ANALYSIS OF INFANT MORTALITY IN THE VICINITY OF JIMMA ZONE

BY

FETENE BEKELE

Addis Ababa University

A Thesis submitted to the School of Graduate Studies of Addis Ababa university in partial fulfillment of the requirements for the degree of Master of Science in Statistics

June, 2000

Addis Ababa, Ethiopia

TABLE OF CONTENTS

Addis Ababa University
Library

ACKNOWLEDGEMENTS

First and for most, I would like to extend my sincere gratitude to my advisor **DR. Tadewos Koroto**, Head department of Statistics, AAU for his kind and ever-present advise and encouragement throughout this work, spending a lot of his time with me.

My special thanks go to the Jimma Infant Survival and Differentials Project (**JISDP**), which permitted me to use the data in this thesis. Especially **DR. Makonen Asefa**, Coordinator of the project and head of Epidimeology and Biostatistics Department, Jimma University (the former Jimma Institute of Health Sciences), and **Ato Fasil Tessema**, Lecturer, Jimma University, for their continuous encouragement and provision of materials. I also benefited from frequent contacts with colleagues studying at Community Health Department of Addis Ababa University, **Drs. Tadale Bogale** and **Eyob Lemma**.

I am also grateful to my friend **Ato Solomon Chefo**, Lecturer in the Department of Statistics, AAU for his all time encouragement and materials throughout the period of the research work.

Last but not least, it is my pleasure to thank my sister **W/t Alemtsehay Maru**, who is always behind me and helped me to get computer at critical times.

Addis Ababa University
Library

ABSTRACT

There has been an explosion of interest in the analysis of survival data in the last 35 years, which is resulting in the development of many new theoretical ideas and useful methods, specially in the study of the relationship between survival times and explanatory variables. The application of survival data analysis is common in different fields ranging from Economics to Engineering. The value of survival analysis techniques is wide in medical statistics. The focus of this thesis is on the application of survival data analysis to the data generated from epidemiological study conducted from 1992 to 1994 in Jimma, Keffa-Sheka and Illubabour zones.

Methodological procedures to handle problems originated from survival data are given briefly. A review on determinants of infant survival from related studies is made. The main objectives of the thesis are to determine the survival pattern up to first birthday and investigate the possible risk factors that contribute most to the early mortality. Demographic, socioeconomic, Environmental status, health service usage and traditional practice indicator variables are considered. The basic Cox regression model is fitted to get the effect of a variable, adjusted for other variables. Results are given both in numerical estimates and graphical presentations where applicable.

The descriptive analysis shows that neonatal mortality rate is 26.6 per 1000 live births, postnatal mortality rate is 73.1 per 1000 live births, and the overall Infant Mortality Rate (IMR) is 101.9 per 1000 live births. The cumulative survival probabilities are

0.9855(s.e=0.0013), 0.9736(s.e=0.0018), 0.9039(s.e.=0.0033) for days 7, 28 and 360, respectively.

The final Cox's regression model shows that breast-feeding practice, vaccination status at birth, weight at birth, family size, death of previous children, sex of an infant, maternal age, marital status, visit to maternal clinic, swallowing butter, place of delivery, and availability of latrine facility are significantly important variables in determining infants' chance of survival. Analysis by age of infants shows that socioeconomic and environmental variables are influential at later ages of life. Also analysis by place of residence shows that the effect of these variables is important only at urban areas.

Finally, relevant discussion is made and possible recommendations are given accordingly to interested policy makers and front-line health workers.

