University of Rajshahi **Rajshahi-6205** **Bangladesh.**

RUCL Institutional Repository **http://rulrepository.ru.ac.bd**

Department of Statistics MPhil Thesis

2002

Modelling of Diabetes Mellitus Data

Sultana, Mst. Papia

University of Rajshahi

http://rulrepository.ru.ac.bd/handle/123456789/127

Copyright to the University of Rajshahi. All rights reserved. Downloaded from RUCL Institutional Repository.

MODELLING OF DIABETES MELLITUS DATA

A Dissertation

Submitted to the University of Rajshahi

in Fulfilment of the Requirements for the Degree of

Master of Philosophy

in Statistics

BY

MST. PAPIA SULTANA

DEPARTMENT OF STATISTICS
UNIVERSITY OF RAJSHAHI
RAJSHAHI, BANGLADESH

AUGUST, 2002

Dedicated

To

My Father

DECLARATION

This thesis entitled "**Modelling of Diabetes Mellitus Data**" submitted by me in the Department of Statistics, University of Rajshahi for the award of the degree of Master of Philosophy is based on my research work carried out under the supervision of **Dr. M. A. Basher Mian**, Professor, Department of Statistics, University of Rajshahi.

To the best of my knowledge, this work neither in part nor in full has been submitted to any other University or Institution for the award of any degree.

Mian M.A. Basher

(Dr. M. A. Basher Mian)
Supervisor

Mst. Papia Sultana
20.08.02

(Mst. Papia Sultana)
Fellow

20.08.02

(Dr. M. Asaduzzaman Shah)
Chairmen
Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh.

ACKNOWLEDGEMENT

First and foremost thanks are to Almighty Allah for giving me strength, courage and ability to accomplish this study.

I would like to express my sincere and heartiest gratitude to my reverent teacher and research supervisor, ***Dr. M. A. Basher Mian****, Professor, Department of statistics, University of Rajshahi for his untiring patience, constant inspiration and constructive guidance throughout the progress of this research work.*

I am very grateful to my honorable teacher, Late Professor K. M. Hossain who was the founder of the Department of Statistics, University of Rajshahi for his continuous encouragement.

I am very much grateful to my honorable teacher Dr. M. Asaduzzaman Shah, Chairmen, Department of statistics, and also to Professor Golam Mostafa, Professor Samad Abedin, Professor M. A. Razzaque, Dr. Nurul Islam, Dr. Sayedur Rahman and all other respective teacher, Department of statistics, University of Rajshahi for their advice and encouragement.

I also acknowledge with thanks the co-operation of the Officers and staff of Department of statistics and the officers and staff of the Computer Unit of this Department.

I would also like to thanks to all the doctors and staffs of Laxmipur Diabetic Center & Talaimeri Diabetic Center, especially to Doctor F. M. Abu Zahid and the staffs DAB for giving informations and journals concerning to diabetes mellitus and for helping to collect data.

Finally I express my regards to my family members, especially to my husband M. Fazlul Haque, Assistant Professor, Department of Islamic History & Culture, University of Rajshahi for their inspiration and encouragement throughout this study.

Mst. Papia Sultana

Abbreviations

AIC	= Akaike's Information Criterion.
BGL	= Blood Glucose Level.
BIC	= Bayesian Information Criterion .
BIRDEM	= Bangladesh Institute of Reserch and Rehabilitation in Diabetics, Endocrine and Metabolic Disorders.
BMI	= Body Mass Index.
BP	= Blood Pressure.
CAD	= Coronary Artery Disease .
CHF	= Cumulative Hazard Function.
CVD	= Cerebro-Vascular Disease.
DAB	= Diabetes Association of Bangladesh.
DM	= Diabetes Mellitus.
GEE	= Generalised Estimating Equations.
GTT	= Glucose Tolerance Test.
HGO	= Hepatic Glucose Output.
IDDM	= Insulin Dependent Diabetes Mellitus.
IGT	= Impaired Glucose Tolerance.
LRT	= Likelihood Ratio Test.
LSE	= Least Square Estimate.
LVD	= Large Vessel Disease.
MC	= Markov Chain.
MCG	= Metabolic Clearance of Glucose.
MLE	= Maximum Likelihood Estimate.
MTD	= Mixture Transition Distribution.
NIDDM	= Non-Insulin Dependent Diabetes Mellitus.
OGTT	= Oral Glucose Tolerance Test.
OHA	= Oral Hypoglycemic Agents.
OLS	= Ordinary Least Square.
OLSE	= Ordinary Least Square Estimate.
OR	= Odds Ratio.
PVD	= Periferal Vascular Disease.
QD	= Quartile Deviation.
SD	= Standard Deviation.
SVD	= Small Vessel Disease.
WHO	= World Health Organization.

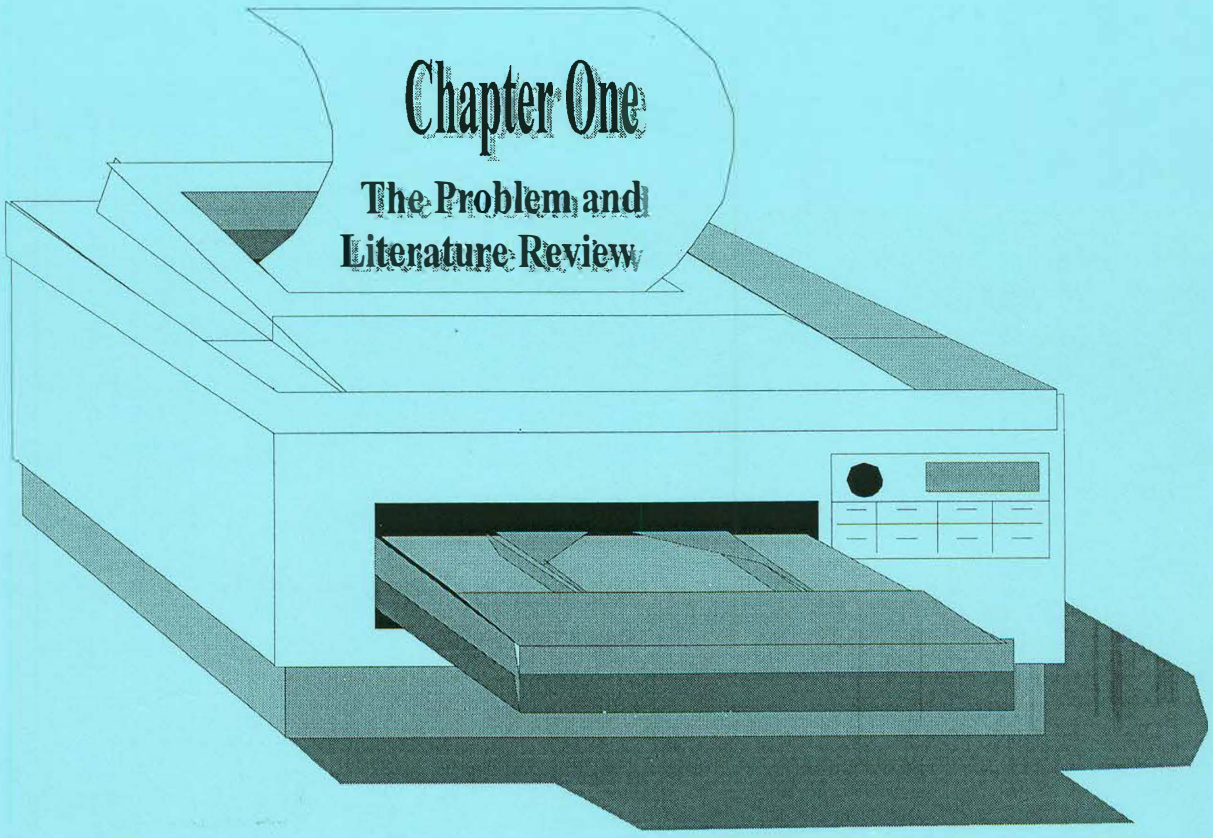
List of Tables

List of Figures

CONTENTS

Chapter One

The Problem and Literature Review



Chapter One

The Problem and Literature Review

1.1 Introduction:

Diabetes afflicts a large number of people of all social conditions throughout the world. So, it is a major health problem of all the countries of the world including Bangladesh. Inspite of increasing advances in the past several years in almost every field of diabetes research and patient care, the personal and public health problem of diabetes, already of vast proportion, continuously increasing day by day. The scale of the problem that diabetes poses to world health is still widely under recognized. Recent estimates predict that if current trends continue the number of persons with diabetes will be more than double, from 140 million to 300 million in the next 25 years. The greater proportion of the increase is likely to occur in the developing countries, which are the communities who can least afford it [Source: http://www.who.int/ncd/dia/index.htm]. Bangladesh also

possess the same picture. Surveys conducted by DAB (1995) reveals that the number of diabetic patients in Bangladesh is 1 to 1.5 percent of the total population. The rate is higher in the urban and industrial areas, perhaps the cause is to adequate physical work of rural areas.

1.2 A Review on Diabetes Mellitus:

Diabetes (people also known this as diabetes mellitus) is a metabolic disease, which is not curable but can be controlled. According to Dr. Latif (1993) diabetes mellitus is not a disease but a heterogeneous group of syndromes. Although knowledge about diabetes is growing, it is difficult to give a satisfactory definition of this disorder. According to Cahil (1985) a generalization is that it is a grouping of anatomic and chemical problems resulting from a number of factors which creates an absolute or relative deficiency of insulin or its function. Also Soracha Mc. Ginnis (May 3, 2002) has said that Diabetes Mellitus is caused by the destruction of insulin-producing beta cells in the pancreas that occurs when the immune system mistakenly attacks the beta cells. Simply we can say that it is a disorder of the chronic system caused by insufficient supply of insulin or inadequate functioning of insulin, which may result from many environmental and genetic factors, most often the two together. As a result, sugar intake by the patients is not metabolized perfectly and the level of the blood sugar increases consequently and also the excessive sugar comes out with urine and sufferers have complications that affect vital organs, including the kidneys, eyes and heart.

The symptoms of the disease are – severe thirst, frequent weakness, weight loss, excessive hunger, profuse urination etc. Sometimes symptoms are totally absent. The exact cause of diabetes is not diagnosed yet perfectly. According to WHO the possible causes of diabetes are – heredity, excessive over weight, lack of physical work, contagious disease, which damages pancreas, obesity, malnutrition etc.

The diagnostic tools of diabetes mellitus are not so rich, rather only blood glucose and urine sugar estimation establishes the diagnosis of the disease. The tools of diagnosis suggested by Macheod (1984) are –

- ❑ Estimation of urine sugar.
- ❑ Estimation of blood sugar and corresponding urine sugar.
- ❑ Oral Glucose Tolerance Test (OGTT).

According to WHO (1985) the diagnosis interpretation of OGTT response is given in the following table:

Table#1.2A: Diagnostic values of OGTT.

Glucose concentration in mmol/l (mg/dl)				
	Whole Blood		Plasma	
	Venous	Capillary	Venous	Capillary
Diabetic Mellitus				
Fasting value	≥6.7 (≥120)	≥6.7 (≥120)	≥7.8 (≥140)	≥7.8 (≥140)
2hrs after glucose load	≥10.0 (≥180)	≥11.1 (≥200)	≥11.1 (≥200)	≥12.2 (≥200)
Impaired Glucose Tolerance				
Fasting value	<6.7 (<120)	<6.7 (<120)	<7.8 (<140)	<7.8 (<140)
2hrs after glucose load	6.7-10.0 (120-180)	6.7-10.0 (120-180)	7.8-11.1 (140-200)	8.9-12.2 (160-200)

[WHO (1985)]

Estimation of blood glucose and urine sugar provide information about the disease, but many complex situations of diabetes mellitus require sophisticated technical maneuver to diagnose. An Oral Glucose Tolerance Test (OGTT) is usually used to establish the status of diabetes mellitus. In this diagnostic test, the blood glucose is measured after a overnight fasting and at two hours after a 75 gms glucose load. For epidemiological studies, the two hours value of blood glucose after a 75 gms glucose load may alone be adequate for diagnosis.

Individuals developing diabetes may pass through the state of Impaired Glucose Tolerance (IGT), which can be defined as the glycaemic response

to a standard glucose challege intermediate between normal stage and diabetic stage. That is, IGT is used to describe the mild degree of glucose tolerance, which can be termed as borderline diabetes, prediabetes, and chemical diabetes. If a patient be alert in this mild stage of diabetes and take care, the picture of the disease could be different. But in Bangladesh we have limited technical assistance to diagnose prestage.

According to WHO the diabetes mellitus can be classified in the following types-

Type-I : Insulin Dependent Diabetes Mellitus (IDDM),
Type-II : Non Insulin Dependent Diabetes Mellitus (NIDDM),
Type-III : Malnutrition-Related Diabetes Mellitus (MRDM),

But Dr. Latif (1993) has said that at present, perhaps diabetes mellitus can be discussed in four groups –

(a) Insulin Dependent Diabetes Mellitus (IDDM),
(b) Non Insulin Dependent Diabetes Mellitus (NIDDM),
(c) Maturity Onset Diabetes in Young (MODY),
(d) Malnutrition-Related Diabetes Mellitus (MRDM).

Causes of Diabetes Mellitus:

Health experts cited, "lifestyle changes and urbanization" as the root causes of the current diabetes crisis. Professor Zimmet has said, "Young people in particular are leading sedentary lifestyles. They are watching a lot of television, playing computer games and turning away from traditional diets in favour of fast foods that are high in carbohydrates and fat." Whereas we know that high cholesterol level, high blood pressure,

smoking and obesity are known to contribute to diabetes (high blood glucose level) which can lead to thickening of the arteries, heart disease and strokes. Type-I diabetes occurs when insulin producing cells in the pancreas are damaged. Type-II diabetes is caused by the body not producing sufficient insulin and what little is produced can be prevented from being used effectively by an individual being overweight.

To Manage the Diabetes Mellitus the aims are –

1. Relieve of Symptoms,
2. Reduction and Maintenance of Standard body weight,
3. Maintenance of englycaemia as far possible,
4. Avoidance of complications both treatment (hypoglycaemia, lactic acidosis) and disease related,
5. Assist in psychological adjustment to improve quality and duration of life, etc.

To attain the above goal we have the following modalities in hand ;

(I) Diet Control,

(II) Exercise,

(III) Drugs –

(a) Oral Hypoglycaemic Agents (OHA),

(b) Insulin, etc.

1.3 Literature Review:

Muenz and Rubinstein (1985) considered heterogeneous group of individuals who were followed over time and at each time point every individual may be in state non-disease (zero) or disease (one). They

assumed that the sequence of states follows a binary Markov Chain and the transition matrix of the chain as

$$M = \begin{bmatrix} \gamma_{00} & 1-\gamma_{00} \\ \gamma_{10} & 1-\gamma_{10} \end{bmatrix}$$

where γ_{00} is the 0 to 0 transition probability and γ_{10} is the 1 to 0 transition probability. They proposed to consider that the transition probabilities γ_{00} and γ_{10} be modeled by two logistic regression as-

$$\gamma_{00}(bx) = \frac{\exp(bx)}{1+\exp(bx)}$$

$$\text{and } \gamma_{10}(dx) = \frac{\exp(dx)}{1+\exp(dx)}$$

where x is the vector of covariates and for the q-th person in the study is equal to $X_q = (1, X_{q1}, \ldots\ldots\ldots\ldots, X_{qp})$. $b = (b_0, b_1, \ldots\ldots\ldots, b_p)$ and $d = (d_0, d_1, \ldots\ldots\ldots, d_p)$ are the parameter vector.

Considering P covariates including the intercept there are 2(P+1) parameters, which are estimated by the maximum likelihood method of estimation.

They also gave an extension of the basic model which allows time dependent covariates and non-stationarey or second-order Markov Chains. But they did not say anything about higher order Markov Chain.

Mallick (1994) worked with the data of diabetes mellitus where he applied the model suggested by Muenz and Rubinstein (1985). But the finding of his study was that the disease process of diabetes does not follow a first order Markov Chain.

Raftary (1994) suggested a model for Makov Chain of order higher than one for the analysis of repeated ordinal data from a disease process which involve only one parameter for each extra log variable. He considered an l-th ($l>1$) order Markov Chain $\{X_t; t = 1,2,.....\}$ on finite set of m states $\{1,2,.....,m\}$ and the transition probabilities are

$$P(i_0|i_1,i_2,.....,i_l) = \Pr\{X_{t+1} = i_0|X_{t+l-1} = i_1,.....,X_t = i_l\};\ t = 1,2,..... \quad ------(1)$$

For $l>1$, the model provides a useful parameter reduction in equation (1) by supposing that

$$P(i_0|i_1,i_2,.....,i_l) = \sum_{j=1}^{l} X_j q(i_0|i_j)$$

where $Q = \{q(i|j)\}$ is a column stochastic matrix satisfying $q(i|j) \geq 0$ and $\sum_{r=1}^{m} q(r|j) = 1, j = 1,2,.....,m$ and $\lambda_1 + \lambda_2 + \cdots\cdots + \lambda_l = 1$ $(X_j \geq 0, j = 1,2,.....,l)$

Here to be noted that he assumed only the transition of present state given all state of different past time points was a linear combination of the contribution from each of the past state. i.e. he did not consider any effect of intermediate state. If such effect is not considered some information will be loosed and accordingly.

Hossain and Islam (1997) fitted a second order Markov model suggested by Raftery (1994) with an extension that first they did not consider the effect of intermediate state at time $t+1$ and next they considered the effect of intermediate state to diabetes mellitus data collected by BIRDEM (Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic disorders) throughout the period 1984-1994 (421 registered individuals). To consider the effect of intermediate state they used Chapman-Kolmogorov equation, where they considered a transition

state E_j to E_k in exactly n steps which occur via different paths $E_j \to E_{j_1} \to E_{j_2} \to \cdots\cdots \to E_{j_{n-1}} \to E_k$. The conditional probability that the system passes through this particular path given that it is at E_j is

$$P_{jj_1} \times P_{j_1 j_1} \times \cdots\cdots \times P_{j_{n-1}k}$$

where $P_{i_r i_s} = \Pr\{X_{t+1} = j_j | X_t = j_i\}$. The sum of the corresponding expression for all possible paths is the probability of finding the system at time $r+n$ in state E_k, given that at time r it was in state E_j. i.e.

$$P_{jk}^{(r+n)} = \sum_{j_i} P_{jj_i}^{(r)} \times P_{j_i k}^{(n)}$$

From the findings they concluded that the probability of the disease process of staying in the same state is high and that of changing the state is low. They also concluded that the state before the immediate past state has considerable contribution in describing the present state behavior of the disease through the immediate past state has much more contribution in performing the same.

Also to identify a Markov chain model of diabetes data of appropriate order they used Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) methods and found that a second order Markov model is reasonable to fit on the diabetes mellitus data.

Ferdousei and Islam (2000) proposed a second order Markov Chain for analyzing covariate dependence pattern of longitudinal data with an application to diabetes mellitus. For the analyze they used the Mixture Transition Distribution (MTD) model by Raftary (1985) for a time homogeneous second order Markov Chain $\{X_t; t = 1,2,.....\}$ modeled by the occupation probability:

$$P(i_0|i_1,i_2) = P[X_{t+1} = i_0|X_t = i_1, X_{t-1} = i_2]; t = 1,2,.....$$
$$= \lambda\gamma_{i_1 i_0} + (1-\lambda)\gamma_{i_2 i_0} \; ; \; i_0, i_1, i_2 = 0,1$$

where $\gamma_{i_1 i_0}$ and $\gamma_{i_2 i_0}$ are the lag variables involving two additional parameters λ and $(1-\lambda)$ respectively, where $\gamma_{i_1 i_0} = q_{i_1 i_0}|i_j$ and $\gamma_{i_2 i_0} = q_{i_2 i_0}|i_j \; ; i_j = 0,1$. The transition probability matrix for this second order Markov Chain at each time point is a two by two matrix

$$Q = \begin{bmatrix} q_{00}|i_j & q_{01}|i_j \\ q_{10}|i_j & q_{11}|i_j \end{bmatrix}, \; i_j = 0,1$$
$$= \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix}$$
$$= \begin{bmatrix} q_{00} & 1-q_{00} \\ q_{10} & 1-q_{10} \end{bmatrix}$$

where $q_{00} = q_{00}|i_j$ is the $0 \to 0$ transition probability and $q_{10} = q_{10}|i_j$ is the $1 \to 0$ transition probability for a second order MC satisfying

$$q_{1i} \geq 0 \; and \; q_{10} = 1 - q_{j1} \; ; \sum \lambda_j = 1$$

In order to relate covariates to the transitions of each state as functions of covariates, the transition probabilities at each time point are modeled by two logistic regression models:

$$q_{00} = \frac{\exp(b'x)}{1+\exp(b'x)}$$

and $$q_{10} = \frac{\exp(d'x)}{1+\exp(d'x)}$$

where X is the vector of covariates and for the i-th person in the study is equal to $X_i = (1, X_{i1}, \ldots\ldots\ldots\ldots, X_{ip})$; $i = 1,2,\ldots\ldots\ldots, n$. $b = (b_0, b_1, \ldots\ldots\ldots, b_p)'$ and $d = (d_0, d_1, \ldots\ldots\ldots, d_p)'$ are the parameter vector for the models.

They estimated the parameters by minimizing the likelihood function:

$$L = \prod_{i=1}^{n} \prod_{i_0, i_1, i_2 = 0}^{1} P(i_0 | i_1, i_2)^{n_{i_2 i_1 i_0 i}}$$

For inference they applied the statistical tools as:

1. Estimating the appropriate order of the fitted MC by AIC and BIC procedure:

   Assumption: The MC is Ergodic.

   H_k : The chain is dependent on k previous occurences, against

   H_m : The chain is dependent on m previous occurences, when k<m

   LR test statistic: ${}_k\eta_m = -2\log\lambda_{k,m}$

   Estimates:

$$AIC\left(\hat{k}_{AIC}\right) = \min_{0 \le k \le m} AIC(k) \text{ where } AIC(k) = {}_k\eta_m - 2(S^m - S^k)(S-1)$$

$$BIC\left(\hat{k}_{AIC}\right) = \min_{0 \le k \le m} BIC(k) \text{ where } BIC(k) = {}_k\eta_m - (S^m - S^k)(S-1)\log_e n$$

2. Testing the time homogegeity of the second order Markov Chain:

$$H_0 : P_{ijk}(t) = P_{ijk} \quad \text{for} \quad i, j, k = 0,1; \quad t = 0,1,\ldots\ldots\ldots, T$$

Test statistic: $\chi^2 = \sum_{t=1}^{T}\sum_{i=1}^{S} n_{ij}(t-1)\frac{\hat{P}_{ijk}(t) - \hat{P}_{ijk}}{\hat{P}_{ijk}} \sim s(s-1)(t-1)$

3. LR test for testing significance of the model:

$$H_0 : b_0 = b_1 = \cdots\cdots = b_p = d_0 = d_1 = \cdots\cdots = d_p = 0$$

Rajshahi University Library
Documentation Section
Document No D-2175
Date...22/6/03.....

$$\text{Test statistic: } \Lambda = -2\log\frac{L\left(\tilde{b},\tilde{d}\right)}{L\left(\hat{b},\hat{d}\right)} \sim \chi^2_{2p}$$

4. Wald test for testing individual parameters:

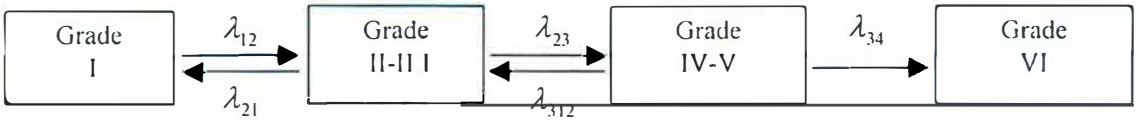
H_0 : *each coefficient is zero.*

$$\text{Test statistic: } Z = \frac{estimated\ coefficient}{S.E.(estimated\ coefficient)} \sim N(0.1)$$

For the application they used the data on diabetes mellitus obtained from BIRDEM from 1984 to 1994. They used the data of 2108 out of 5481 registered patients having symptom of diabetes mellitus at different time points during their follow-ups since registration. They also compared the obtained results from the proposed model with a first order MC. Lastly they concluded that their proposed second order Markov Chain appears more reasonable than Meunz and Rubinstein's first order Markov Chain for analyzing diabetes data. But in their study they also did not consider the effects of the intermediate states while estimating the occupation probabilities.

Marshall and Jones (1995) discussed the application of a multi-state model to diabetic retinopathy under assumption that a continuous time Markov process determines the transition times between disease stages. They considered a four-state Markov model for the analysis which include three transient disease states: grade I, grades II-III and grades IV-V of early retinopathy and one absorbing state grade VI. In their model the transient

states are ordered according to j $(j = 1,2,3,4)$ and instantaneous transition, represented by the intensities λ, can occur from state j to the adjoining states $j-1$ or $j+1$ as shown in the figure:



No direct transitions are allowed from an early stage of retinopathy to the absorbing stage of (except from the state IV-V) and if transitions like this occurred, the model assumes that unobserved transitions have occurred before the final transition.

Assuming that the underlying process is a Markov process, they represented the transition intensity matrix Λ as

$$\Lambda = \begin{bmatrix} -\lambda_{12} & \lambda_{12} & 0 & 0 \\ \lambda_{21} & -(\lambda_{21}+\lambda_{23}) & \lambda_{23} & 0 \\ 0 & \lambda_{32} & -(\lambda_{32}+\lambda_{34}) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

or the transition probability matrix $P(t)$. They established the relation between the transition probability matrix $P(t)$ and the transition intensity matrix Λ with the Kolmogorov forward differential equations

$$\frac{\partial P(t)}{\partial t} = P(t)\Lambda$$

where the (i,j)-th element of the matrix $P(t)$ represents the probability of a transition from state i to j in a time interval t, denoted as $P_{ij}(t)$. The solution of the system of equations is given by

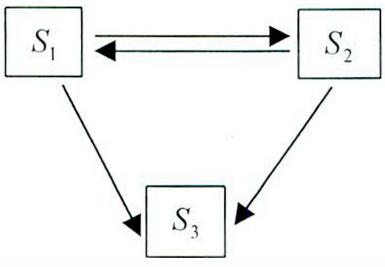
$$P(t) = \Lambda diag\{\exp(P_1 t), \exp(P_2 t), \ldots\ldots\ldots, \exp(P_k t)\}\Lambda^{-1}$$

They extended the model by introducing covariates as a proportional fraction in the baseline transition intensities λ's. They represented the regression for the (i,j)-th element of the transition intensity matrix Λ as

$$\lambda_i(z) = \lambda_{ij} \exp\left(\beta_{ij}' z\right)$$

where β_{ij} is the vector of regression co-efficients associated with the vector of covariates z for the transition between i and j. The resulting transition intensity matrix $\Lambda(z)$ can be used to compute the transition probability matrix $P(t|z)$. In the context of the multi-state Markov model they considered two types of model selection procedure – first is the selection of covariates associated significantly with the progression of the process and second, the selection of the association between each covariate and the disease process. They also mentioned that the maximum likelihood estimates for λ and β can be obtained by maximizing the likelihood function w.r.t. the parameters, where the full liklihood function is the product of all individual contributions. For the selection of the covariate the likelihood ratio test was suggested and Wald test for testing the null hypothesis.

Khan and Islam (1998) used an extension of Marshall and Jones's (1995) multi-state progression and regression models with an application to diabetic complication data. For the purpose they collected data on 200 patients out of 412 patients who were registered at different points in time in BIRDEM and he considered two intercommunicating state S1, indicating the IGT category and S2, indicating DM and one absorbing state S3, indicating dental complication as a result of DM. Thus the possible transition that can take place are $S_1 \rightarrow S_2$, $S_2 \rightarrow S_1$, $S_1 \rightarrow S_3$, and $S_2 \rightarrow S_3$, i.e.



Also two covariates were considered in their study: age and sex. From the obtaining results he observed that when age was considered as a time-independent covariate it was significantly related on the transitions IGT→DM, IGT→dental complication and DM→dental complication, and it has no significant effect on the transition DM→IGT. But when they considered age as time dependent covariate, they found that it was significantly related to all the transitions.

Albert and Waclawiw (1998) developed a Markov model of heterogeneous transitional data under the assumption that N subjects are followed over m equally spaced time points. Let an individual is observed for m occasions and suppose that there are N individuals. Thus the random vector of responses for the i-th individual is $m \times 1$ where the response variable is dichotomous. i.e.

$$Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{im} \end{bmatrix}, \quad \text{i=1,2,3,} \cdots N$$

Let N_{i00} = the number of transition for the i-th subject from state 0 to 0

N_{i01} = the number of transition for the i-th subject from state 0 to 1

N_{i10} = the number of transition for the i-th subject from state 1 to 0

N_{i11} = the number of transition for the i-th subject from state 1 to 1

Let N_i =($N_{i00}, N_{i01}, N_{i10}$) be a vector of the number of 0-0,0-1,and 1-0 transition with $\delta = E\ (N_i)$=$(\delta_{00}, \delta_{01}, \delta_{10})'$ .Therefore the transition matrix for i-th subject is given by

$$\begin{bmatrix} N_{i00} & N_{i01} \\ N_{i10} & N_{i11} \end{bmatrix}$$

Let $P_i = (P_{i01}, P_{i10})'$, where the i-th subject's transition matrix is

$$\begin{bmatrix} 1-P_{i01} & P_{i01} \\ P_{i10} & 1-P_{i10} \end{bmatrix}$$

where P_{i01} = the probability of having transition from 0 to 1 state

P_{i10} = the probability of having transition from 1 to 0 state

P_{i00} = the probability of having transition from 0 to 0 state

P_{i11} = the probability of having transition from 1 to 1 state

Let $E[P_{i01}] = \mu_{01}$, $E[P_{i10}] = \mu_{01}$, $V[P_{i01}] = \sigma_{01}^2$, $V[P_{i10}] = \sigma_{10}^2$ and $Cor(P_{i01}, P_{i10}) = \rho$.

The model allows for heterogeneous transitions and reduces to one in which all subjects follow the same Markov chain. The specification of heterogeneity distribution generalize the works by Meshkani and Billard (1992) and Cole et. al. (1995) by specifying only the first two moments in the random effect distribution and not intire distributional form.

Albert and Waclawiw proposed a Generalised Estimating Equations (GEE) approach to estimate the parameters of the suggested two-state quasi-likelihood model. Following quasi-likelohood approach, the GEE for $\mu = (\mu_{01}, \mu_{10})'$ and $\theta = (\sigma_{01}^2, \sigma_{10}^2, \rho)'$ are of the form

$$V(\mu) = \sum_{i=1}^{N} D' V^{-1} (N_i - \delta) = 0$$

$$\text{and } V(\theta)=\sum_{i=1}^{N} E'W^{-1}(S_i-\gamma)=0$$

where $\delta = E(N_i)$ and $\gamma = E(S_i)$.

They used a second order Taylor series expansions to approximate δ and γ. But in the suggested GEE approach of Albert and Waclawiw, correlation between the transitions in each subject was not taken into account.

