wald test            = 5.54  on 3 df,   p=0.1363
Score (logrank) test = 5.76  on 3 df,   p=0.1241
```

Figure 5.3 : Result of Cox Proportional Hazard Model

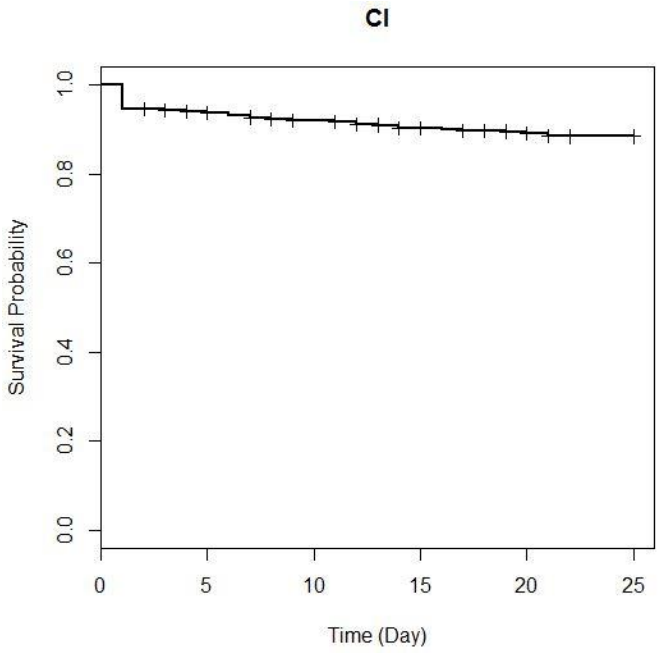


Figure 5.4 : Survival curve based on Cox PHM

The database is divided into training and testing data sets for verification. 30% of the total data is sampled for testing.

Train data set

```
                 Name          Time          Status           Sex
MEHADI HASAN       : 10   Min.   : 1.0   Min.   :0.0000   Min.   :1.000
Md. Fazle Rabbi    :  6   1st Qu.:25.0   1st Qu.:0.0000   1st Qu.:1.000
Md.Suman Matbar    :  6   Median :25.0   Median :0.0000   Median :1.000
Md.Tafazzal Hossain:  5   Mean   :21.8   Mean   :0.1329   Mean   :1.081
Md. RUHUL AMIN     :  4   3rd Qu.:25.0   3rd Qu.:0.0000   3rd Qu.:1.000
   Farhana Yesmin  :  3   Max.   :25.0   Max.   :1.0000   Max.   :2.000
(Other)            :410
  Firm.Size   Working.Condition
Min.   :1.0   Min.   : 1.00
1st Qu.:1.0   1st Qu.:13.00
Median :1.0   Median :22.00
Mean   :1.2   Mean   :19.95
3rd Qu.:1.0   3rd Qu.:27.00
Max.   :2.0   Max.   :29.00
```

Figure 5.5 : The Train Data Set

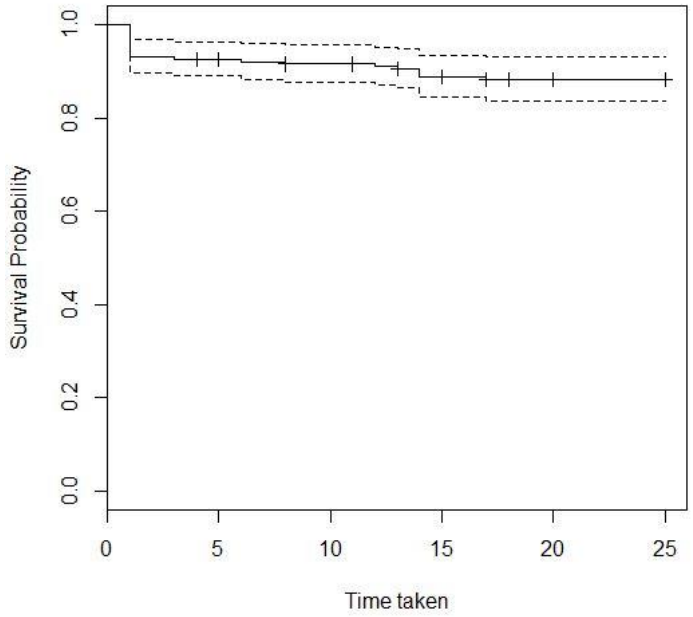


Figure 5.6 : Survival curve of the train data set

Table 5.2 : Survival probability table of train data set

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	190	13	0.932	0.0183	0.896	0.968
3	177	1	0.926	0.0190	0.890	0.964
6	172	1	0.921	0.0196	0.883	0.960
8	171	1	0.916	0.0202	0.877	0.956
12	168	1	0.910	0.0208	0.870	0.952
13	167	1	0.905	0.0214	0.864	0.948
14	165	3	0.888	0.0230	0.844	0.934
17	161	1	0.883	0.0235	0.838	0.930

```
Call:
survreg(formula = Surv(Time, Status) ~ Sex + Firm.Size + Working.Condition,
    data = data1, dist = "weibull")
                   Value Std. Error       z        p
(Intercept)       5.7832      1.271  4.5486 5.40e-06
Sex               0.0366      0.908  0.0403 9.68e-01
Firm.Size         1.7815      0.798  2.2334 2.55e-02
Working.Condition -0.0405      0.033 -1.2283 2.19e-01
Log(scale)        0.6430      0.112  5.7180 1.08e-08

Scale= 1.9

Weibull distribution
Loglik(model)= -443.1   Loglik(intercept only)= -446.2
        Chisq= 6.3 on 3 degrees of freedom, p= 0.098
Number of Newton-Raphson Iterations: 8
n=633 (1 observation deleted due to missingness)
```

Figure 5.7 : Result of Weibull PHM on Train data set

```
Call:
coxph(formula = Surv(Time, Status) ~ Sex + Firm.Size + Working.Condition,
    data = data1)

  n= 633, number of events= 75
   (1 observation deleted due to missingness)

                     coef exp(coef) se(coef)      z Pr(>|z|)
Sex               -0.01769   0.98246  0.47762 -0.037   0.9704
Firm.Size         -0.92747   0.39555  0.40727 -2.277   0.0228 *
Working.Condition  0.02113   1.02136  0.01718  1.230   0.2186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
Sex                  0.9825     1.0179    0.3853    2.5053
Firm.Size            0.3956     2.5281    0.1780    0.8788
Working.Condition    1.0214     0.9791    0.9875    1.0563

Concordance= 0.588  (se = 0.034 )
Rsquare= 0.01   (max possible= 0.779 )
Likelihood ratio test= 6.17  on 3 df,   p=0.1036
Wald test            = 5.54  on 3 df,   p=0.1363
Score (logrank) test = 5.76  on 3 df,   p=0.1241
```

Figure 5.8 : Result of Cox PHM on Train data set

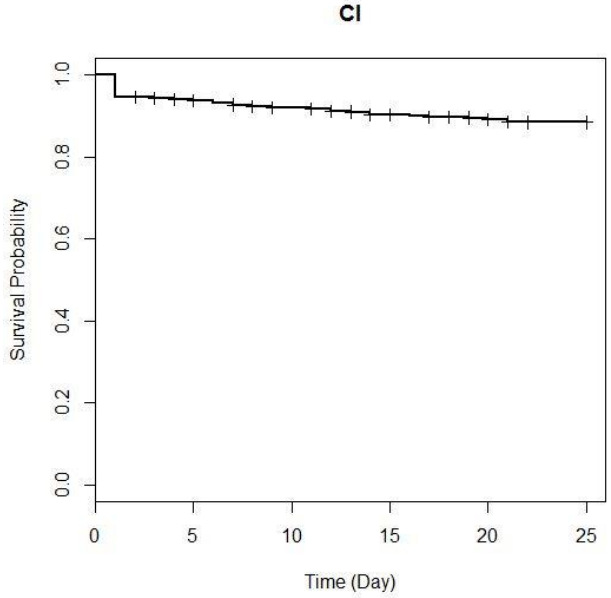


Figure 5.9 : Cox PHM Survival curve of Train data set

Test Data Set