CHAPTER 1

INTRODUCTION

1.1. GENERAL

Survival analysis is the analysis of data that correspond to the time from a well- defined time origin until the occurrence of some particular event or end-point. It is one area of statistical application in medical field that has been drawing greater attention of recognition. The present state of the art in survival analysis is the product of a long process, which has undergone a great impulse in the last 65 years. Researchers from Biomedical fields, as well as form industry, have stimulated this process with their scientific problems. Specially, the last 35 years have seen widespread application of the methods of survival analysis to clinical data [1]. In the clinical context, survival time is used to indicate not only 'time to death' but also time to any event, generally defined as failure (example, disease progression, metastasis or toxic events). The methodology can also be applied to data from other application areas such as the survival times of animals in an experimental study, the time taken by an individual to complete a task in a psychological experiment, the storage times of seeds being kept in a seed bank, or the life times of individual or electronic components [2]. In general survival analysis is useful whenever the researcher is interested not only in the frequency of occurrence of a certain type of event, but also in the time process underlying such occurrence. The focus of this thesis is on the application of survival analysis to the data arising from epidemiological study conducted from 1992 to 1994 on infants from parts of South West Ethiopia.

Effective planning, management and evaluation of health services to any segment of the population require adequate, reliable and timely health information. Such information is

not readily available in many developing countries, including Ethiopia. Current data deficiencies are large and serious in their implications for rational health planning and health research [3]. The Jimma setting community based birth cohort study is one of rare well-defined studies in developing countries to fill the gap for health planners and researchers. The main aim of the study was to describe infant growth in the cohort and to investigate some of its determinant [4]. The analysis of the data from this study is being undertaken by different researchers for the purpose of getting feed backs from different disciplines for the final intervention program. The main aim of my work is to contribute findings on the infant survival pattern and the risk factors affecting the chance of survival up to the first year of life to the research output of the Jimma research team.

Death in the first year of life has long been used as a marker of the socioeconomic development of a nation. Within Africa in 1996 only as high as 55 infant deaths per 1000 live births has been observed in Algeria, Morocco, South Africa and Egypt, whereas rates of over 110 per 1000 live births have been reported in Sierra Leone, Mali, Malawi and Chad [5]. The summary information on world population data sheet [6], 1997 shows that the infant mortality rate is 9 per 1000 live births for more developed countries whereas it is 64 for less developed. For North African countries it is 60 per 1000 live births and 93 per 1000 live births, for Sub- Saharan Africa. These figures indicate a negative relationship between infant mortality and socioeconomic development.

A projection for 1997, using data from 1994 census in Ethiopia shows, the health service coverage is 48.5%, the infant mortality rate is 105 per 1000 live births. Under–1 year population is 3.37% of the 57.27 Million total population. The percentage of women aged 15-49 years is 23.26% (4.05% of it is pregnant women). The GDP (per capita at

constant factors cost) is 253.4 Million Ethiopian Birr for 1996/97. Concerning the environmental health service, access to safe water is 27% (19% and 80% for rural and urban respectively). The availability of latrine facilities is 10% for national and surprisingly only 1% in rural and 60% in urban areas (which is dominated by Addis Ababa) [7]. Having this background information (expected to hold for the study area) the thesis attempts to see the pattern of infant survival, a study that had not been attempted in Ethiopia.

1.2. HOW SURVIVAL DATA ARISE?

The phrase "survival data" is being used in the context of measuring the duration of time until some event. Survival data are collected when the objective of an undertaking is to study the time elapsed from some particular starting point to the occurrence of an event. In biomedical studies the starting point may be:-

- a medical intervention such as first diagnosis of a given disease,
- a surgical intervention or the beginning of a treatment,
- specifically, for epidemiological studies it may be birth or the beginning of exposure to some risk factor.

The final event may be death or a pre-specified event of interest (occurrence of a certain response or recurrence of disease). The time taken from the starting point to the occurrence of an event is represented by the random variable T. And T is called the "*survival time*" or sometimes "*failure time*" which is assumed to be a non-negative random variable. The realizations generated for the random variable T are referred to as *survival data*.

A distinctive characteristic of survival data is that the event of interest may not be observed on every experimental unit/ individuals, that is, one should not wait until an event of interest occurs on all observations entered to the study at the starting point. This feature leads to truncation of data and it is known as *censoring*. Censoring can arise because of time limits and other restrictions depending on the nature of the experiment/study. There are two types of censoring

"Type I censoring" – The study terminates after a pre-specified period.

"Type II censoring"- The study terminates after a pre-specified r events occur.