Kabir and Islam (2000) also analysis with diabetes mellitus data. Their study was based on the data set of size 1453 collected at BIRDEM from patients registered in 1984. With the data they analyzed heterogeneous transitions using GEE for a two-state Markov model proposed by Albert and Waclawiw (1998). For the purpose four consecutive follow-up visits on each of the 1453 individuals were considered across a ten-year period (1984-1994). Estimating the parameters by applying GEE proposed by Albert and Waclawiw, they observed that a small amount of between subjects heterogeneity between the non-diabetic to confirmed diabetic transition existed, but large amount of that of confirmed to diabetic state. They also used GEE assuming the pairwise correlation as

$$(R_1(\alpha_1))_{st} = corr(N_{is}, N_{it}) = \alpha_{1st}, s \neq t$$

$$\text{and } (R_2(\alpha_2))_{st} = corr(S_{is}, S_{it}) = \alpha_{2st}, s \neq t$$

where $R_1(\alpha_1)$ and $R_2(\alpha_2)$ can be estimated by

$$\hat{R}_1(\alpha_1)=\frac{1}{n}\sum_{i=1}^{N} A_1^{-\frac{1}{2}}(N_i-\delta)(N_i-\delta)' A_1^{-\frac{1}{2}}$$

$$\text{and } \hat{R}_2(\alpha_2)=\frac{1}{n}\sum_{i=1}^{N} A_2^{-\frac{1}{2}}(S_i-\gamma)(S_i-\gamma)' A_2^{-\frac{1}{2}}$$

respectively and observed that the estimates of the average probability of making a transition from one state to another were more efficient obtained by GEE for transition counts assuming the pairwise correlation as

correlation structure, than that of obtained by GEE of Albert and Waclawiw. Lastly they concluded that for heterogeneous transition data the two-state quasi-likelihood model incorporates the heterogeneity by allowing the transition probabilities to vary randomly across subjects.

Gulshan and Islam (2000) proposed an extension of GEE for repeated measures with polytomous responses. Let each of N individuals is observed for T occations. Then the random vector of responses for the i-th individual is $T\times 1$. For k independent variables, the matrix of covariates for i-th individual is $T\times(k+1)$. For logits $g_j(x)=\log\left[\frac{P(Y=j|X)}{P(Y=0|X)}\right]=X'\beta_j$; $\beta_j'=(\beta_{j0},\beta_{j1},.....,\beta_{jp})$, j=1,2; the mean vectors μ_{1j} and μ_{2j} are of order $T\times 1$ with $\mu_{1ij}=P_{1ij}=P[Y_{ij}=1|X_{ij}]$ and $\mu_{2ij}=P_{2ij}=P[Y_{ij}=2|X_{ij}$, i=1,2,.....,$N$ and j=1,2,.....,T. Then GEEs for β's are

$$U(\beta_1)=\sum_{i=1}^{N}D_{1i}^{T}V_{1i}^{-1}(Y_{1i}-\mu_{1i})=0$$

$$\text{and } U(\beta_2)=\sum_{i=1}^{N}D_{2i}^{T}V_{2i}^{-1}(Y_{2i}-\mu_{2i})=0$$

where $D_{1i}=\frac{\delta\mu_{1i}}{\delta\beta_1^T}=X_i^TA_{1i}$ and $D_{2i}=\frac{\delta\mu_{2i}}{\delta\beta_2^T}=X_i^TA_{2i}$; A_{1i} and A_{2i} being $T\times T$ diagonal matrices of variance of the responses and V_{1i} and V_{2i} being working covariance matrices which can be expressed as

$$V_{1i}=A_{1i}^{\frac{1}{2}}R_{1i}(\alpha)A_{1i}^{\frac{1}{2}} \text{ and } V_{2i}=A_{2i}^{\frac{1}{2}}R_{2i}(\alpha)A_{2i}^{\frac{1}{2}}$$

where $R_{1i}(\alpha)=Corr(Y_i)$ and $R_{2i}(\alpha)=Corr(Y_i)$ working correlation matrices of order $T\times T$.

Thus GEE's are

$$U(\beta_1)=\sum_{i=1}^{N}X_i^TA_{1i}V_{1i}^{-1}(Y_i-\mu_{1i})=0$$

$$\text{and } U(\beta_2) = \sum_{i=1}^{N} X_i^T A_{2i} V_{2i}^{-1} (Y_i - \mu_{2i}) = 0$$

They solved these equations simultaneously using Newton-Raphson iteration procedure.

They also used some specifications for $Corr(Y_i)$ as : (i) identity matrix, (ii) exchangeable correlation and (iii) pairwise correlation (as used by Kabir and Islam (2000)).

In the study they used the data of 480 diabetic patients each with 2 follow up visits registered at BIRDEM during 1984-1994. At each time of follow up visits the dependent variable is defined in terms of the observed glucose level after two hours of 75 gms glucose load and is considered to have three categories: controlled, borderline and confirmed diabetes coded as 0, 1 and 2 respectively. Also they included age, sex, education level, area, family history of father and mother (FHFM) and time since registration as independent variables. They also compare the estimates with respect to their efficiencies and observed that the estimates obtained under the assumption of exchangeable correlation and pairwise correlation within the repeated observations are much more efficient as compared to the estimates obtained under the assumption of working independence. Between them the estimates under exchangeable correlation assumption provide more efficient estimates as compared to the previous one for the problem.

Sherif and Islam (2000) studied with the diabetes mellitus data from 197 diabetic patients collected by BIRDEM during 1984-1996. They used two popular forms of kernel with local bandwidth 1 and 2 and global bandwidth 10 and 8. Also five different case weights were being used for the purpose, $w = w_{j0} = 1$; $w1 = w_{j1} = n_{j-1}$, $(j = 1,2,\ldots\ldots\ldots, p)$;

$w2 = w_{j1}^{c} = \frac{n_{j-1} + n_j}{2}$, $(j = 1,2,\ldots\ldots\ldots, p)$; $w3 = w_{j2} = \frac{n_{j-1}}{\Delta\tilde{q}(t_j)\{1 - \Delta\tilde{q}(t_j)\}}$ and

$w4 = w_{j2}^{c} = \frac{n_{j-1} + n_j}{2\Delta\tilde{q}_c(t_j)\{1 - \Delta\tilde{q}_c(t_j)\}}$, $(j = 1,2,\ldots\ldots\ldots, p)$. They observed that the weight $w1 = n_{j-1}$, $(j = 1,2,\ldots\ldots\ldots, p)$ gives the smaller error sum of squares in all case among all the other case weights. This weight shows sizeable gain when case weights are employed which adjust for the declining number of subjects at risk and accompanying increased variance. Then $w2 = w_{j1}^{c} = \frac{n_{j-1} + n_j}{2}$, $(j = 1,2,\ldots\ldots\ldots, p)$ the average of the two consecutive numbers of individuals at risk performs better. They also observed that applying cross-validation method for bandwidth selection and omitting 6-th observation they got the minimum error sum of squares irrespective of the case weights. Also (global) bandwidth 10 gave more precise estimate than (global) bandwidth 8 and they concluded that larger bandwidth produced more precise estimate than smaller bandwidth. Lastly they observed that for the weights smoothed error sum of squares are always smaller than that of weighted linear estimates.

Sharmin (2000) also work with diabetes mellitus data. For the purpose she collected information on 406 patients till December 1994, among 4382 patients who were registered at BIRDEM in the year 1984. She has used Vecek's (1997) model for examining the effects of intensity on exposure-response analysis when intensities vary over time and the used model proposed by Vecek is

$$\lambda_i(u) = \lambda_{yi}(u)\exp[\beta_1 Z_i(u) + \beta_2 S_i]$$

where $Z(u)$ is the exposure metric, S is the smoking status, $\lambda_i(u)$ is the probability that the i-th subject will die during year u and $\lambda_{si}(u)$ is the probability of death during year u for an unexposed subject born in the same year as subject i. And the proposed functions are-

1. Uniform function:

$$g(x_i) = (x_i - \mu), \text{ where } \mu = \text{mean age of the respondents.}$$

2. Linear function:

$$g(x_i) = (x_i - \nu), \ if \ x_i > \nu$$
$$= 0, \quad \text{otherwise}$$

3. Logarithmic function:

$$g(x_i) = \ln\{(x_i + \alpha)/(\gamma + \alpha)\}, \ if \ x_i > \gamma$$
$$= 0, \quad \text{otherwise}$$

4. Normal cumulative distribution function:

$$g(x_i) = \phi\{(x_i - \mu)/\sigma\}, \ if \ x_i > \mu$$
$$= 0, \quad \text{otherwise}$$

Shee has also used a simple statistic for testing the goodness-of-fit of the proposed model suggested by Rao (1965):

$$\chi^2 = \frac{\left[\sum_{j=1}^{J} E_j\left(O_j - E_j^*\right)\right]^2}{\sum_{j=1}^{J} Z_j^2 E_j^* - \left(\sum_{j=1}^{J} Z_j E_j^*\right)^2 \Big/ OBS}$$

where $E_i = \sum_{i=1}^{J} \int_{[z_i(u) \in S_j]} \lambda_i^*(u)du$, $E_i^* = E_i . OBS/EXP$ and $\sum_{i=1}^{J} E_i^* = OBS$.

For the analysis she has considered two distinct cases of the disease, case A: disease progression of the respondents is considered as upward direction, i.e. controlled diabetic to confirmed diabetic and case B: disease progression is taken to be backward direction, confirmed diabetic to controlled diabetic. The results obtained from his analysis were that the

smaller deviances obtained for use of exposure metrics based on non-linear functions of intensity indicating the possibility that the cumulative exposure index might not be the most appropriate metric for assessing the relationship between the exposure and the risk of the disease. She fitted the proposed model for case A and case B separately and observed that in case A, logarithmic function and in case B, normal cumulative distribution function yield smaller deviance than the others. So, he concluded that exposure metric based on non-linear functions of intensity appear to fit the data much better than that of linear functions. She also concluded that age appeared to have no significant affect for the transition from controlled to confirmed diabetic state and the sex was found to be positively associated with the disease from controlled to confirmed state and the transition was greater among male than female in Bangladesh. In case B, she found that both the variables age and sex showed negative association with the disease and also variable age has significant and sex has insignificant affect on the transition of the disease from confirmed to controlled diabetes state.

Mallick and Islam (2000) estimated and tested the Markov Chain based logistic regression model for longitudinal data. For the purpose they considered the model proposed by Singh and Sutradhar (1989). Accordingly, the elements of one-step transition probability matrix is

$$P_{00} = \pi_0 + \varphi\pi_1,\ P_{01} = \pi_1 + (1-\varphi),\ P_{10} = \pi_0 + (1-\varphi) \text{ and } P_{11} = \pi_1 + \varphi\pi_0$$

where φ is the dependence parameter defined by the first order autocorrelation , i.e. $Corr(X_i, X_j) = \varphi^{|i+j|}$, $(1 \le i, j \le n)$ with the restriction $0 < \pi_1 < 1$, $\max(-\pi_0/\pi_1, -\pi_1/\pi_0) \le \varphi \le 1$.

Then π_1 and φ are the functions of P_{ij}'s :

$$\pi_1 = P'_{01}(P_{01} + P_{10}) \text{ and } 1 - \varphi = P_{10} + P_{01}$$

Also for the extension of the test they used a logistic regression model employed with covariate dependence proposed by Muenz and Rubinstein (1985) which takes into account the transition probabilities P_{00} and P_{10} by two logistic regressions:

$$P_{00} = \frac{e^{\beta' X}}{1+e^{\beta' X}} \text{ and } P_{10} = \frac{e^{d' X}}{1+e^{d' X}}$$

where the vector X contains covariates and for the i-th person in the study is equal to $X_i' = (1, X_{i1}, \ldots\ldots\ldots\ldots, X_{ip})$, $i = 1,2,3,\ldots\ldots\ldots, n$ and $\beta = (\beta_0, \beta_1, \ldots\ldots\ldots, \beta_p)'$ and $d = (d_0, d_1, \ldots\ldots\ldots, d_p)'$ are parameter vectors.

For the specified stationary probability test for both the models they considered Billingsley's (1961) χ^2_B test and Tavare Altham's (1983) χ^2_{TA} test. They are-

$$\chi^2_B = \frac{\sum_{i=0}^{1}\sum_{j=0}^{1}\left(y_{ij} - \bar{m}_{ij}\right)^2}{m_{ij}} \sim \chi^2_{(1)}$$

where y_{ij} is the one step transition count from state i to j, $\bar{m}_{ij}$ is the estimated expected counts when $\pi_1 = \pi_1^*$ and $\varphi = \bar{\varphi}$, $\bar{\varphi}$ being the restricted maximum likelihood estimate of φ under null hypothesis when modified likelihood is employed which is obtained from the solution

$$\pi_0^*\left(\pi_1^* + \varphi\pi_0^*\right)^{-1} y_{11} + \pi_1^*\left(\pi_0^* + \varphi\pi_1^*\right)^{-1} y_{00} - (1-\varphi)^{-1}\left(y_{10} + y_{01}\right) = 0$$

$$\text{and } \chi^2_{TA} = \frac{1-\tilde{\varphi}}{1+\tilde{\varphi}} \times \frac{\left(n_1 - n\pi_1^*\right)}{n\pi_1^*\pi_0^*} \sim \chi^2_{(1)}$$

According to Billingsley as $n \to \infty$ the asymptotic behavior of one step transition counts $\{y_{ij}$ for a stationary and ergodic Markov Chain coincide with the product multinomial case, thus ignoring the first term

$\{\pi_1^{X_1}\pi_0^{1-X_1}W(x_1,x_n,y)$, W being Whittle's formula, the modified function of the full likelihood is

$$L_{\text{mod}} = \binom{Y_{0+}}{Y_{00}} P_{00}^{Y_{00}} P_{01}^{Y_{01}} \binom{Y_{1+}}{Y_{10}} P_{10}^{Y_{10}} P_{11}^{Y_{11}}$$

where $Y_{0+} = Y_{00} + Y_{01}$ and $Y_{1+} = Y_{10} + Y_{11}$

Considering the modified likelihood function the score functions of the model without covariate become

$$S(\pi_1) = (1-\varphi)(D_0 + D_1)$$

and $$S(\varphi) = \pi_0 D_1 - \pi_1 D_0$$

where $D_k = (Y_{k1}/P_{k1} - Y_{k0}/P_{k0})$, k=0,1 and for the logistic regression model the scores become

$$S(\beta_0) = \sum[Y_{00i} - (Y_{00i} + Y_{01i})P_{00}]$$

$$S(\beta_j) = \sum X_{ij}[Y_{00i} - (Y_{00i} + Y_{01i})P_{00}]$$

$$S(d_0) = \sum[Y_{10i} - (Y_{10i} + Y_{11i})P_{10}]$$

$$S(d_j) = \sum X_{ij}[Y_{10i} - (Y_{10i} + Y_{11i})P_{10}]$$

An alternative expression of χ_B^2 and χ_{TA}^2 can be obtained as a score test as given by Cox and Hinkley (1974) using modified likelihood. The score test rejects null hypothesis for large values of W_w defined by

$$W_w = \{S(\pi_1) - \beta S(\varphi)\}^2 / Var\{S(\pi_1)\} - \beta Cov\{S(\pi_1), S(\varphi)\}$$

where $\beta = Cov\{S(\pi_1), S(\varphi)\}/Var\{S(\varphi)\}$

If we let $\tau = (P_{11}P_{01}/Y_1 + P_{00}P_{10}/Y_{0+})$ and $\psi = \tau(Y_{0+}Y_{1+}/n)$, then the alternative expressions become

$$\chi_B^2 = W_w(\pi_1^*, \bar{\varphi}) \cong \psi^{-1}W_1^2 \text{ and } \chi_{TA}^2 = (\pi_1^*, \bar{\varphi}) \cong \psi^{-1}W_2^2$$

with $W_1 = n^{-\frac{1}{2}}\{(Y_{11} - Y_{1+}P_{11})(Y_{0+}/n\pi_0^*) + (Y_{01} - Y_{0+}P_{01})(Y_{1+}/n\pi_1^*)$

$$\text{and } W_2 = n^{-\frac{1}{2}}\left[\{(Y_{11} - Y_{1+}P_{11}) + (Y_{01} - Y_{0+}P_{01})\} + \{(X_n + X_1)P_{11} + (1 - X_1)P_{01}\}\right]$$

For the analysis they selected 3474 patients out of 5481 from BIRDEM for the period 1984-1994. The logistic regression model employed three independent variables: age sex and disease status of patients. Again disease status had been partitioned into two states interms of Blood Glucose Level (BGL): state 0 for controlled diabetic, i.e. BGL<11.1 m.ml./lr and state 1 for confirmed diabetic, i.e. BGL≥11.1 m.ml./lr. From the analysis they observed that increasing age playes important role for changing disease status from non-diabetic to diabetic state and the tendency of changing state from diabetic to non-diabetic state decreases with the increase of age. They also observed that the patient having diabetic father has a greater chance to switch from diabetic to controlled diabetic state. For testing a specified probability they found that both χ^2_B and χ^2_{TA} produce the same result giving evidence about the probability of being in the diabetic state eventually differs from 0.5 while considering any model with or without covariates. Hence they concluded that the model without covariates is more useful because of it's simple computation.

Population survey conducted in Michigan on a series of 50-79 years old diabetic patients reported by Barrett-Connev et al (1981) confirmed that the tendency to develop hypertension is more among diabetic women than diabetic men.

Zimmet P. et al (1983) did rural-urban and ethnic comparisons of IGT and DM in the biracial population of Fiji in 1980 and observed that no

statistically significant differences existed in age standardized IGT prevalence between rural and urban groups or between Melanesians and Indians.

Omar M. A. K. et al (1985) also worked with diabetes mellitus and the prevalence of diabetes mellitus and IGT among 866 Indians living in the Chatswoth area of Durban were determined. The study group was selected by cluster sampling and the participants underwent a modified GTT. On the basis of the revised WHO criteria the overall prevalence of diabetes mellitus was 11% and of IGT 5.8%. Of the 368 men, 7.6% were found to have diabetes mellitus and 7.1% IGT, the prevalence of diabetes mellitus was much greater among women (13.5%), while there was less IGT (4.8%). Subjects with diabetes mellitus were significantly older (mean 50.7 years) than those with a normal GTT (mean 30.9 years), but of similar age distribution compared with the IGT group (mean 46 years). Subjects with a normal GTT had a significantly lower mean body mass index compared with diabetic subjects or the IGT group. Obesity was commonly associated with both diabetes mellitus and IGT, particularly among women.

Kelleher et al (1988) studied with diabetic patient to see the effect of hypertension and in the study they showed that out of 386 patients attending a diabetic clinic 38% of type II diabetics (NIDDM) and 15% of type I diabetic had hypertension; also the percentage of hypertension among 255 control subjects (nondiabetic) was 15%.

Prevalence of hypertension among IGT and diabetic subjects had been reviewed separately by Simonson (1988) and Christlieb (1981). In both of

this reviews it was concluded that hypertension among IGT and NIDDM subjects were 1.5 to 2 times more prevalent irrespective of age, sex and body weight. In type I diabetics the prevalence is same as that of NIDDM or IGT but hypertension develops later.

Ramchandran A., Jali M. V., Mohan V., Snehalatha C. and Viswanathan M. (1988) analysed with an urban population in a township in south India for diabetes with an oral glucose tolerance test. Every fifth person aged 20 and over registered at the local iron ore company's hospital was screened. Of 678 people (346men and 332 women) who were tested, 34 (5%; 20 men and 14 women) had diabetes and 14 (2%; 8 men and 7 women) had impaired glucose tolerance. Thirteen subjects were already known to be diabetic. Diabetes was present in 21% of people aged over 40. The peak prevalence (41%) was in the group aged 55-64. A family history of diabetes was present in 16 of the 34 subjects with diabetes and 9 of the 15 with impaired tolerance. Diabetes was significantly related to obesity in women but not in men (57% v 5%). The plasma glucose concentration two hours after glucose loading was correlated to body mass index, age and income in both sexes. The prevalence of diabetes was significantly higher in subjects whose income was above the mean. They also observed that when the prevalence of diabetes was adjusted to the age distribution of the Indians living in Southall, London, and in Fiji, it increased to 19% and 9% respectively. The prevalence of diabetes was high among urban Indians and was comparable with the high prevalence seen in migrate Indian populations.

McKeigul P. M. et al (1989) worked with South Asians (Indians, Pakistanis and Bangladeshis) and Europeans. For the purpose 567 men

aged 40-64 in industrial workforces in West-London were examined for insulin resistance. Prevalence of NIDDM was 14.8% in South Asians compared with 3.8% in Europeans. Serum insulin levels in South Asian men compared with European men were 20% higher in the fasting state and 66% higher at 2 hr. after a glucose load. Plasma triglyceride levels were 17% higher and HDL cholesterol levels were 0.1 mmol/L lower in South Asian than Europeans ($p<0.01$). South Asians showed a striking tendency to central obesity average waist hip circumference ratio 0.97 compared with 0.92 in Europeans which accounted for most of the ethnic difference in insulin levels. The results confirmed that the high prevalence of diabetes in South Asian was part of a pattern of metabolic disturbances related to insulin resistance. For the first time they observed the association of central obesity.

Ambrosio G. B. et al (1990) worked with a random sample of 1,903 subjects (50.1% men) aged 20 to 59 of Mirano (Venice) Northern Italy, where 55 were diabetic (fasting plasma glucose $\geq$ 140 mg/dl or diagnosed by a physician). In this paper an assessment was made on the more frequent occurrence of coronary risk factors (serum cholesterol and triglycerides, BMI, systolic blood pressure, cigarette smoking) and, in particular, of their aggregation in diabetic patients as compared with non-diabetic control subjects. The occurrence of any one of the coronary risk factors studied was more frequent in diabetic subjects and significantly so for triglyceridemia in both genders and for systolic blood pressure and BMI in men. The aggregation of two or more risk factors was also more frequent in diabetic subjects than in control subjects. Finally, the combined sore of coronary risk as calculated by multiple logistic function showed higher values for diabetic subjects.

Although died therapy is considered the cornerstone of therapy for obese patients with NIDDM, losing weight is often difficult, and the plasma glucose concentration does not always improve after weight loss. Watts N. B. et al (1990) looked for predictors of improvement in plasma glucose levels after weight loss in 135 obese patients with NIDDM who had lost at least 9.1 kg of body weight. After weight loss there was a bimodal distribution of plasma glucose levels, which made it possible to identify patients as responders and non-responders according to whether a random plasma glucose level was greater than or less than 10.0 mmol/l after a 9.1 kg weight loss. 55 (41%) of 135 patients were responders. Many responders had improved plasma glucose levels after only slight weight loss. 80 patients were non-responders. Lastly the authors concluded that, in contrast to conventional teaching, many patients with NIDDM would not have any improvement in plasma glucose levels after a 9.1 kg weight loss. However, a substantial minority (approximately 40%) of obese patients with NIDDM had much lower plasma glucose levels with a weight loss of greater than or equal to 9.1 kg. The success or failure of diet therapy could be predicted from the plasma glucose level after a weight loss of only 2.3 to 4.5 kg. Mild pr moderately obese patients with NIDDM who remain hyperglycemic after a weight loss of 2.3 to 9.1 kg were unlikely to improve with further weight loss and should be considered for treatment with insulin or hyperglycemic agents.

Sprafka J. M. et al (1990) anlaysed with prevalence of undiagnosed eye disease in high-risk diabetic individuals. A total of 533 diabetic individuals using the Marshall, Minn, medical care system were identified as potential subjects for the study of unrecognized eye disease. Ophthalmic examination was performed on 145 of these high-risk individuals and

revealed that 61% had clinical characteristics consistent with diabetic retinopathy, glaucoma, cataract, or other eye abnormalities. 25 of these subjects presented with eye disease that required immediate treatment, referral, or accelerated follow-up. Of those indicating they had an ophthalmologist, approximately 35% reported a time since last visit of greater than or equal to 2 years. Hence they concluded that a high prevalence of ocular morbidity among diabetic individuals who were not under routine ophthalmic surveillance and suggest that improvements in both patients and professional compliance with recommended guidelines for eye care were warranted.

Walker W. G. et al (1990) completed a prospective study of the impact of hypertension upon kidney function in diabetes mellitus. Longitudinal data were obtained on 131 diabetic subjects enrolled in a study designed to evaluate the impact of persistent elevation of BP upon progression of renal damage in diabetes mellitus. For both IDDM and NIDDM, serum creatinine exhibited a more rapid rise in those individuals whose BP remained elevated above 140 mmHg despite therapy. From the analysis they concluded that persistent elevation of the BP adds significantly to the risk of renal damage in both insulin-dependent and non-insulin-dependent diabetes, with more rapid decline occurring in non-insulin-dependent diabetes. Hypertensive subjects exhibited higher levels of plasma angiotensin II during the follow-up period.

Osei K. (1990) measured the rates of basal hepatic glucose output (HGO), glucose disappearance, and metabolic clearance of glucose (MCR) in 27 non diabetic first degree relatives of NIDDM patients 16 age, gender, and weight matched healthy control subjects with no family history of

NIDDM. Mean fasting plasma glucose was significantly lower ($p<0.05$) in control subjects than in relatives. Mean basal insulin levels were not significantly different between relatives and control subjects. Mean basal HGO was significantly lower in control subjects compared with relatives. Mean MCR was similar in relatives and control subjects. Hence he concluded that impaired basal hepatic glucose regulation rather than glucose disposal is present as an early defect in glucose tolerant first degree relatives of NIDDM patients.

Johnston C. (1990) analysis with islet function and insulin sensitivity in non-diabetic offspring of conjugal type 2 diabetic patients and concluded that the genetic predisposition to type 2 diabetes is not associated in young adults with any major premorbid impairment in insulin secretion or insulin action but the relationship between the two may be abnormal. A-cell function appears to be normal.

In a large study done by Reaven et al (1990) on 2480 men and women of age 50-89 years in Southern California, showed that in both sexes adults with IGT or NIDDM had increased mean blood pressure as compared with those of normal glucose tolerance. The difference were statistically ($p<0.05$) as well as clinically (3-12mmHg) significant and were independent of age and obesity.

Mitchell B. D. et al (1990) studied whether lifetime cigarette smoking is associated with the presence of diabetic neuropathy. The research design consisted of a case-control study conducted from a referral-based diabetes clinic at a major medical center. The patients were a 65% sample (163 IDDM and 166NIDDM patients) of all patients admitted during a 26-

months period. Neuropathy was diagnosed on the basis of signs and symptoms. Smoking history was obtained by mailed questionnaire (66% response rate). Diabetes duration, HbA_1, age gender, peripheral vascular disease, hypertension history, and lifetime alcohol consumption were measured as covariates. The prevalence of neuropathy was 49% and 38% in IDDM and NIDDM patients respectively. In IDDM, but not NIDDM, current or ex-smokers were significantly more likely to have neuropathy than individuals who have never smoked (odds ratio 2.46, $p= 0.02$), and the prevalence of neuropathy increased with increasing number of pack-years smoked ($p<0.001$). After adjustment for covariates, IDDM patients smoking ≥30 pack/year were 3.32 times more likely to have neuropathy than patients smoking less than this amount. Cigarette smoking was associated with the presence of neuropathy in this clinic-based population of IDDM patients.

Dowse G. K., Gareeboo H., Zimmet P. Z., Alberti K. G., Tuomilehto J., Fareed D., Brissonnette L. G., and Finch C. F. (1990) worked in Mauritus, a multiethnic island nation in the southwestern Indian Ocean, has one of the world's highest diabetes mortality rates. The prevalence of both impaired glucose tolerance (IGT) and noninsulin dependent diabetes mellitus (NIDDM) was investigated in 5080 Muslim and Hindu Indian, Creole (mixed African, European, and Indian origin), and Chinese Mauritan adults aged 20 to 74 years who were selected by random cluster sampling. Based on a 75 gm OGTT and WHO criteria, the age standardized prevalence of IGT was significantly greater in women (19.7%) than in men (11.7%). By contrast, the prevalence of NIDDM was similar in men (12.1%) and women (11.7%) for all ethnic groups combined. The gender difference in IGT prevalence was seen in all ethnic

groups, but for NIDDM, the gender difference was not consistent across ethnic groups. However, age and gender standardized prevalence of IGT and NIDDM was remarkably similar across ethnic groups (16.2% and 12.4% in Hindu Indians, 15.3% and 13.3% in Muslim Indians, 17.5% and 10.4% in Creoles, and 16.6% and 11.9% in Chinese, respectively). Three new cases of diabetes were diagnosed for every two known cases. The high prevalence of abnormal glucose tolerance in Indian subjects is consistent with studies of other migrant Indian communities, but the findings in Creole and, in particular, Chinese subjects were unexpected. Potent environmental factors shared between ethnic groups in Mauritus may be responsible for the epidemic of glucose intolerance.

Warram J. H. et al. (1991) worked with the risk of IDDM in children of diabetic mothers. They tested the association with maternal age in an independent set of families (n=103) in which the mother had at least one pregnancy before and after the onset of IDDM. In the 304 offspring, the mean ± SE risk of IDDM by age 20 was 6.0 ± 2.4% for those born at maternal ages<25 years, whereas, the risk was significantly lower (0.7 ± 0.7%) for those born at older maternal ages (p=0.03). These 304 offspring were combined with a sample of 1391 offspring previously reported for a multivariate analysis of other factors related to pregnancy. In the combined analysis, the risk of IDDM in offspring born at maternal age >25 years was one-fifth that for offspring born to younger mothers. The risk of IDDM in the offspring was not significantly related to birth order, mother's age at first pregnancy or the interval between pregnancies for subsequent ones. The risk for children born before the mother's onset of diabetes was higher than that for those exposed in outer to her diabetes, but the difference did not reach statistical significance. Lastly they

concluded that although genetic factors are important determinants of susceptibility to IDDM, exposure to maternal diabetes protects offspring from IDDM during the first two decades of life.

Ahuja, M. M. S. and Shah, P. (`1991) analysed with 4625 subjects (2776 males and 1849 females) with NIDDM. In the cohort, age of onset of diabetes was less than 25 years in 14.2% (onset of NIDDM is often after 35 years age in India). In 12.5% duration of diabetes was of more than 14-year duration. Body mass index was >25 in 32% of subjects, the obesity was present in only one-third of the NIDDM group. Severe group of hyperglycemia i.e. HbA1c>12% was present in 30.6% and insulin therapy was being followed in only 12.7% while OHA were being prescribed in 61%. The vascular disease was evaluated based on the following criteria:

Large Vessel Disease (LVD)

1. Coronary Artery Disease (CAD): ECG-Minnesota code, Probable/Possible.
2. Cerebro-Vascular Disease (CVD): TIA and stroke.
3. Peripheral Vascular Disease (PVD): Intermittent claudication, absent pulses or gangrene.

Small Vessel Disease (SVD)

1. Ratinopathy: Fundoscopy (hemorrhages and exudates more than 3, or proliferative retinopathy).
2. Nepropathy: Significant proteinuria (two plus) or serum creatinine>2.5mg/dl.

They obtained the results as

LVD	Percent	
CAD	Male	Female
Probable	8.9	5.2
Possible	15.8	21.3
CVD	2.6	3.0
PVD	0.4	0.2
SVD		
Ratinopathy	16.3	14.3
Nepropathy	15.4	13.9

Lastly they concluded that in males, CAD probable was higher in diabetics with duration>14 years, while SVD was significantly increased after a 7-year duration. While coronary probable was more frequent in males, females predominate for coronary possible. Similarly, gangrene and amputation were more frequent in males, while in females intermittent claudication was observed more frequently. Raised blood pressure (systolic and diastolic) was observed to be related to SVD. Small vessel disease (background retinopathy) was more in males while for proteinuria, there was no sex prediction. Also, BMI, HbA_{1c} and lipid values did not seem to influence the vascular complications in the diabetics in their study. Five year follow-up data was being analyzed.

Sayeed (BIRDEM) (1993) analyzed with three thousand nine hundred sixty nine subjects of age >20yrs were investigated for blood glucose in different areas and 2030 subjects were investigated for height, weight and blood pressure together with blood glucose 2hrs after glucose drink. The waist/hip ratio (WHR) and blood glucose were measured in 1523 subjects.

Overall prevalence rates of NIDDM and IGT were 3% and 10.5% respectively. Highest prevalence was observed in rich (6.2%) and lowest in poor socio-economic groups (1.4%) in rural areas. Older age group >40yrs and high BMI >22.1 were associated with glucose intolerance in both sexes. The male to female ratio was 2:1. WHR was strongly associated with blood glucose value >11.1 mmol/lr in male, mildly with blood glucose >8.1 mmol/lr in female. The subjects with abnormal glucose tolerance (i.e., blood glucose >8.1 mmol/lr) of height <164cm in male and <155cm in female showed a significant association with glucose intolerance in male (chi=6.7, $p<0.01$) but not in female (chi=0.04, $p>0.1$) when compared to those AGT subjects with higher height of male (>165cm) and female (>156cm).