```
                  Name          Time           Status           Sex
MEHADI HASAN        :  5   Min.   : 1.00   Min.   :0.00000   Min.   :1.00
Md. Alimuzzaman     :  3   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:1.00
Golam Kader         :  2   Median :25.00   Median :0.00000   Median :1.00
Md. Kamal Uddin     :  2   Mean   :22.61   Mean   :0.08421   Mean   :1.09
Md. Masud Rana      :  2   3rd Qu.:25.00   3rd Qu.:0.00000   3rd Qu.:1.00
Md. Mizanur Rahman  :  2   Max.   :25.00   Max.   :1.00000   Max.   :2.00
(Other)             :174                                     NA's   :1
  Firm.Size     Working.Condition
Min.   :1.000   Min.   : 1.00
1st Qu.:1.000   1st Qu.:13.00
Median :1.000   Median :21.00
Mean   :1.222   Mean   :19.24
3rd Qu.:1.000   3rd Qu.:28.00
Max.   :2.000   Max.   :29.00
NA's   :1
```

Figure 5.10 : The Test Data Set

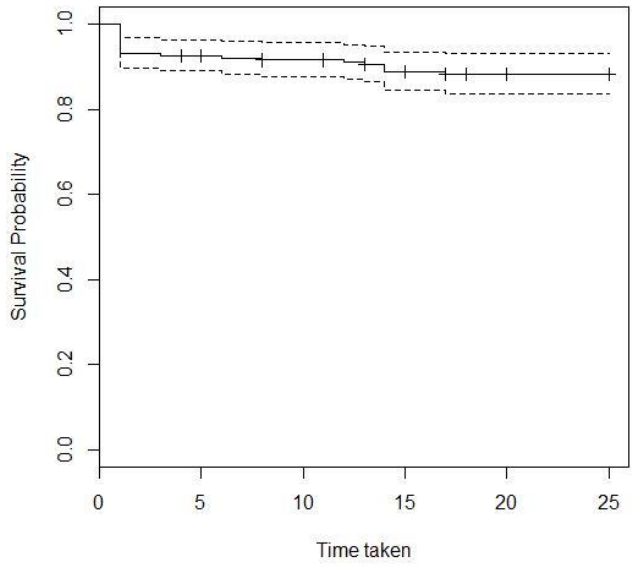


Figure 5.11 : Survival Curve of Test Data Set

Table 5.3 : Survival Probability Table of Test Data Set

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	190	13	0.932	0.0183	0.896	0.968
3	177	1	0.926	0.0190	0.890	0.964
6	172	1	0.921	0.0196	0.883	0.960
8	171	1	0.916	0.0202	0.877	0.956
12	168	1	0.910	0.0208	0.870	0.952
13	167	1	0.905	0.0214	0.864	0.948
14	165	3	0.888	0.0230	0.844	0.934
17	161	1	0.883	0.0235	0.838	0.930

```
Call:
survreg(formula = Surv(Time, Status) ~ Sex + Firm.Size + Working.Condition,
    data = data1, dist = "weibull")
                   Value Std. Error       z        p
(Intercept)       5.7832      1.271  4.5486 5.40e-06
Sex               0.0366      0.908  0.0403 9.68e-01
Firm.Size         1.7815      0.798  2.2334 2.55e-02
Working.Condition -0.0405      0.033 -1.2283 2.19e-01
Log(scale)        0.6430      0.112  5.7180 1.08e-08

Scale= 1.9

Weibull distribution
Loglik(model)= -443.1   Loglik(intercept only)= -446.2
        Chisq= 6.3 on 3 degrees of freedom, p= 0.098
Number of Newton-Raphson Iterations: 8
n=633 (1 observation deleted due to missingness)
```

Figure 5.12 : Result of Weibull PHM on test data set

```
Call:
coxph(formula = Surv(Time, Status) ~ Sex + Firm.Size + Working.Condition,
    data = data1)

  n= 633, number of events= 75
   (1 observation deleted due to missingness)

                     coef exp(coef) se(coef)      z Pr(>|z|)
Sex               -0.01769   0.98246  0.47762 -0.037   0.9704
Firm.Size         -0.92747   0.39555  0.40727 -2.277   0.0228 *
Working.Condition  0.02113   1.02136  0.01718  1.230   0.2186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
Sex                  0.9825     1.0179    0.3853    2.5053
Firm.Size            0.3956     2.5281    0.1780    0.8788
Working.Condition    1.0214     0.9791    0.9875    1.0563