Let`s consider an experiment on N individuals for a continuous treatment. To carry out the experiment until all individuals respond to the treatment could be unethical and may be costly, especially, when there is an adverse effect. Therefore, the experiment could be modified in two ways: First, the experiment is terminated when a pre-specified time T^* is reached. This induces the so-called "type I censoring", in which the response time is known exactly only for individuals who respond to the treatment before T^*. In this case the number of individuals responded is a random variable. Second, the experiment terminates when a number of responded individuals, r is reached, with $r \leq N$. This determines the so-called "type II censoring". In this design, the number of failures is not a random variable.

In epidemiological studies, censoring is mainly caused by a time restriction, and therefore of type I [1]. For example, in risk factor study time and economic reasons suggest that the study continue until a pre-specified time point (called cut-off date) starting from exposure date. Then the time to the event of interest is known precisely only on those subjects who develop the event before that time point. For the remaining individuals (those who do not develop the event during the period from starting to cut-off date) it is only known that the time to the event is greater than the pre-specified time. The incomplete data for the latter groups is called "*right censored*" and the subjects providing them are called *withdrawn alive* from the study. However, in medical and epidemiological studies, the censoring process is caused not only due to predetermined duration of the study. Enrolled individuals may be unwilling or unable, for some reasons, to continue participating in the

study and providing follow-up information up to the pre-specified time. These subjects are *lost to follow up or dropouts*. As those withdrawn alive, they give incomplete data (information is available only as far as they are under follow up), which are "right censored", indicating that the event has not yet occurred before they leave the study.

For a sample of n independent individuals, the observed data in survival analysis are often represented by the form (t_j, δ_j) j=1, 2, . . ., n, where $t_j \in (0,\infty)$ is the time an individual is known to have survived before the event, $\delta_j \in \{0, 1\}$ is an indicator of failure, assuming values of δ_j=1 if the event of interest is observed at t_j and δ_j=0 if the time is right censored at t_j (event on j^{th} individual has not occurred up to time t_j, and the individual is either drop-out or withdrawn alive at t_j).

If data were also collected on additional variables for each individual enrolled in the study, the data would be represented in the form $(t_j, \delta_j, \underline{X}_j)$ where the additional $\underline{X}_j$ is a vector of known covariates for j^{th} individual. When covariates are considered in a survival analysis there may be covariates which are time dependent, that their value may vary during the study period. This must be carefully considered specially in long-term studies [1].

1.3 STATISTICAL PROBLEMS ON SURVIVAL DATA

Statistical procedures on survival data, specifically generated from epidemiological studies, can be applied to deal with the following main problems:-

- Estimation of survival time (failure time) distributions

- Identification of risk factors, jointly or singularly considered.

In subsequent sections we will briefly discuss how statistical procedures would handle these problems.

1.4 SURVIVAL DISTRIBUTIONS

In most areas of statistics, stochastic models for random variables are expressed in terms of densities and distribution functions. In survival analysis the equivalent functions are *survivorship* and *hazard functions*.

1.4.1 THE SURVIVORSHIP FUNCTION

Let T denotes the continuous random variable specifying time until the event of interest. The realizations of T are survival time. The survivorship function denoted by S(t), is defined as the probability that an individual survives longer than t , and is given by

$$S(t)=P_r\{T>t\} \quad =1-F(t) \qquad (1)$$

where F(t) is a distribution function for T.

S(t) has the following properties:

- $\lim_{t \downarrow 0} S(t) = 1$
- $\lim_{t \uparrow \infty} S(t) = 0$
- S(t) is non-increasing function.
- a graphical presentation of S(t) is called the survival curve, which is a continuous function defining the survival rate at any time after time zero [8].

1.4.2 THE HAZARD FUNCTION

The hazard function of survival time T, h(t), represents the instantaneous or immediate risk of death or failure for an individual who has survived to time t. It is defined as the probability of failure during a very small time interval, or as the limit of the probability that the individual has survived to the beginning of the interval, or as the limit of the probability that an individual fails in a very short interval, t to t+Δt, given that the individual has survived to time t. The hazard function h(t) is given as

$$h(t) = \lim_{\Delta t \downarrow 0} P\left\{\frac{\beta_t}{\Delta t}\right\}$$

where β_t is the event that an individual of age t fails in the interval (t, t+Δt).