Egger, M. et al (1997) analysis with risk of adverse effects of intensified treatment in insulin-dependent diabetes mellitus (a meta-analysis) and they concluded that a substantial risk of severe adverse effects was associated with intensified insulin treatment. Mortality from acute metabolic causes was increased; however, largely counterbalance occurred by a reduction in cardiovascular mortality.

Muggeo, M. et al (1997) also worked on long-term instability of fasting plasma glucose, a novel predictor of cardiovascular mortality in elderly patients with non-insulin-dependent diabetes. 566 elderly patients with type 2 diabetes were followed up for five years to assess mortality and causes of death and after completing the analysis they concluded that fasting plasma glucose instability was a predictor of cardiovascular-related mortality in elderly patients with type 2 diabetes. Mathiesen, B. et al (1997) worked on "diabetes and accident insurance: a 3-year follow-up of

7,599 insured diabetic individuals". Hamada, Y. et al (1997) worked with effects of glycemic control on plasma 3-deoxyglucosone levels in NIDDM. Waldron, S. et al (1997) worked with the dietary management of Children's diabetes. Harris, S. B. et al (1997) analyzed with the epidemiology of diabetes in pregnant native Canadians. Van Belle, E. et al (1997) compared between restenosis rates of coronary stenting and balloon angioplasty in native coronary vessels in diabetic. John, J. N. and John, K. L. (1998) also worked on early-onset insulin-resistant diabetes in Mexican-American adolescents. Hansen, T. (1998) worked on insulin sensitivity index, acute insulin response and glucose effectiveness in a population-based sample of 380 young (18-32 years), healthy Caucasians. Ijff G. A. (1991) studied on cold and warm cutaneous sensation in diabetic patients. Barth R. (1991) analyzed whether intensive education improves Knowledge, compliance and foot problems in type-II diabetes. Chandalia H. B. (1991) studied with a series of 5831 new diabetics registered in their clinic from 1982 to 1989.

1.4 Aim and Objective of the Study:

As revealed from literature review lots of works has been done so far on diabetes mellitus, both clinical and statistical. Most of the statistical works are done on the transition of state of the disease to non-disease (control) and control to disease based on Markov Chain models followed by the analysis of prognostic factors, popularly known as covariates, that might have association with diabetes mellitus. A few studies also have investigated the influence of other diseases on diabetes mellitus. No works

so far is observed to investigate the time trend of control time for diabetes mellitus from the date of diagnosis.

Diabetes is a disease which is not curable but can be controlled. The aim of modeling diabetes mellitus is to investigate how many days on an average would require to control it if the present trends are continued and certain measures are applied by using best available informations. Furthermore, if individual covariates are taken into consideration, in the model, probabilistic statement for each individuals control time requirement may be given. Influence of other diseases on the control time of diabetes mellitus also may be incorporated in such models that may enrich the probabilistic statements. Such modeling also may help in assessing the rate of controlling. This information together with the rate of registration may help the diabetes centers to plan their future activities.

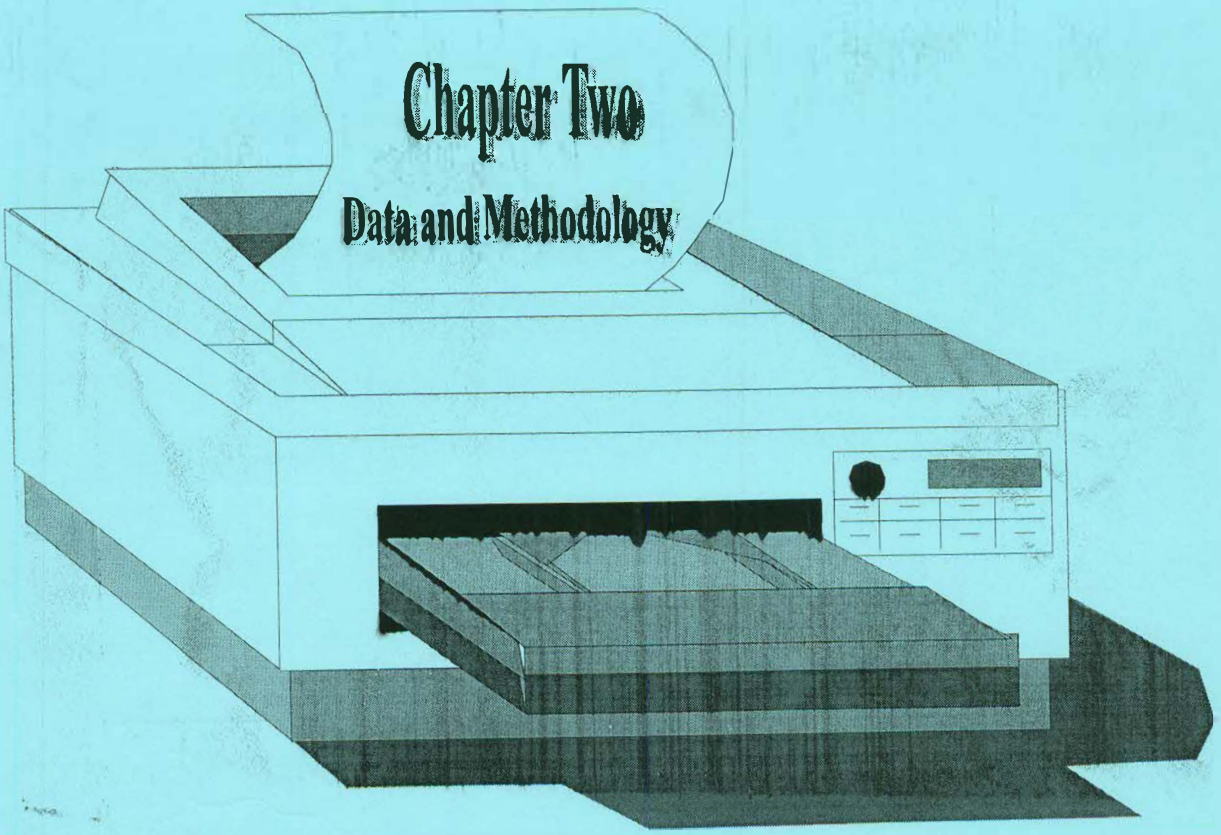
Keeping in mind the above points, this study will investigate to find out a suitable statistical probability distribution to graduate the control time for diabetes mellitus data.

Prognostic factors that are likely to influence the control time would be investigated for significance. Those found statistically significant would be incorporated in the model through parametric regression/proportional hazards modeling to infer the individual control time requirement.

Diseases, which have possibilities to influence the control time of diabetes mellitus, would be investigated by logistic regression model. Those found statistically significant would be incorporated in the model if possible.

Chapter Two

Data and Methodology



Chapter Two

Data and Methodology

2.1 The Data :

In applied research, data plays a vital role. Unless a data is reliable, findings based on such a data is not acceptable. Generally most of the data for statistical investigations are collected on sampling basis. Reliability of a data is related with the mode of collection of data and the accuracy of informations collected. If our aim is to acquire knowledge about some unknown things then accuracy of informations are more vital than the mode of collecting informations. But if we like to infer about some characteristics of a larger population based on a small sample informations then mode of collecting data and sampling design become important along with the accuracy of collected informations. A brief description of data used in the study is given in the next two subchapters.

2.2 Sources of data :

We have collected data from Laxmipur Diabetic Center [a sub-organization of Diabetes Association of Bangladesh (DAB)], Laxmipur, Rajshahi, through a set questionnaire by interview method. During the registration at the center a social officer and doctor interview each patient to fill-up the case history sheet of the record (guide) book, which is given to the patient. They also check some clinical test of the patient and record them on the book. Also a follow-up record is maintained on the book from patients at each visit to center after registration for further need. Our data is based on these information recorded on the book. Now we take a glance of the structure (i.e. information available) of the guide book:

Laboratory Test

Date:	Blood	Urine		
OGTT	mm/l	Sugar	Albumin	S.Gravity
Fasting:				
After Glucose Load:				
½ hour				
1 ½ hours				
2 hours				

Hemoglobin (gm%) SGPT IU (/l)
PCV (%) Cholesterol (mg%)
DC

Other Tests:

ECG
Chest X-ray (PA View)
Diagnosis of Disease:
Other Disease(s):

Plasma Glucose in blood (Oxidedge method) 1mm/l=18mg%

Case History

Date:

Symptoms of the disease:

How old the disease is:

Treatment taken before:

Age: ,Male/Female, Profession:

Height: ,Weight: ,Needed weight:

Temperature: Pulse:

BP: Lying: Standing:

Heart:

Lunge:

Lever:

Spleen:

Nervous System:

Fudascopic Test:

Allergy tendency to medicine:

Progress Report

Date	Blood Glucose & others	Urine			Weight	BP	Advise
		Sugar	Acetone	Albumin			

The study is based on the set of follow up data of size 615 diabetic patients registered at different point of time in the Laxmipur Diabetic Center, Laxmipur, Rajshahi.

2.3 Description of data:

For the analysis we have collected data on 25 points (see appendix) from 615 diabetic patients registered in the center at different time who have come to visit to the center during the period 25 July 2001 to 30 August 2001. The data has been collected by interview method setting a questionnaire. Among the points on the questionnaire data on most of the information have been collected from the guidebook of the patients and some from the patient himself (herself) directly, e.g. food habit of the patient and in some cases the diseases of the patient before the diabetes mellitus.

From the WHO reports and many other research work we can get idea about the factors related to the disease. But do these factors influence the control time? We cannot give answer of this question yet now, but we can only say that these factors may influence the control time. Hence getting idea about the influential factors of the disease we have prepared the questionnaire and then collect the data.

2.4 Methodology:

Methodology used in an applied research are equally important as the data. Every methodology is not suitable for analysing every set of data. Matching of an appropriate methodology for graduating and analyzing a

set of data is a difficult task for a researcher. For this reason, in most of the times, researchers use alternative methodology to graduate and analyze a set of data. Finally they compare the findings obtained from different methodology and support the most logical one as compared with the reality. Brief description of methodologies used in this study are given in the following sections.

2.5 Product Limit Estimate:

The product Limit (PL) estimate, derived by Kaplan and Meier (1958), possesses a number of important properties, a main one being that it is a consistent estimate under quite general conditions. It is a kind of nonparametric maximum likelihood estimate of $S(t)$.

Let there are k distinct lifetimes $t_1 < t_2 < \cdots\cdots < t_k$, with d_j deaths at t_j and n_j individuals at risk at t_j. Also let in the interval $[t_{j-1}, t_j)$ there are λ_j observed censoring times $L_i^j \ (i = 1,\ldots\ldots,\lambda_j)$. If $t_0 = 0, t_{k+1} = \alpha$, and j ranges over $1,\ldots\ldots,k$ then the observed likelihood function is of the form

$$L = \prod_{j=1}^{k}\left[\left(\prod_{i=1}^{\lambda_j} S(L_i^j)\right)[S(t_j) - S(t_j + 0)]^{d_j}\right]\prod_{i=1}^{\lambda_{k+1}} S(L_i^{k+1}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.5.1)$$

To maximize L with respect to $S(t)$, $\hat{S}(t)$ must be discontinuous at the t_j's or else $L=0$. Also the $\hat{S}(L_i^j)$'s should be as large as possible, in accordance with the restrictions on $S(t)$ which implies that

$$\hat{S}(t_1) = \hat{S}(L_i^1) = 1 \qquad i = 1,\ldots\ldots,\lambda_1$$

$\hat{S}(L_i^{j+1}) = \hat{S}(t_j + 0) = \hat{S}(t_{j+1})$ $\quad j = 1,\ldots\ldots\ldots,k$ and $i = 1,\ldots\ldots\ldots,\lambda_{j+1}$

Let us write $S(t_j + 0) = P_j$ $\quad (j = 1,\ldots\ldots\ldots,k)$ and define $P_0 = 1$ then we see from (2.5.1) that we need to maximize

$$L_1 = \prod_{j=1}^{k} (P_{j-1} - P_j)^{d_j} P_j^{\lambda_j + 1} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.5.2)$$

with respect to $P_1,\ldots\ldots\ldots,P_k$. Let $p_j = P_j / P_{j-1}$ and $q_j = 1 - p_j$ $\quad (j = 1,\ldots\ldots\ldots,k)$ then (2.5.2) becomes

$$L_1 = \prod_{j=1}^{k} (p_1 \cdots p_{j-1} q_j)^{d_j} (p_1 \cdots p_j)^{\lambda_j + 1} = \prod_{j=1}^{k} q_j^{d_j} p_j^{n_j - d_j}$$

which is maximized for $\hat{p}_j = (n_j - d_j)/n_j$. Hence L is maximum with

$$\hat{S}(t) = \prod_{j:t_j < t} \frac{n_j - d_j}{n_j} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.5.3)$$

which is known as the Product Limit (PL) estimate of $S(t)$.

2.6 Ordinary Least Square Method:

Regression analysis is a statistical technique for investigating and modeling the relationship between variables and is routinely applied in almost every field of quantitative research. In fact it may be the most widely used statistical technique. Out of many possible regression techniques the Ordinary Least Square (OLS) method has been generally adopted because of tradition and ease of calculation.

The underlying principles of the OLS method is to estimate the unknown parameters of a regression such that

$$\phi = U'U = (Y - X\beta)'(Y - X\beta)$$

be minimum.

This is satisfied when

$$\frac{\partial \phi}{\partial \beta} = 0$$

$$\Rightarrow 2X'(Y - X\beta) = 0$$

$$\Rightarrow X'Y - X'X\beta = 0$$

The least square estimator of β is thus given by the vector

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad \text{.....(2.6.1)}$$

Generally t-test is used for testing the significance of an observed regression coefficient and F-test is used for testing the significance of over all regression.

2.7 Maximum Likelihood Method:

Many statistical problems require the determination of maximum likelihood estimates, that is, point $\theta = (\theta_1, \cdots\cdots, \theta_k)'$ at which a (log) likelihood function is maximized. Sometimes computation of maximum likelihood estimators creates numerical problems, especially when it needs iteration to get the estimates from the maximum likelihood equations obtained. Also if initial values are not close to the true values of parameters, the equations may not coverage. Inspite of this, maximum likelihood gives consistent estimators.

Let there are *n* observations $x_1, x_2, \cdots\cdots, x_n$ with p.d.f. $f(x;\theta)$. Then the likelihood function would be

$$L(\theta) = \prod_{i=1}^{n} f(x_i) \quad \text{.....(2.7.1)}$$

Now if the lifetimes are grouped into $k+1$ intervals $I_j = [a_{j-1} - a_j)$, $j = 1,2,\cdots\cdots, k+1$, where $0 = a_0 < a_1 < \cdots\cdots < a_k < a_{k+1} = \alpha$ and let d_j be the number of lifetimes observed to fall into I_j. If no censoring can occur, except possibly in the last interval, then $(d_1, \cdots\cdots, d_k)$ has a multinomial probability function

$$\frac{n!}{d_1! \cdots\cdots d_k! d_{k+1}!} P_1^{d_1} \cdots\cdots P_k^{d_k} P_{k+1}^{d_{k+1}}$$

where P_j, a function of unknown parameters, is the probability of unconditional failure in the j-th interval. The likelihood function for the unknown parameters θ in the parametric model can thus be taken as

$$L(\theta) = \prod_{j=1}^{k+1} P_j^{d_j} \quad \text{.....(2.7.2)}$$

Maximizing $L(\theta)$ we can get estimates of the underlying parameters. We will gate the same result by maximizing $\log(L(\theta))$ which is more simple and easier to handle. Hence taking *log* on both sides of (2.6.1) we get

$$\log(L(\theta)) = d_j \sum_{j=1}^{i+1} P_j$$

To test the overall significance of parameters Likelihood Ratio Test (LRT) is used and t-test for single case.

2.8 Chi-Square Test as a Goodness of Fit Test :

With grouped uncensored data, tests of fit can be based on the multinomial model, the best known test of fit procedures being the classical Pearson's chi-square test.

Consider the hypothesis

$$H_o : P_j = P_{jo},\ j = 1, \cdots\cdots, k+1$$

where the P_{jo}'s are specified but may involve unknown parameters. Let $\tilde{P}_{jo}$ be the m.l.e. of P_j under H_o, or some other asymptotically fully efficient estimator and let $e_j = n\tilde{P}_{jo}$. The Pearson statistic for testing H_o is

$$\chi^2 = \sum_{j=1}^{k+1} \frac{(d_j - e_j)^2}{e_j} \qquad \text{.....(2.8.1)}$$

where P_{jo}'s are Known constants, $e_j = nP_{jo}$.

The limiting distribution of χ^2 is $\chi^2_{(k-s)}$, s is the number of unknown constants.

χ^2-test is also used to test the homogeneity of correlation coefficients and also to test the association of attributes.

2.9 Likelohood Ratio Test:

Likelihood Ratio Test (LRT), introduced by Neyman and Pearson, is used for testing a hypothesis, simple or composite, against a simple or composite hypothesis. This test is related to the maximum likelihood estimates.

Let $x_1, x_2, \ldots\ldots\ldots, x_n$ be a random sample of size n>1 from a population with p.d.f. $f(x;\theta)$, where $\theta \in \Omega$ is a vector of parameters of order $k \times 1$. We want to test the null hypothesis

$$H_o : \theta = \theta_o$$

against alternative hypothesis of the type

$$H_o : \theta \neq \theta_o$$

The likelihood function of the sample observations is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i : \theta) \qquad \text{.....(2.9.1)}$$

According to the principle of maximum likelihood, the likelihood equation for estimating any parameter θ_i is given by

$$\frac{\partial L}{\partial \theta_i} = 0, \quad (i = 1,2,\text{.....},k) \qquad \text{.....(2.9.2)}$$

Substituting the m.l.e. in (2.8.1) we will get a maximum value of the likelihood function. Also we can get another value for θ_0. Then the criterion for the likelihood ratio test is defined as the quotient these two maxima and is given as

$$\lambda = \frac{L(\theta_0)}{L\left(\hat{\theta}\right)} \qquad \text{.....(2.9.3)}$$

where $L(\theta_0)$ and $L\left(\hat{\theta}\right)$ are the maxima of the likelihood function (2.9.1) with respect to θ_0 and m.l.e. of θ.

2.10 Newton-Raphson Iteration:

Let us denote the likelihood function as $L(\theta_1,\text{.....},\theta_k) = L(\theta)$, define over the parameter space Ω. We consider situations in which the point $\hat{\theta}$ at which $L(\theta)$, and also $\log L(\theta)$, is maximized in a local maximum and satisfies the likelihood equations

$$U_i(\theta) = \frac{\partial \log L}{\partial \theta_i} = 0, \qquad i = 1,2,\text{.....},k$$

Suppose that θ_0 is a first guess at $\hat{\theta}$ and expand each of the functions $U_i(\theta)$ in a Taylor series about θ_0. To first order this gives

$$U(\theta) \approx U(\theta_0) + G(\theta_0)(\theta - \theta_0) \qquad \text{.....(2.10.1)}$$

$G(\theta)$ is the $k \times k$ matrix with entries

$$G_{ij}(\theta) = \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}$$

Since $\hat{\theta}$ satisfies $U\left(\hat{\theta}\right) = 0$, (2.7.1) yields the approximation

$$\hat{\theta} = \theta - G(\theta_0)^{-1} U(\theta_0) \quad \text{.....(2.10.2)}$$

The Newton-Raphson procedure involves the following steps:

1. Obtain an initial estimate θ_0 of $\hat{\theta}$.
2. Calculate $U(\theta_0)$ and $G(\theta_0)$.
3. Calculate the next approximation θ_1 of $\hat{\theta}$ using (2) as

$$\theta_1 = \theta_0 - G(\theta_0)^{-1} U(\theta_0)$$

4. Repeat steps 2 and 3, replacing θ_0 with θ_1. Continue repeating this until convergence is (hopefully) achieved. One stops when θ_i and θ_{i+1} are close together and $U(\theta_{i+1})$ is close to 0.

2.11 Score Test:

Cox and Hinkly (1974) showed that under the following regularity conditions the score vector $U(\beta)$ is asymptotically normally distributed with mean 0 and variance-covariance matrix $I(\beta)$, where $U(\beta) = [\beta_i]_{1 \times K}$ and

$$I(\beta) = \left[\frac{\partial \log L}{\partial \beta_i \partial \beta_j}\right]_{K \times K}.$$

Conditions:

1. The order of integration and differentiation are interchangeable.
2. The dimension of parameter space Ω is finite and the value of the parameters are interior to Ω.
3. The probability distributions for different values of β are distinct.
4. The first two derivatives of log-likelihood with respect to β exist in the neighborhood of the true parameter value.

Let β be the *(k+1)* vector of parameters, then the hypothesis can be written as

$$H_0 : \beta = \beta_0 \quad v.s. \quad H_1 : \beta \neq \beta_0$$

Under the null hypothesis the score statistic $U\left(\hat{\beta}-\beta\right)$ is asymptotically normally distributed with mean vector 0 and variance –covariance matrix $I\left(\hat{\beta}-\beta\right)$. Hence we can define the test statistic for score test as

$$\chi^2 = \left[U\left(\hat{\beta}-\beta\right)\right]'\left[I\left(\hat{\beta}-\beta\right)\right]^{-1}\left[U\left(\hat{\beta}-\beta\right)\right] \quad \text{.....(2.11.1)}$$

which follows chi-square distribution with *k* degrees of freedom.

2.12 Parametric models:

Several parametric families have been suggested as lifetime distributions. Among these the normal (as a model for log-lifetimes), exponential and weibul distributions are worthy to mention. These and other models are discussed at length by Johnson and Kotz (1972), Barlow and Proschan (1975), David and Moeschberger (1978) and others. When one of these

models is assumed, tests and estimation of parameters is usually straightforward.

An important class of models is that for which the log lifetime $Y = \log T$, given X, has a distribution with a location parameter $\mu(x)$ and a constant scale parameter σ. These models can be written in the form

$$Y = \mu(x) + \sigma e \qquad \text{.....(2.12.1)}$$

where $\sigma > 0$ and e has a distribution that is independent of X. The family of models for which e has a standard normal distribution is familiar, model in which e has some other distribution are also important.

The survivor function for Y, given X, is of the form $G\{[y - \mu(x)]/\sigma\}$, where $G(e)$ is the survivor function for e. The survivor function for $T = exp\ Y$ is thus of the form

$$S(t/x) = G\left\{\frac{y - \mu(x)}{\sigma}\right\}$$

$$= S_1\left[\left(\frac{t}{\alpha(x)}\right)^{\delta}\right] \qquad \text{.....(2.12.2)}$$

where $\alpha(x) = \exp[\mu(x)]$, $\delta = 1/\sigma$, and $S_1(w) = G(\log w)$. Alternatively, this can be written as

$$S(t/x) = S_0\left\{\frac{t}{\alpha(x)}\right\} \qquad \text{.....(2.12.3)}$$

where $S_0(w) = S_1(w)$.

2.13 Logistic Regression Method:

The logistic regression model has been used in statistical analysis and frequently used in survival analysis for many years. It is more applicable because of its distribution free assumption of the categorical independent variable and obviously a powerful analytic tool for epidemiologic research. For the analysis of dichotomous outcome the logistic distribution is preferred for two primary reasons:

(i) From a mathematical point of view, it is an extremely flexible and easily used function.

(ii) It lands itself to a biologically meaningful interpretation.

2.13.1 The model:

The logistic regression model allows a categorical variable (dichotomous or polytomous variable) as dependent variable. Let Y is a dichotomous dependent variable, which takes values 0 and 1. i.e.

$$Y_i = \begin{cases} 1, & \textit{if the individual leaves the que during the study} \\ 0, & \textit{otherwise} \end{cases} \quad , i = 1,2,3,\cdots\cdots\cdots,n$$

Also let there is a collection of k independent variables which will be denoted by the vector $X' = (x_1, x_2, x_3, \cdots\cdots, x_k)$ and β be a $(k+1)\times 1$ vector of unknown parameters.

For simplification, we can use the quantity $\pi(X) = P(Y = 1|X)$ the probability that the event occurs conditional on the value of X. Hence

$$\pi(x_i) = P(Y=1|X) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}} = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \quad \text{.....(2.13.1.1)}$$

and

$$1-\pi(x_i) = P(Y=0|X) == \frac{1}{1+e^{x_i\beta}} \quad \text{.....(2.13.1.2)}$$

Hence

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{x_i\beta} \quad \text{.....(2.13.1.3)}$$

The central part of logistic regression in a transformation of $\pi(X)$ is known as logit transformation, which is defined in terms of $\pi(X)$, is as follows:

$$g(x_i) = \text{logit } \pi(x_i) = \log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = x_i\beta$$

$$\Rightarrow g(x_i) = \beta_0 + \beta_1 x_{1i} + \cdots\cdots + \beta_k x_{ki} \quad \text{.....(2.13.1.4)}$$

which is the logit of the multiple logistic regression models. The logit, *g(X)* is linear in its parameters and has many of the desirable properties of linear regression model.

With a two-category response variable, we will examine models for $\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]$. It is to be remarked that when these models are regression type models, they are called logistic regression models. When these models are ANOVA type models, they are often referred to as logit models. The two terms "logit" and "logistic regression" as applied to models, are essentially two names for the same idea. Technically, the terms logit and logistic are names for transformations. The logit transformation takes a number $\pi(x)$ between 0 and 1 and transforms it to

$\log\left[\frac{\pi(x)}{1-\pi(x)}\right]$. The logistic transformation takes a number x on the real line and transforms it to $\frac{e^{\beta x}}{1-e^{\beta x}}$. Note that the logit transformation and the logistic transformation are inverses to each other.

2.13.2 Estimation of Parameters:

Although the most common method used to estimate unknown parameters in linear regression is the least squares method, we will not use this method for the logistic regression method. Because, under usual assumptions, least squares estimations have some desirable properties. But when this method is applied to estimate a model with dichotomous outcome the estimators no longer have these same properties. In such situation, the general method for estimating the parameters of logistic regression models is the method of maximum likelihood and for this we will use a very effective and well known iterative method, Newton-Raphson method as the likelihood equation is nonlinear and explicit function of unknown parameters. For this model the likelihood function is

$$L(\beta_0,\beta_1,.....,\beta_k)=\prod_{i=1}^{n}\frac{\exp\left(y_i\sum_{j=0}^{k}x_{ij}\beta_j\right)}{1+\exp\left(y_i\sum_{j=0}^{k}x_{ij}\beta_j\right)}$$

$$=\frac{\exp\left(\sum_{i=1}^{n}y_i\sum_{j=0}^{k}x_{ij}\beta_j\right)}{\prod_{i=1}^{n}\left\{1+\exp\left(y_i\sum_{j=0}^{k}x_{ij}\beta_j\right)\right\}}$$

$$\Rightarrow \log L(\beta_0,\beta_1,.....,\beta_k)=\sum_{i=1}^{n}y_i\sum_{j=0}^{k}x_{ij}\beta_j-\log\left\{n+\exp\left(\sum_{i=1}^{n}y_i\sum_{j=0}^{k}x_{ij}\beta_j\right)\right\}$$

In order to get the estimate of the parameters by maximizing this equation the computer package SPSS for windows based 7.5 version may be used.

2. 13.3 Interpretation of Parameters:

As in linear regression model interpretation of parameters in logistic regression model is not so straightforward. Interpretation of parameters in logistic regression model can be done in two ways – interms of logit and interms of odd ratio.

In previous we have defined that the logit transformation of logistic regression model is called the logit and it is defined as

$$\pi(x)=\frac{e^{\sum_{i=0}^{k}\beta_i x_i}}{1+e^{\sum_{i=0}^{k}\beta_i x_i}}$$

$$\Rightarrow \log it[\pi(x)]=g(x)=\log\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$=\beta_0+\beta_1 x_1+\cdots\cdots+\beta_k x_k$$

which is linear in parameters.

Hence we can interpret the parameters now using the arguments of linear regression that the parameter β_j $(j=0,1,\ldots\ldots,k)$ represents the rate of change in $\log it[\pi(x)]$ for one unit changes in X_j, given other variables remaining constant.

The other way of interpreting the parameters is interpretation interms of odds ratio, which is an interesting aspect. Infect, this is the proper

interpretation for the parameters of qualitative variable coefficients. It also gives the conceptual foundation for all other situation.

From logistic regression model we have,

$$\pi(x)=\frac{\exp\left(\beta_0+\beta_1 x_1+\cdots\cdots+\beta_k x_k\right)}{1+\exp\left(\beta_0+\beta_1 x_1+\cdots\cdots+\beta_k x_k\right)}$$

and $$1-\pi(x)=\frac{1}{1+\exp\left(\beta_0+\beta_1 x_1+\cdots\cdots+\beta_k x_k\right)}$$

Then the odds of outcome being present for given X is defined as

$$\frac{\pi(x)}{1-\pi(x)}=\exp\left(\beta_0+\beta_1 x_1+\cdots\cdots+\beta_k x_k\right)$$

Now we consider the situation where the independent variable is dichotomous. We assume that x_i takes values 0 and 1, then the odds ratio, denoted by OR, is defined as the ratio of odds for $x_i=1$ to the odds for $x_i=0$ and is given by,

$$OR=\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

$$=\frac{\exp\left(\beta_0+\beta_1 x_1+\cdots+\beta_i\times 1\cdots+\beta_k x_k\right)}{\exp\left(\beta_0+\beta_1 x_1+\cdots+\beta_i\times 0\cdots+\beta_k x_k\right)}$$

$$=\exp\left(\beta_i\right)$$

So we can directly estimate the coefficients of a logistic regression model as $\log OR$ and hence can interpret. If a qualitative independent variable has m categories, we introduce only $(m-1)$ dummy variables and the remaining one is taken as reference category.

2.13.4 Test of Significance of Parameters:

To test the significance of the parameters of logistic regression model we can use the following procedures:

(i) Likelihood ratio test
(ii) Wald test
(iii) Score test.

Likelihood ratio test:

In logistic regression the likelihood ratio test is used for testing the significance of coefficient for the parameters. The test is based on the function of ratio of two likelihood functions as follows:

$$D = -\sum \ln(\lambda)$$

where $\lambda = \frac{L_0}{L_1}$, the ratio of two likelihood functions: the likelihood function for the current model and the likelihood function for the standard model.

For assessing the significance of an independent variable we compare the value of D with and without the independent variable in the equation. This can be written as

G=D(for the model without the variable)-D(for the model with the variable)

i.e. G measure the change in D due to inclusion of the independent variable in the model. G can also be expressed as

$$G = -\sum \ln \left[\frac{Likelihood\ without\ the\ variable}{Likelihood\ with\ the\ variable} \right]$$

Under the null hypothesis that β_i's $(i=1,2,.....,p)$ are equal to zero, the statistic G follows chi-square distribution with p degrees of freedom.