Concordance= 0.588  (se = 0.034 )
Rsquare= 0.01   (max possible= 0.779 )
Likelihood ratio test= 6.17  on 3 df,   p=0.1036
Wald test            = 5.54  on 3 df,   p=0.1363
Score (logrank) test = 5.76  on 3 df,   p=0.1241
```

Figure 5.13 : Result of Cox PHM on Test Data Set

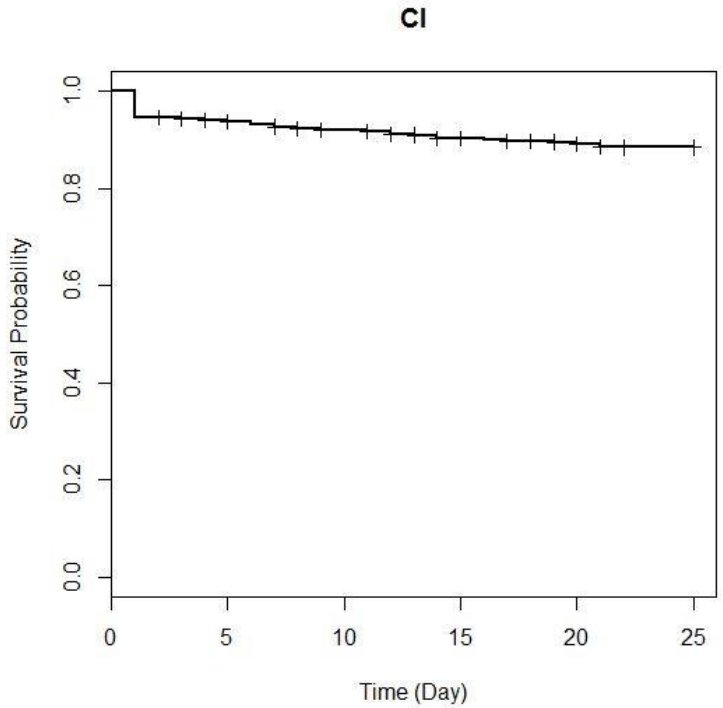


Figure 5.14 : Cox PHM Survival Curve of Test Data Set

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1 CONCLUSION

The Cox PH model is the most widely used way of analyzing survival data in clinical research. In a review paper of survival analysis published in cancer journals, it was found that only Five percent of all studies using the Cox PH models check PH assumption. However PH assumption is not always satisfied in the data. If this assumption does not hold there are various solutions to consider. One solution is to include the time-dependent variable for the predictors with non-proportional hazards. When this approach is used to account for a variable with non-proportionality, different results may be obtained from different choices of time-dependent variables. It is hard to choose between models. Alter-natively we can use a model where we stratify on the non-proportional predictors. The stratified Cox model is not appropriate when the covariate with non-proportionality is continuous or of direct interests. And both ways are still based on comparison of hazards

In this study, the data was measured over one month of employee records. Three covariates were considered which were taken to be the prime factor causing individual absenteeism in a workplace. From the original data set, it can be seen in Fig 5.1 that the survival probability of employee absenteeism is decreasing with time which is in accordance to the theoretical shape of survival curve. From the Table 5.1, survival probability at any particular time can also be read. From example, the probability that any employee will pass at least two days before asking for absence is 0.940. In other words, his probablility of taking leave after two consecutive days of work is 0.006. This result was obtained with Kaplan-Meier estimator.

The data set and the results were refined by using a parametric distribution, Weibull distribution in this case. Figure 5.2 shows the Weibull distribution results from which p value signifies the probability of the event (absenteeism) occurring and the value of the log likelihood function as the value of goodness fit test. However, none of these models take in consideration the effects of the covariates. From Figure 5.3, we can analyse the effects of each of the covariates from the

Cox Proportional Hazards Model. We can see that "Working Condition" is the factor or covariate that has the highest influence on absenteeism since it has the highest coefficient value. From the figure, we can see that "Working Condition" has a more than 1 value of exp(coef) which signifies that it has the greatest hazard ratio.

Morevover, when the original data set is divided into test data and train data, consistent results are obtained which can be seen from the figures and tables corresponding to the test data and train data sets. The predict function applied on the test data sets were not completely randomly distributed, however there were some regions were the data showed randomness (some positive and some negative) proving that the model requires a larger set of data.

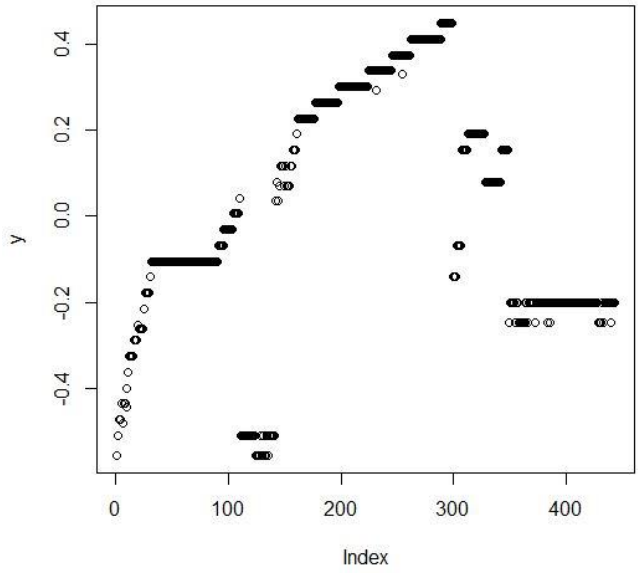


Fig 6.1: Validation of the model

6.2 RECOMMENDATIONS

- The study can be performed over a longer duration. In this paper, the study has been performed for one month. The study can be expanded to include data for one year to verify the models.
- The study can be performed in an industry with a greater number of workers.
- Exponential or Gompertz distribution model can be applied as part of the parametric estimator
- The Cox Proportional Hazard Model can be solved by the "efron" method in R rather than "breslow" to compare the results.

REFERENCES

1. Abeysekera, W. W. M., & Sooriyarachchi, R. (2009). Use of Schoenfeld's global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study.

2. Baram, D., Daroowalla, F., Garcia, R., Zhang, G., Chen, J. J., Healy, E., ... & Richman, P. (2008). Use of the All Patient Refined-Diagnosis Related Group (APR-DRG) risk of mortality score as a severity adjustor in the medical ICU.*Clinical medicine. Circulatory, respiratory and pulmonary medicine*, *2*, 19.

3. Barmby, T. A., Ercolani, M. G., & Treble, J. G. (2002). Sickness absence: an international comparison. *The Economic Journal*, *112*(480), F315-F331.

4. Barmby, T., Sessions, J., & Treble, J. (1994). Absenteeism, efficiency wages and shirking. *The Scandinavian Journal of Economics*, 561-566.

5. Biørn, E., Gaure, S., Markussen, S., & Røed, K. (2013). The rise in absenteeism: disentangling the impacts of cohort, age and time. *Journal of Population Economics*, *26*(4), 1585-1608.

6. Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis Part III: multivariate data analysis–choosing a model and assessing its adequacy and fit. *British journal of cancer*, *89*(4), 605.

7. Bradley, S., Green, C., & Leeves, G. (2007). Worker absence and shirking: Evidence from matched teacher-school data. *Labour Economics*, *14*(3), 319-334.
8. Brooke, P. P., & Price, J. L. (1989). The determinants of employee absenteeism: An empirical test of a causal model*. *Journal of Occupational Psychology*, *62*(1), 1-19.