From the earlier discussion we can give the hazard function as

$$h(t) = \lim_{\Delta t \downarrow 0} P\left\{\frac{C_t \mid D_t}{\Delta t}\right\}$$

where C_t is the event that an individual fails in the interval (t, t+Δt)

D_t is the event that an individual survived up to t.

It follows that

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta_t} p(C_t) \Big/ p(D_t).$$

Hence,

$$h(t) = \frac{f(t)}{S(t)} \quad - - - - - - - - - - \quad (2)$$

The hazard function, h(t), is also known as the instantaneous failure rate, force of mortality, conditional mortality rate, age specific failure rate. It is a measure of the proneness to failure as a function of age of the individual. Thus h(t) gives the risk of

failure per unit time during the aging process.

1.4.3. RELATIONSHIP OF THE SURVIVAL FUNCTIONS

The two functions discussed above have a relationship which can be shown as follows

From (1)

$$F(t)=1-S(t) \quad \Rightarrow F'(t)=-S'(t)$$

Therefore, $f(t)= -S'(t)$ (since $f(t)= F'(t)$)

Using this in (2) we obtain

$$h(t) = -\frac{d}{dt}\log\ S(t)$$

By integrating the hazard function from 0 to t

$$-\int_0^t h(u)\,du = \log\ S(t)$$

from which follows

$$S(t) = \exp\{\ -\int_0^t h(u)\,du\ \}$$

Usually this last equation is expressed as

$$S(t)=\exp\{-H(t)\} \qquad \text{--------(3)}$$

where $H(t) = \int_0^t h(u)\,du$

Thus, from the relationship between S(t) and h(t) is given above, one of them is obtained if the other is given and vise versa. That is why most of the survival analysis focuses only on one of these functions. Mostly the hazard function is preferred for its convenience in

incorporating the covariates in regression models [8, 9].

1.5 ESTIMATION OF THE SURVIVAL FUNCTIONS

Due to the nature of the data (see Section 1.1) a method of estimation of the survival data functions should:

- Accommodate censoring; information on drop-out or withdrawn alive should be used as the individual is under observation,
- account for different periods of observation, life spans, on each individual,
- account for the time at which events occur.

The methods of estimation which are used in this study do not require the form of the function of T to be specified. The procedures use the simple idea of subdividing the observation time, age, into a sequence of time intervals. Then, the proportion of survivors at each time unit will be used to describe the data in terms of survival curve.

1.5.1 SURVIVORSHIP PROBABILITY ESTIMATES BASED ON PRODUCT LIMIT-METHOD

Suppose failure (death) and censored times $\{t_1, t_2, \ldots, t_n\}$ are known on n independent individuals in a random sample, and we wish to estimate the probability of surviving one year from birth. The product limit (PL) method is based on the simple consideration that in order to survive one year from the beginning of the observation (birth), the individual/infant has to survive every day from the first to 360^{th} . We know that the estimate of the probability of surviving a given day is a proportion of individuals/infants, among those alive just before the day, who survive until the next day. For example at day

28 from birth, none of the infants who had a failure (died) or were censored on one of the previous days convey any information on the probability of surviving day 28. Only infants who are alive and under observation (non-censored) just before day 28, including those whose failure time is 28, are useful for estimating, the probability of survival at day 28, p(28). These are called *infants at risk*, and the probability p(28) is estimated by

$$\hat{p}(28) = \frac{\text{number of infants at risk in day 28 - no. of failures on day 28}}{\text{number of infants at risk in day 28}}$$

Now, by definition these probabilities, for each day, are conditional probabilities. Therefore, to obtain the probability of surviving one full year, we accumulate step by step the probabilities of surviving each day by taking their product.