Wald Test:

In logistic analysis due to the nature of maximum likelihood estimation Wald test, introduced by Wald (1943), has a definite advantages over the likelihood test. Wald test is obtained by comparing the maximum likelihood estimate of any parameter to the estimate of its standard error. i.e. for testing

$$H_o : \beta_i = 0 \text{ vs } H_o : \beta_i \neq 0$$

the Wald statistic is defined as

$$W_i = \frac{\hat{\beta}_i}{SE\left(\hat{\beta}_i\right)}$$

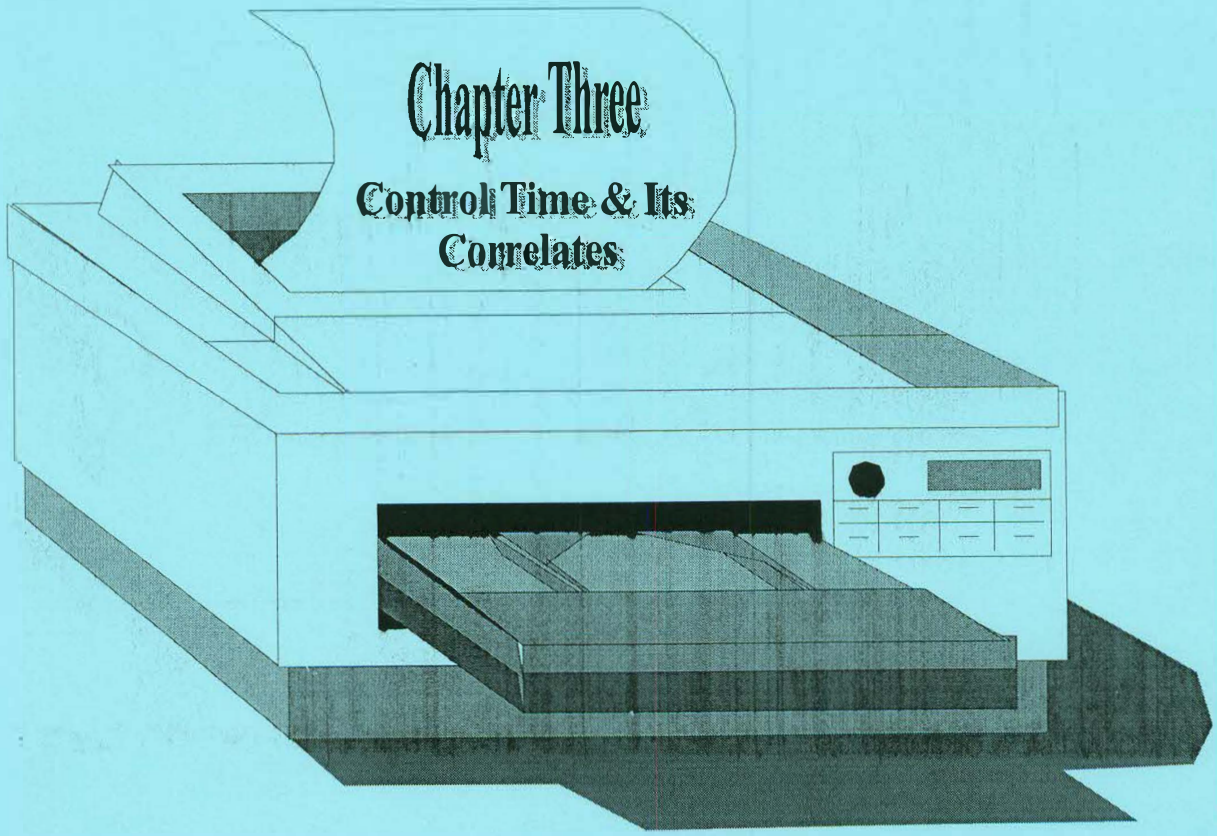
which follows a standard normal distribution.

Score Test:

For testing whether the overall effect is significant or not the score test is used. About the score test is described before in section 2.11 .

Chapter Three

Control Time & Its Correlates



Chapter Three

The Control Time & Its Correlates

3.1 Introduction:

In Bangladesh it is very much hard to go through study with data related to disease. Because the population of our country is not conscious about their health, where they have not yet enough facility of food, housing, clothing and education. Although the health service centers, both government and private, are increasing, the picture is not changed accordingly, especially in the rural area. Many persons after knowing about their disease don't take any treatment until they face and feel some problem. For the disease diabetes mellitus it is very much true for both poor and rich. Diabetes is a very complex disease which may attack all the organs of the body. Moreover, if a person is attacked by this disease once he (she) will never get rid of it, but can control the disease by dieting, exercising, medicine therapy etc. Hence in this study we have given emphasis on control time-

the time that a patient take to control the disease after registration as a diabetic patient at the center and started to take treatment along with other regularities.

In previous chapter we have given a discussion about the data. In this chapter we have depicted distribution of the control time as a whole to get idea about different measures of the distribution and also with respect to different covariates followed by check for association of control time with those covariates.

3.2 Univariate Distribution of Control Time:

How many days on an average a patient would require to come to control state is a fundamental question. 50% of the registered patient would come to control state in how many days? Answer to such questions may be searched in the univariate distribution, popularly known as frequency distribution, of control times. Hence first we have constructed a frequency distribution of control times, the main component of the study. Here we have taken 15 days interval, since after registration as a diabetic patient at the center is generally advised to report to the center at an interval of 15 days.

Table#3.2.1: Univariate distribution of control time.

Class Interval	Frequency
<15	**108**
15-30	**126**
30-45	**92**
45-60	**41**
60-75	**38**
75-90	**19**
90-105	**17**
105-120	**5**
120-135	**13**
135-150	**8**
150-165	**5**
165-180	**10**
180-195	**7**
195-210	**6**
210-225	**9**
225+	**111**

**lower limit excluded.

Here we observe that at the beginning the frequency distribution shows a Poission pattern with a decreasing trend from 2nd to 7th group, but after that it shows a zigzag nature. Moreover it is a continuous distribution. The mean, median, mode, standard deviation, quartile deviation, skewness and kurtosis of this distribution are as follows:

Table#3.2.2: Location & dispersion measures of the distribution of control times

Mean	90.7561
Median	41.9837
Mode	20.1923
S.D.	830.046
Q.D.	60
Skewness	0.936628
Kurtosis	2.228267

From the table also it is observed that mean>median>mode for this distribution, and the distribution is platykurtic and right skewed. About 25 percent of the cases become control within three weeks, while 50 percent of the patients takes about 42 days to be in control state. On an average, patients take three months to be in control state. More or less, 1.1018-1.1973 patient return to control state per day per 100 patients registered.

3.3 Bivariate distribution:

Literature on diabetes mellitus reveals that a number of correlates like age, sex, food habit, body mass index etc. of individuals are associated with the incidence of diabetes mellitus. Naturally, question may arise whether such correlates exist for control time for diabetes. Such prognostic factors are investigated in the following subsections for association with control times of diabetes in the form of bivariate tables followed by Pearsonian chi-square and likelihood ratio tests.

The following correlates are investigated in the study:

1. Age = Exact age in the last birthday as is recorded in the patient history.
2. Sex = $\begin{cases} 1, & \textit{if male} \\ 0, & \textit{if female} \end{cases}$
3. Marital status = $\begin{cases} 1, & \textit{if married} \\ 0, & \textit{otherwise} \end{cases}$
4. Food habit (rice) = $\begin{cases} 1, & \textit{if takes rice thrice a day} \\ 0, & \textit{otherwise} \end{cases}$
5. Food habit (vegetable) = $\begin{cases} 1, & \textit{if vegetarian} \\ 0, & \textit{otherwise} \end{cases}$
6. Food habit (sweet) = $\begin{cases} 1, & \textit{if the patient fonds of sweet or eats more sweet} \\ 0, & \textit{otherwise} \end{cases}$
7. Body Mass Index (BMI) = Exact BMI calculated from height & weight of patient.
8. Blood Pressure (BP) = Exact systolic blood pressure as was recorded in patient book at the time of registration.
9. Blood Glucose Level (BGL) = Exact blood glucose level after 2-hours of 75 gm glucose load as was recorded in the patient book.
10. Heredity = $\begin{cases} 1, & \textit{if parents or blood connected relatives have / had diabetes} \\ 0, & \textit{otherwise} \end{cases}$
11. Length of suffering = Exact length of sufferings in days before registration and as was recorded in the patient book.

3.3.1 Distribution of Control Time by Age of Patient:

Age is an important factor for particular type of diseases and hence it is also a point to get importance for diabetes. Generally it is observed that after a certain age (generally forty) the incidence of diabetes mellitus occurs, although there are some malnutrition cases which do not follow this argument. To group the age we have used two remarkable points- forty and sixty, because after sixty years a person is considered as an aged person. The following table shows the distribution of control time by age.

Table#3.3.1: Distribution of control time by age.

		Age			Total
		<40	40-60	60+	
Control Time (days)	<15	36	54	18	108
	15-30	42	75	9	126
	30-45	28	56	8	92
	45-60	17	20	4	41
	60-75	16	17	5	38
	75-90	10	6	3	19
	90-105	6	10	1	17
	105-120	4	1	0	5
	120-135	2	9	2	13
	135-150	2	6	0	8
	150-165	2	2	1	5
	165-180	3	5	2	10
	180-195	4	2	1	7
	195-210	3	3	0	6
	210-225	3	5	1	9
	225+	36	69	6	111
Total		214	340	61	**615**
Pearson Chi-Square Test: Value=32.221, df =30, Asymp. Sig. (2-sided)=0.357 Likelihood RatioTest: Value=33.296, df =30, Asymp. Sig. (2-sided)=0.310					
Mean		90.1121	93.57353	74.5082	**90.7561**
Median		45.8824	40.9844	36.528	**41.9837**
Mode		19.5	22.875	---	**20.1923**

Apparently it seems that mean & median control times for 60+ age group are comparatively smaller than other two groups. But none is observed to have a noticeable difference with that of the mean & median of the total study population. Both Chi-square test and likelihood ratio test give insignificant results, i.e, there is no association between control time and age. A paired *t*-test would be done if the distribution of control times would be normal, at least asymptotically. But they are extremely skewed.

3.3.2 Distribution of Control Time by Sex:

Since in Bangladesh, society is male dominated, the shade of sex differentiation is observed at every stage of life. The following table shows the distribution of control time by sex.

Table#3.3.2: Distribution of control time by Sex.

		Sex		Total
		0	1	
Control Time (day)	<15	61	47	108
	15-30	67	59	126
	30-45	50	42	92
	45-60	29	12	41
	60-75	17	21	38
	75-90	11	8	19
	90-105	5	12	17
	105-120	5	0	5
	120-135	5	8	13
	135-150	1	7	8
	150-165	3	2	5
	165-180	3	7	10
	180-195	4	3	7
	195-210	3	3	6
	210-225	7	2	9
	225+	56	55	111

Total	327	288	**615**
Pearson Chi-Square Test: Value=26.399, df=15, Asymp. Sig. (2-sided)=0.034 Likelihood RatioTest: Value=29.317, df=15, Asymp. Sig. (2-sided)=0.015			
Mean	87.43119	94.53125	**90.7561**
Median	40.6500	43.5714	**41.9837**
Mode	18.91	21.207	**20.1923**

Here we observe that neither modal nor median control times show remarkable difference between the sexes and also with the total study population, but the mean control time shows a somewhat difference. Both chi-square test & likelihood ratio test dictates that there exists association between sex and control time. Females are observed to take smaller time to control diabetes mellitus than males.

3.3.3 Distribution of Control Time by Marital Status:

Marital status is an important factor for the ailment/control of some diseases. So it may be analysed also in case of diabetes mellitus. The following table shows the distribution of control time by marital status.

Table#3.3.3: Distribution of control time by Marital Status.

		Marital Status		Total
		0	1	
Control Time (day)	<15	2	106	108
	15-30	2	124	126
	30-45	1	91	92
	45-60	1	40	41
	60-75	1	37	38
	75-90	0	19	19
	90-105	1	16	17
	105-120	0	5	5
	120-135	0	13	13
	135-150	1	7	8
	150-165	1	4	5
	165-180	0	10	10
	180-195	0	7	7
	195-210	0	6	6
	210-225	0	9	9
	225+	1	110	111
Total		11	604	**615**
Pearson Chi-Square Test: Value=18.584, df=15, Asymp. Sig. (2-sided)=0.233 Likelihood RatioTest: Value=10.132, df=15, Asymp. Sig. (2-sided)=0.811				
Mean		84.54545	90.86921	**90.7561**
Median		52.5000	41.8681	**41.9837**
Mode		15	20.29	**20.1923**

Unmarried patients are observed to have smaller mean & modal control time than their counterparts-married patients while they have larger median control time than their counterparts. The differences are not recognized by the chi-square test & likelihood ratio test because of extremely smaller sample size of unmarried patients in comparison to married group leading to the conclusion that there is no association between the marital status & control time of diabetes mellitus.

3.3.4 Distribution of Control Time by Food Habit:

Diet therapy is usually used for the treatment of diabetes and it often dictates that individuals limit their intake of foods high in fats or sugars. The addition of fibrous foods may be appealing to this population. During the past two decades, clinical investigations have attempted to characterize the benefits of dietary fiber for diabetes. Evidence that soluble dietary fibers are food components associated with health benefits is also supported by recent investigations. Tieyen J. (1989) in his paper had written that a flexible approach for inclusion of dietary fiber in the diets of individuals with diabetes might be used. Toeller M. (1991) in his paper said that a 'diabetic diet' can not only normalize blood glucose, but also can help reduce such diabetic-related risks as overweight, insulin resistance, dyslipoproteinemia and hypertension. The following tables show the distribution of control time by food habit (rice), food habit (vegetable) and food habit (sweet) respectively.

Table#3.3.4(a): Distribution of control time by food habit (rice).

		Food Habit Rice		
		0	1	Total
Control Time (day)	<15	57	51	108
	15-30	52	74	126
	30-45	43	49	92
	45-60	13	28	41
	60-75	13	25	38
	75-90	4	15	19
	90-105	9	8	17
	105-120	2	3	5
	120-135	7	6	13
	135-150	2	6	8
	150-165	1	4	5
	165-180	6	4	10
	180-195	4	3	7
	195-210	4	2	6
	210-225	2	7	9
	225+	40	71	111
Total		259	356	**615**
Pearson Chi-Square Test: Value=22.247, df=15, Asymp. Sig. (2-sided)=0.102 Likelihood RatioTest: Value=22.777, df=15, Asymp. Sig. (2-sided)=0.089				
Mean		84.20849	95.51966	**90.7561**
Median		37.1511	47.1429	**41.9837**
Mode		0	22.1875	**20.1923**

Both mean & median control times show remarkable difference between those patients who take rice thrice a day and those who take once or twice. The control times are observed to be larger for who take rice thrice a day. But chi-square test fails to identify this difference because in 12 cells the expected frequency is less than 5. The likelihood ratio test dictates this difference at about 9% level of significance. Based on these findings, we may conclude that there exists an association between food habit (rice) & control time, but it may be a weak one.

Table#3.3.4(b): Distribution of control time by food habit (vegetable).

		Food Habit (Vegetable)		
		0	1	Total
Control Time (day)	<15	94	14	108
	15-30	107	19	126
	30-45	76	16	92
	45-60	28	13	41
	60-75	30	8	38
	75-90	19	0	19
	90-105	14	3	17
	105-120	5	0	5
	120-135	12	1	13
	135-150	8	0	8
	150-165	5	0	5
	165-180	6	4	10
	180-195	6	1	7
	195-210	5	1	6
	210-225	8	1	9
	225+	91	20	111
Total		514	101	**615**
Pearson Chi-Square Test: Value=21.217, df=15, Asymp. Sig. (2-sided)=0.130 Likelihood RatioTest: Value=25.318, df=15, Asymp. Sig. (2-sided)=0.046				
Mean		90.14591	93.86139	**90.7561**
Median		41.0526	46.7308	**41.9837**
Mode		19.43	24.375	**20.1923**

The observed result dictates that control times for vegetarians are larger than that of non-vegetarians at mean, median and modal points. But chi-square test fails to identify this fact due to the reason that a large number of cells have expected frequency less than 5. The likelihood ratio test identifies this difference at 4.6% level of significance. Thus we conclude that association exists between vegetarian food habit & control time of diabetes mellitus.

Table#3.3.4(c): Distribution of control time by food habit (sweet).

		Food Habit (Sweet)		
		0	1	Total
Control Time (day)	<15	72	36	108
	15-30	71	55	126
	30-45	54	38	92
	45-60	28	13	41
	60-75	23	15	38
	75-90	6	13	19
	90-105	4	13	17
	105-120	3	2	5
	120-135	6	7	13
	135-150	6	2	8
	150-165	2	3	5
	165-180	6	4	10
	180-195	3	4	7
	195-210	4	2	6
	210-225	6	3	9
	225+	65	46	111
Total		359	256	**615**
Pearson Chi-Square Test: Value=22.654, df=15, Asymp. Sig. (2-sided)=0.092 Likelihood RatioTest: Value=22.793, df=15, Asymp. Sig. (2-sided)=0.089				
Mean		88.74652	93.57422	**90.7561**
Median		40.1389	44.6053	**41.9837**
Mode		0	22.92	**20.1923**

An inspection of the findings of this table show that patients who prefer sweets in his food habit have larger control times both at mean & median points. Both chi-square & likelihood ratio tests confirm this fact at 9.2% & 8.9% level of significance which indicates an weak association between control time and preference of sweet.

An overview of the results of table # 3.3.4(a), 3.3.4(b) & 3.3.4(c) show that all the three categories of food habit have association with diabetes mellitus and the patient takes less time on an average to control when he

(she) does not take rice thrice a day and does not prefer sweet. But in case of vegetarian the observed result does not match with reality.

3.3.5 Distribution of Control Time by BMI:

Overweight is an important factor for diseases. In our country there is proverb that fatty body is the house of diseases. Also in case of diabetes all the doctors suggest an overweighed person first to control his weight. Hence BMI is a main factor for diabetes mellitus. In our collected data the minimum value of BMI is 7.598877 and maximum is 37.999625. Mean, median and mode of the data are 23.6022, 23.83045 &22.499637, respectively. The following table shows the distribution of control time by BMI which are classified according to their median.

Table#3.3.5: Distribution of control time by BMI.

		BMI		Total
		<24	24+	
Control Time (day)	<15	51	57	108
	15-30	69	57	126
	30-45	47	45	92
	45-60	22	19	41
	60-75	20	18	38
	75-90	11	8	19
	90-105	12	5	17
	105-120	2	3	5
	120-135	9	4	13
	135-150	4	4	8
	150-165	3	2	5
	165-180	8	2	10
	180-195	2	5	7
	195-210	4	2	6
	210-225	4	5	9
	225+	56	55	111

Total	324	291	**615**
Pearson Chi-Square Test: Value=12.233, df=15, Asymp. Sig. (2-sided)= .719 Likelihood RatioTest: Value=12.918, df=15, Asymp. Sig. (2-sided)= .688			
Mean	91.2963	90.15464	**90.7561**
Median	43.40425	40.5	**41.9837**
Mode	21.75	---	**20.1923**

Here we observe that the patient with BMI <24 takes more time to control at mean and median point, but the difference is not noticeable with that of the total study population. The results of chi-square test and likelihood ratio test reveal that BMI does not have association with control time of diabetes mellitus.

3.3.6 Distribution of Control Time by BP:

Blood Pressure (BP) is an important factor in human body. If BP is abnormal (both high and low) there may happen any accident, especially in case of diabetes, heart disease, any type of operation etc. Hence BP can be a noted point for the control time of diabetes mellitus. Here we have considered for our study only the systolic blood pressure. . In our collected data mean, median and mode of BP are124.38, 120 &120, respectively. The following table shows the distribution of control time by BP which are classified according to their median.

Table#3.3.6: Distribution of control time by BP.

		BP		Total
		<120	120+	
Control Time (day)	<15	70	38	108
	15-30	79	47	126
	30-45	57	35	92
	45-60	25	16	41
	60-75	22	16	38
	75-90	13	6	19
	90-105	11	6	17
	105-120	2	3	5
	120-135	7	6	13
	135-150	4	4	8
	150-165	4	1	5
	165-180	5	5	10
	180-195	6	1	7
	195-210	5	1	6
	210-225	5	4	9
	225+	64	47	111
Total		379	236	**615**
Pearson Chi-Square Test: Value=6.706, df=15, Asymp. Sig. (2-sided)= .921				
Likelihood RatioTest: Value=6.598, df=15, Asymp. Sig. (2-sided)= 0.903				
Mean		88.41689	94.51271	**90.7561**
Median		40.65789	44.14286	**41.9837**
Mode		19.3548	21.42857	**20.1923**

The time to control at mean, median & modal points of both the groups are not observed to have a noticeable difference with that of total study population. Also from the results of chi-square test and likelihood ratio test we may conclude that BP does not have association with control time of diabetes mellitus.

3.3.7 Distribution of Control Time by BGL:

As we stated in the first chapter that Blood Glucose Level (BGL) is the main pivotal of the disease and there we have described in details the

relation of BGL with diabetes. We have also stated there that if the BGL of a person, 2 hours after 75 grms glucose load, is greater than or equal to 10 mmol/l, we can say that he (she) has diabetes mellitus. In our collected data, we have observed that the minimum value of BGL is 10 and maximum is 31.70. Mean, median & mode of the BGL are 15.4674, 15 & 10, respectively. The following table shows the distribution of control time by BGL, which are classified according to their median.

Table#3.3.7: Distribution of control time by BGL.

		BGL		Total
		<15	15+	
Control Time (day)	<15	56	52	108
	15-30	71	55	126
	30-45	51	41	92
	45-60	17	24	41
	60-75	27	11	38
	75-90	13	6	19
	90-105	6	11	17
	105-120	2	3	5
	120-135	8	5	13
	135-150	3	5	8
	150-165	4	1	5
	165-180	6	4	10
	180-195	3	4	7
	195-210	2	4	6
	210-225	4	5	9
	225+	51	60	111
Total		324	291	
Pearson Chi-Square Test: Value=39.702, df=30, Asymp. Sig. (2-sided)=0.111 Likelihood RatioTest: Value=41.363, df=30, Asymp. Sig. (2-sided)=0.081				
Mean		85.69444	96.39175	**90.7561**
Median		40.29412	44.08536	**41.9837**
Mode		21.43	17.647	**20.1923**

Here we observe that the patients of the group 15+ BGL takes larger control time than the patients of the group BGL<15. Chi-square test fails

to identify this difference due to the fact that a large number of cells have expected frequency less than 5. Likelihood ratio test identifies this difference at 8% level of significance. Based on these findings we may conclude that BGL has association with control time of diabetes mellitus, but it may be a weak one.

3.3.8 Distribution of Control Time by Heredity:

Heredity is another important factor for the disease. It is observed from recent investigations that a person having parents with diabetes have more chance to be on the risk set. The following table shows the distribution of control time by heredity.

Table#3.3.8: Distribution of control time by Heredity.

		Heredity		Total
		0	1	
Control Time (day)	<15	58	50	108
	15-30	67	59	126
	30-45	58	34	92
	45-60	24	17	41
	60-75	22	16	38
	75-90	8	11	19
	90-105	10	7	17
	105-120	1	4	5
	120-135	8	5	13
	135-150	4	4	8
	150-165	1	4	5
	165-180	9	1	10
	180-195	4	3	7
	195-210	3	3	6
	210-225	3	6	9
	225+	55	56	111

Total	335	280	**615**
Pearson Chi-Square Test: Value=17.573, df=15, Asymp. Sig. (2-sided)=0.286 Likelihood RatioTest: Value=18.757, df=15, Asymp. Sig. (2-sided)=0.225			
Mean	87.22388	94.98214	**90.7561**
Median	40.99137	43.67647	**41.9837**
Mode	22.5	18.97	**20.1923**

Here we observe that the patients having parents or blood connected relatives with diabetes mellitus takes larger control time than that of those without diabetes mellitus. But from the results of chi-square test and likelihood ratio test we may conclude that heredity is not associated with control time of diabetes mellitus.

3.3.9 Distribution of Control Time by Length of Sufferings:

Before taking the registration as a diabetic patient and starting the treatment how many days the patient suffer with the disease is termed as suffering time. The doctors find out this time by interviewing the patients about their symptoms. It may an important factor for the control time of diabetes mellitus. In our collected data, we have observed that mean, median & mode of the length of sufferings are 267.39, 7 & 0 days, respectively. The quartiles of the length of sufferings are 0, 7 & 120 days. We have grouped suffering time according to their quartiles. The following table shows the distribution of control time by suffering time.

Table#3.3.9: Distribution of control time by length of sufferings.

		length of sufferings (day)			Total
		<7	7-120	120+	
Control Time (day)	<15	72	16	20	108
	15-30	68	36	22	126
	30-45	51	22	19	92
	45-60	30	9	2	41
	60-75	21	8	9	38
	75-90	12	2	5	19
	90-105	5	7	5	17
	105-120	3	1	1	5
	120-135	4	3	6	13
	135-150	5	1	2	8
	150-165	3	1	1	5
	165-180	2	2	6	10
	180-195	5	0	2	7
	195-210	2	3	1	6
	210-225	6	0	3	9
	225+	43	22	46	111
Total		332	133	150	615
Pearson Chi-Square Test: Value=66.112, df=30, Asymp. Sig. (2-sided)=0.000 Likelihood RatioTest: Value=69.207, df=30, Asymp. Sig. (2-sided)=0.000					
Mean		78.20783	86.2782	122.5	**90.7561**
Median		37.64706	39.88636	74	**41.9837**
Mode		---	23.82	21	**20.1923**

From the above table we observe that the patient with larger suffering time takes larger control time at mean & median point. Also from the results of chi-square test and likelihood ratio test we may conclude that suffering time is highly associated with control time of diabetes mellitus.

3.4 Association of Other Diseases with Control Time:

The following diseases which are assumed to affect the control time of diabetes mellitus are investigated in this section:

1. Heart Problem (B) $= \begin{cases} 1, & \textit{if the patient is attacked by the disease before diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

2. Heart Problem (A) $= \begin{cases} 1, & \textit{if the patient is attacked by the disease after diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

3. Dental Problem (B)$= \begin{cases} 1, & \textit{if the patient is attacked by the disease before diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

4. Dental Problem (A) $= \begin{cases} 1, & \textit{if the patient is attacked by the disease after diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

5. Kidney Problem (B) $= \begin{cases} 1, & \textit{if the patient is attacked by the disease before diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

6. Kidney Problem (A)

$= \begin{cases} 1, & \textit{if the patient is attacked by the disease after diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

7. Eye Problem (B) $= \begin{cases} 1, & \textit{if the patient is attacked by the disease before diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

8. Eye Problem (A) $= \begin{cases} 1, & \textit{if the patient is attacked by the disease after diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

9. Other Problems (B) $= \begin{cases} 1, & \textit{if the patient is attacked by any one of the other diseases before diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

10. Other Problems (A)$= \begin{cases} 1, & \textit{if the patient is attacked by any one of the other diseases after diagnosis of diabetes mellitus} \\ 0, & \textit{if not} \end{cases}$

3.4.1 Distribution of Control Time by Heart Problem:

Heart problem has a high association with diabetes mellitus. The association of non-fasting blood glucose levels with the incidence of cardiovascular disease (CVD) was determined prospectively in 1382 men and 2094 women aged 45-84 years participating in the Framingham heart study conducted by Wilson P. W. F. et al. (1991) and the study showed that hyperglycemia in the original Framingham cohort is an independent risk factor for CVD in non-diabetic women, but not among men. Gries F. A. and Koschinsky T. (1991) also said that although arterial disease is not unique to diabetes, but there exists a relationship with the diabetic condition. Hence there may also exists a relation of heart problem with control time of diabetes mellitus. The following tables show the distribution of heart problem.

Table#3.4.1A: Distribution of Control Time by Heart Problem (B).

		Heart Problem (B)		Total
		0	1	
Control Time (day)	<15	100	8	108
	15-30	116	10	126
	30-45	88	4	92
	45-60	36	5	41
	60-75	38	0	38
	75-90	17	2	19
	90-105	17	0	17
	105-120	5	0	5
	120-135	13	0	13
	135-150	7	1	8
	150-165	4	1	5
	165-180	9	1	10
	180-195	7	0	7
	195-210	6	0	6
	210-225	8	1	9
	225+	104	7	111

Total	575	40	615
Pearson Chi-Square Test: Value =12.438, df =15, Asymp. Sig. (2-sided) =0.646 Likelihood RatioTest: Value =16.914, df =15, Asymp. Sig. (2-sided) =0.324			
Mean	97.01739	87	**90.7561**
Median	42.1875	37.5	**41.9837**
Mode	20.4545	18.75	**20.1923**

Here we observe that patients with heart problem (B) take smaller control time at mean, median & modal points than those without heart problem (B). The differences are not recognized by the chi-square test and likelihood ratio test because extremely smaller sample size with heart problem (B) patients in comparison to their counterparts leading to the conclusion that there is no association between heart problem (B) and control time of diabetes mellitus.

Table#3.4.1B: Distribution of Control Time by Heart Problem (A).

		Heart Problem (A)		Total
		0	1	
Control Time (day)	<15	99	9	108
	15-30	118	8	126
	30-45	87	5	92
	45-60	38	3	41
	60-75	37	1	38
	75-90	15	4	19
	90-105	14	3	17
	105-120	5	0	5
	120-135	12	1	13
	135-150	8	0	8
	150-165	4	1	5
	165-180	10	0	10
	180-195	7	0	7
	195-210	6	0	6
	210-225	7	2	9
	225+	94	17	111
Total		561	54	615

Pearson Chi-Square Test: Value =21.606, df =15, Asymp. Sig. (2-sided) =0.119 Likelihood RatioTest: Value =22.694, df =15, Asymp. Sig. (2-sided) =0.091			
Mean	88.155508	117.7778	**90.7561**
Median	40.94827	78.75	**41.9837**
Mode	20.7	---	**20.1923**

From the above table we observe that patients with heart problem (A) take larger control time at mean & median points than those without heart problem (A). Chi-square test fails to find-out this difference due to the fact that a large number of expected cell frequencies are smaller than 5. The likelihood ratio test dictates this at 9.1% level of significance. Based on these findings we may conclude that there exists association between control time and heart disease, when patients are attacked by this problem after the diagnosis of diabetes mellitus, but this association may be a weak one.

3.4.2 Distribution of Control Time by Dental Problem:

Dental problem also may have association with control time of diabetes mellitus. The following table shows the distribution of control time by dental problem (B).

Table#3.4.2A: Distribution of Control Time by Dental Problem (B).

		Dental Problem (B)		Total
		0	1	
Control Time (day)	<15	68	40	108
	15-30	96	30	126
	30-45	72	20	92
	45-60	29	12	41
	60-75	24	14	38
	75-90	18	1	19
	90-105	13	4	17
	105-120	5	0	5
	120-135	12	1	13
	135-150	6	2	8
	150-165	4	1	5
	165-180	7	3	10
	180-195	5	2	7
	195-210	5	1	6
	210-225	6	3	9
	225+	86	25	111
Total		456	159	615
Pearson Chi-Square Test: Value =20.385, df =15, Asymp. Sig. (2-sided) =0.158 Likelihood RatioTest: Value =23.030, df =15, Asymp. Sig. (2-sided) =0.084				
Mean		93.68421	82.35849	**90.7561**
Median		43.333	37.125	**41.9837**
Mode		23.0769	---	**20.1923**

From the above table we observe that patients with dental problem (B) take smaller control time at mean & median points than those without dental problem (B). Chi-square test fails to find-out this difference due to the fact that a large number of cells have expected frequencies smaller than 5. The likelihood ratio test dictates this at 8.4% level of significance. Based on these findings we may conclude that there exists association between control time and dental problem (B), but this association may be a weak one.

Table#3.4.2B: Distribution of control time by Dental Problem (A).

		Dental Problem (A)		Total
		0	1	
Control Time (day)	<15	97	11	108
	15-30	97	29	126
	30-45	74	18	92
	45-60	29	12	41
	60-75	34	4	38
	75-90	13	6	19
	90-105	11	6	17
	105-120	2	3	5
	120-135	9	4	13
	135-150	5	3	8
	150-165	4	1	5
	165-180	10	0	10
	180-195	6	1	7
	195-210	6	0	6
	210-225	8	1	9
	225+	80	31	111
Total		485	130	615
Pearson Chi-Square Test: Value =30.294, df =15, Asymp. Sig. (2-sided) =0.011 Likelihood RatioTest: Value =33.628, df =15, Asymp. Sig. (2-sided) =0.004				
Mean		87.43299	103.1538	**90.7561**
Median		39.8311	53.75	**41.9837**
Mode		---	24.31034	**20.1923**

Here we observe that the patients with dental complication after the diagnosis of diabetes mellitus take larger control time at mean and median points. Both chi-square test and likelihood ratio test identify this fact at 1.1% & 0.4% level of significance which indicates an strong association between control time of diabetes mellitus and dental problem after diagnosis of diabetes mellitus.