9. Burt, J. M. (1984). *Relationships between leadership behavior and employee absenteeism and turnover in community hospitals' departments* (Doctoral dissertation, George Peabody College for Teachers of Vanderbilt University).

10. Chatterji, M., & Tilley, C. J. (2002). Sickness, absenteeism, presenteeism, and sick pay. *Oxford Economic Papers*, 669-687.

11. Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *British journal of cancer*,*89*(2), 232.

12. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187-220.

13. Dale-Olsen, H. (2013). Absenteeism, efficiency wages, and marginal taxes.*The Scandinavian Journal of Economics*, *115*(4), 1158-1185.

14. De Paola, M. (2010). Absenteeism and peer interaction effects: evidence from an Italian public institute. *The Journal of Socio-Economics*, *39*(3), 420-428.

15. Drago, R., & Wooden, M. (1992). The determinants of labor absence: Economic factors and workgroup norms across countries. *Industrial & Labor Relations Review*, *45*(4), 764-778.

16. Driver, R. W., & Watson, C. J. (1989). Construct validity of voluntary and involuntary absenteeism. *Journal of Business and Psychology*, *4*(1), 109-118.

17. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). Model-building strategies and methods for logistic regression. *Applied Logistic Regression, Third Edition*, 89-151.

18. Jensen, S., & McIntosh, J. (2007). Absenteeism in the workplace: results from Danish sample survey data. *Empirical Economics*, *32*(1), 125-139.

19. Johansson, P., & Palme, M. (1996). Do economic incentives affect work absence? Empirical evidence using Swedish micro data. *Journal of Public Economics*, *59*(2), 195-218.

20. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.

21. Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

22. Kul, S. (2010). The use of survival analysis for clinical pathways. *International Journal of Care Pathways*, *14*(1), 23-26.

23. Liu, Y. A. (1996). *Incentives perceived by management dietitians to reduce absenteeism rate of foodservice personnel in health care systems* (Doctoral dissertation, Oklahoma State University).

24. Markussen, S., Røed, K., Røgeberg, O. J., & Gaure, S. (2011). The anatomy of absenteeism. *Journal of health economics*, *30*(2), 277-292.

25. Meier, E. N. (2012). *A sensitivity analysis for clinical trials with informatively censored survival endpoints* (Doctoral dissertation, University of Washington).

26. Ose, S. O. (2005). Working conditions, compensation and absenteeism.*Journal of health economics*, *24*(1), 161-188.

27. Peterson, A. V. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, *73*(1), 11-13.

28. Qi, J. (2009). *Comparison of proportional hazards and accelerated failure time models* (Doctoral dissertation, University of Saskatchewan Saskatoon).

29. Satten, G. A., Datta, S., & Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics & probability letters*, *54*(4), 397-403.

30. Scott, D., & Markham, S. E. (1982). Absenteeism control methods: A survey of practices and results. *Personnel Administrator*, *27*(6), 73-84.

31. Sherbert, E. G. (2001). *The Impact of Work Redesign on Job Satisfaction, Organizational Commitment, Employee Absenteeism and Turnover: A Longitudinal Study*.

32. Smith, T., & Smith, B. (2001, April). Survival analysis and the application of Cox's proportional hazards modeling using SAS. In *Proceedings of the twenty-sixth annual SAS user's group international conference. SAS Institute Inc, Cary, NC* (pp. 244-246).

33. Smith, T., Smith, B., & Ryan, M. A. (2003). Survival analysis using Cox proportional hazards modeling for single and multiple event time data. In*Proceedings of the twenty-eighth annual SAS users group international conference, SAS Institute, Inc, Cary, paper* (Vol. 2003, pp. 254-228).

34. Steers, R. M., & Rhodes, S. R. (1978). Major influences on employee attendance: A process model. *Journal of Applied psychology*, *63*(4), 391.

35. Williams, J. S., & Lagakos, S. W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. *Biometrika*, *64*(2), 215-224.

36. Zheng, M., & Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, *82*(1), 127-138.

APPENDIX A

Name	Time	Status	Sex	Firm Size	Working Condition
Mr. Sudesh	10	1	1	1	1
Mr. Sudesh	15	0	1	1	1
Md. Abdul Kader Bepary	8	1	1	1	1
Md. Abdul Kader Bepary	17	0	1	1	1
Most. Ayesha	25	0	2	1	2
Mr. Sumon Kumar Das	25	0	1	1	2
K.M Nesar Uddin	25	0	1	1	3
Baten Miah	25	0	1	1	3
Md. Mainuddin Mizi	25	0	1	1	3
Iftey Khayrul Alam Shiplu	25	0	1	1	4
Rabeya Begum Lubna	25	0	2	1	4
Md. Kamrul Islam	25	0	1	1	4
Md. Jahiddur Rahman	25	0	1	1	4
Md. Abdur Rahim	25	0	1	1	4
Md. Abul Hossain	25	0	1	1	4
Mrs. Sahera Begum	25	0	2	1	4
Mrs. Rahima Begum	25	0	2	1	5
Mrs. Momtaz Begum	25	0	2	1	5
Md. Mahabub Alam Manik	25	0	1	1	5
Md. Monirul Islam Parvez	25	0	1	1	5
Md. Sayem	25	0	1	1	6
Hasan Rinku	25	0	1	1	6
Md. Nur Hossain	25	0	1	1	7
Md. Kamrul Hasan	25	0	1	1	7
Md. Julfiker Haydar	25	0	1	1	7
Md. Tareq Aziz Khan	25	0	1	1	7
Shofiqul Islam	6	1	1	1	7
Shofiqul Islam	17	1	1	1	7
Shofiqul Islam	2	0	1	1	8
Md. Bashir Rayhan	25	0	1	1	8
Al-Amin	25	0	1	1	8
Md. Habibullah	25	0	1	1	8
Md. Samsuddin	25	0	1	1	9
Md. Monjur Hossain	25	0	1	1	9
Mr. Bitu Chakma	25	0	1	1	9

Mr. Roton Kumar Roy	25	0	1	1	9
Dr. Tahmina Nasreen	21	1	2	1	9
Dr. Tahmina Nasreen	4	0	2	1	10
Dr. Habiba Shirin	25	0	2	1	10
Miss Rubi	25	0	2	1	10
Mrs. Shikha Mazumder	25	0	2	1	10
Md. Dudu Miah	25	0	1	1	10
Md. Khorshed Alam Jahangir	25	0	1	1	11
Md. Shahjahan Khan	25	0	1	1	11
Md. Anowar Hossain	25	0	1	1	11
Md. Abdur Rahim	25	0	1	1	11
Md. Shafiqul Islam	25	0	1	1	12
Md. Abu Nasher	25	0	1	1	12
Khandakar Mijanur Rahman	25	0	1	1	12
Mahedi Hasan	25	0	1	1	13
Razaur Rahman Bhiayan	25	0	1	1	13