Let $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, $r \leq n$, be the distinct ordered failure times observed among n infants, (here (j) indicates the j^{th} ordered time not j^{th} infant)

Let d_j (j=1, 2 . . ., r) be the corresponding number of failures (deaths) with $d_j \geq 1$. The number n_j of infants at risk is the number of infants actually observed when deaths occur, that is, the number of infants who have either failure or censored times greater than or equal to $t_{(j)}$, then $p(t_{(j)})$ is estimated by

$$\hat{p}(t_{(j)}) = \hat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \frac{d_j}{n_j} = 1 - \hat{q}_j \quad \text{----------} (4)$$

where $\hat{q}_j = \frac{d_j}{n_j}$ is the estimated conditional probability of death at $t_{(j)}$.

Then the survival probability S(t) is estimated by the product

$$\hat{S}(t) = \prod_{j \mid t_{(j)} \leq t} \hat{p}_j \quad \text{- - - - - - -} (5)$$

The conditional probability in (4) is 1 if there is no death at $t_{(j)}$. Therefore, $\hat{S}(t)$ in (5) only changes at time points $t_{(j)}$, j = 1, 2, . . ., r, when at least a death occurs. The corresponding graph of $\hat{S}(t)$ versus t is a step function. Since the value of $\hat{S}(t)$ at each death time $t_{(j)}$ is the one on the right of $t_{(j)}$, it is right continuous.

The estimator of S(t) given by (5) is called the Product Limit estimator. Since it was derived by Kaplan and Meier in 1958, it is often called the Kaplan-Meier(K-M) estimator.

1.5.2. STANDARD ERROR OF THE KAPLAN-MEIER ESTIMATES

As any statistic it is subjected to a random variation, it is desirable to report $\hat{S}(t)$ with its standard error. The derivation of the standard error of $\hat{S}(t)$ is given below.

Taking logarithim of (5), that is,

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j$$

from which follows $\operatorname{var}\{\log \hat{S}(t)\} = \sum_{j=1}^{k} \operatorname{var}\{\log \hat{p}_j\}$

Now let ℓ be the number of individuals who survive through the time interval $(t_{(j)}, t_{(j+1)})$. Thus, ℓ ~bin (n_j, p_j) where p_j is the true probability of survival through that interval. Here the observed number of survivors is $\ell = n_j - d_j$. It follows then that

$$\operatorname{var}(\ell) = \operatorname{var}(n_j - d_j) = n_j p_j (1 - p_j)$$

out of which we get $\operatorname{var}(\hat{p}_j) = \dfrac{\operatorname{var}(n_j - d_j)}{n_j^2} = \dfrac{p_j(1-p_j)}{n_j}$.

Thus the variance of $\hat{p}_j$ is estimated by $\frac{\hat{p}_j(1-\hat{p}_j)}{n_j}$ (6)

Now let

$$g(t) = \log \hat{p}_j .$$

Then the Taylor series approximation to the variance of a function of a random variable X is

$$\text{var}\{g(x)\} \approx \left[\frac{dg(x)}{dx}\right]^2 \text{var(x)} \qquad (7).$$

It follows that

$$\text{var}\{\log \hat{p}_j) \approx \frac{\text{var}(\hat{p}_j)}{\hat{p}_j^2} = \frac{\hat{p}_j(1-\hat{p}_j)}{n_j p_j^2} = \frac{1-\hat{p}_j}{n_j \hat{p}_j}$$

Then substituting $\hat{p}_j$ by $\frac{n_j - d_j}{n_j}$ we get,

$$\text{var}\{\log \hat{p}_j\} \approx \frac{d_j}{n_j(n_j - d_j)},$$

from which we get

$$\text{var}\{\log \hat{S}(t)\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \qquad (8)$$

but from (7) $\text{var}\{\log \hat{s}(t)\} \approx \frac{1}{[\hat{s}(t)]^2} \text{var}\{\hat{s}(t)\}$.