3.4.3 Distribution of Control Time by Kidney Problem:

Kidney is another important organ of human body and ofcourse for diabetes mellitus. Weir M.R. & Bakris G.L. (1992) gave special attention to the pathophysiology of the hypertensive disease and how it affects the kidneys and they said that pre-existing accentuated vascular reactivity to vasoconstrictive growth factors in diabetic patients stimulates maladaptive compensatory responses in the kidney. Schimitz A. and et al. (1990) compared a group of 19 microalbuminuric NIDDM patients, of mean age 65 ± 4 yrs, with 19 randomly selected matched normoalbuminuric patients and observed that kidney volume was enhanced in NIDDM patients. Hence kidney may be affected or damaged that may have impact on control time of diabetes mellitus.

Table#3.4.3A: Distribution of Control Time by Kidney Problem (B).

		Kidney Problem (B)		Total
		0	1	
Control Time (day)	<15	108	0	108
	15-30	126	0	126
	30-45	91	1	92
	45-60	41	0	41
	60-75	38	0	38
	75-90	19	0	19
	90-105	17	0	17
	105-120	5	0	5
	120-135	13	0	13
	135-150	6	2	8
	150-165	5	0	5
	165-180	10	0	10
	180-195	7	0	7
	195-210	6	0	6
	210-225	9	0	9
	225+	111	0	111
Total		612	3	615

Pearson Chi-Square Test: Value =102.227, df =15, Asymp. Sig. (2-sided) =0.000 Likelihood RatioTest: Value =17.893, df =15, Asymp. Sig. (2-sided) =0.268			
Mean	90.63725	115	**90.7561**
Median	41.8681	---	**41.9837**
Mode	20.0943	---	**20.1923**

From the above table we observe that one category has so small number of observation which is negligible in contest to the other category. Hence the test reults may not be acceptable.

Table#3.4.3B: Distribution of control time by Kidney Problem (A).

		Kidney Problem (A)		Total
		0	1	
Control Time (day)	<15	106	2	108
	15-30	126	0	126
	30-45	92	0	92
	45-60	40	1	41
	60-75	38	0	38
	75-90	19	0	19
	90-105	17	0	17
	105-120	5	0	5
	120-135	13	0	13
	135-150	7	1	8
	150-165	5	0	5
	165-180	10	0	10
	180-195	7	0	7
	195-210	6	0	6
	210-225	9	0	9
	225+	109	2	111
Total		609	6	615
Pearson Chi-Square Test: Value =16.968, df =15, Asymp. Sig. (2-sided) =0.321 Likelihood RatioTest: Value =12.121, df =15, Asymp. Sig. (2-sided) =0.670				
Mean		90.46798	120	**90.7561**
Median		41.8206	60	**41.9837**
Mode		20.5556	---	**20.1923**

Here we also observe that one category takes a very small size of sample and we can not draw any conclusion on the basis of these results.

3.4.4 Distribution of Control Time by Eye Problem:

After conducting a survey on diabetes in UK Nabarro J. D. N (1991) observed that blindness occurred in 0.28% of patients with type-I diabetes per year and 0.097% per year in type-II diabetes and he had said that if the mean survival of patients with retinopathy going blind is 7.5 years, this would mean 7500 people in the UK would be blind from diabetic retinopathy. Hence eye problem may have association with control time of diabetes mellitus. We have used the code 1, if the patient have the problem and 0, if not.

Table#3.4.4A: Distribution of control time by Eye Problem (B).

		Eye Problem (B)		Total
		0	1	
Control Time (day)	<15	55	53	108
	15-30	78	48	126
	30-45	57	25	92
	45-60	23	18	41
	60-75	22	16	38
	75-90	15	4	19
	90-105	12	5	17
	105-120	5	0	5
	120-135	9	4	13
	135-150	6	2	8
	150-165	4	1	5
	165-180	8	2	10
	180-195	6	1	7
	195-210	3	3	6
	210-225	3	6	9
	225+	70	41	111

Total	376	239	615
Pearson Chi-Square Test: Value =20.204, df =15, Asymp. Sig. (2-sided) =0.164 Likelihood RatioTest: Value =22.528, df =15, Asymp. Sig. (2-sided) =0.095			
Mean	94.30851	85.16736	**90.7561**
Median	44.4737	37.9286	**41.9837**
Mode	22.8409	---	**20.1923**

Here we observe that the patients with eye problem before diagnosis diabetes mellitus take smaller control time at mean & modal points. Chi-square test fails to identify this fact because a large number of cells having expected frequency less than 5. Likelihood ratio test dictates this at 9.5% level of significance. Based on these results we may conclude that eye problem (B) has association with control time of diabetes mellitus, but this association may be a weak one.

Table#3.4.4B: Distribution of control time by Eye Problem (A).

		Eye Problem (A)		Total
		0	1	
Control Time (day)	<15	92	16	108
	15-30	87	39	126
	30-45	61	31	92
	45-60	31	10	41
	60-75	28	10	38
	75-90	6	13	19
	90-105	12	5	17
	105-120	4	1	5
	120-135	7	6	13
	135-150	3	5	8
	150-165	2	3	5
	165-180	5	5	10
	180-195	3	4	7
	195-210	6	0	6
	210-225	8	1	9
	225+	69	42	111
Total		424	191	615

Pearson Chi-Square Test: Value =45.245, df =15, Asymp. Sig. (2-sided) =0.000 Likelihood RatioTest: Value =47.089, df =15, Asymp. Sig. (2-sided) =0.000			
Mean	84.44575	104.7644	**90.7561**
Median	38.1147	59.25	**41.9837**
Mode	---	26.129	**20.1923**

Here we observe that the patients with eye problem after the diagnosis of diabetes mellitus take larger control time at mean and median points. Both chi-square test and likelihood ratio test dictate this fact at 0.0% & 0.0% level of significance which indicates existence of strong association between control time of diabetes mellitus and eye problem after diagnosis of diabetes mellitus.

3.4.5 Distribution of Control Time by Other Problems:

Among other diseases foot problem, ulcer, various type of allergy, hypertension, neurological complication, gynecological problem, nephropathy etc. are observed to have association with diabetes. All these are included in the 'other problems'.

Table#3.4.5A: Distribution of control time by Other Problems (B).

		Other Problems (B)		Total
		0	1	
Control Time (day)	<15	67	41	108
	15-30	70	56	126
	30-45	48	44	92
	45-60	17	24	41
	60-75	23	15	38
	75-90	14	5	19
	90-105	13	4	17
	105-120	3	2	5
	120-135	9	4	13
	135-150	7	1	8
	150-165	2	3	5
	165-180	7	3	10
	180-195	2	5	7
	195-210	3	3	6
	210-225	4	5	9
	225+	70	41	111
Total		359	256	615
Pearson Chi-Square Test: Value =20.620, df =15, Asymp. Sig. (2-sided) =0.149 Likelihood RatioTest: Value =21.264, df =15, Asymp. Sig. (2-sided) =0.129				
Mean		93.50975	86.89453	**90.7561**
Median		43.28125	40.5682	**41.9837**
Mode		15.6	23.3333	**20.1923**

Here we observe that patients with other problems (B) take smaller control time at mean & median points than those of without problems, but the converse is observed at modal point. From the results of chi-square test and likelihood ratio test we may conclude that there is no association between other problems (B) and control time of diabetes mellitus.

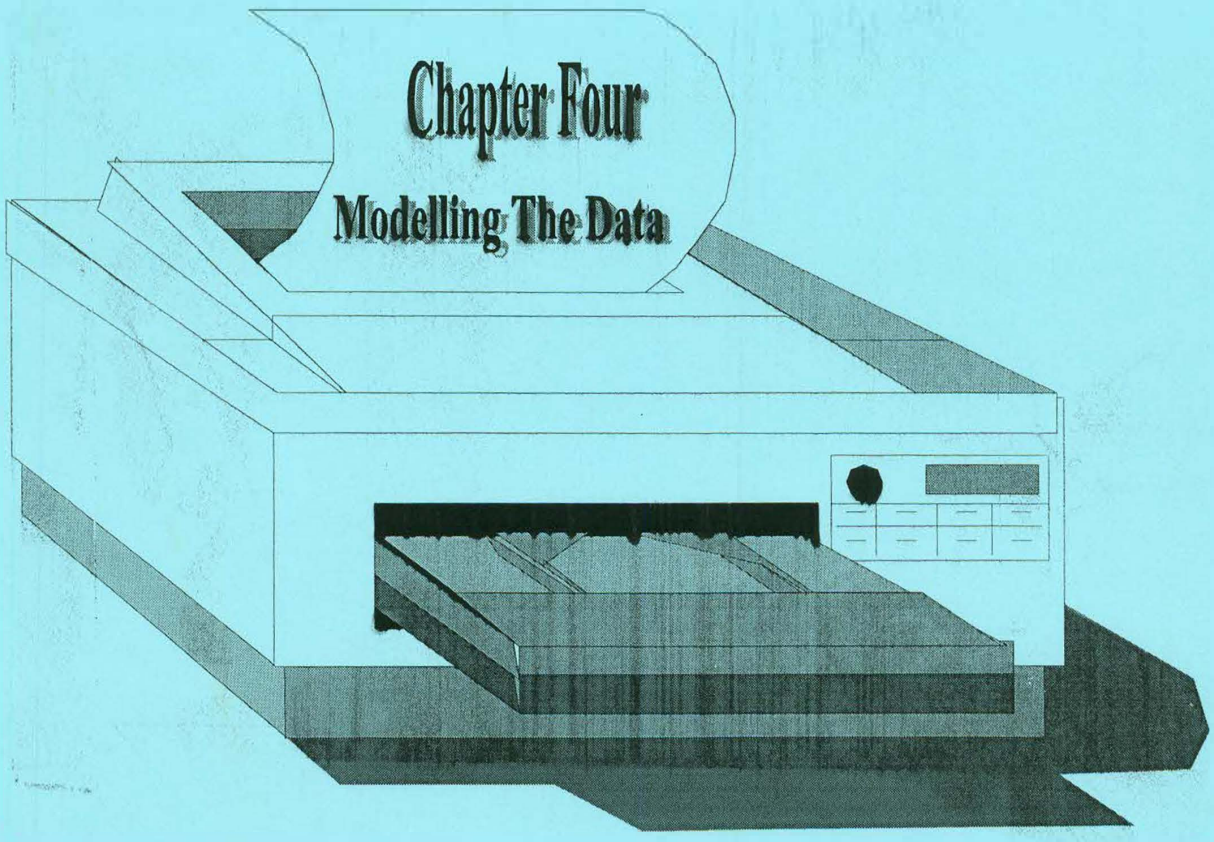
Table#3.4.5B: Distribution of control time by Other Problems (A).

		Other Problems (A)		Total
		0	1	
Control Time (day)	<15	73	35	108
	15-30	85	41	126
	30-45	65	27	92
	45-60	26	15	41
	60-75	27	11	38
	75-90	9	10	19
	90-105	13	4	17
	105-120	3	2	5
	120-135	8	5	13
	135-150	4	4	8
	150-165	4	1	5
	165-180	8	2	10
	180-195	4	3	7
	195-210	4	2	6
	210-225	8	1	9
	225+	70	41	111
Total		411	204	615
Pearson Chi-Square Test: Value =10.585, df =15, Asymp. Sig. (2-sided) =0.781 Likelihood RatioTest: Value =10.855, df =15, Asymp. Sig. (2-sided) =0.763				
Mean		89.27007	93.75	**90.7561**
Median		40.9615	44.4444	**41.9837**
Mode		20.625	19.5	**20.1923**

From the above table we observe that patients with other problems (A) take larger control time at mean & median points than those of without problems, but at modal point the situation is converse. From the results of chi-square test and likelihood ratio test we may conclude that there is no association between other problems (A) and control time of diabetes mellitus.

Chapter Four

Modelling The Data



Chapter Four

Modelling the Data

4.1 Introduction:

We know that when a patient is affected by a disease he (she) tries to get cure of it and the time that the disease takes to be cured from the date of diagnosis may be considered as the lifetime of the disease. Diabetes is not curable but can be controlled. So, a person getting registered as a diabetic patient in a diabetic center may be considered alive in the queue waiting to be in the control state and when he (she) is in control state of the disease, may be considered to be dead in the queue. The waiting time in the queue to be in the control state may be considered as his lifetime of diabetes state. If a suitable statistical probability distribution for these lifetimes can be explored, probabilistic statement for expected time to reach the control state can be given in advance. In addition to that the rate at which patients

are reaching the control state can be inferred. These informations help both the patient and diabetic centers to administer the disease.

An overview of the frequency distribution displayed in table#3.2 of chapter three, it may be inferred that the data may fit a distribution from the Exponential family of distributions. In the exponential family three distributions are most usable and prominent life distributions: Negative Exponential distribution, Weibull distribution and Gamma distribution. The Exponential distribution occupies an important position in graduating lifetime data. It is the first lifetime model for which statistical methods were extensively developed. The Weibull distribution is also an important life distribution and a large body of literature on statistical methods has evolved for it. One reason that so many works have been done on this distribution concern is its statistical properties. Although the distributions of most estimators and other statistics associated with the Weibull distribution are mathematically intractable and has led to extensive work in producing tables for carrying out inferences, but good statistical procedures that are relatively easy to use are now available

After the exponential and Weibull distributions, the most frequently used model is the gamma distribution of the exponential family. This model has some features not possessed by the more common parametric distributions, e.g., the flexibility to give a U-shaped hazard function. With uncensored data, some inference procedures for the gamma distribution are fairly straightforward. However, when they are censored, or when it is desired to obtain interval estimates of quantities such as quantiles or the survivor function of the distribution, matters are more complicated. This is one reason why the gamma distribution is less widely used as a lifetime

distribution than is the exponential or Weibull distribution. However, the gamma distribution is nevertheless a useful model. In this chapter, we have investigated these three distributions to single out the closest fit one to our data.

4.2 Graphical Plot:

Plots of estimated survivor or cumulative hazard functions provide useful pictures of lifetime data, as well as information on the underlying life distribution. They can also be used for informal checks on the appropriateness of a model and for obtaining parameter estimates of the assumed model. The basic idea is to make plots that should be roughly linear if the proposed model is appropriate. For this one can use cumulative hazard function H(t)=-logS(t), where $S(t)=\prod_{J=1}^{K+1}\frac{N_J - D_J}{N_J}$, *Dj* representing the number of subjects getting control of diabetes and leaving the queue at time *tj* and *Nj* be the corresponding number of subjects suffering from diabetes and remaining in the queue just prior to *tj*, is the Kaplan-Meier estimate, sometimes called "product-limit" (PL) estimate.

From the frequency distribution of the previous chapter we have observed that the data may fit a member of exponential family. Negative Exponential, Weibull and Gamma distributions are the three probable important members of this family which may graduate our data. Now for exponential distribution with survivor function

$$S(t)=\exp\left\{-\left(\frac{t-\mu}{\theta}\right)\right\} \quad , \quad t \geq \mu,\ \theta > 0$$

$$\Rightarrow -\log_e S(t) = -\frac{\mu}{\theta}+\frac{t}{\theta}$$

$$\Rightarrow H(t) = -\frac{\mu}{\theta} + \frac{t}{\theta}$$

Hence $H(t)$ is linearly related to t and so if the graphical plot of $H(t)$ against t shows roughly linear, then we can say that the data may fit Exponential distribution. On the other hand for Weibull distribution with survivor function

$$S(t) = \exp\left[-(\lambda t)^{\beta}\right],\ t > 0$$

$$\Rightarrow H(t) = (\lambda t)^{\beta}$$

$$\Rightarrow \log_e H(t) = \beta \log_e \lambda + \beta \log_e t$$

Where $\log_e H(t)$is linearly related to $\log_e t$. Hence if the graphical plot of $\log_e H(t)$ against $\log_e t$ shows roughly straight line then we can say that the data may fit Weibull distribution.

In case of Gamma distribution, $H(t)$ is approximately a third degree polynomial in t having no intercept term (Mian, 1987) giving an elongated S-shaped curve. If a significant intercept term is observed, then we may suspect the presence of covariates in the data.

Hence before going to model we first plot H(t)=-logS(t), against t, logH(t) against log t to obtain the primary information about the distribution of the lifetime and the corresponding plots are shown in figure-1 and figure-2, which is done by the computer program Microsoft Excel. For the purpose necessary informations are shown in the following table.

Table#4.2.1: Survivor probability estimate.

Class Interval	Tj	Dj	Nj	(Nj-Dj)/Nj	S(Tj)	H(Tj)	LN(Tj)	LN(H(Tj))
<15	15	108	615	0.82439	0.82439	0.193111	2.70805	-1.64449
16-30	30	126	507	0.751479	0.619512	0.478823	3.401197	-0.73642
31-45	45	92	381	0.75853	0.469919	0.755196	3.806662	-0.28078
46-60	60	41	289	0.858131	0.403252	0.908194	4.094345	-0.0963
61-75	75	38	248	0.846774	0.341463	1.074515	4.317488	0.071869
76-90	90	19	210	0.909524	0.310569	1.169349	4.49981	0.156447
91-105	105	17	191	0.910995	0.282927	1.262567	4.65396	0.233147
106-120	120	5	174	0.971264	0.274797	1.291724	4.787492	0.255977
121-135	135	13	169	0.923077	0.253659	1.371766	4.905275	0.316099
136-150	150	8	156	0.948718	0.24065	1.42441	5.010635	0.353758
151-165	165	5	148	0.966216	0.23252	1.458778	5.105945	0.377599
166-180	180	10	143	0.93007	0.21626	1.531273	5.192957	0.4261
181-195	195	7	133	0.947368	0.204878	1.58534	5.273	0.460799
196-210	210	6	126	0.952381	0.195122	1.634131	5.347108	0.491111
211-225	225	9	120	0.925	0.180488	1.712092	5.4161	0.537716
226+	∞	111	111	0	0		0	

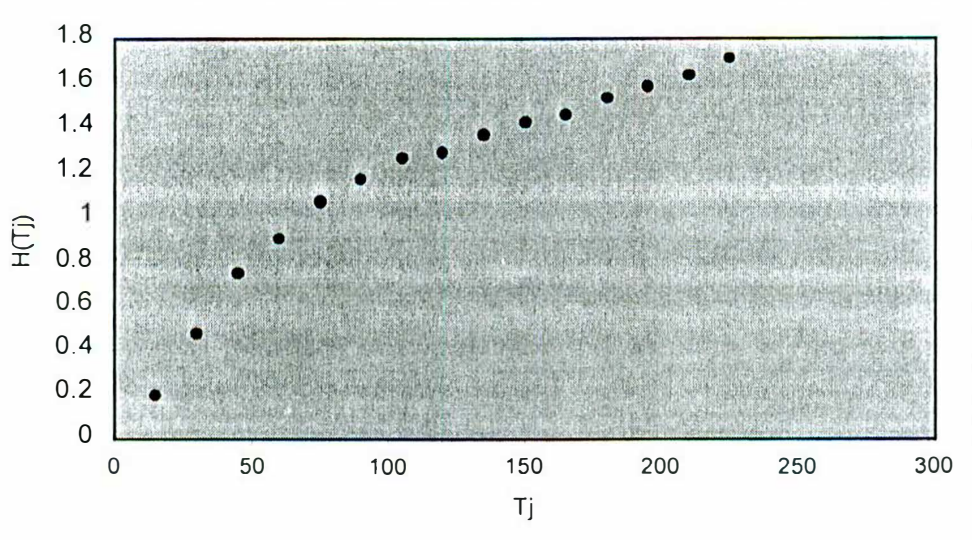


Figure-1

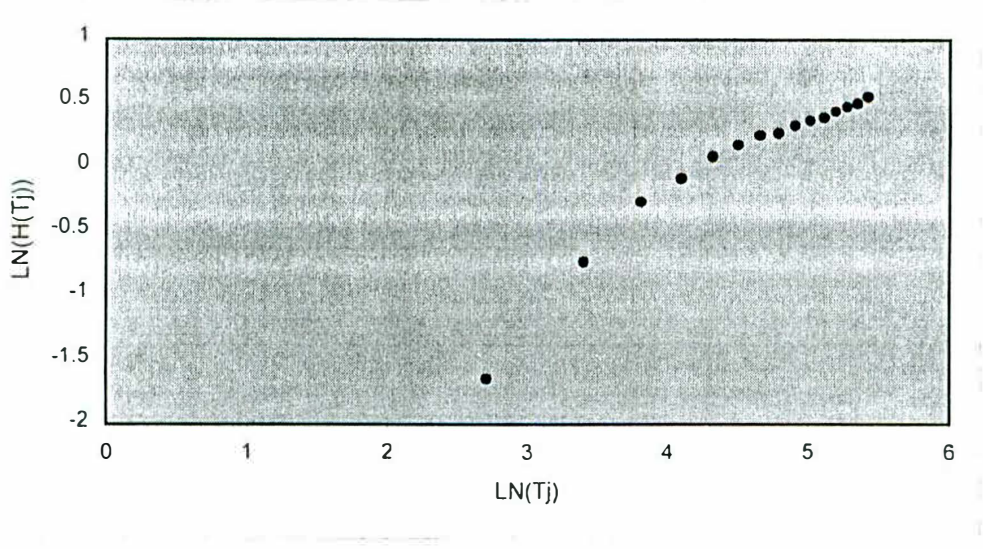


Figure-2

From the above plots we observe that figure-2 shows to be more linear and more concentrated than figuer-1. That is, the data may fit Weibull distribution better than the negative exponential distribution. The possibility of a gamma distribution is by far the least.

4.3 Least Squares Estimation:

Least square estimation of a model parameter is possible iff the survivor function or any transformation of it can be expressed as a linear function of time or function of time or a linear function of model parameter.

4.3.1 The Exponential Distribution:

The p.d.f. of the Exponential distribution is of the form

$$f(t;\mu,\theta)=\frac{1}{\theta}\exp\left\{-\left(\frac{t-\mu}{\theta}\right)\right\} \quad , \quad t\geq\mu,\ \theta>0$$

with survivor function

$$S(t) = \exp\left\{-\left(\frac{t-\mu}{\theta}\right)\right\} \quad , \quad t \geq \mu,\ \theta > 0$$

$$\Rightarrow H(t) = -\frac{\mu}{\theta} + \frac{t}{\theta}$$

$$\Rightarrow \hat{H}(t) = -\frac{\mu}{\theta} + \frac{t}{\theta} + U \text{, where } \hat{H}(t) \text{ is the estimated CHF and U is a random error.}$$

$$\Rightarrow Y = A + BT + U$$

where $A = -\frac{\mu}{\theta}$, $B = \frac{1}{\theta}$ and $Y = \hat{H}(t)$.

Applying the OLS method we can get the estimates as

$$\hat{B} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(T_i - \bar{T})}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \quad \text{and} \quad \hat{A} = \bar{Y} - \hat{B}\bar{X}$$

and for grouped data

$$\hat{B} = \frac{\sum_{i=1}^{n} D_i(Y_i - \bar{Y})(T_i - \bar{T})}{\sum_{i=1}^{n} D_i(T_i - \bar{T})^2} \quad \text{and} \quad \hat{A} = \bar{Y} - \hat{B}\bar{X}$$

Hence $\hat{\theta} = \frac{1}{\hat{B}}$ and $\hat{\mu} = -\frac{\hat{A}}{\hat{B}}$

4.3.2 The Weibull Distribution:

The p.d.f. of Weibull distribution is

$$f(t;\lambda,\beta) = \lambda\beta(\lambda t)^{\beta-1}\exp\left[-(\lambda t)^{\beta}\right],\ t > 0$$

with survivor function

$$S(t) = \exp\left[-(\lambda t)^{\beta}\right],\ t > 0$$

$$\Rightarrow H(t) = (\lambda t)^{\beta}$$

$$\Rightarrow \log_e H(t) = \beta \log_e \lambda + \beta \log_e t$$

$$\Rightarrow \log_e \hat{H}(t) = \beta \log_e \lambda + \beta \log_e t + \in$$

$$\Rightarrow Y = A + BT + \in$$

where $A = \beta \log_e \lambda$, $B = \beta$, $T = \log_e t$ and $Y = \log_e \hat{H}(t)$ and $\in$ is a random error.

Applying the OLS method we can get the estimates as

$$\hat{B} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(T_i - \bar{T})}{\sum_{i=1}^{n} (T_i - \bar{T})^2} \quad \text{and} \quad \hat{A} = \bar{Y} - \hat{B}\bar{X}$$

and for grouped data

$$\hat{B} = \frac{\sum_{i=1}^{n} D_i (Y_i - \bar{Y})(T_i - \bar{T})}{\sum_{i=1}^{n} D_i (T_i - \bar{T})^2} \quad \text{and} \quad \hat{A} = \bar{Y} - \hat{B}\bar{X}$$

Hence $\hat{\beta} = \hat{B}$ and $\hat{\lambda} = e^{\hat{A}}$.

4.3.3 The Gamma Distribution:

The p.d.f. of Gamma distribution is

$$\frac{1}{\alpha \Gamma(k)} \left(\frac{t}{\alpha}\right)^{k-1} \exp\left(-\frac{t}{\alpha}\right), \qquad t > 0$$

where $\alpha > 0$ and $k > 0$ are unknown scale and shape parameters, respectively.

According to Mian (1987) we can write

$$H(t) \approx \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$$

where β_i's are functions of α and k.

For Gamma distribution the parameters are estimated from the raw data using the program SPSS for windows. For Exponential and Weibull distribution we have used grouped data and values of t_i are taken as the upper limit of each interval. Grouped data is also used to test the goodness-of-fit of the fitted distributions. Results of the least square estimates of the parameters of exponential and Weibull distrbution with coefficient of determinant R^2 are computed by using the computer program Microsoft Excel which are as follows:

Exponential distribution:	Weibull distribution:	Gamma distribution:
$R^2 = .893$	$R^2 = .927$	$R^2 = .901$
$Adjusted\ R^2 = .884$	$Adjusted\ R^2 = .922$	$Adjusted\ R^2 = .901$
$\hat{A} = .269261^{**}$	$\hat{A} = -3.7951^{***}$	$\hat{\beta}_0 = 0.323^{***}$
$\hat{B} = 0.007928^{***}$	$\hat{B} = .871393^{***}$	$\hat{\beta}_1 = 5876E-03^{***}$
$\hat{\mu} = -33.9616$	$\hat{\lambda} = 0.01284$	$\hat{\beta}_2 = -3.667E-06^{***}$
$and\ \hat{\theta} = 126.1287$	$and\ \hat{\beta} = .871393$	$and\ \hat{\beta}_3 = 8.002E-10^{***}$

** = 0.1% level of significance.

*** = 0.0% level of significance.

Here we observe that the value of coefficient of determinant R^2 of exponential model, weibull model and gamma model are 0.893, 0.927 & 0.901, respectively. We also observe that in gamma model, there exists an intercept term with p = 0.000. Hence we can say that there exists covariates in the data.

4.4 Maximum Likelihood Estimation:

We know that the likelihood function for the grouped data is given by

$$L = \prod_{j=1}^{k+1} P_j^{d_j}$$

where P_j is the probability of death or failure in the j-th interval and d_j is the number of failures or death in I_j. Let $S(a_j)$ is the corresponding probability of survival. Then we can write

$$L = \prod_{j=1}^{k+1} [S(a_{j-1}) - S(a_j)]^{d_j}$$

4.4.1 The Exponential Distribution:

In case of Exponential distribution with survivor function

$$S(t) = \exp\left\{-\left(\frac{t-\mu}{\theta}\right)\right\} \quad , \quad t \geq \mu, \theta > 0$$

$$L(\mu,\theta) = \prod_{j=1}^{k+1}\left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right) - \exp\left(-\frac{t_j-\mu}{\theta}\right)\right]^{d_j}$$

$$\Rightarrow \log L = \sum_{j=1}^{k+1} d_j \log\left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right) - \exp\left(-\frac{t_j-\mu}{\theta}\right)\right]$$

$$\Rightarrow \frac{\partial \log L}{\partial \mu} = \sum_{j=1}^{k+1} d_j \left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right) - \exp\left(-\frac{t_j-\mu}{\theta}\right)\right]^{-1} \frac{1}{\theta}\left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right) - \exp\left(-\frac{t_j-\mu}{\theta}\right)\right]$$

$$= \frac{\sum d_j}{\theta}$$

$$= \frac{n}{\theta}$$

and $$\frac{\partial \log L}{\partial \theta} = \sum_{j=1}^{k+1} d_j \left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right) - \exp\left(-\frac{t_j-\mu}{\theta}\right)\right]^{-1}\left[\left(\frac{t_{j-1}-\mu}{\theta^2}\right)\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right)\right.$$

$$-\left(\frac{t_j-\mu}{\theta^2}\right)\exp\left(-\frac{t_j-\mu}{\theta}\right)\Bigg]$$

Hence the estimating equations are

$$\mu:\ \frac{n}{\theta}=0 \text{ and}$$

$$\theta:\ \sum_{j=1}^{k+1} d_j\left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right)-\exp\left(-\frac{t_j-\mu}{\theta}\right)\right]^{-1}\left[\left(\frac{t_{j-1}-\mu}{\theta^2}\right)\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right)\right.$$

$$\left.-\left(\frac{t_j-\mu}{\theta^2}\right)\exp\left(-\frac{t_j-\mu}{\theta}\right)\right]=0$$

We observe that ML equation for μ does not contain μ. It is due to the fact that μ lies in the lower limit of the range of the distribution of t and thus violating the regularity condition of ML method. In such a case $L(\mu,\theta)$ will be maximum for $\hat{\mu}=t_{(1)}$, the lowest order statistic of the observations. Thus the MLE of μ is

$$\hat{\mu}=t_{(1)}$$

Now solving ML equation for θ by Newton-Raphson iteration we can get the estimate easily, where

$$f=\sum_{j=1}^{k+1} d_j\left[\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right)-\exp\left(-\frac{t_j-\mu}{\theta}\right)\right]^{-1}\left[\left(\frac{t_{j-1}-\mu}{\theta^2}\right)\exp\left(-\frac{t_{j-1}-\mu}{\theta}\right)\right.$$

$$\left.-\left(\frac{t_j-\mu}{\theta^2}\right)\exp\left(-\frac{t_j-\mu}{\theta}\right)\right]$$

and

$$f' = \sum_{j=1}^{k+1} d_j \left[\left[\exp\left(-\frac{t_{j-1}-\mu}{\theta} \right) - \exp\left(-\frac{t_j-\mu}{\theta} \right) \right]^{-1} \left[\left(\frac{t_{j-1}-\mu}{\theta^2} \right)^2 \exp\left(-\frac{t_{j-1}-\mu}{\theta} \right) \right.\right.$$

$$\left. - 2\left(\frac{t_{j-1}-\mu}{\theta^3} \right) \exp\left(-\frac{t_{j-1}-\mu}{\theta} \right) - \left(\frac{t_j-\mu}{\theta^2} \right)^2 \exp\left(-\frac{t_j-\mu}{\theta} \right) + 2\left(\frac{t_j-\mu}{\theta^3} \right) \exp\left(-\frac{t_j-\mu}{\theta} \right) \right]$$

$$- \left[\exp\left(-\frac{t_{j-1}-\mu}{\theta} \right) - \exp\left(-\frac{t_j-\mu}{\theta} \right) \right]^{-2} \left[\left(\frac{t_{j-1}-\mu}{\theta^2} \right) \exp\left(-\frac{t_{j-1}-\mu}{\theta} \right) \right.$$

$$\left. \left. - \left(\frac{t_j-\mu}{\theta^2} \right) \exp\left(-\frac{t_j-\mu}{\theta} \right) \right]^2 \right]$$