Md. Habibur Rahman	25	0	1	1	13
Sheikh Md. Ramjan Ali	4	1	1	1	13
Sheikh Md. Ramjan Ali	21	0	1	1	13
Md. Mohir Uddin	25	0	1	1	13
Md. Rubel Miah	25	0	1	1	13
Md. Al-Mamun	25	0	1	1	13
Md. Jalal Uddin	25	0	1	1	13
Md.Akhteruzzaman	25	0	1	1	13
Chunnu Mia	25	0	1	1	13
Shimul Ahamed Belayet	25	0	1	1	13
Razu Miah	25	0	1	1	13
Nazrul Islam	12	1	1	1	13
Nazrul Islam	13	0	1	1	13
Md. Belayet Hossain	25	0	1	1	13
Md.Fazlur Rahman	25	0	1	1	13
Suman Mahmud	25	0	1	1	13
Md.Jashim Uddin	25	0	1	1	13
Md.Monir	25	0	1	1	13
Md.Abul Kashem	25	0	1	1	13
Md. Shakil Islam	25	0	1	1	13
Md.Abdul Jalil-1	25	0	1	1	13
Md.Masud	25	0	1	1	13
Monwar Hossain Shamim	25	0	1	1	13

Md.Masum	25	0	1	1	13
Ashaduzzaman Ripon	25	0	1	1	13
Zannat Chowkider Enayet	25	0	1	1	13
Jashim Uddin -4	25	0	1	1	13
Md.Jashim Mallik	25	0	1	1	13
Md. Rostom Dali -2	25	0	1	1	13
Abu Bakar Siddique	25	0	1	1	13
Md.Shahin	25	0	1	1	13
Sogir Hossain	25	0	1	1	13
Anarul Islam	25	0	1	1	13
Md. Raju Ahmed	25	0	1	1	13
Md. Hazrat Ali	25	0	1	1	13
Md. Kamal Hossain	25	0	1	1	13
Md. Mizanur Rahman	25	0	1	1	13
Md Fazlul Haque	25	0	1	1	13
Abdul Jabbar	25	0	1	1	13
Md. Rafiqul Islam-2	25	0	1	1	13
Md. Shamsul Haque	25	0	1	1	13
Nur Mohammed	25	0	1	1	13
Md. Liton	25	0	1	1	13
Md. Harun Talukder	25	0	1	1	13
Md. Abdur Rahman	25	0	1	1	13
Md. Kabul Shikder	25	0	1	1	13
Md. Noyon Gazi	25	0	1	1	13
Md. Abdul Kader-2	25	0	1	1	13
Md. Sayedul Rahman	25	0	1	1	13
Md. Shazahan Sarder	25	0	1	1	13
Md. Robin	25	0	1	1	13
Md. Rashed	25	0	1	1	13
Md. Marfudul Islam	25	0	1	1	13
Md. Shohag	25	0	1	1	13
Md. Mafizul Khan	25	0	1	1	13
Md. Abdus Salam Bacchu	25	0	1	1	13
Md. Hasan Mizi	25	0	1	1	13
Md. Shariful Islam	25	0	1	1	13
Md. Zahirul Islam	25	0	1	1	13
Md. Forhad	25	0	1	1	13
Md. Minul Islam	25	0	1	1	13
Md. Nazrul Islam	25	0	1	1	13
Md. Ibrahim	25	0	1	1	13

Md.Dulal	25	0	1	1	13
Md. Forkan	8	1	1	1	13
Md. Forkan	17	0	1	1	13
Md. Abdul Mannan	25	0	1	1	13
Md. Hanif Mia	25	0	1	1	13
Md. Rony	25	0	1	1	13
Manojit Biswas (Sajib)	25	0	1	1	13
Md. Jahangir Alam	25	0	1	1	13
Nijam Uddin	25	0	1	1	13
Abdur Rouf	25	0	1	1	13
Maidul	25	0	1	1	13
Md. Elias Kanchon	25	0	1	1	13
Md. Shahed Miah	25	0	1	1	13
Md. Ripon Hossain	25	0	1	1	13
Md. Ali Azom Tarafder	25	0	1	1	13
Md. Rafiqul Islam-3	25	0	1	1	13
Nur Mohammad Shohag	25	0	1	1	13
Md. Ashraf	25	0	1	1	13
Md. Ali Ahamed	25	0	1	1	13
Md. Mahbub	25	0	1	1	13
Saiful Islam-1	25	0	1	1	13
Md. Saiful Islam-2	15	1	1	1	13
Md. Saiful Islam-2	1	1	1	1	13
Md. Saiful Islam-2	9	0	1	1	13
Rasel Khan	25	0	1	1	13
Md. Nazrul Islam	25	0	1	1	13
Helal Uddin	25	0	1	1	13
Mahfuzur Rahman	25	0	1	1	13
Md. Halim	25	0	1	1	13
Md. Mohon	25	0	1	1	13
Tanvir Anzum Tanzir	25	0	1	1	13
Md. Soloyman	20	1	1	1	13
Md. Soloyman	5	0			14
Md. Alamgir Hossain	25	0	1	1	14
Md. Fazle Rabbi	4	1	1	1	14
Md. Fazle Rabbi	1	1	1	1	14
Md. Fazle Rabbi	1	1	1	1	14
Md. Fazle Rabbi	12	1	1	1	14
Md. Fazle Rabbi	1	1	1	1	14
Md. Fazle Rabbi	1	1	1	1	14
Md. Fazle Rabbi	5	0	1	1	15

Md.Golam Mostafa	25	0	1	1	15
Md. Rasel Kabir	25	0	1	1	15
Md. Ashrafuzzaman Ratan	25	0	1	1	15
Shahinur Alam	25	0	1	1	15
Shahedur Jaman Badol	25	0	1	1	15
Md. Hasan Ullah	25	0	1	1	15
Md. Mazharul Islam	25	0	1	1	15
Md.Obaydul Haque	25	0	1	1	15
Md. Nurul Alam	25	0	1	1	15
Obaidur Rahman	25	0	1	1	15
Md. Mamunur Rashid	25	0	1	1	15
Md. Alimuzzaman	14	1	1	1	16
Md. Alimuzzaman	3	1	1	1	16
Md. Alimuzzaman	8	0	1	1	16
Md. Mir Hossain	25	0	1	1	16
Md. Shahed Miah	25	0	1	1	16
Md. Amzad Hossain	25	0	1	1	16
Md. Ziaur Rahman	25	0	1	1	16
Mr. Jhanardhan Punnoumoni	25	0	1	1	17
Md. Zahidul Islam	25	0	1	1	17
Md. Harun ur Rashid	25	0	1	1	17
Md.Sirajul Islam	25	0	1	1	17
Nasir Hossain	25	0	1	1	17
Md. Liton Hawlader	25	0	1	1	2
Md. Iqbal Hossain	25	0	1	1	2
Md. Ashadul Islam	25	0	1	1	2
Mohammed Leaket Ali	25	0	1	1	2
Md. Masud Ali	25	0	1	1	2
Shahidul Islam	25	0	1	1	2
Saidur Rahman	25	0	1	1	2
Md. Abu Taher	25	0	1	1	2
Md.Monjur Hasan (Nayan)	25	0	1	1	2
Md.Nuru Miah	25	0	1	1	2
Md.SahIdul Islam	25	0	1	1	2
Md.Sharif Ahmed Khan	25	0	1	1	2
ASM Mottasim Billah	25	0	1	1	2
Shahab Uddin	25	0	1	1	2
Kamrun Nahar Shilpi	25	0	1	1	2
Khairun Nahar Shampa	25	0	2	1	2
Nazma Khatun	25	0	2	1	2

Afroza Akter Mukta	25	0	2	1	2
Sajeda Yesmin	25	0	2	1	2
Mrs.Baby	25	0	2	1	2
Md. Feroz Alam	25	0	1	1	2
Peara Begum	25	0	2	1	2
Md.Jahangir	25	0	1	1	2
Mrs. Sufia Begum	25	0	2	1	2
Mrs. Sufia Motaleb	25	0	2	1	2
Md. Mosarraf Hossain	25	0	1	1	2
Mrs. Nazma Begum	25	0	2	1	2
Main Uddin Ahmmed	25	0	1	1	2
Md.Tafazzal Hossain	6	1	1	1	2
Md.Tafazzal Hossain	1	1	1	1	2
Md.Tafazzal Hossain	1	1	1	1	2
Md.Tafazzal Hossain	1	1	1	1	2
Md.Tafazzal Hossain	1	1	1	1	2
Md.Tafazzal Hossain	15	0	1	1	2
Nur Muhammad	25	0	1	1	2
Md. Shariful Hossain	25	0	1	1	2
Rubina Begum	25	0	2	1	2
Md. Abdur Rahman	25	0	1	1	2
Miss Shewly Parven	25	0	2	1	18
MONWARA	25	0	2	1	18
Md. Mizanur Rahman	25	0	1	1	18
Sree Sabitri Rani	25	0	2	1	18
Ayrin Sultana	25	0	2	1	19
Md. Tariqul Islam	25	0	1	1	19
Md. RAFIQ	25	0	1	1	19
Md. Roich Uddin	25	0	1	1	19
Md. Rafikur Rahman	25	0	1	1	19
S.M Fazlul Haque	25	0	1	1	19
Md. Kamal Uddin	25	0	1	1	19
Md. Nazrul Islam	25	0	1	1	19
Md. Nurun Nabi	25	0	2	1	19
Mr. JOJ MIAH	25	0	1	1	19
Ferdousi Begum	25	0	2	1	19
Md.Alam Miah	25	0	1	1	19
Mrs. JOSHNA BEGUM	25	0	2	1	19
SHIRIN SULTANA	25	0	2	1	19
Md. Hafizur Rahman	25	0	1	1	19
Md. Joynal Abedin	25	0	1	1	19

Md. RAMZAN ALI	25	0	1	1	20
Md. NADIM	25	0	1	1	20
Md. SANAULLAH	25	0	1	1	20
Md. HARUNUR RASHID	20	1	1	1	20
Md. HARUNUR RASHID	5	0	1	1	20
Md.Monir Hossain	25	0	1	1	20
Md.Monir Hossain	25	0	1	1	20
Md. POLASH	25	0	1	1	21
Md.Yousuf Shiekh	25	0	1	1	21
Md.Abul Kashem	25	0	1	1	22
Md.Mamun Khan	25	0	1	1	22
Md.Ashaduzzaman	25	0	1	1	22
Md.Kabir Hossain	25	0	1	1	22
Amirul Islam	25	0	1	1	22
Alaul Huda	25	0	1	1	22
Md.Nesar Uddin	25	0	1	1	22