Using this in (8) we get

$$\frac{\text{var}\{\hat{s}(t)\}}{[\hat{S}(t)]^2} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

Hence,

$$\text{var}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}.$$

Therefore, the standard error of $\hat{S}(t)$ can be approximated by

$$s.e\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right] \sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}} \quad \text{for } t_{(k)} \le t \le t_{(k+1)}.$$

For detail information about the proof see Reference 2.

1.5.3. ESTIMATING THE HAZARD FUNCTION

There are two ways of estimating the hazard function, which show the dependence of the instantaneous risk of death on time. These are Method of Life Table Estimates and the Kaplan-Meier methods. The Kaplan-Meier Method which will be used in this analysis is discussed briefly.

The Kaplan-Meier Method OF ESTIMATION OF THE HAZRD FUNCTION

Assuming the hazard function is constant between successive times, the hazard per unit time (risk of death for that time) can be found by dividing the ratio of deaths at that time to the number of individuals at risk at that time by the time interval. Thus, if there are d_j deaths at the j^{th} death time $t_{(j)}$, j=1,2, . . . r, and n_j at risk at time $t_{(j)}$, the hazard function in the interval from $t_{(j)}$ to $t_{(j+1)}$ can be estimated by

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \quad \text{for } t_{(j)} \le t < t_{(j+1)} \text{ ----------9}$$

where $\tau_j = t_{(j+1)} - t_{(j)}$, and this estimate is referred to as the Kaplan-Meier (K-M) estimate.

1.5.4. ESTIMATION OF THE CUMULATIVE HAZARD FUNCTION

The cumulative hazard function at time t, H(t) . from the relation (3) is

$$H(t) = -\log(s(t))$$

Therefore, using the K-M estimate for S(t) given in (4), the estimate of H(t) is given by

$$\hat{H}(t) = -\log \hat{S}(t) = -\sum_{i=1}^{k} \log\left(\frac{n_i - d_i}{n_i}\right) \quad \text{for } t_{(k)} \le t < t_{(k+1)} \qquad k=1,2,\ldots r.$$

Using the series expansion of log (1-x), and ignoring the higher order terms

$$\log\left(\frac{n_j - d_j}{n_j}\right) = \log\left(1 - \frac{d_j}{n_j}\right) \approx \frac{d_j}{n_j}$$

Therefore, $$\hat{H}(t) \approx \sum_{j=1}^{k} \frac{d_j}{n_j}$$

which is the cumulative sum of the estimated probabilities of death from the first to the k^{th} time interval, k=1,2, . . . , r .

1.6 COMPARISION OF SURVIVAL FUNCTIONS FOR GROUPS: A BIVARIATE ANALYSIS

In this section we compare the survivor function of two or more groups of subjects that differ by a given characteristic, that is, to compare a survivor function of infants in different categories of a given factor.

If there are g categories for a given factor, we estimate the survivor function for each category using the K-M procedure discussed under Section 1.5.1. Now the problem is that of testing whether these functions are the same or not. For this we are going to check whether the different g categories of the factor affect the survival or not. Thus the overall null hypothesis to be tested is

H_O: $S_1(t) = S_2(t) = \ldots = S_g(t)$. for all t.

Against the composite alternative hypothesis that there are at least two groups with different survival curves.

The test which could be used is the log-rank test or the Breslow or the Tarone-Ware test. These tests will provide the same result. The log-rank test is used in this analysis.

Log-rank test

Suppose there are r distinct death times, $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, across the g categories, and at time $t_{(j)}$, d_{kj} infants in group k die. For k = 1, 2, . . ., g, j = 1, 2, . . ., r. Suppose n_{kj} infants are at risk of death in k^{th} category just before time $t_{(j)}$, and consequently, at time $t_{(j)}$. There are $d_j = \sum_{k=1}^{g} d_{kj}$ deaths in total out of $n_j = \sum_{k=1}^{g} n_{kj}$ individuals at risk. The summary information at the j^{th} death time can be given as follows:

Category	Number at risk just before $t_{(j)}$	Number of deaths at $t_{(j)}$
1	n_{1j}	d_{1j}
2	n_{2j}	d_{2j}
.	.	.
.	.	.
.	.	.
g	n_{gj}	d_{gj}
Total	n_j	d_j