4.4.2 The Weibull Distribution:

In case of Weibull distribution with survivor function $S(t) = \exp\left[-(\lambda t)^{\beta} \right]$, $t > 0$ the maximum likelihood function is

$$L(\lambda, \beta) = \prod_{j=1}^{k+1} \left[\exp\left[-(\lambda t_{j-1})^{\beta} \right] - \exp\left[-(\lambda t_j)^{\beta} \right] \right]^{d_j}$$

$$\Rightarrow \log L = \sum_{j=1}^{k+1} d_j \log\left[\exp\left[-(\lambda t_{j-1})^{\beta} \right] - \exp\left[-(\lambda t_j)^{\beta} \right] \right]$$

$$\Rightarrow \frac{\partial \log L}{\partial \lambda} = \sum_{j=1}^{k+1} d_j \left[\exp\left[-(\lambda t_{j-1})^{\beta} \right] - \exp\left[-(\lambda t_j)^{\beta} \right] \right]^{-1} \left[-\frac{\beta}{\lambda} (\lambda t_{j-1})^{\beta} \exp\left[-(\lambda t_{j-1})^{\beta} \right] \right.$$

$$\left. + \frac{\beta}{\lambda} (\lambda t_j)^{\beta} \exp\left[-(\lambda t_j)^{\beta} \right] \right]$$

and

$$\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^{k+1} d_j \left[\exp\left[-(\lambda t_{j-1})^{\beta} \right] - \exp\left[-(\lambda t_j)^{\beta} \right] \right]^{-1} \left[-(\lambda t_{j-1})^{\beta} \log(\lambda t_{j-1}) \exp\left[-(\lambda t_{j-1})^{\beta} \right] \right.$$

$$\left. + (\lambda t_j)^{\beta} \log(\lambda t_j) \exp\left[-(\lambda t_j)^{\beta} \right] \right]$$

Hence the estimating equations are

$$\sum_{j=1}^{k+1} d_j \left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-1} \left[-\frac{\beta}{\lambda}(\lambda t_{j-1})^{\beta} \exp\left[-(\lambda t_{j-1})^{\beta}\right] + \frac{\beta}{\lambda}(\lambda t_j)^{\beta} \exp\left[-(\lambda t_j)^{\beta}\right]\right] = 0$$

and

$$\sum_{j=1}^{k+1} d_j \left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-1} \left[-(\lambda t_{j-1})^{\beta} \log(\lambda t_{j-1}) \exp\left[-(\lambda t_{j-1})^{\beta}\right] + (\lambda t_j)^{\beta} \log(\lambda t_j) \exp\left[-(\lambda t_j)^{\beta}\right]\right] = 0$$

The estimates will be obtained by solving these equations by iteration simultaneously, where

$$f = \sum_{j=1}^{k+1} d_j \left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-1} \left[-\frac{\beta}{\lambda}(\lambda t_{j-1})^{\beta} \exp\left[-(\lambda t_{j-1})^{\beta}\right] + \frac{\beta}{\lambda}(\lambda t_j)^{\beta} \exp\left[-(\lambda t_j)^{\beta}\right]\right]$$

$$f' = \sum_{j=1}^{k+1} d_j \Bigg(\left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-1} \left[\left(\frac{\beta}{\lambda}\right)^2 (\lambda t_{j-1})^{2\beta} \exp\left[-(\lambda t_{j-1})^{\beta}\right] - \frac{\beta(\beta-1)}{\lambda^2}(\lambda t_{j-1})^{\beta} \exp\left[-(\lambda t_{j-1})^{\beta}\right] - \left(\frac{\beta}{\lambda}\right)^2 (\lambda t_j)^{2\beta} \exp\left[-(\lambda t_j)^{\beta}\right] + \frac{\beta(\beta-1)}{\lambda^2}(\lambda t_j)^{\beta} \exp\left[-(\lambda t_j)^{\beta}\right]\right] - \left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-2} \times \left[-(\lambda t_{j-1})^{\beta} \log(\lambda t_{j-1}) \exp\left[-(\lambda t_{j-1})^{\beta}\right] + \frac{\beta}{\lambda}(\lambda t_j)^{\beta} \exp\left[-(\lambda t_j)^{\beta}\right]\right]^2 \Bigg)$$

and

$$g = \sum_{j=1}^{k+1} d_j \left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-1} \left[-(\lambda t_{j-1})^{\beta} \log(\lambda t_{j-1}) \exp\left[-(\lambda t_{j-1})^{\beta}\right] + (\lambda t_j)^{\beta} \log(\lambda t_j) \exp\left[-(\lambda t_j)^{\beta}\right]\right]$$

$$g' = \sum_{j=1}^{k+1} d_j \Big(\left[\exp\left[-(\lambda t_{j-1})^{\beta}\right] - \exp\left[-(\lambda t_j)^{\beta}\right]\right]^{-1} \left[-(\lambda t_{j-1})^{\beta} (\log(\lambda t_{j-1}))^2 \exp\left[-(\lambda t_{j-1})^{\beta}\right] + (\lambda t_{j-1})^{2\beta} (\log(\lambda t_{j-1}))^2 \exp\left[-(\lambda t_{j-1})^{\beta}\right] + (\lambda t_j)^{\beta} (\log(\lambda t_j))^2 \exp\left[-(\lambda t_j)^{\beta}\right]\right]$$

$$-(\lambda t_i)^{2\beta}(\log(\lambda t_i))^2 \exp\left[-(\lambda t_i)^{\beta}\right] - \left[\exp\left[-(\lambda t_{i-1})^{\beta}\right] - \exp\left[-(\lambda t_i)^{\beta}\right]\right]^{-2}$$
$$\times\left[-(\lambda t_{i-1})^{\beta}\log(\lambda t_{i-1})\exp\left[-(\lambda t_{i-1})^{\beta}\right] + (\lambda t_i)^{\beta}\log(\lambda t_i)\exp\left[-(\lambda t_i)^{\beta}\right]\right]^{-2}\Bigg)$$

4.4.3 The Gamma Distribution:

Now for the two-parameter gamma distribution with p.d.f.

$$\frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right), \qquad t>0$$

where $\alpha>0$ and $k>0$ are unknown scale and shape parameters, respectively; the likelihood function is

$$L(k,\alpha)=\frac{1}{\alpha^{nk}\Gamma(k)^n}\left(\prod_{i=1}^{n}t_i^{k-1}\right)\exp\left(-\sum_{i=1}^{n}\frac{t_i}{\alpha}\right)$$

Let

$$\bar{t}=\sum_{i=1}^{n}\frac{t_i}{n} \text{ and } t=\left(\prod_{i=1}^{n}t_i\right)^{1/n}$$

represents the arithmetic and geometric means, and $L(k,\alpha)$ can be written as

$$L(k,\alpha)=\frac{t^{n(k-1)}}{\alpha^{nk}\Gamma(k)^n}\exp\left(-\frac{n\bar{t}}{\alpha}\right)$$

$$\Rightarrow \log L(k,\alpha)=-nk\log\alpha-n\log\Gamma(k)+n(k-1)\log t-\frac{n\bar{t}}{\alpha}$$

Setting $\partial\log L/\partial k$ and $\partial\log L/\partial\alpha$ equal to zero and rearranging slightly we get the likelihood equations

$$\hat{k}\,\hat{\alpha} = \bar{t}$$

$$\text{and } \log\hat{k} - \Psi\left(\hat{k}\right) = \log\left(\frac{\bar{t}}{t}\right)$$

where $\Psi(k) = d\log\Gamma(k)/dk = \Gamma'(k)/\Gamma(k)$ is the digamma function.

A very close approximation to $\hat{k}$ is given by the empirically determined formulas

$$\hat{k} \doteq S^{-1}\left(0.5000876 + 0.1648852S - 0.0544274S^2\right), \qquad 0 < S \le 0.5772$$

$$\hat{k} \doteq S^{-1}\left(17.79728 + 11.968477S + S^2\right)^{-1} \times \left(8.898919 + 9.059950S + 0.9775373S^2\right),$$
$$0 < S \le 0.5772$$

where $S = \log\left(\bar{t}/t\right)$

For gamma distribution the estimates are calculated by using windows based program Maple-V4.0 from the raw data, and for exponential and Weibull distribution the maximum likelihood estimates are computed from grouped data by the language program FORTRAN77 (program # 4.4.1, 4.4.2 and 4.4.3) considering the initial value of the parameters as their least square estimates, where the number of intervals is 16. The obtained results are given in table#4.4.1.

Program # 4.4.1: Program to estimate MLE of mue.

```
DIMENSION T(0:50),D(50)
OPEN(UNIT=10,FILE='DIA.DAT',STATUS='OLD')
OPEN(UNIT=11,FILE='LEM.OUT',STATUS='UNKNOWN')
A=2
B=115
EPS=1E-10
T(0)=0.0
```

```
      DO 20 J=1,16
      READ(10,*)T(J),D(J)
20    CONTINUE
      ITER=0.0
      F=0.0
      DELF=0.0
30    DO 40 J=1,16
      K=J-1
      A1=(T(K)-A)/B
      B1=(T(J)-A)/B
      A2=EXP(-A1)
      B2=EXP(-B1)
      A3=A1/B
      B3=B1/B
      A4=A2*A3
      B4=B2*B3
      C1=A2-B2
      C2=1/C1
      C3=A4-B4
      AA1=D(J)*C2*C3
      F=F+AA1
      A5=A2*A3*A3
      A6=2*A2*A3/B
      B5=B2*B3*B3
      B6=2*B2*B3/B
      C4=A5-A6-B5+B6
      C5=C2*C4
      C6=C2*C2*C3*C3
      C7=C5-C6
      BB1=D(J)*C7
      DELF=DELF+BB1
40    CONTINUE
      BOLD=B
      BNEW=BOLD-F/DELF
      DIF=BNEW-BOLD
      IF(ABS(DIF) .LE. EPS) GO TO 50
      B=BNEW
      ITER=ITER+1
      GO TO 30
50    WRITE(11,*)'ITERATION=',ITER
      WRITE(11,*)'THETA=',B
      STOP
      END
```

Program # 4.4.2: Program to estimate MLE of lemda.

```
      DIMENSION T(0:50),D(50)
      OPEN(UNIT=10,FILE='DIA.DAT',STATUS='OLD')
      OPEN(UNIT=11,FILE='LEM.OUT',STATUS='UNKNOWN')
      A=0.011304917
      B=.715
      EPS=1E-10
      T(0)=0.0
      DO 20 J=1,16
      READ(10,*)T(J),D(J)
20    CONTINUE
      ITER=0.0
      F=0.0
      DELF=0.0
30    DO 40 J=1,16
      K=J-1
      A1=(A*T(K))**B
      B1=(A*T(J))**B
      A2=EXP(-A1)
      B2=EXP(-B1)
      IF(T(J).EQ.T(16)) B2=0.0
      C1=B/A
      C2=A2-B2
      A3=C1*A1*A2
      B3=C1*B1*B2
      C3=1/C2
      C4=-A3+B3
      AA1=D(J)*C3*C4
      F=F+AA1
      C5=C1*(B-1)/A
      A4=C1*C1*A1*A1*A2
      A5=C5*A1*A2
      B4=C1*C1*B1*B1*B2
      B5=C5*B1*B2
      C6=A4-A5-B4+B5
      C7=C3*C6
      C8=C3*C3*C4*C4
      C9=C7-C8
      BB1=D(J)*C9
      DELF=DELF+BB1
40    CONTINUE
      AOLD=A
      ANEW=AOLD-F/DELF
      DIF=ANEW-AOLD
      IF(ABS(DIF) .LE. EPS) GO TO 50
```

```
      A=ANEW
      ITER=ITER+1
      GO TO 30
50    WRITE(11,*)'ITERATION=',ITER
      WRITE(11,*)'LEMDA=',A
      STOP
      END
```

Program # 4.4.3: Program to estimate MLE of theta.

```
      DIMENSION T(0:50),D(50)
      OPEN(UNIT=10,FILE='DIA.DAT',STATUS='OLD')
      OPEN(UNIT=11,FILE='LEM.OUT',STATUS='UNKNOWN')
      A=0.0108834
      B=.715
      EPS=1E-10
      T(0)=0.0
      DO 20 J=1,16
      READ(10,*)T(J),D(J)
20    CONTINUE
      ITER=0.0
      F=0.0
      DELF=0.0
30    DO 40 J=1,16
      K=J-1
      A1=(A*T(K))**B
      B1=(A*T(J))**B
      A2=EXP(-A1)
      B2=EXP(-B1)
      IF(T(J).EQ.T(16)) B2=0.0
      C2=A2-B2
      IF(T(K) .EQ. 0.0)THEN
      A3=0.0
      ELSE
      A3=LOG(A*T(K))
      ENDIF
      B3=LOG(A*T(J))
      A4=A1*A3*A2
      B4=B1*B3*B2
      C3=1/C2
      C4=-A4+B4
      AA1=D(J)*C3*C4
      F=F+AA1
      A5=A3*A4
      A6=A1*A5
```

```
      B5=B3*B4
      B6=B1*B5
      C5=-A5+A6+B5-B6
      C6=C3*C5
      C7=C3*C3*C4*C4
      C8=C6-C7
      BB1=D(J)*C8
      DELF=DELF+BB1
40    CONTINUE
      BOLD=B
      BNEW=BOLD-F/DELF
      DIF=BNEW-BOLD
      IF(ABS(DIF) .LE. EPS) GO TO 50
      B=BNEW
      ITER=ITER+1
      GO TO 30
50    WRITE(11,*)'ITERATION=',ITER
      WRITE(11,*)'BETA=',B
      STOP
      END
```

Table # 4.4.1: MLE of parameters.

Exponential distribution:	Weibull distribution:	Gamma distribution:
$\hat{\mu} = 2$	$\hat{\lambda} = 0.0108834$	$\hat{\alpha} = 366.5703$
$\hat{\beta} = 95.9640800$	$\hat{\beta} = 0.7186241$	$\hat{k} = 0.528507$

4.5 Goodness of fit test:

To check between the two distributions, which fit the data better we have applied goodness of fit test. With grouped uncensored data the best known test of fit procedures being the classical Pearson (χ^2) test.

For exponential distribution we know that

$$\int_{t_i}^{\infty} \frac{1}{\theta} \exp\left[-\frac{t-\mu}{\theta}\right] dt$$

$$= \exp\left[-\frac{t_i-\mu}{\theta}\right]$$

Hence we can calculate the probability as

$$p_1 = 1 - \exp\left[-\frac{t_1-\mu}{\theta}\right]$$

$$p_2 = \exp\left[-\frac{t_1-\mu}{\theta}\right] - \exp\left[-\frac{t_2-\mu}{\theta}\right]$$

$$p_3 = \exp\left[-\frac{t_2-\mu}{\theta}\right] - \exp\left[-\frac{t_3-\mu}{\theta}\right]$$

$$\vdots$$

$$p_{n-1} = \exp\left[-\frac{t_{n-2}-\mu}{\theta}\right] - \exp\left[-\frac{t_{n-1}-\mu}{\theta}\right]$$

$$p_n = \exp\left[-\frac{t_{n-1}-\mu}{\theta}\right]$$

For Weibull distribution we know that

$$\int_{t_i}^{\infty} \beta\lambda(\lambda t)^{\beta-1} \exp\left[-(\lambda t)^{\beta}\right] dt$$

$$= \exp\left[-(\lambda t_i)^{\beta}\right]$$

Hence we can calculate the probability as

$$p_1 = 1 - \exp\left[-(\lambda t_1)^{\beta}\right]$$

$$p_2 = \exp\left[-(\lambda t_1)^{\beta}\right] - \exp\left[-(\lambda t_2)^{\beta}\right]$$

$$p_3 = \exp\left[-(\lambda t_2)^{\beta}\right] - \exp\left[-(\lambda t_3)^{\beta}\right]$$

$$\vdots$$

$$p_{n-1} = \exp\left[-(\lambda t_{n-2})^{\beta}\right] - \exp\left[-(\lambda t_{n-1})^{\beta}\right]$$

$$p_n = \exp\left[-(\lambda t_{n-1})^{\beta}\right]$$

Also for Gamma distribution with p.d.f.

$$\frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right), \qquad t > 0$$

$$S(t_i) = \int_{t_1}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt$$

Hence we can calculate the probability as

$$p_1 = 1 - \int_{t_1}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt$$

$$p_2 = \int_{t_1}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt - \int_{t_2}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt$$

$$p_3 = \int_{t_2}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt - \int_{t_3}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt$$

$$\vdots$$

$$p_{n-1} = \int_{t_{n-2}}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt - \int_{t_{n-1}}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt$$

$$p_n = \int_{t_{n-1}}^{\infty} \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(-\frac{t}{\alpha}\right) dt$$

Consider the hypothesis

$$H_o : P_j = P_{jo}, \; j = 1, \cdots\cdots, 16$$

where the P_{jo}'s are specified. Let $\tilde{P}_{jo}$ be the m.l.e. of P_j under H_o, or some other asymptotically fully efficient estimator and let $e_j = n\tilde{P}_{jo}$. The Pearson statistic for testing H_o is

$$\chi^2 = \sum_{j=1}^{k+1} \frac{(d_j - e_j)^2}{e_j}$$

The limiting distribution of χ^2 is $\chi^2_{(k-s)}$.

The results of the goodness of fit test are carried out by Microsoft Excel and the results are summarised in table# 4.5(a) and 4.5(b), respectively.

Table#4.5(a): Goodness of fit test with LSE.

		Exponential distribution		Weibull distribution		Gamma distribution	
I_j	d_j	$S(t_j)$	P_j	$S(t_j)$	P_j	$S(t_j)$	P_j
<15	108	0.902064	0.097936	0.788169	0.211831	0.663439	0.336561
15-30	126	0.800919	0.101145	0.646953	0.141216	0.608961	0.054479
31-45	92	0.711115	0.089804	0.537927	0.109026	0.559861	0.0491
46-60	41	0.63138	0.079735	0.450822	0.087105	0.515545	0.044316
61-75	38	0.560585	0.070794	0.379963	0.070859	0.475489	0.040055
76-90	19	0.497729	0.062857	0.321645	0.058318	0.439235	0.036254
91-105	17	0.44192	0.055809	0.273247	0.048399	0.406376	0.03286
106-120	5	0.392369	0.049551	0.232824	0.040423	0.376552	0.029823
121-135	13	0.348374	0.043995	0.198889	0.033935	0.349449	0.027103
136-150	8	0.309312	0.039062	0.170281	0.028608	0.324785	0.024664
151-165	5	0.27463	0.034682	0.146079	0.024203	0.302311	0.022474
166-180	10	0.243837	0.030793	0.12554	0.020539	0.281806	0.020504
181-195	7	0.216496	0.027341	0.108064	0.017476	0.263075	0.018731
196-210	6	0.192221	0.024275	0.093158	0.014906	0.245943	0.017132
211-225	9	0.170668	0.021553	0.080417	0.012741	0.230254	0.015689
226+	111	0	0.170668	0	0.080417	0	0.230254
Value of χ^2		217.3415		158.1476		474.8196	

Table#4.5(b): Goodness of fit test with MLE.

		Exponential distribution		Weibull distribution		Gamma distribution	
I_j	d_j	$S(t_j)$	P_j	$S(t_j)$	P_j	$S(t_j)$	P_j
<15	108	0.873308	0.126692	0.761964	0.238036	0.794825	0.205175
15-30	126	0.746936	0.126372	0.639307	0.122656	0.708141	0.086684
31-45	92	0.638851	0.108085	0.549525	0.089783	0.643339	0.064802
46-60	41	0.546406	0.092445	0.478932	0.070593	0.590394	0.052945
61-75	38	0.467339	0.079068	0.42137	0.057561	0.545297	0.045096
76-90	19	0.399712	0.067626	0.373353	0.048017	0.505938	0.039359
91-105	17	0.341872	0.05784	0.332657	0.040697	0.47103	0.034908
106-120	5	0.292402	0.049471	0.297752	0.034905	0.439714	0.031316
121-135	13	0.25009	0.042312	0.267536	0.030216	0.411381	0.028334
136-150	8	0.213901	0.036189	0.241181	0.026355	0.385575	0.025806
151-165	5	0.182948	0.030952	0.21805	0.023131	0.361948	0.023627
166-180	10	0.156475	0.026473	0.19764	0.020409	0.340221	0.021727
181-195	7	0.133832	0.022643	0.179551	0.01809	0.32017	0.020051
196-210	6	0.114466	0.019366	0.163452	0.016098	0.30161	0.01856
211-225	9	0.097902	0.016564	0.149076	0.014376	0.284385	0.017225
226+	111	0	0.097902	0	0.149076	0	0.284385
Value of χ^2		179.1484		107.232		196.8873	

From the above tables we observe that the chi-square value for Weibull distribution with maximum likelihood estimates is the smallest. Hence we can say that Weibull distribution fits the data better than Exponential and Gamma distribution, but we can not draw any final conclusion about the distribution using this result. For that we need further investigation.

4.6 Test of the shape parameter of Weibull distribution:

For further investigation we want to test the shape parameter of Weibull distribution, because if it is equal to one then there is no difference between Exponential and Weibull distributions. But if it significantly differ from one then we can say that the data will fit Weibull distribution. The significance test can be based on the large-sample normal

approximation $\hat{U} \sim N(U, I^{-1})$, where $U = (\lambda, \beta)$ and I^{-1} is the inverse of Fisher (or "expected") information, which can be calculated from

$$I = \begin{bmatrix} E\left(-\frac{\partial^2 \log L}{\partial \lambda^2}\right) & E\left(-\frac{\partial^2 \log L}{\partial \lambda \partial \beta}\right) \\ E\left(-\frac{\partial^2 \log L}{\partial \beta \partial \lambda}\right) & E\left(-\frac{\partial^2 \log L}{\partial \beta^2}\right) \end{bmatrix}$$

$$= -\begin{bmatrix} \frac{\partial^2 \log L}{\partial \lambda^2} & \frac{\partial^2 \log L}{\partial \lambda \partial \beta} \\ \frac{\partial^2 \log L}{\partial \beta \partial \lambda} & \frac{\partial^2 \log L}{\partial \beta^2} \end{bmatrix}_{\hat{\lambda}, \hat{\beta}}$$

Let

$$I^{-1} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

Hence the test procedure of $H_o : \beta = \beta_o$ is based on the test criterion

$$\frac{\hat{\beta} - \beta_o}{\sqrt{I_{22}}} \sim N(0,1)$$

The significance test of the shape parameter of the Weibull distribution, i.e. $H_o : \beta = 1$ is completed by computer program FORTRAN77 (program#4.6.1) and the obtained results are given bellow.

Program # 4.6.1: Program to calculate information matrix.

```
DIMENSION T(0:50),D(50)
OPEN(UNIT=10,FILE='DIA.DAT',STATUS='OLD')
OPEN(UNIT=11,FILE='LEM.OUT',STATUS='UNKNOWN')
A=0.0108834
B=.7186241
EPS=1E-10
T(0)=0.0
DO 20 J=1,16
READ(10,*)T(J),D(J)
```

```
20    CONTINUE
      AA1=0.0
      BB1=0.0
      CC1=0.0
30    DO 40 J=1,16
      K=J-1
      A1=(A*T(K))
      B1=(A*T(J))
      A2=A1**B
      B2=B1**B
      A3=EXP(-A2)
      B3=EXP(-B2)
      IF(T(J).EQ.T(16)) B3=0.0
      IF(T(K) .EQ. 0.0)THEN
      A4=0.0
      ELSE
      A4=LOG(A1)
      ENDIF
      B4=LOG(B1)
      C1=A3-B3
      A5=A3*A2*A4
      B5=B3*B2*B4
      C2=1/C1
      C3=-A5+B5
      A6=A5*A4
      A7=A6*A2
      B6=B5*B4
      B7=B6*B2
      C4=-A6+A7+B6-B7
      C5=C2*C2*C3*C3
      C6=C2*C4
      C7=-C5+C6
      BB1=BB1+D(J)*C7
      D1=B/A
      D2=(B-1)/A
      D3=D1*D1
      D4=D1*D2
      A8=D3*A2*A3
      A9=D4*A2*A3
      B8=D3*B2*B3
      B9=D4*B2*B3
      A10=D1*A2*A3
      B10=D1*B2*B3
      C8=A8-A9-B8+B9
      C9=-A10+B10
      C10=C2*C8-C2*C2*C9*C9
```

```
      AA1=AA1+D(J)*C10
      A11=A5*D1
      B11=B5*D1
      A12=A11*A2*D1
      B12=B11*B2*D1
      A13=A2*A3/A
      B13=B2*B3/A
      C11=A12-A13-A11-B12+B13+B11
      C12=-C2*C2*C3*C9+C1*C11
      CC1=CC1+D(J)*C12
40    CONTINUE
      AA2=-AA1
      BB2=-BB1
      CC2=-CC1
      DD=AA2*BB2-CC2*CC2
      AA=BB2/DD
      BB=AA2/DD
      AB=-CC2/DD
      DDD=SQRT(BB)
      Z=(B-1)/DDD
      WRITE(11,*)'I11=',AA
      WRITE(11,*)'I12=',AB
      WRITE(11,*)'I22=',BB
      WRITE(11,*)'AA1=',AA1
      WRITE(11,*)'AB1=',CC1
      WRITE(11,*)'BB1=',BB1
      WRITE(11,*)'Z=',Z
      STOP
      END
```

Using the program we get the results as

$$I = \begin{bmatrix} 2527884 & -27897.6 \\ -27897.6 & 1155.663 \end{bmatrix}$$

$$I^{-1} = \begin{bmatrix} 5.3924670E-007 & 1.3017370E-005 \\ 1.3017370E-005 & 0.0011795 \end{bmatrix}$$

Hence the value of the test criteria is-

$$|Z| = \left| \frac{0.7186241 - 1}{\sqrt{0.0011795}} \right| = 8.1927570$$

The calculated value of the $|Z|$-statistic is significantly larger than $Z_{0.05}$=1.96 and $Z_{0.01}$=2.58 leading to the decision in favour of the Weibull distribution.

4.6 Results and Discussion:

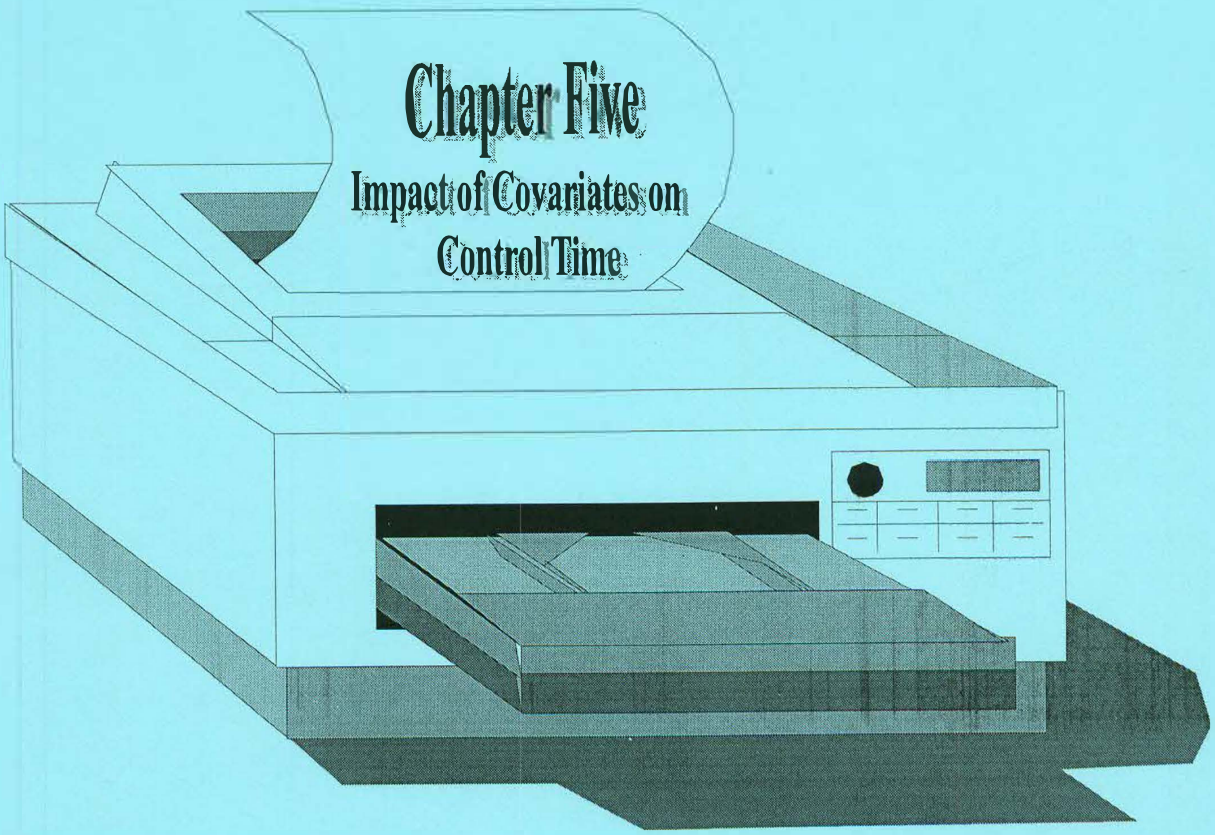
In the quest of an appropriate probability model for the control time distribution of diabetes mellitus we have constructed a frequency distribution in the previous chapter (table#3.2.1). An overview of this table gives us preliminary knowledge of the concerning probability model that it may fit a negative exponential or weibull or gamma distribution. Survival plots also concentrated the preliminary idea. Fitting of the distributions-negative exponential, weibull and gamma followed calculation of coefficient of determination R^2 and goodness-of-fit test by frequency chi-square led us to conclude that weibull distribution is the best fit to the data. This finding is confirmed by testing the shape parameter of weibull distribution embedding exponential in weibull distribution. Summery of the findings is as below:

Table#4.6: Summery of the findings.

		Exponential distribution	Weibull distribution	Gamma distribution
Coefficient of determinant, R^2		0.893	0.927	0.901
goodness-of-fit	Value of χ^2 with LSE	179.1484	158.1476	360.0298
	Value of χ^2 with MLE	179.1484	107.232	196.8873
Test of Weibull vs. Exponential $\|Z\| = 8.1927$				

Chapter Five

Impact of Covariates on Control Time



Chapter Five

Impact of Covariates on Control Time of Diabetes Mellitus

5.1 Introduction:

In practice many situations involve heterogeneous populations, and it is important to consider the relationship of lifetime to other factors. One way to do this is through a regression model in which lifetime has a distribution that depends upon the concomitant, or regressor, variables. This involves specifying a model for the distribution of lifetimes, given a vector of regressor variables for individuals. There are two approaches to regression: one employs parametric families of lifetime distributions and extends models such as the Exponential, Weibull, and Log-normal models to include regressor variables; the second approach is distribution-free and

assumes less about the underlying distributions than do the parametric methods. In this chapter we will study the impact of various co-factors on the control time using parametric regression method.

5.2 Simple Linear Regression:

As stated in chapter two we have collected data on 25 points among which 13 points can be considered as covariate which can influence the control time. Let us define these covariates as

X_1 = Age,

X_2 = Sex,

X_3 = Marital status,

X_4 = Food habit (rice),

X_5 = Food habit (vegetable),

X_6 = Food habit (Sweet),

X_7 = BMI,

X_8 = BP,

X_9 = BGL,

X_{10} = Heredity

X_{11} = Length of suffering.

But it is complicated to handle a parametric regression model with this large number of covariates. For this to avoid complexity first we observe the association of these covariates with control time of diabetes mellitus which is already done in chapter three. There we observe that only sex, food habit (rice), food habit (vegetable), food habit (sweet), BGL and suffering time shows significant association with control time. Now we will reduce the no. of covariates by examining the relationship of these factors: sex, food habit (rice, vegetable, sweet), BGL, and length of

suffering to logarithm of control time (using multivariate approach). That is,

$$Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_9 X_9 + \beta_{11} X_{11}$$

where Y= logarithm of control time. The results are completed using the computer program SPSS for windows and are given bellow.

Table # 5.2A: Model test using multivariate approach.

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	43.052	6	7.176	3.628	0.002
Residual	1202.392	608	1.978		
Total	1245.445	614			

Coefficient of determinant, R^2 = 0.035

Table # 5.2B: Coefficient test using multivariate approach.