Md. Ramjan Hossain	25	0	1	1	22
Md. Abu Hanif	25	0	1	1	22
Md.Faruq Hossain	25	0	1	1	22
Md. ABU BAKKAR SIDDIQE	25	0	1	1	22
Md. Ismail Hossain	25	0	1	1	22
Md. SALEH MUSA	25	0	1	1	22
Md. Nazmul Huda	25	0	1	1	22
Md. CHAN MIA	25	0	1	1	22
Md. SAIFUL ISLAM	25	0	1	1	22
MD. ABUL KALAM	25	0	1	1	22
Md. Mowdud Ahmmed	25	0	1	1	22
Md. Rana Islam	25	0	1	1	23
Md. Jamshed Sarker	25	0	1	1	23
Md. ZAHID	25	0	1	1	23
MOIN UDDIN AHMED	25	0	1	1	23
ABDUR RAHMAN	25	0	1	1	23
Md.Ashiqur Rahman	25	0	1	1	23
Md.Suman Matbar	13	1	1	1	23
Md.Suman Matbar	1	1	1	1	23
Md.Suman Matbar	1	1	1	1	23
Md.Suman Matbar	1	1	1	1	23
Md.Suman Matbar	1	1	1	1	23
Md.Suman Matbar	1	1	1	1	23
Md.Suman Matbar	7	0	1	1	23
Mr.Belal Hossain	25	0	1	1	23

Mr.Poritosh Babu	25	0	1	1	23
Siddiqur Rahman	25	0	1	1	23
Md.Babul Miah	25	0	1	1	23
Saif Uddin	25	0	1	1	23
Sohrab Hossain	25	0	1	1	23
Md.Bacchu Miah	25	0	1	1	23
Md.Saiful Islam	25	0	1	1	23
Md.Ripon Miah	14	1	1	1	23
Md.Ripon Miah	11	0	1	1	23
Md.Saju Miah	25	0	1	1	23
Md.Khorshed Alam	25	0	1	1	23
Md. MONIR HOSSAIN	25	0	1	1	23
Md.Enayet Mollah	25	0	1	1	23
Md.Mohiuddin Badsha	25	0	1	1	24
MEHADI HASAN	8	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	1	1	1	1	24
MEHADI HASAN	4	0	1	1	24
Md.Jahangir Alam	25	0	1	1	24
Md.Anisur Rahman	25	0	1	1	24
SHAMSUL HAQUE	25	0	1	1	24
Md. Shajahan	25	0	1	1	24
Md. Abdul Haque	25	0	1	1	24
Md. Monsur Ali	25	0	1	1	24
Md. Mizanur Rahman	25	0	1	1	24
Md. Badrudduza Chowdhury	25	0	1	1	24
Md. Zahidur Rahman Shahin	25	0	1	1	24
Md. Mazidul Islam	25	0	1	1	24
Md. Mohin Uddin	25	0	1	1	24
Md. Azizul Haque	25	0	1	1	24

Md. Sheikh Bahauddin	25	0	1	1	24
Md. Jamir Uddin	25	0	1	1	24
Md. Rubel Hossain	25	0	1	1	24
Md. Masud Rana	25	0	1	1	24
Md. Mamun Hossain	25	0	1	1	24
Md. Aminul Islam	25	0	1	1	24
Md. ROFIQUL ISLAM RONY	25	0	1	1	24
Md. ABDUR RAHIM	25	0	1	1	24
Md. BILLAL HOSSAIN	25	0	1	1	24
Md. SUMON MIA	25	0	1	1	24
Md. KAIYUM	25	0	1	1	24
Md. SHEIKH MUSTAFIZUR RAHMAN	25	0	1	1	24
Md. ZAHID MOLLA	25	0	1	1	24
Md. FARUK	25	0	1	1	25
Md. TOHIDUL ISLAM	25	0	1	1	25
Md. MASUD PARVEZ	25	0	1	1	25
Md. SAGAR	25	0	1	1	25
Md. IQBAL HOSSAIN	25	0	1	1	25
Mr.Mostafizur Rahman	25	0	1	1	25
Walid Hossain	25	0	1	1	25
Md.Kamal Hossain	25	0	1	1	25
Md. A.K.M Aftab Uddin	25	0	1	1	25
Altaf Hossain	25	0	1	1	25
Md.Almash Hossain	17	1	1	1	25
Md.Almash Hossain	8	0	1	1	25
Md.Monir Hossain	25	0	1	1	25
Mr.Harun-or-Rashid	25	0	1	1	25
Biplob Hossain Badal	25	0	1	1	25
LITON	25	0	1	1	25
Mrs.Mita	25	0	2	1	25
Md.Salim Howlader	6	1	1	1	25
Md.Salim Howlader	19	0	1	1	25
Mr.Nizam Uddin	25	0	1	1	25
Md.Abdul Halim Bhu.	25	0	1	1	25
Md.Biplob Hossain	25	0	1	1	25
Md. ANWAR HOSSAIN	25	0	1	1	25
Md.Khokon	25	0	1	1	25
Md.Abu Hanif	25	0	1	1	25
Md.Bahadur	25	0	1	1	25
MD. BAZLUR RASHID	25	0	1	1	25
FARUK HOSSAIN	25	0	1	1	25

Md.Jahid Hasan	25	0	1	1	25
Nur Islam	25	0	1	1	25
Rahul Biswas	25	0	1	1	25
Md.Iqbal Kabir	25	0	1	1	25
Md.Hossain	25	0	1	1	26
Md.Helal Uddin	25	0	1	1	26
AYNAL	25	0	1	1	26
Abul Hossain Mukul	25	0	1	1	26
Md. Abdullahel hadi	21	1	1	1	26
Md. Abdullahel hadi	4	0	1	1	26
ABDUR RAHMAN	25	0	1	1	26
RASEL	25	0	1	1	26
Md. Delowar Hossain	25	0	1	1	26
Md. JAHIDUL ISLAM	25	0	1	1	26
KHADIZA BEGUM	25	0	2	1	26
Md. RIPON MOLLHA	25	0	1	1	26
Md. RIAZ	25	0	1	1	26
Md. SAHIN MOLLA	25	0	1	1	26
Md. ANOWAR HOSSAIN	25	0	1	1	26
Md. SOFIQUL ISLAM MOLLA	25	0	1	1	26
Md. RUHUL AMIN	19	1	1	1	26
Md. RUHUL AMIN	1	1	1	1	26
Md. RUHUL AMIN	1	1	1	1	26
Md. RUHUL AMIN	1	1	1	1	26
Md. RUHUL AMIN	3	0	1	1	27
Syed Arif Ullah	25	0	1	1	27
Mr.Monirul Islam	25	0	1	1	27
Md. Motibul Haque	25	0	1	1	27
Md.Shahin Miah	25	0	1	1	27
Md.Bulbul Sikdar	25	0	1	1	27
Md. Rakiul Islam	25	0	1	1	27
Mr.Somendra	25	0	1	1	27
Md.Musa Miah	25	0	1	1	27
Md.Farhad Hossain Akondh	25	0	1	1	27
Md. Azizul Islam	25	0	1	1	27
Md. Mahbubur Rahman	25	0	1	1	27
Md. Anwar Hossain	25	0	1	1	27
Md.Shafi Ullah	25	0	1	1	27
S.M.A Kader	3	1	1	1	27
S.M.A Kader	22	0	1	1	27
Md. Md.Mahmudul Hasan	25	0	1	1	27

Md. Nazrul Islam Mirbahar Mirbahar	25	0	1	1	27
Md.Saidul Islam	25	0	1	1	27
Md. SAROWAR HOSSAIN	11	1	1	1	27
Md. SAROWAR HOSSAIN	1	1	1	1	27
Md. SAROWAR HOSSAIN	13	0	1	1	27
Md. ABBAS UDDIN	9	1	1	1	27
Md. ABBAS UDDIN	14	0	1	1	27
TAUHIDUL ISLAM	25	0	1	1	27
AMINUL	25	0	1	1	27
Md. SHAFIQUL ISLAM	25	0	1	1	27
Md.Saiful Islam	25	0	1	1	27
Md. SALIMUL HAQ HAQ	25	0	1	1	27
Md.Khorshed Alam	25	0	1	1	27
ARIF HOSSAIN	25	0	1	1	27
Md. MURAD HOSSAIN	25	0	1	1	27
PROFULLO RAY	25	0	1	1	27
Md. Akhter uz Zaman	25	0	1	1	27
Md. Shakib Hossain	25	0	1	1	27
Md. JAHIDUL ISLAM	25	0	1	1	28
Md. MOHAMMAD SHAMIM AHMED	25	0	1	1	28
Md. Mohsin Hazari	25	0	1	1	28
Mr.Abdur Rajib	25	0	1	1	28
Mr. Sree Uttom Kumar	25	0	1	1	28
Md.Nazmul Haque	25	0	1	1	28
Mr. NONI GOPAL HAWLADER	25	0	1	1	28
S.M Yousuf	25	0	1	1	28
Md. Zakir Hossain	25	0	1	1	28
Md.Zahidul Islam	25	0	1	1	28
Golam Kader	7	1	1	1	28
Golam Kader	1	1	1	1	28
Golam Kader	17	0	1	1	28
LIAJUR RAHMAN	25	0	1	1	28
Md.Belal Hossain	25	0	1	1	28
S.M Akramul Haque	25	0	1	1	28
Md. Rafiqul Islam	25	0	1	1	28
Md. HAFIJUR RAHMAN	25	0	1	1	28
Abdul Mannan	25	0	1	1	28
Md. Nur Khoda	25	0	1	1	12
ALIM AL RAZI	25	0	1	1	12
Md. RASEL ALOM	25	0	1	1	12
Md. Okul Uddin	25	0	1	1	12

NEHAR RANJAN SAHA	25	0	1	1	12
Md.Khorsed Alam	25	0	1	1	12
Md.Nazim Uddin	25	0	1	1	12
Md.Dulal Miah	25	0	1	1	12
Abdur Rashid Mollah	25	0	1	1	14
Dudu Miah	25	0	1	1	14
Obaidur Rahman	25	0	1	1	14
Md.Rafiqul Islam	25	0	1	1	14
Lutfullahil Mahmud	22	1	1	1	14
Lutfullahil Mahmud	3	0	1	1	14
NASIR UDDIN	25	0	1	1	20
Md. Khorshed Alam	25	0	1	1	20
Md.Jalal	25	0	1	1	20
Md.Afzal Hossain	25	0	1	1	20
Md. Bellal Miah	25	0	1	1	20
Md.Nasir Mollah	25	0	1	1	20
Md. Monir Hossain	25	0	1	1	21
Md. Zakir Hossain	25	0	1	1	21
Md. Abdul Barek	25	0	1	1	21
Zahid Hasan	25	0	1	1	21
Md. Sabuz Alam	25	0	1	1	21