Thus, under the null hypothesis, the probability of death at time $t_{(j)}$ does not depend on the category that an individual belongs. The probability of death at $t_{(j)}$ is d_j/n_j. Multiplying this by n_{kj} gives the expected number of deaths in category k under the null

hypothesis, that is,

$$E(d_{kj}) = e_{kj} = n_{kj}\frac{d_j}{n_j} \quad \text{for k=1,2, ... g}$$

Now to combine the information from the death times and to get an overall measure of deviation of the observed values of d_{kj} from their expected values, we sum the differences d_{kj}-e_{kj} over the total number of death times, r and the resulting statistics is given by

$$U_k = \sum_{j=1}^{r} \left(d_{kj} - e_{kj}\right) \qquad \text{for k=1,2,..., g-1.}$$

These quantities are expressed in the form of a vector with g-1 components, which is denoted by $\mathbf{u}_L$ with coresponding variance-covariance matrix $\mathbf{v}_L$, where $(k,k')^{th}$ element of $\mathbf{v}_L$ is given by [2]

$$Var\{U_k, U_{k'}\} = \sum_{j=1}^{r} \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\varphi_{kk'} - \frac{n_{kj}}{n_j}\right)$$

where $\varphi_{kk'} = \begin{cases} 1, & \text{if } k = k' \\ 0, & \text{otherwise} \end{cases}$ $\qquad k,k' = 1,2,...\,g-1.$

Thus, to test of the null hypothesis, we use the test statistic

$$\mathbf{u}'_L \mathbf{v}^{-1}_L \mathbf{u}_L$$

which has a central chi-square distribution with g-1 degrees of freedom H_0.

1.7. MODELLING SURVIVAL DATA: MULTIVARIATE ANALYSIS

In most studies which give rise to survival data, supplementary information would be recorded on each individual. Data may be recorded on demographic, socioeconomic, environmental, biomedical, etc characteristics for each individual. And the aim of the

study may be to explore the relationship between these characteristics and the survival pattern. In regression analysis these variables are referred to as explanatory variables, whereas the failure time is referred to as dependent variable. In survival data analysis, the approach based on statistical modelling can be extended to the hazard function. The hazard function is modelled directly by incórporating the explanatory variables, and then, using the relationship (see Section 1.4.3) an estimate of the survivor function would be obtained. Therefore, the auxilary objective of the modelling process is to determine which potential explanatory variables affect the form of the hazard function.

The basic model for survival data, the proportional hazards (PH) model would be used. This model was first proposed by Cox (1972) [10] and is also known as the Cox regression model. Although, the model is based on the assumption of proportional hazard, it does not assume any particular form of the probability distribution of survival times. Therefore, the model is referred to as a *semi-parametric model.*

1.7.1. THE COX REGRESSION MODEL

Definition Let N be the number of individuals in the study at the beginning, each with observed vector (t_i, δ_i, $\underline{X}_i$). The basic model assumes that the hazard function for failure time t for an individual i with covariate vector $\mathbf{X}_i^{'}$ = (X_{1i}, X_{2i}, …, X_{pi}) is

$$h\left(t, \mathbf{x}_i^{'}\right) = h_o(t)\exp\left(\beta' \ \mathbf{x}_i\right) \quad \text{for i=1,2,… ,N} \quad \text{----------(10)}$$

where $h_o(t)$ is a function of time only, and assumed to be the same for all subjects (usually called the baseline hazard function), and

β is a px1 vector of regression coefficients.

The above is the Cox regression model.

The Cox regression model given in (10) is not a fully parametric model since it does not specify the form of $h_o(t)$. It does, however, specify the hazard ratio for any two individuals with covariate vectors $\underline{X}_1$ and $\underline{X}_2$, and for this reason it is defined as a semi-parametric model. The function $\exp(\beta' x_i)$ is a function of the values of the vector of explanatory variables for the i^{th} individual. This function can be interpreted as *the hazard at time t for an individual whose vector of explanatory variables is $\underline{X}_i$, relative to the hazard for an individual for whom $\underline{X} = 0$.*

1.7.2 ESTIMATION OF THE PARAMETERS

Since the functional form of $h_o(t)$ is unspecified, it is not possible to use an ordinary likelihood to estimate the regression coefficients. Therefore, the Cox's partial likelihood function which will be used in the analysis will be breifly discussed below.