Dependent Variables	Estimated Coefficients	t	Sig.
constant	3.507	16.500	.000
X_2=Sex	0.206	1.798	0.073
X_4=Food Habit (Rice)	0.268	2.316	0.021
X_5=Food Habit (Vege.)	0.106	0.683	0.495
X_6=Food Habit (Sweet)	0.09215	0.793	0.428
X_9=BGL	0.01209	1.020	0.308
X_{11}=Length of Suffering	0.0002769	3.440	0.001

From the above table we observe that only three covariates sex, food habit (rice) and length of suffering show significance to control time in multivariate regression.

5.3 Logistic Regression Model:

For survival data the most commonly used model is logistic regression model. This model is most frequently employed the relationship between a dichotomous (binary) or polytomous (with some modification) outcome variable and a set of covariates. In most health science applications, time of an event is not exactly known but only to have occurred in an interval. In such situations this model is frequently used. The logistic model assumes that the logit of the probability of death of a subject in an interval, conditional to death has not occurred prior to the interval, is a linear function of the covariates. Let us consider that the dependent variable is control time. Since in logistic regression model the response variable must be dichotomous (or polytomous) we have used the code 1, if the patient come into control within 42 days (median of control times) and 0, otherwise, i.e.

$$Y_i = \begin{cases} 1, & \textit{if the patient came into control within 42 days} \\ 0, & \textit{otherwise} \end{cases}, \; i = 1,2,3,\cdots\cdots\cdots,615$$

For the covariates we have a collection of 11 independent variables as stated in section 5.2. They will be denoted by the vector $X' = (x_1, x_2, x_3, \cdots\cdots, x_{11})$ and β be a $(11+1)\times 1$ vector of unknown parameters.

Hence $\pi(X) = P(Y = 1|X)$ is the probability that the patient came into control within 42 days conditional on the value of X and $1 - \pi(X) = P(Y = 0|X)$ is the probability that the patient do not. Hence

$$\pi(x_i) = P(Y = 1|X) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}} = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

and

$$1-\pi(x_i)=P(Y=0|X)==\frac{1}{1+e^{x_i\beta}}$$

Hence

$$\frac{\pi(x_i)}{1-\pi(x_i)}=e^{x_i\beta}$$

$$\Rightarrow g(x_i)=\log it\ \pi(x_i)=\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]=x_i\beta$$

$$\Rightarrow g(x_i)=\beta_0+\beta_1x_{1i}+\cdots\cdots+\beta_{11}x_{11i}$$

The analysis is carried out by using the computer program SPSS for windows and the results are given in table # 5.3.1A.

Table # 5.3.1A: Impact of covariates on control time.

Model Summery

-2log likelihood	Cox & Snell R Square
833.371	.031

Coefficient test

Covariates	B	S.E.	Wald	df	Sig.	Exp(B)
X_1 = Age	-.012	.008	2.330	1	.127	.988
X_2 = Sex	.244	.176	1.919	1	.166	1.276
X_3= Marital status	.158	.688	.052	1	.819	1.171
X_4 = Food habit (rice)	.449	.170	6.990	1	.008	1.567
X_5 = Food habit (vege.)	.212	.230	.850	1	.357	1.236
X_6 = Food habit (sweet)	.074	.169	.194	1	.660	1.077
X_7 = BMI	.012	.023	.280	1	.597	1.012
X_8 = BP	.005	.004	1.884	1	.170	1.005
X_9 = BGL	.028	.018	2.575	1	.109	1.029
X_{10} = Heredity	.085	.169	.253	1	.615	1.089
X_{11} = Length of suffering	.000	.000	6.502	1	.011	1.000
Constant	-1.537	.919	2.796	1	.095	.215

Here we observe that only food habit (rice) and length of suffering are significant.

Now if we categorise the dependent variable at the mean then we get,

Table # 5.3.1B: Impact of covariates on control time.

Model Summery

-2log likelihood	Cox & Snell R Square	Nagelkerke R Square
737.546	.039	.055

Coefficient test

covariates	B	S.E.	Wald	df	Sig.	Exp(B)
X_1 = Age	-.007	.009	.707	1	.400	.993
X_2 = Sex	.391	.191	4.203	1	.040	1.479
X_3= Marital status	-.138	.718	.037	1	.848	.871
X_4 = Food habit (rice)	.202	.185	1.187	1	.276	1.224
X_5 = Food habit (vege.)	.179	.250	.514	1	.473	1.197
X_6 = Food habit (sweet)	.205	.183	1.255	1	.263	1.227
X_7 = BMI	.025	.025	1.011	1	.315	1.025
X_8 = BP	.004	.004	1.219	1	.270	1.004
X_9 = BGL	.034	.019	3.101	1	.078	1.034
X_{10} = Heredity	.253	.183	1.917	1	.166	1.288
X_{11} = Length of suffering	.000	.000	13.22	1	.000	1.000
Constant	-2.629	.982	7.172	1	.007	.072

Here we observe that sex and length of suffering are highly significant and BGL gives a poor (p=.078) significant impact.

Hence with three covariates – sex, food habit rice and length of suffering we will go through the parametric regression methods.

5.4 Parametric Regression Model:

The most important parametric regression models are extensions of the distributions – Exponential, Weibull, and Log-normal. In chapter three we

have observed that our data fits better the Weibull distribution. Hence our main target is the Weibull Regression Model. We know that the two-parameter Weibull distribution has a scale parameter α and a shape parameter β and can be extended to regression model by allowing α and β to depend on X, where X is a vector of regressor variables, or covariates. The most commonly used Weibull Regression Models are those for which just α, and not β, depends on X. The fact that β does not depend on X implies proportional hazards for lifetimes and constant variance for log-lifetimes of individuals. These have survivor function for T (lifetime), given the vector X of regressor variables, of the form

$$S(t/x)=\exp\left[-\left(\frac{t}{\alpha(x)}\right)^{\beta}\right] \quad , \quad t\geq 0$$

where $\alpha(x)$ is a function of x that typically involves unknown parameters. Hence the p.d.f. of lifetime, given X, is of the form

$$\frac{\beta}{\alpha(x)}\left(\frac{t}{\alpha(x)}\right)^{\beta-1}\exp\left[-\left(\frac{t}{\alpha(x)}\right)^{\beta}\right] \quad , \quad t\geq 0$$

Rather than this we shall work with log lifetime : the p.d.f. of $Y=logT$, given X, is

$$f(y/x)=\frac{1}{\sigma}\exp\left[\frac{y-\mu(x)}{\sigma}-\exp\left(\frac{y-\mu(x)}{\sigma}\right)\right] \quad , \quad -\infty<y<\infty$$

where $\sigma=\beta^{-1}$, $\mu(x)=\log\alpha(x)$. The most frequently used model is the linear one with

$$\mu(x) = X\delta$$

where $X = (x_2, x_4, x_{11})$ and $\delta = (\delta_2, \delta_4, \delta_{11})'$.

5.4.1 Maximum Likelihood Estimation:

Let y_i is either a log control time or a log censoring time; D and C denote the sets of individuals for which y_i is a log lifetime and a log censoring time, respectively. The likelihood function is then

$$L(\delta,\sigma) = \prod_{i\in D} \frac{1}{\sigma} \exp\left[\frac{y-\mu(x)}{\sigma} - \exp\left(\frac{y-\mu(x)}{\sigma}\right)\right] \prod_{i\in C} \frac{1}{\sigma} \exp\left[-\exp\left(\frac{y-\mu(x)}{\sigma}\right)\right]$$

$$\Rightarrow \log L = -r\log\sigma + \sum_{i\in D} \frac{y - x_i\delta}{\sigma} - \sum_{i=1}^{n} \exp\left(\frac{y - x_i\delta}{\sigma}\right)$$

where r is the observed number of lifetimes and is equal to 504. If we let

$$Z_i = \frac{y - x_i\delta}{\sigma}$$

Then

$$\log L = -r\log\sigma + \sum_{i\in D} Z_i - \sum_{i=1}^{n} e^{Z_i}$$

Hence the estimating equations are

$$\frac{\partial \log L}{\partial \delta_2} = -\frac{1}{\sigma}\sum_{i\in D} x_{2i} + \frac{1}{\sigma}\sum_{i=1}^{n} x_{2i}e^{Z_i} = 0$$

$$\frac{\partial \log L}{\partial \delta_4} = -\frac{1}{\sigma}\sum_{i\in D} x_{4i} + \frac{1}{\sigma}\sum_{i=1}^{n} x_{4i}e^{Z_i} = 0$$

$$\frac{\partial \log L}{\partial \delta_{11}} = -\frac{1}{\sigma}\sum_{i\in D} x_{11i} + \frac{1}{\sigma}\sum_{i=1}^{n} x_{11i}e^{Z_i} = 0$$

and

$$\frac{\partial \log L}{\partial \sigma} = -\frac{r}{\sigma} - \frac{1}{\sigma}\sum_{i\in D} Z_i + \frac{1}{\sigma}\sum_{i=1}^{n} Z_i e^{Z_i} = 0$$

These equations can usually be solved by the Newton-Raphson method or some other iterative procedure. For that other equations are

$$\frac{\partial^2 \log L}{\partial \delta_2^{\,2}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{2i}^{\,2} e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_4^{\,2}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_4^{\,2} e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_{11}^{\,2}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{11i}^{\,2} e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \sigma^2} = \frac{r}{\sigma^2} + \frac{2}{\sigma^2}\sum_{i\in D} Z_i - \frac{1}{\sigma^2}\sum_{i=1}^{n} Z_i e^{Z_i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} Z_i^{\,2} e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_2 \partial \delta_4} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{2i}x_{4i}e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_2 \partial \delta_{11}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{2i}x_{11i}e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_4 \partial \delta_{11}} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{4i}x_{11i}e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_2 \partial \sigma} = \frac{1}{\sigma^2}\sum_{i\in D} x_{2i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{2i}e^{Z_i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{2i}Z_i e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_4 \partial \sigma} = \frac{1}{\sigma^2}\sum_{i\in D} x_{4i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{4i}e^{Z_i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{4i}Z_i e^{Z_i}$$

$$\frac{\partial^2 \log L}{\partial \delta_{11} \partial \sigma} = \frac{1}{\sigma^2}\sum_{i\in D} x_{11i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{11i}e^{Z_i} - \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{11i}Z_i e^{Z_i}$$

In this case the initial values of δ_2, δ_4 and δ_{11} are considered as the least square estimates of $Y = \delta_2 x_2 + \delta_4 x_4 + \delta_{11} x_{11}$. The results are computed using the language program FORTRAN77 (program # 5.4.1.1) and the results are given in table # 5.4.1.1 . Also to be noted that here we have considered the censoring time as t=226.

5.4.2 Test of Covariates:

With these estimates to test whether covariates have any impact on control time, i.e. (i) $H_o : \delta_2 = 0$, (ii) $H_o : \delta_4 = 0$ & (iii) $H_o : \delta_{11} = 0$ and also to test (iv) $H_o : \sigma = 1$ we can use the large-sample normal approximation as used in section 3.5, where the Fisher information matrix is

$$I = \begin{bmatrix} E\left[-\frac{\partial^2 \log L}{\partial \delta_2^{\,2}}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_2 \partial \delta_4}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_2 \partial \delta_{11}}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_2 \partial \sigma}\right] \\ E\left[-\frac{\partial^2 \log L}{\partial \delta_4 \partial_2}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_4^{\,2}}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_4 \partial \delta_{11}}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_4 \partial \sigma}\right] \\ E\left[-\frac{\partial^2 \log L}{\partial \delta_{11} \partial \delta_2}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_{11} \partial \delta_4}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_{11}^{\,2}}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_{11} \partial \sigma}\right] \\ E\left[-\frac{\partial^2 \log L}{\partial \delta_2 \partial \sigma}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_4 \partial \sigma}\right] & E\left[-\frac{\partial^2 \log L}{\partial \delta_{11} \partial \sigma}\right] & E\left[-\frac{\partial^2 \log L}{\partial \sigma^2}\right] \end{bmatrix}$$

$$= -\begin{bmatrix} \frac{\partial^2 \log L}{\partial \delta_2^{\,2}} & \frac{\partial^2 \log L}{\partial \delta_2 \partial \delta_4} & \frac{\partial^2 \log L}{\partial \delta_2 \partial \delta_{11}} & \frac{\partial^2 \log L}{\partial \delta_2 \partial \sigma} \\ \frac{\partial^2 \log L}{\partial \delta_4 \partial_2} & \frac{\partial^2 \log L}{\partial \delta_4^{\,2}} & \frac{\partial^2 \log L}{\partial \delta_4 \partial \delta_{11}} & \frac{\partial^2 \log L}{\partial \delta_4 \partial \sigma} \\ \frac{\partial^2 \log L}{\partial \delta_{11} \partial \delta_2} & \frac{\partial^2 \log L}{\partial \delta_{11} \partial \delta_4} & \frac{\partial^2 \log L}{\partial \delta_{11}^{\,2}} & \frac{\partial^2 \log L}{\partial \delta_{11} \partial \sigma} \\ \frac{\partial^2 \log L}{\partial \delta_2 \partial \sigma} & \frac{\partial^2 \log L}{\partial \delta_4 \partial \sigma} & \frac{\partial^2 \log L}{\partial \delta_{11} \partial \sigma} & \frac{\partial^2 \log L}{\partial \sigma^2} \end{bmatrix}_{\hat{\delta}_1,\ \hat{\delta}_2,\ \hat{\delta}_3,\ \hat{\sigma}}$$

Let

$$I^{-1} = \begin{bmatrix} I_{11} & I_{12} & I_{13} & I_{14} \\ I_{21} & I_{22} & I_{23} & I_{24} \\ I_{31} & I_{32} & I_{33} & I_{34} \\ I_{41} & I_{42} & I_{43} & I_{44} \end{bmatrix}$$

Then the test criterions are

$$(\text{i})\ Z_1 = \frac{\hat{\delta}_2}{\sqrt{I_{11}}} \sim N(0,1)$$

$$(\text{ii})\ Z_2 = \frac{\hat{\delta}_4}{\sqrt{I_{22}}} \sim N(0,1)$$

$$(\text{iii})\ Z_3 = \frac{\hat{\delta}_{11}}{\sqrt{I_{33}}} \sim N(0,1)$$

and

$$(\text{iv})\ Z_4 = \frac{\hat{\sigma} - 1}{\sqrt{I_{44}}} \sim N(0,1)$$

respectively.

The results are carried out by the language program FORTRAN77 (program # 5.4.1.1) and the results are given in table # 5.4.1.1

Program # 5.4.1.1: Estimation of parameters & Information Matrix by Newton-Raphson iteration.

```
      DIMENSION T(1000),Y(1000),X1(1000),X2(1000),Z(1000),EZ(1000)
      DIMENSION X1NEW(1000),X2NEW(1000),X3(1000),X3NEW(1000)
      DIMENSION ZNEW(1000)
      INTEGER X1,X2,ITER
      OPEN(UNIT=10,FILE='DIA.DAT',STATUS='OLD')
      OPEN(UNIT=11,FILE='EST.OUT',STATUS='UNKNOWN')
      B1=0.0002769
      B2=0.206
      B3=0.268
      B=0.871393
      S1=0.0
      S2=0.0
      S3=0.0
      S4=0.0
      S5=0.0
      S6=0.0
      S7=0.0
      S8=0.0
      S9=0.0
      S10=0.0
      S11=0.0
      S12=0.0
      S13=0.0
      S14=0.0
      S15=0.0
      S16=0.0
      S17=0.0
      S18=0.0
      R=504
      ITER=0.0
      DO 20 I=1,615
      READ(10,*)T(I)
      Y(I)=LOG(T(I))
20    CONTINUE
      DO 30 I=1,615
      READ(10,*)X1(I)
30    CONTINUE
      DO 40 I=1,615
      READ(10,*)X2(I)
40    CONTINUE
      DO 41 I=1,615
      READ(10,*)X3(I)
41    CONTINUE
45    DO 50 I=1,615
      Z(I)=(Y(I)-X1(I)*B1-X2(I)*B2-X3(I)*B3)*B
      EZ(I)=EXP(Z(I))
50    CONTINUE
      DO 60 I=1,615
      IF(T(I) .LT. 226)THEN
      X1NEW(I)=X1(I)
```

```
      X2NEW(I)=X2(I)
      X3NEW(I)=X3(I)
      ZNEW(I)=Z(I)
      ELSE
      X1NEW(I)=0.0
      X2NEW(I)=0.0
      X3NEW(I)=0.0
      ZNEW(I)=0.0
      ENDIF
60    CONTINUE
C     *******DLOGL/DB1*******
      DO 61 I=1,615
      S1=S1+X1NEW(I)
      S2=S2+(X1(I)*EZ(I))
61    CONTINUE
      F1=(-S1+S2)*B
C     *******DLOGL/DB2*******
      DO 62 I=1,615
      S3=S3+X2NEW(I)
      S4=S4+(X2(I)*EZ(I))
62    CONTINUE
      F2=(-S3+S4)*B
C     *******DLOGL/DB3*******
      DO 51 I=1,615
      S5=S5+X3NEW(I)
      S6=S6+(X3(I)*EZ(I))
51    CONTINUE
      F3=(-S5+S6)*B
C     *******DLOGL/DB4*******
      DO 63 I=1,615
      S7=S7+ZNEW(I)
      S8=S8+(Z(I)*EZ(I))
63    CONTINUE
      F4=(-R-S7+S8)*B
C     *******DLOGL/DB11******
      DO 64 I=1,615
      S9=S9+(X1(I)*X1(I)*EZ(I))
64    CONTINUE
      F11=(-S9)*(B*B)
C     *******DLOGL/DB22******
      DO 65 I=1,615
      S10=S10+(X2(I)*X2(I)*EZ(I))
65    CONTINUE
      F22=(-S10)*(B*B)
C     *******DLOGL/DB33******
      DO 66 I=1,615
      S11=S11+(X3(I)*X3(I)*EZ(I))
66    CONTINUE
      F33=(-S11)*(B*B)
C     *******DLOGL/DB44******
      DO 67 I=1,615
      S12=S12+(Z(I)*Z(I)*EZ(I))
67    CONTINUE
      F44=(R+2*S7-2*S8-S12)*(B*B)
C     *******DLOGL/DB12******
      DO 68 I=1,615
      S13=S13+(X1(I)*X2(I)*EZ(I))
68    CONTINUE
      F12=(-S13)*(B*B)
C     *******DLOGL/DB13******
      DO 69 I=1,615
```

```
      S14=S14+(X1(I)*X3(I)*EZ(I))
69    CONTINUE
      F13=(-S14)*(B*B)
C     *******DLOGL/DB23******
      DO 70 I=1,615
      S15=S15+(X2(I)*X3(I)*EZ(I))
70    CONTINUE
      F23=(-S15)*(B*B)
C     *******DLOGL/DB14******
      DO 71 I=1,615
      S16=S16+(X1(I)*Z(I)*EZ(I))
71    CONTINUE
      F14=(S1-S2-S16)*(B*B)
C     *******DLOGL/DB24******
      DO 72 I=1,615
      S17=S17+(X2(I)*Z(I)*EZ(I))
72    CONTINUE
      F24=(S3-S4-S17)*(B*B)
C     *******DLOGL/DB34******
      DO 73 I=1,615
      S18=S18+(X3(I)*Z(I)*EZ(I))
73    CONTINUE
      F34=(S5-S6-S18)*(B*B)
C     *******SCORE-MATRIX***
      A11=-F11
      A12=-F12
      A13=-F13
      A14=-F14
      A22=-F22
      A23=-F23
      A24=-F24
      A33=-F33
      A34=-F34
      A44=-F44
C     *******INVERSE********
      CC1=A24*(A23*A34-A33*A24)
      C11=A22*(A33*A44-A23*A23)-A23*(A23*A44-A24*A34)+CC1
      CC2=A24*(A13*A34-A14*A33)
      C12=A12*(A33*A44-A34*A34)-A23*(A13*A44-A14*A34)+CC2
      CC3=A24*(A13*A24-A14*A23)
      C13=A12*(A23*A44-A24*A34)-A22*(A13*A44-A14*A34)+CC3
      CC4=A23*(A13*A24-A14*A23)
      C14=A12*(A23*A34-A33*A24)-A22*(A13*A34-A14*A33)+CC4
      CC5=A14*(A13*A34-A14*A33)
      C22=A11*(A33*A44-A34*A34)-A13*(A13*A44-A14*A34)+CC5
      CC6=A14*(A13*A24-A14*A23)
      C23=A11*(A23*A44-A34*A24)-A12*(A13*A44-A14*A34)+CC6
      CC7=A13*(A13*A24-A14*A23)
      C24=A11*(A23*A34-A33*A24)-A12*(A13*A34-A14*A33)+CC7
      CC8=A14*(A12*A24-A14*A22)
      C33=A11*(A22*A44-A24*A24)-A12*(A12*A44-A14*A24)+CC8
      CC9=A13*(A12*A24-A22*A14)
      C34=A11*(A22*A34-A23*A24)-A12*(A12*A34-A23*A14)+CC9
      CC10=A13*(A12*A23-A22*A13)
      C44=A11*(A22*A33-A23*A23)-A12*(A12*A33-A13*A23)+CC10
      DS=A11*C11-A12*C12+A13*C13-A14*C14
      B11=C11/DS
      B12=C12/DS
      B13=C13/DS
      B14=C14/DS
      B22=C22/DS
```

```
      B23=C23/DS
      B24=C24/DS
      B33=C33/DS
      B34=C34/DS
      B44=C44/DS
      DF1=B11*F1+B12*F2+B13*F3+B14*F4
      DF2=B12*F1+B22*F2+B23*F3+B24*F4
      DF3=B13*F1+B23*F2+B33*F3+B34*F4
      DF4=B14*F1+B24*F2+B34*F3+B44*F4
      BN1=B1+DF1
      BN2=B2+DF2
      BN3=B3+DF3
      BN=B+DF4
      B1=BN1
      B2=BN2
      B3=BN3
      B=BN
      U=0.000000000001
      ITER=ITER+1
      IF(DF1.LE.U.AND.DF2.LE.U.AND.DF3.LE.U.AND.DF4.LE.U)GO TO 90
      GO TO 45
90    WRITE(11,*)'ITER=',ITER
      WRITE(11,*) 'COEFFICIENTS=',B1,B2,B3,B
      WRITE(11,*)'VAR=',B11,B22,B33,B44
      D11=SQRT(B11)
      D22=SQRT(B22)
      D33=SQRT(B33)
      D44=SQRT(B44)
      Z1=B1/D11
      Z2=B2/D22
      Z3=B3/D33
      Z4=(1/B-1)/D44
      WRITE(11,*)'STD=',D11,D22,D33,D44
      WRITE(11,*)'Z=',Z1,Z2,Z3,Z4
      STOP
      END
```

Table # 5.4.1.1: Test of covariates with censoring.

covariate	estimated coefficient	standard error	test statistic (Z)
X_2 = sex	7.577468	0.0992275	32.14252
X_4 =food habit (rice)	6.084215	0.1891501	29.95816
X_{11} =suffering time	0.0261756	0.0013624	17.07202
sigma	1.2550800	0.0061761	46.03728

But if we consider no censoring then the figure of table # 5.4.1.1 becomes as follows:

Table # 5.4.1.1E: Test of covariates without censoring.

covariate	variance	standard error	test statistic (Z)
X_2 = sex	0.0388916	0.1972096	30.85152
X_4 =food habit (rice)	1.1903480E-006	0.001091	23.99162
X_{11} =suffering time	0.2236646	0.472932	16.02232
sigma	1.732015E-005	0.0041617	61.29154

5.5 Goodness of Fit Test With Covariates:

As usual we will use the best Known test of fit procedures being the Classical Pearson (χ^2) test.

For parametric Weibull distribution we know that

$$\int_{t_i}^{\infty} \beta\lambda e^{\theta X}\left(\lambda t e^{\theta X}\right)^{\beta-1} \exp\left[-\left(\lambda t e^{\theta X}\right)^{\beta}\right] dt$$

$$S(t_i) = \exp\left[-\left(\lambda t_i e^{\theta X}\right)^{\beta}\right.$$

$$\approx \exp\left[-\left(\lambda t_i e^{\theta\sum_{i_1}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right]$$

where $\bar{X}_i$ is the mean vector of covariates for the individuals in the i-th interval. Hence we can calculate the probability as

$$p_1 = 1-\exp\left[-\left(\lambda t_1 e^{\theta\sum_{i_1}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right]$$

$$p_2 = \exp\left[-\left(\lambda t_1 e^{\theta\sum_{i_1}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right] - \exp\left[-\left(\lambda t_2 e^{\theta\sum_{i_2}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right]$$

$$p_3 = \exp\left[-\left(\lambda t_2 e^{\theta\sum_{i_2}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right] - \exp\left[-\left(\lambda t_3 e^{\theta\sum_{i_3}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right]$$

$$\vdots$$

$$p_{n-1} = \exp\left[-\left(\lambda t_{n-2} e^{\theta\sum_{i_{n-2}}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right] - \exp\left[-\left(\lambda t_{n-1} e^{\theta\sum_{i_{n-1}}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right]$$

$$p_n = \exp\left[-\left(\lambda t_{n-1} e^{\theta\sum_{i_{n-1}}^{n}(\bar{X}_i-\bar{X})}\right)^{\beta}\right]$$

Consider the hypothesis

$$H_o : P_i = P_{io}\,,\ i = 1, \cdots\cdots, k+1$$

where the P_{io}'s are specified but may involve unknown parameters. Let $\tilde{P}_{io}$ be the m.l.e. of P_i under $E_i = n\tilde{P}_{io}$. The Pearson statistic for testing H_o is

$$\chi^2 = \sum_{i=1}^{k+1} \frac{(O_i - E_i)^2}{E_i}$$

The limiting distribution of χ^2 is $\chi^2_{(k-s)}$.

The results of the goodness of fit test are carried out by Microsoft Excel and the results are summarised in table# 5.5.1.

Table#5.5.1: Goodness of fit test of Weibull Distribution with covariates.

I_i	t_i	O_i	$S(t_i)$	Pi	Ei	$(Oi-Ei)^2/Ei$
<15	15	108	0.812753	0.187247	115.1569	0.444792
15-30	30	126	0.635344	0.177409	109.1064	2.615752
31-45	45	92	0.540858	0.094486	58.10905	19.76623
46-60	60	41	0.465769	0.075089	46.17973	0.580981
61-75	75	38	0.40566	0.06011	36.96736	0.028845
76-90	90	19	0.357719	0.047941	29.48363	3.727715
91-105	105	17	0.316582	0.041136	25.29879	2.722262
106-120	120	5	0.280677	0.035905	22.08163	13.2138
121-135	135	13	0.25082	0.029858	18.3625	1.566039
136-150	150	8	0.225322	0.025498	15.68132	3.762607
151-165	165	5	0.204267	0.021055	12.94882	4.879499
166-180	180	10	0.183493	0.020773	12.77552	0.602991
181-195	195	7	0.16508	0.018414	11.32442	1.651352
196-210	210	6	0.147604	0.017476	10.74778	2.09731
211-225	225	9	0.132345	0.015259	9.384328	0.01574
226+	---	111	---	0.132345	81.39188	10.77061
Total		615		1	615	68.44652

Here we have observed that the value of χ^2 statistic is smaller than that of obtained before in chapter four with MLE estimates. Hence we can say that the covariates have impact on control time of diabetes mellitus.

5.6 Impact of Other Diseases on Control Time:

In this section we will find out the impact of other diseases on the control time of diabetes mellitus using logistic regression model. We have used 10 variables as explanatory variable. Let us define them as

X_{12} = Heart problem (B),

X_{13} = Heart problem (A),

X_{14} = Dental problem (B),

X_{15} = Dental problem (A),

X_{16} = Kidney problem (B),

X_{17} = Kidney problem (A),

X_{18} = Eye problem (B),

X_{19} = Eye problem (A),

X_{20} = Other problems (B),

X_{21} = Other problems (A),

Then the model (5.3.1) reduces to the form

$$g(x_i) = \beta_0 + \beta_{12}x_{12i} + \cdots\cdots + \beta_{21}x_{21i}$$

In order to get the estimate of the parameters we have used the computer program SPSS for windows based 7.5 version. To test the significance of parameters we have used Wald test procedure, as usual.

We have categorise the dependent variable according to the median and mean control time and the results are given in table # 5.6.3.1A & 5.6.3.1B respectively.

Table # 5.6.3.1A: Impact of cofactors on control time.

Model Summary

-2log likelihood	Cox & Snell R Square
838.143	.023

Coefficient test

Cofactors	B	S.E.	Wald	df	Sig.	Exp(B)
Heart Problem (B)	-.140	.336	.173	1	.677	.869
Heart Problem (A)	.253	.300	.712	1	.399	1.288
Dental Problem (B)	-.122	.199	.376	1	.540	.885
Dental Problem (A)	.277	.222	1.565	1	.211	1.320
Kidney Problem (B)	.881	1.246	.500	1	.480	2.414
Kidney Problem (A)	.627	.889	.497	1	.481	1.872
Eye Problem (B)	-.043	.202	.046	1	.831	.958
Eye Problem (A)	.427	.218	3.849	1	.050	1.533
Other Problems (B)	.081	.174	.217	1	.641	1.085
Other Problems (A)	.037	.180	.043	1	.836	1.038
Constant	-.207	.183	1.272	1	.259	.813

Table # 5.6.3.1B: Impact of cofactors on control time.

Model Summary

-2log likelihood	Cox & Snell R Square	Nagelkerke R Square
746.281	0.025	0.036

Coefficient test

Cofactors	B	S.E.	Wald	df	Sig.	Exp(B)
Heart Problem (B)	-.046	.373	.015	1	.901	.955
Heart Problem (A)	.539	.300	3.231	1	.072	1.714
Dental Problem (B)	-.202	.222	.824	1	.364	.817
Dental Problem (A)	.223	.231	.928	1	.335	1.250
Kidney Problem (B)	1.623	1.248	1.692	1	.193	5.068
Kidney Problem (A)	.640	.837	.585	1	.444	1.897
Eye Problem (B)	-.087	.225	.149	1	.699	.917
Eye Problem (A)	.264	.232	1.295	1	.255	1.302
Other Problems (B)	-.196	.190	1.066	1	.302	.822
Other Problems (A)	-.068	.195	.122	1	.727	.934
Constant	-.814	.199	16.787	1	.000	.443

Here we observe that if the patient is affected by the disease eye problem after diagnose the diabetes mellitus then it is significant to the control time

at 5% level of significance and all other problems are insignificant when the dependent variable is categorized according to their median, but when it is categorized according to their mean then only heart problem after diagnosis of the diabetes mellitus, is significant at 7.2% level of significant.

5.7 Analysis of all cofactors together:

Now we want to investigate among all cofactors which has (have) impact on the control time by using logistic regression model. The results are carried out by the same program. Here we also categorize the dependent variable first according to their median and then according to their mean and the results are given table # 5.7.1A and 5.7.1B, respectively.

Table # 5.7.1A: Results of logistic model with all cofactors.

Model Summary

-2Log likelihood	Cox & Snell R Square
820.600	.051

Coefficient test

Cofactors	B	S.E.	Wald	df	Sig.	Exp(B)
X_1 = Age	-.014	.009	2.457	1	.117	.987
X_2 = Sex	.286	.184	2.405	1	.121	1.331
X_3 = Marital status	.226	.708	.102	1	.750	1.253
X_4 = Food habit (rice)	.395	.174	5.136	1	.023	1.485
X_5 = Food habit (vege)	.298	.235	1.606	1	.205	1.347
X_6 = Food habit (sweet)	.092	.172	.285	1	.593	1.096
X_7 = BMI	.006	.004	2.540	1	.111	1.006
X_8 = BP	.011	.023	.235	1	.628	1.011
X_9 = BGL	.031	.018	3.029	1	.082	1.032
X_{10} = Heredity	.113	.171	.432	1	.511	1.119
X_{11} = Length of suffering	.000	.000	4.496	1	.034	1.000
X_{12} = Heart Problem (B)	-.179	.343	.274	1	.601	.836
X_{13} = Heart Problem (A)	.231	.306	.570	1	.450	1.260
X_{14} = Dental Problem (B)	-.048	.211	.052	1	.820	.953
X_{15} = Dental Problem (A)	.339	.227	2.219	1	.136	1.403
X_{16} = Kidney Problem (B)	.754	1.259	.359	1	.549	2.126
X_{17} = Kidney Problem (A)	.595	.942	.398	1	.528	1.813
X_{18} = Eye Problem (B)	.020	.217	.009	1	.926	1.020
X_{19} = Eye Problem (A)	.432	.228	3.580	1	.058	1.541
X_{20} = Other Problems (B)	.122	.180	.460	1	.497	1.130
X_{21} = Other Problems (A)	.046	.188	.060	1	.807	1.047
Constant	-1.977	.947	4.356	1	.037	.138

Here we observe that food habit (rice), BGL, length of suffering and eye problem (A) gives significant impact on control time of diabetes mellitus.