Md. Shiraj Biswas	25	0	1	1	21
Md. Ibrahim Miah	25	0	1	1	21
Md. House Mollah	25	0	1	1	21
Md.Shafiqul Islam	25	0	1	1	21
ABDUL HALIM	25	0	1	1	21
Md. TIPU BISWAS	25	0	1	1	21
Md. MIZANUR RAHMAN	25	0	1	1	21
Md. SABUJ SARKAR	25	0	1	1	21
Md. FAKHRUL ISLAM	25	0	1	1	21
Md. NASIR UDDIN	25	0	1	1	21
Md. Faisal Nadim	25	0	1	1	21
Mr.Abdur Rahim	25	0	1	1	21
Md.Roni Patowary	25	0	1	1	21
Shakya Kishore Chakma	25	0	1	1	21
Md.Riajul Hossain	25	0	1	1	21
Abu Siddik	25	0	1	1	21
Md. Kamrul Hasan	25	0	1	1	21
Md. Farhad Uddin	14	1	1	1	18
Md. Farhad Uddin	11	0	1	1	18
Mr. Kishor Kumar Dwari	5	1	1	1	18

Mr. Kishor Kumar Dwari	20	0	1	1	18
Md. Mohi Uddin	25	0	1	1	18
Md. Tanvir Ahmmed	16	1	1	1	18
Md. Tanvir Ahmmed	9	0	1	1	18
Md. Masud Rana	25	0	1	1	18
Md. Shohidul Islam	12	1	1	1	18
Md. Shohidul Islam	13	0	1	1	18
Mr. Subrata Paul	20	1	1	1	18
Mr. Subrata Paul	5	0	1	1	18
Md. Sayed Anamul Hoque	25	0	1	1	18
Md. Kamal Uddin	25	0	1	1	18
Md. MOKHLESAR RAHMAN	25	0	1	1	18
Md. Zakaria	25	0	1	1	18
Md. Aminul Islam	25	0	1	1	18
Ahsan Nasid	25	0	1	1	20
Md. ROBIUL ISLAM	25	0	1	1	20
Md. HANIF	25	0	1	1	20
Md. ABDUR RASID	7	1	1	1	20
Md. ABDUR RASID	18	0	1	1	20
Md. ZAHANGIR ALOM	25	0	1	1	20
MUHAMMAD FARUQ CHOWDHURY	25	0	1	1	20
Mr. Mohd. Shahidullah	25	0	1	2	29
Mrs. Umme Aasma Rahman	25	0	2	2	29
Mr. Firoz Ahmed Faruk	25	0	1	2	29
Mr. Golam Mohammad Tarique	25	0	1	2	29
Mr. Md. Moshiur Rahman	12	1	1	2	29
Mr. Md. Moshiur Rahman	1	1	1	2	29
Mr. Md. Moshiur Rahman	12	0	1	2	29
Mr. Md. Mohsin Mian	25	0	1	2	29
Mr. Iqbal Ahmed (Nayon)	25	0	1	2	29
Ms. Ammatul Amina Hussain	25	0	2	2	29
Mr. Mirza Zahid Hassan	25	0	1	2	29
Mr. Mahbub Ali Faisal	25	0	1	2	29
Mr. Mohammad Shaiful Islam	25	0	1	2	29
Ms. Shahana Pervin	25	0	2	2	29
Ms. Qurratul Ayn	25	0	2	2	29
Mr. Kazi Rabiul Karim	25	0	1	2	29
Ms. Sadia Jahan	25	0	2	2	29
Ms. Prianka Zaman	25	0	2	2	29
Mrs. Farzana Faruque Naz	25	0	2	2	29
Mr. K. Rashedur Rahman	25	0	1	2	29

Mrs. Laila Arjuman Banu	25	0	2	2	29
Mr. Shahrier Amin	25	0	1	2	29
Mr. Abu Hena Mostafa Rahman	25	0	1	2	29
Ms. Bushra Jabin	25	0	2	2	29
Mr. Rashed Ahmed	25	0	1	2	29
Mr. Md. Ashikur Rahman	25	0	1	2	29
Mr. Md. Abdul Wahid	25	0	1	2	29
Mr. Md. Shafiul Azam	25	0	1	2	29
Ms. Faria Islam	25	0	2	2	29
Mr. Mizanur Rahman	25	0	1	2	29
Mr. Khondker Quadir Hossain	25	0	1	2	29
Ms. Syeda Shahrazad Rahman	25	0	2	2	29
Mr. Junaidul Haque	25	0	1	2	29
Mr. Shyamal K. Chakrabortty	25	0	1	2	29
Mr. Md. Akhter Hamid	25	0	1	2	29
Mr. Upen Mitra	25	0	1	2	29
Mr. Md. Nazrul Islam	25	0	1	2	29
Mr. A.N.M. Shaiful Hasan Khan	25	0	1	2	29
Mr. Md. Insan Ali	25	0	1	2	29
Mr. Mohammed Anisur Rahman	25	0	1	2	29
Mr. Ratan Kumar Nath	25	0	1	2	29
Mr. Mohd. Mahbub-ul-Hoque	25	0	1	2	29
Mr. Ahmed Hussain Mansoor	25	0	1	2	29
Mr. Mobarok Hossain	6	1	1	2	29
Mr. Mobarok Hossain	1	1	1	2	29
Mr. Mobarok Hossain	18	0	1	2	29
Mr. Md. Harun-Or-Rashid	25	0	1	2	29
Ms. Shaila Arjuman Banu	25	0	2	2	29
Mr. Mohd. Emran Hossain	25	0	1	2	29
Ms. Hosne Ara	25	0	2	2	29
Mr. Mohd. Malik Masud Khan	25	0	1	2	29
Ms. Khadiza Akter Pushpo	25	0	2	2	29
Mr. Md. Rafiqul Islam	25	0	1	2	29
Mr. Md. Abdul Haque	25	0	1	2	29
Mr. K.M. Misbahul Alam	25	0	1	2	29
Mr. Pradip K. Das	25	0	1	2	29
Mr. Ahmed Jamal	25	0	1	2	29
Mr. Enamul Haque Bhuiya	25	0	1	2	29
Mr. Md. Abul Khair	25	0	1	2	29
Mr. Nadim Ahmed	25	0	1	2	29
Mr. Kamruzzaman	25	0	1	2	29

Mr. Shams Imran	25	0	1	2	29
Mr. Md. Mobarak Hossain	25	0	1	2	29
Mr. Mahmood Hasan	25	0	1	2	29
Mr. Sudipta Paul	25	0	1	2	29
Mr. Jashim Uddin	25	0	1	2	29
Mr. Kishor Kumar Chandra	25	0	1	2	29
Mr. Shakawat Hossain	25	0	1	2	29
Mr. Mohd. Osman Ghani	25	0	1	2	29
Mr. Mohd. Faruque Ahmed	25	0	1	2	29
Mr. Mohd. Bahauddin Ahmed (Sujan)	25	0	1	2	29
Mr. Md. Mizanur Rahman	25	0	1	2	29
Mr. Mohd. Mustafa Kamal	25	0	1	2	29
Mr. Md. Serajul Islam	25	0	1	2	29
Mr. Md. Rezaul Karim	25	0	1	2	29
Mr. Mohd. Mofazzal Hossain	25	0	1	2	29
Mr. Ranjit Chandra Mallik	25	0	1	2	29
Mr. Md. Sayeduzzaman	25	0	1	2	29
Mr. Md. Kuddusur Rahman Tahim	25	0	1	2	29
Mr. Md. Mahbubur Rahman	25	0	1	2	29
Mr. Jaher Mia	5	1	1	2	29
Mr. Jaher Mia	20	0	1	2	29
Mr. Ratan Majumder	25	0	1	2	29
Mr. Goljar Hossain	25	0	1	2	29
Mr. Satendra Chandra Mallik	25	0	1	2	29
Mr. Dawan Jarif	25	0	1	2	29
Mr. Md. Abul Kalam Azad	25	0	1	2	29
Mr. Shakil Ahmed	25	0	1	2	29
Mr. Nafiz Imran	25	0	1	2	29
Mr. Mushfiqur Rahman	25	0	1	2	29
Mr. Aga Ekram Chowdhury	25	0	1	2	29
Mr. Shaker Ahmed Khan Chow.	25	0	1	2	29
Mr. Mohammed Rakibul Islam	25	0	1	2	29
Mr. Mirza Khalid Hasan	25	0	1	2	29
Mr. Md. Nazmul Haider	25	0	1	2	29
Mr. Md. Abul Basar	25	0	1	2	29
Mr. Md. Rafiqul Huq	25	0	1	2	29
Mr. Mukul Chicham	25	0	1	2	29
Mr. Md. Salah Uddin Bhuiyan	25	0	1	2	29
Mr. Md. Jamal Uddin	25	0	1	2	29
Mr. Ripon Miah	25	0	1	2	29
Mr. Md. Abdul Malek	25	0	1	2	29

Mr. Liton Bazi	25	0	1	2	29
Mr. F.M. Abdullah	25	0	1	2	29
Mr. Md. Riyazuddin	25	0	1	2	29
Mr. Mohd. Omar Faruq	25	0	1	2	29
Mr. Kanan Barua	25	0	1	2	29
Mr. Mikhail Snal	25	0	1	2	29
Mr. Bishnu Roy Manik	25	0	1	2	29
Mr. Md. Mokhlesur Rahman	25	0	1	2	29
Farhana Yesmin	14	1	2	2	29
Farhana Yesmin	1	1	2	2	29
Farhana Yesmin	1	1	2	2	29
Farhana Yesmin	9	0	2	2	29
Rezwana Shahid	25	0	2	2	29
Ahmed Riyadh Jamal	25	0	1	2	29
Ishrat Jahan	25	0	2	2	29
Habibur Rahman	25	0	1	2	29
Kamruzzaman Sardar	25	0	1	2	29
Sheikh Shoel Imtiaz	25	0	1	2	29
Farhana Ahmed	25	0	1	2	29
Sayeem Hossain	25	0	1	2	29
Muhammad Yusuf Mehedee	25	0	1	2	29
Mohd. Masudur Rahman	25	0	1	2	29
Md. Nurul Abrar Sadi	25	0	1	2	29
Nowrin Nawaz	21	1	2	2	29
Nowrin Nawaz	4	0	2	2	29
Rezina Parveen Reema	25	0	2	2	29
Salauddin Sarker	25	0	1	2	29
Mr. Mohd. Kamal Hossain	25	0	1	2	29
Mr. Nawshad Nur Newaz Khan Lodhi	25	0	1	2	29

Table A1 : Original Data Collected

APPENDIX B

R codes are given below:

```
library(splines)
library(survival)
library(KMsurv)


data1<-read.csv(file.choose(),header=TRUE)
attach(data1)

#Kaplan-Meier on the Time and Status for every employee
my.surv1<-survfit(Surv(Time,Status)~1,conf.int=0.95)

my.surv1


#Kaplan-Meier on Sex and Status
my.surv2<-survfit(Surv(Sex,Status)~1,conf.int=0.95)

my.surv2

#Kaplan-Meier on Firm Size and Status
my.surv3<-survfit(Surv(Firm.Size)~1,conf.int=0.95)
```

```
my.surv3


#Kaplan-Meier on Working Condition and Status
my.surv4<-survfit(Surv(Working.Condition,Status)~1,conf.int=0.95)

my.surv4

#Weibull on Time and Status
p.surv1=survreg(Surv(Time,Status)~1,data1,dist='weibull',scale=0)

p.surv1


#Weibull on Sex and Status
p.surv2=survreg(Surv(Sex,Status)~1,data1,dist='weibull',scale=0)

p.surv2


#Weibull on Firm Size and Status
p.surv3=survreg(Surv(Firm.Size,Status)~1,data1,dist='weibull',scale=0)

p.surv3
```

```
#Weibull on Working Condition and Status

p.surv4=survreg(Surv(Working.Condition,Status)~1,data1,dist='weibull',scale=0)


p.surv4



#Cox proportional hazards model bearing all the covariates

c.surv<-Surv(Time,Status)

coxph(c.surv~Sex+as.factor(Firm.Size)+as.factor(Working.Condition),method='breslow')



##############################################################################
###########

##############################################################################
###########


#weibull proportional hazards model bearing all the covariates



WeiPHM = survreg(Surv(Time,Status)~Sex+Firm.Size+Working.Condition,dist="weibull",
data=data1)

summary(WeiPHM)


#Cox proportional hazards model bearing all the covariates
```

```
CoxPHM = coxph(Surv(Time,Status)~Sex+Firm.Size+Working.Condition, data=data1)

summary(CoxPHM)




################################################################################

######## Now divide the dataset based on sex, firm size and working condition and find out their

### effect. From the values, try to find out the influencial variables..


########### Then identify the survival plots and find out which survival plot is telling which story


# Sample survival curves for models

plot(survfit(CoxPHM), main="CI",xlab="Time (Day)", ylab="Survival Probability" ,lwd=2,conf=F)

legend("topright", bty="n",c("CoxPHM","WeiPHM"), col=c(1,2), lwd=c(2,2))



kap=WeiPHM$scale

lada=exp(-(1/kap)*(WeiPHM$coef[1]+WeiPHM$coef[2]+WeiPHM$coef[3]+WeiPHM$coef[4]))^kap

ze=seq(0,30,1)

W1=exp(-(lada*ze^kap))

lines(W1,col=2, main="CI",xlab="Time (Day)", ylab="Survival Probability" ,lwd=2)


### Then devide the database into trainging and testing
```

```
# Sample splitting of data randomly, 30% testing data


r.N=sample(1:nrow(data1), 0.3*nrow(data1), replace=F)# randomly sample 30% of the data to be the test data set

train.N=data1[-r.N,] # train data set

test.N=data1[r.N,] # test data set

summary(train.N)

nrow(train.N)

summary(test.N)

nrow(test.N)


model<-coxph(Surv(Time,Status)~Sex+Firm.Size+Working.Condition, data=train.N)

y<-predict(model,response="TRUE",data=test.N)
```