Table # 5.7.1B:Results of logistic model with all cofactors.

Model Summary

-2Log likelihood	Cox & Snell R Square	Nagelkerke R Square
724.680	0.059	0.083

Coefficient test

Cofactors	B	S.E.	Wald	df	Sig.	Exp(B)
X_1 = Age	-.005	.009	.325	1	.569	.995
X_2 = Sex	.373	.199	3.503	1	.061	1.452
X_3 = Marital status	-.053	.733	.005	1	.943	.949
X_4 = Food habit (rice)	.139	.191	.534	1	.465	1.149
X_5 = Food habit (vege)	.252	.257	.967	1	.326	1.287
X_6 = Food habit (sweet)	.215	.187	1.327	1	.249	1.240
X_7 = BMI	.027	.025	1.149	1	.284	1.028
X_8 = BP	.004	.004	1.319	1	.251	1.004
X_9 = BGL	.038	.019	3.838	1	.050	1.039
X_{10} = Heredity	.279	.186	2.257	1	.133	1.322
X_{11} = Length of suffering	.000	.000	10.198	1	.001	1.000
X_{12} = Heart Problem (B)	-.074	.382	.037	1	.847	.929
X_{13} = Heart Problem (A)	.496	.308	2.591	1	.107	1.643
X_{14} = Dental Problem (B)	-.149	.235	1.170	1	.526	.862
X_{15} = Dental Problem (A)	.260	.240	1.752	1	.279	1.296
X_{16} = Kidney Problem (B)	1.664	1.257	.130	1	.186	5.278
X_{17} = Kidney Problem (A)	.327	.908	.137	1	.718	1.387
X_{18} = Eye Problem (B)	-.090	.241	1.031	1	.711	.914
X_{19} = Eye Problem (A)	.249	.246	.694	1	.310	1.283
X_{20} = Other Problems (B)	-.165	.198	.070	1	.405	.848
X_{21} = Other Problems (A)	-.054	.205	8.674	1	.791	.947
Constant	-2.963	1.006		1	.003	.052

Here we also observe that among the diseases all are insignificant. Among others sex, BGL and length of suffering may be considered as significant.

Now we introduce these cofactors in the weibull model and go through the analysis as before. Here we also used the computer program SPSS for windows to obtain the estimated values of coefficients of variables and their significance, which will be used as the initial value of the parameters in the model for Newton-Raphson iteration.

Table # 5.7.2A: Linear Regression Model.

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	103.266	21	4.917	2.553	.000
Residual	1142.179	593	1.926		
Total	1245.445	614			

Coefficient of determinant $R^2 = 0.083$

Table # 5.7.2B: Coefficient test.

Dependent variables	Estimated coefficients	t	Sig.
Constant	2.090	3.342	.001
X_1 = Age	-4.831E-03	-.833	.405
X_2 = Sex	.320	2.575	.010
X_3 = Marital status	7.202E-02	.153	.879
X_4 = Food habit (rice)	.253	2.147	.032
X_5 = Food habit (vege)	.271	1.703	.089
X_6 = Food habit (sweet)	.122	1.046	.296
X_7 = BMI	2.584E-02	1.653	.099
X_8 = BP	3.748E-03	1.541	.124
X_9 = BGL	2.169E-02	1.790	.074
X_{10} = Heredity	.180	1.550	.122
X_{11} = Length of suffering	2.476E-04	3.026	.003
X_{12} = Heart Problem (B)	-5.295E-03	-.023	.982
X_{13} = Heart Problem (A)	.432	2.115	.035
X_{14} = Dental Problem (B)	-.147	-1.028	.304
X_{15} = Dental Problem (A)	.220	1.438	.151

X_{16} = Kidney Problem (B)	.414	.508	.611
X_{17} = Kidney Problem (A)	.154	.262	.793
X_{18} = Eye Problem (B)	1.925E-02	.130	.896
X_{19} = Eye Problem (A)	.352	2.276	.023
X_{20} = Other Problems (B)	-1.672E-03	-.014	.989
X_{21} = Other Problems (A)	5.817E-02	.458	.647

Here we observe that sex, food habit (rice), BGL, length of suffering, heart problem (A) and eye problem (A) are significant. Using these cofactors we will go through the parametric model (weibull model) using the language program FORTRAN77 as stated in section 5.3.

Table # 5.7.3: Parametric regression model.

Coefficients	Estimated coefficient	standard error	Value of Z-statistic
X_2 = sex	6.645276	0.408923366	16.25066
X_4 = food habit (rice)	6.627613	0.123507623	53.66157
X_{11} = length of suffering	0.023081	1.9175505E-03	12.03681
X_{13} = heart problem (A)	3.422771	0.468130498	11.63771
X_{19} = eye problem (A)	5.447967	0.502160376	10.84905
sigma	1.20778	7.443587E-03	162.25777

From the above tables we observe that for the parametric regression model sex, food habit (rice), length of suffering, heart problem (A), eye problem (A) and sigma are significant. Whereas, in the logistic regression model only sex, BGL and length of suffering gives significant impact on control time of diabetes mellitus.

5.8 Goodness of fit test of the model with cofactors

Now using these five cofactors we have tested the goodness of fit test as described in section 5.5.

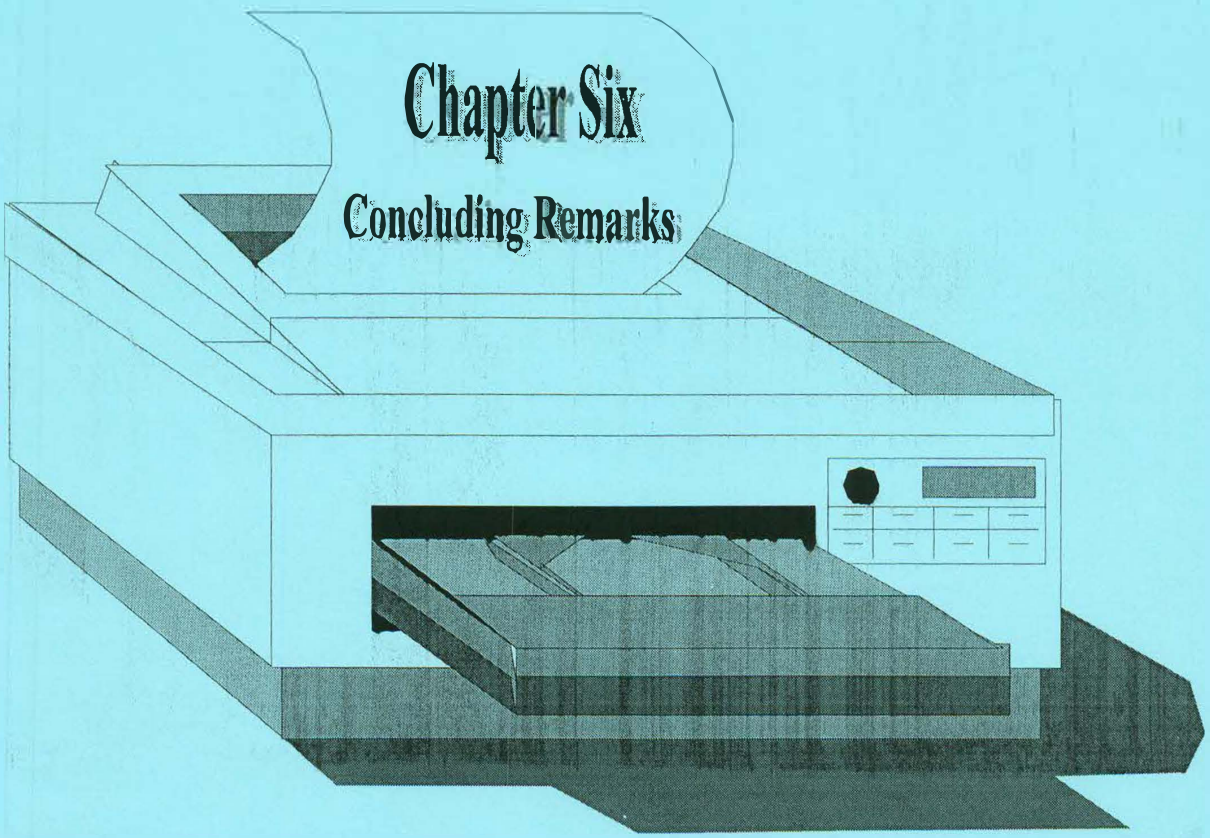
Table#5.8.1: Goodness of fit test of Weibull model with all cofactors.

I_i	t_i	Oi	$S(t_i)$	Pi	Ei	$(Oi-Ei)^2/Ei$
<15	15	108	0.811164	0.188836	116.134	0.569704
15-30	30	126	0.632828	0.178336	109.6765	2.429474
31-45	45	92	0.538844	0.093984	57.80016	20.23573
46-60	60	41	0.462531	0.076313	46.93246	0.749887
61-75	75	38	0.4022	0.060332	37.10392	0.021641
76-90	90	19	0.356073	0.046127	28.36796	3.093583
91-105	105	17	0.313548	0.042525	26.1529	3.203298
106-120	120	5	0.277612	0.035936	22.10081	13.23199
121-135	135	13	0.248228	0.029384	18.0709	1.42295
136-150	150	8	0.224613	0.023616	14.52369	2.930286
151-165	165	5	0.203993	0.02062	12.68123	4.652648
166-180	180	10	0.184972	0.019021	11.69777	0.246407
181-195	195	7	0.168724	0.016248	9.992542	0.896199
196-210	210	6	0.150509	0.018215	11.20215	2.415819
211-225	225	9	0.132697	0.017812	10.9541	0.348593
226+		111	---	0.132697	81.60891	10.58507
Total		615		1	615	67.03329

Here we have observed that the value of χ^2 statistic is smaller than that of obtained before in chapter four with MLE estimates and also that of in section 5.5. Hence we can say that the diseases have impact on control time of diabetes mellitus.

Chapter Six

Concluding Remarks



Chapter Six

Concluding Remarks

Diabetes is a very complex problem that leads to many difficulties when not properly handled. The people of all countries of the world - developed or underdeveloped or developing countries, are victims of this problem. The number of diabetic patients of the world is growing rapidly. BIRDEM reports the same results. In 1956 it starts with only 39 patients. In 2001 the number of registered patients of BIRDEM is 15188, which is alarming situation for us. Hence we need more research and more investigation on this disease. An overview of the disease and intensity of the problem is given in chapter one. Only basic materials have been discussed and supplemented with references leaving room for indepth & detailed informations, wherever it is necessary. The fringes of role and the significance of covariates & cofactors have been touched in literature review in order to keep the size of the dissertation to an optimum.

The most important part of a research is it's methodological aspects. Appropriate tools and techniques should amenable to build up models and to analysis. Otherwise the analysis may lead us to wrong results. Chapter two of this dissertation is contributed to the discussion of methodological aspects. Another important point for statistical analysis is the mode of collection and size of data. We know that a large size of data may also lead to a wrong conclusion if it is not collected systematically and from reliable source. Mode of collection and source of data is also appended in this chapter.

We know that diabetes is a serious problem of all countries of the world including Bagladesh. The only cause of this is that the preventive method of diabetes mellitus is not discovered yet now and if a person is attacked by this disease once, he (she) will never get rid of it, but can control it by treatment therapy along with other regularities. Hence we have given the emphasis most on control time – the time that a patient take to control the disease after registration as a diabetic patients at the center. From our collected data we have observed that about 25% of the cases become control within three weeks, while 50% of the patient take about 42 days to be in control state and on an average, patients take three months to be in control state. Association of this control time with covariates – age, sex, marital status, food habit (rice, vegetable, sweet), BMI, BP, BGL, heredity and length of suffering are also investigated in chapter three of this dissertation. Here we have used Pearson chi-square test and likelihood ratio test as test criterions. After investigation we have observed that the length of suffering is highly associated with control time of diabetes mellitus. Sex, food habit (rice) and BGL are also observed to have association with control time of diabetes mellitus. Here to be noted that we

have also collected data on the patients employment and initial urine sugar, two important covariates of diabetes mellitus. But because of incomplete informations we can not include these two factors in our analysis. Among other diseases heart problem, dental problem and eye problem are observed to have association with control time of diabetes mellitus.

Modelling the data is the main target of this study. Without an appropriate model no statistical analysis can be done. For seeking an appropriate model of control time of diabetes mellitus we have studied three most important life distributions – negative exponential distribution, weibull distribution and gamma distribution of exponential family as dictated by the frequency distribution of control time displayed in table # 3.2.1 of chapter three. We have checked these three distribution models by graphical method and test of goodness of fit procedure (pearson chi-square test) which leads us to a result that the data fits the most weibull distribution. The test of shape parameter of weibull distribution embedding exponential in weibull distribution confirmed this finding. These results are summerised in chapter four.

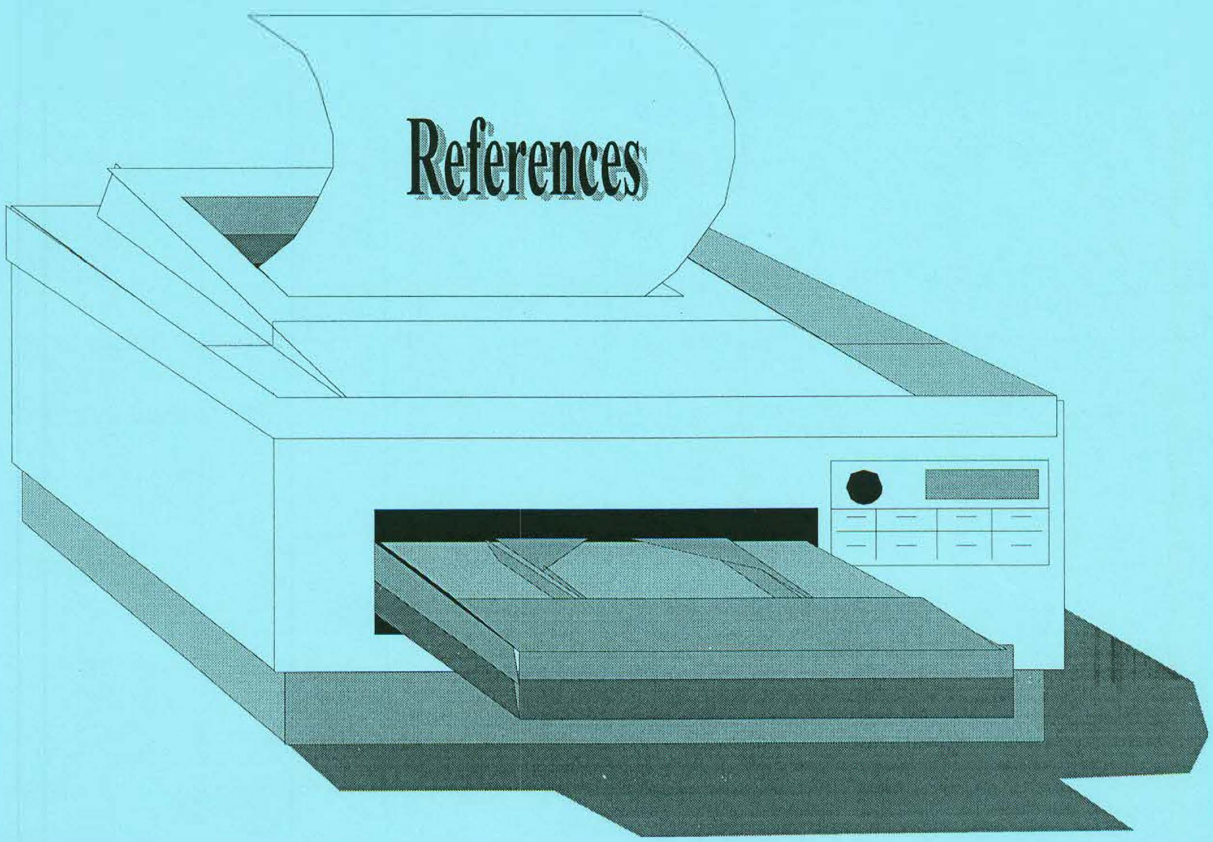
Next introducing all these associated covariates to the weibull model we have checked the impact of covariates on control time of diabetes mellitus in chapter five. There we observed that sex, food habit (rice) and suffering time have significant impact on control time of diabetes mellitus. After that we have also introduced other diseases in to the model and carried out the analysis. There we have observed that along with three covariates heart problem (A) and eye problem (A) have also significant impact on control time of diabetes mellitus. Logistic regression model is an interesting model

and the special case in which the response variable has only two categories is of particular interest and lends itself to an specially nice treatment. This is because, with only two categories, there is essentially only one way to define the odds. If p_1 is the probability in the first category and p_2 is the probability in the second category, then the odds of getting category one and two are $\frac{p_1}{p_2}$ and $\frac{p_2}{p_1}$. The important point is that either of these numbers, together with the fact that $p_1 + p_2 = 1$, completely determine both p_1 and p_2. So with two categories, the two choices for the odds lead to the same results. We have also checked the impact of the covariates and other diseases with logistic regression model. In that case we have categorised the dependent variable according to mean and median control time. We have also checked the goodness of fit of the parametric regression model (weibull model) and observed that the value of the chi-square statistic has become significantly smaller than that observed in table # 4.5(b) of chapter four. It is observed that sex, food habit (rice), suffering time, heart problem (A) and eye problem (A) have significant impact on control time of diabetes mellitus.

Some covariates like educational level, income, employment status, urine suger etc of patients may also be included in the analysis. Also the analysis may give better results if the patients are classified according to IDDM, NIDDM, MODY & MRDM, at least IDDM & NIDDM. Hence these all can be considered in further works on diabetes mellitus.

Finally, we may conclude that a Weibull parametric regression model is the best to graduate and analyse control time of diabetes mellitus. Apart from limitations mentioned above, findings of this study may help the diabetes centers to plan their future course of activities and also may help in assessing the rate of controlling.

References



References

Ahuja, M. M. S., and Shah, P. (1991). Profile of non-insulin dependen diabetes mellitus in India. *International Journal of Diabetes in India*, Vol-II, pp 3-4.

Barth, R. et al. (1991). Intensive education improves knowledge, compliance and foot problems in type-II diabetes. *Diabetic Med*, 8/2, pp 111-117 (*International Diabetes Monitor*, Vol 3. No. 4).

Buchanan, T. A. et al. (1990). Accelerated starvation in late pregnancy: a comparison between obese women with and without gestational diabetes mellitus. *Am. J. Obstet. Gynecol*, 162/4, pp 1015-1020 (*International Diabetes Monitor*, Vol 2, No. 5).

Chandalia, H. B. (1991). Non-insulin dependent diabetes in India. *International Journal of Diabetes in India*, Vol-II, p 1.

Dowse, G. K. et al., (1990). High prevalence of NIDDM and impaired glucose tolerance in Indian, Creole and Chinese Mauritians. *Diabetes*, 39, pp 390-396.

Ferdousei and Islam, (2000). A Second Order Markov Chain for Analyzing Covariate Dependence Pattern of Longitudinal Data with an Application to Diabetes Mellitus. *7th National Statistical Conference*, Bangladesh Statistical Association.

Gehan, E. A., and M. M. Siddiqui. (1973). Simple regression methods for survival time studies. *Journal of the American Statistical Association,* 68, pp 848-856.

Gries F. A. and Koschinsky T. (1991) Diabetes and arterial disease. *Diabetic Med*, 8-S, pp 82-87.

Gujaraty, Damodar N., (1995). *Basic Econometrics*. Third edition, McGraw- Hill Book Company.

Gulshan and Islam, (2000). An Extension of GEE for Repeated Measures with Polytomous Responses. *7th National Statistical Conference*, Bangladesh Statistical Association.

Gupta, S. C., and V. K. Kapoor. (1994). *Fundamentals of Mathematical Statistics*. Ninth edition, Delhi : Sultan Chand & Sons.

Hamada, Y. et al. (1997). Effects of Glycemic Control on Plasma 3-Deoxyglucosone Levels in NIDDM Patients. *Diabetes Care*, 20/9, pp 1466-1475.

Hanson, T. (1998). Insulin Sensitivity Index, Acute Insulin Response and Glucose Effectiveness in a Population-Based Sample of 380 Young, Healthy Caucasians. *International Diabetes Monitor*, vol, 10. p A-7.

Hinkley, D. V. (1978). Likelihood Inference About Location and Scale Parameters. *Biometrika*, 65, pp 253-262.

John, J. N. and Jones, K. L. (1998). Early- Onset Insulin- Resistant Diabetes in Mexican- American Adolescents. *International Diabetes Monitor*, vol, 10, p A-5.

Johnston, C. (1990). Islet Function and Insulin Sensitivity in the Non-Offspring of Conjugal Type 2 Diabetic Patients. *Diabetes Med*, 7, pp 119-244.

Kabir and Islam, (2000). Analyzing Heterogeneous Transitions Using GEE for a Two-State Markov Model wth an Application to Diabetes Mellitus. *7th National Statistical Conference*, Bangladesh Statistical Association.

Knuth, S. A. (1969). *The Airt of Computer Programming.* Vol, 2, Reading Mass : Addison-Wesley.

Kuzuya, T. (1997). Diagnosis and classification of Diabetes Mellitus. *Asian Medical Journal,* 40(6), 271-277.

Latif, Z. A. (1993). Pathophisiology of Diabetes Mellitus. *The 1st Novo Nordisk Diabetes Update Seminar in the Memory of Late Prof. M. Ibrahim*, Dhaka.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, New York : John Wiley and Sons, Inc.

Lipschutz, S., and Arthurpoe. (1982). *Theory and Problems of Programming with FORTRAN,* Singapore : McGraw-Hill Book Company.

Mahtab, H. (1993). Diabetic Care in Bangladesh Present and Future Strategies. *The 1st Novo Nordisk Diabetes Update Seminar in the Memory of Late Prof. M. Ibrahim*, Dhaka.

Mallick and Islam, (2000). Estimation and Test for Markov Chain Based Logistic Regression Model for Longitudinal Data. *7th National Statistical Conference*, Bangladesh Statistical Association.

Mathiesen, B. (1997). Diabetes and Accident Insurance. A 3-year follow-up of 7,599 Insured Diabetic Individuals. *Diabetes Care*, 20/11, pp 1781-1785.

Mc Keigul, P. M. et al (1989). Diabetes, Insulin Resistance and Central Obesity in South Asians and Europeans. *Diad. Med,* 6 suppl 1, pp A41-42.

Mian, A. K., and Ansari, A. J. (1993). Hypertension and Diabetes Mellitus. *The 1st Novo Nordisk Diabetes Update Seminar in the Memory of Late Prof. M. Ibrahim*. Dhaka.

Mian, M.A.B. (1987). On the selection of a life testing model. *Assam StatisticalReview,* Vol. 1, No.2.

Mitchell, B. D. et al. (1990). Cigarette smoking and neuropathy in diabetic patients. *Diabetes Care*, 13/4, pp 434-437 (*International Diabetes Monitor*, Vol. 2, No. 5).

Montogomary, D. C., and E. Peck. (1982). *An Introduction to Linear Regression Analysis*, New York : John Wiley & Sons.

Mood, A. M., and F. A. Graybill, and D. C. Boes. (1974). *Introduction to the theory of statistics*, McGraw-Hill, Inc.

Muggeo, M. et al. (1997). Long-Term Instability of Fasting Plasma Glucose, a Novel Predictor of Cardiovascular Mortality in Elderly Patients With Non-Insulin- Dependent Diabetes Mellitus: the Verona Diabetes Study. *Circulation*, 96/6, pp 1750-1754.

Nabarro J. D. N (1991). (Diabetes in UK/; a personal series, *Diabetic Med*, 8/1, pp 59-68.

Omar M. A. K. et al (1985). The Prevalence of Diabetes Mellitus in a Large Group of South African Indians. *S. Afr. Med. J,* 67, pp 924-930.

Osei, K. (1990). Increased Basal Glucose Production and Utilization in Nondiabetic First Degree Relative of Patients with NIDDM. *Diabetes*, 39, pp 597-601.

Ramchandran, A. (1988). High Prevalence of Diabetes in an Urban Population in South India. *Br. Med. J,* 293, pp 589-679.

Rohatgi, V. K. (1975). *An Introduction to Probability and Mathematical Statistics*. New York : John Wiley and Sons.

Sayeed, M. A. (1993). Epidemiological Study of Diabetes Mellitus in Bangladesh. *The 1st Novo Nordisk Diabetes Update Seminar in the Memory of Late Prof. M. Ibrahim,* Dhaka.

Sharmin, (2000). Assesment of Time Varying Exposure in a Hazards Model. *7th National Statistical Conference*, Bangladesh Statistical Association.

Sprent, P. (1969). *Models in Regression and Related Topics.* Third edition, London : Methuen.

Tietyen, J. (1989). Dietary fiber in foods: options for diabetes education. *Diabetes Education*, 15/6, pp 523-529 (*International Diabetes Monitor*, vol. 2, No. 5, 1990).

Toeller M. (1991). Dietary treatment of diabetes mellitus, *Fortschr. Med*, 109/2, pp 41-45.

Walker, W. G. (1990). Prospective study of the impact of hypertension upon kidney function in diabetes mellitus. *Nephron*, 55, suppl 1, pp 21-26 (*International Diabetes Monitor*, Vol. 2, No. 5).

Warram, J. H. et al. (1991). Risk of IDDM in children of diabetic mothers decreases with increasing maternal age at pregnancy. *Diabetes*, 40/12, pp 1679-1684 (*International Diabetes Monitor*, Vol. 4, No. 3).

Watts, N. B. (1990). Prediction of glucose response to weight loss in patients with non-insulin-dependent diabetes mellitus. *Arch. Intern. Med*, 150/4, pp 803-806 (*International Diabetes Monitor*, Vol. 2, No. 5).

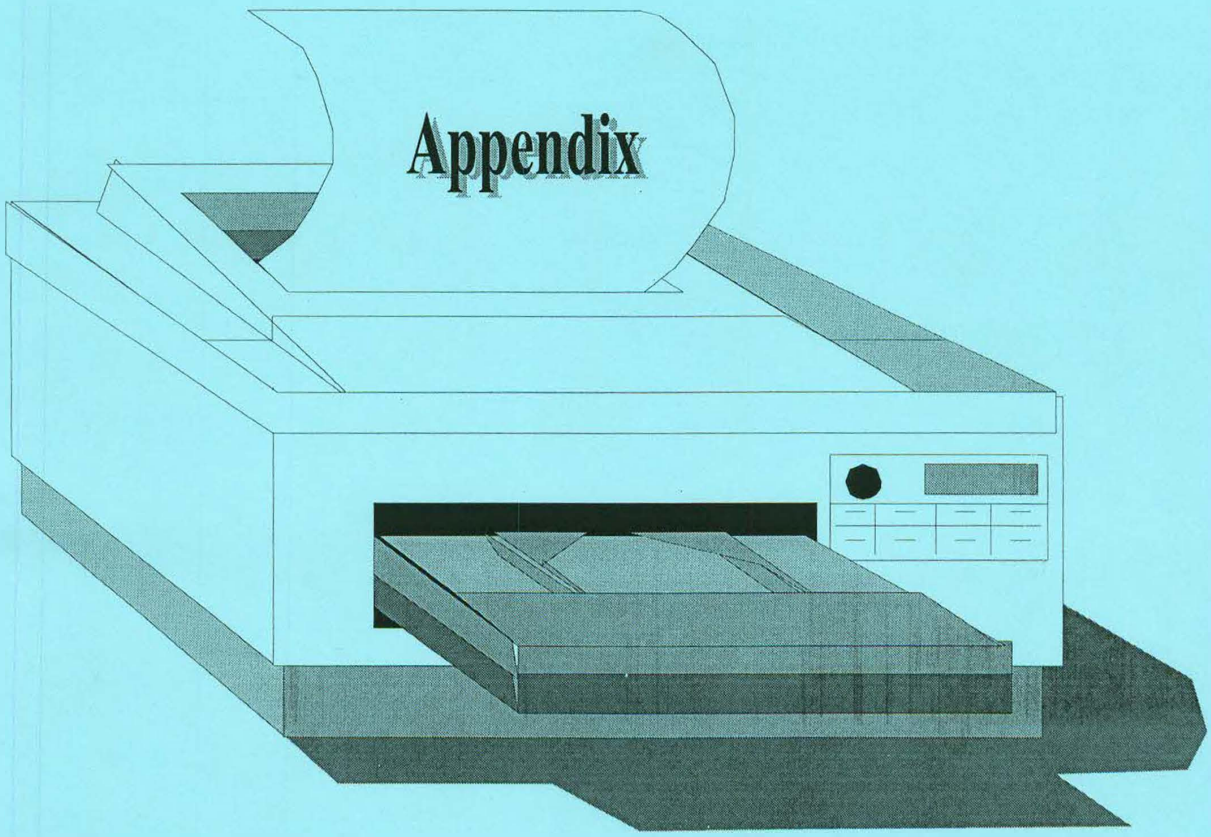
Weir M.R. & Bakris G.L. (1992). Risk for renal injury in diabetic hypertensive patients; the physiologic basis for blood pressure control. *Postgrad, Med./91/3,* pp 83-87 (*International Diabetes Monitor*, Vol.4, No.4, 1992).

Wilson P. W. F. et al. (1991). Is hyperglycemia associated with cardiovascular disease? The Framingham study, *Am. Heart J*, 121/2-I, pp 586-590.

WHO Study Group. (1985). Diabetes Mellitus, WHO Technical Report Series, 727, Geneva.

Zimmet, P. (1983). Prevalence of Diabetes and Impaired Glucose Tolerance in the Biracial (Melanesian and Indian) Population of Fiji : a Rural-Urban Comparison. *Am. J. Epidemiol*, 118, pp 673-167.

Appendix



Appendix

Department of Statistics
University of Rajshahi
Research Project: Modelling of Diabetes Mellitus Data
Fellow: Mst. Papia Sultana
Supervisor: Dr. M. Aabul Basher Mian

(Information obtained from this questionnaire would be kept secret and used for research purpose)

Date: .....

Questionnaire

1. Name of the patient:
2. Address:
3. Card No.:
4. Profession:
5. Marital Status: Married / Unmarried.
6. Food habit :
   - ☐ Rice-rice-rice / Bread-rice-rice / Bread-rice-bread.
   - ☐ Meat / Fish / Vegetable.
   - ☐ Sweet / Sour / Hot.
7. Age:
8. Sex: Male / Female
9. Height:
10. Weight:
11. Needed Weight:
12. Blood Pressure:
13. BGL:

14. Urine Sugar:

15. Heredity: Yes / No / Unknown.

16. How long had you been suffering from the disease before taking the treatment?

.....

17. When have you come to take the treatment?

.....

18. When have you come to a controlled state?

.....

19. Other Complication and from when (before / after the disease), if any. :

- ❑ Heart disease, ...
- ❑ Dental complication, ...
- ❑ Kidney problem, ...
- ❑ Eye problem, ...
- ❑ Others, ...

Rajshahi University Library
Documentation Section
Document No ...D.-.2175
Date....22/6/03.....