Consider a sample of n subjects where r failures occur, $r < n$, due to the presence of censoring. Let $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ be the r distinct ordered failure times observed and let R(t) be the set of subjects at risk at time t, who are alive and under observation just before t. Let X_j be a vector of covariates of the subject who fails at $t_{(j)}$ labeled j^{th} subject. The probability that an individual with covariates $\underline{X}$ fails in a small interval (t, t+dt), given the set at risk at t is $h(t,\underline{X})dt$. Thus, conditional on the fact that one individual is observed to fail/die at $t_{(j)}$, the probability that it is an individual with covariates $\underline{X}_j$ is

$$h(t_{(j)}, \mathbf{x}_j)dt \Big/ \sum_{l \in R(t_{(j)})} h(t_{(j)}, \mathbf{x}_l)dt$$

It follows that the function describing the failure pattern is the product of r terms (see Reference 11).

$$L(\beta) = \prod_{j=1}^{r} \left[h(t_{(j)}, \mathbf{x}_j)dt \Big/ \sum_{l \in R(t_{(j)})} h(t_{(j)}, \mathbf{x}_l)dt \right] \text{----------(11).}$$

Now for the data consisting n observed survival times, including the failure and censored times, the likelihood function can be expressed in the form

$$L(\beta) = \prod_{j=1}^{n} \left[h(t_{(j)}, \mathbf{x}_j)dt \Big/ \sum_{l \in R(t_{(j)})} h(t_{(j)}, \mathbf{x}_l)dt \right]^{\delta_j}$$

$$= \prod_{j=1}^{n} \left[\frac{\exp(\beta' \mathbf{x}_j)}{\sum_{l \in R(t_{(j)})} \exp(\beta' \mathbf{x}_l)} \right]^{\delta_j}$$

where δ_j is an indicator of censoring.

And the corresponding log-likelihood function is given by

$$LL(\beta) = \log L(\beta) = \sum_{j=1}^{n} \delta_j \left\{ \beta' \mathbf{x}_j - \log \sum_{l \in R(t_{(j)})} \exp(\beta' \mathbf{x}_l) \right\} \text{----(12)}$$

Thus, the maximum likelihood estimates of the β parameters in the proportional hazards model can be found by maximizing this log-likelihood function using a numerical method, namly the Newton Raphson procedure and this is handeled by the SPSS statistical package.

1.7.3 STANDARD ERROR OF THE PARAMETER ESTIMATES

Let $U(\beta)$ be the px1 vector of first derivatives of the log-likelihood function in (11) with respect to the β parameters. That is,

$$u_{\eta}(\beta) = \frac{\partial LL(\beta)}{\partial \beta_{\eta}}, \quad \eta=1,2,\ldots, p.$$

Let the pxp matrix $I(\beta)$ be the matrix of negative second derivatives of the log-likelihood, so that the $(j,k)^{th}$ element of $I(\beta)$ is

$$\frac{\partial^2 LL(\beta)}{\partial \beta_j \partial \beta_k} \qquad j,k=1,2,\ldots,p$$

The matrix $I(\beta)$ is known as the observed information matrix. Thus, according to the Newton-Raphson procedure, an estimate of the vector of β parameters at the $(s+1)^{th}$ step of the iterative procedure, $\hat{\beta}_{s+1}$, is

$$\hat{\beta}_{s+1} = \hat{\beta}_s + I^{-1}(\hat{\beta}_s)U(\hat{\beta}_s) \quad \text{for } s = 0, 1, 2, \ldots$$

where $I^{-1}(.)$ is the inverse of the information matrix.

The iterative process can begin by taking $\hat{\beta}_0 = 0$.

The process terminates when the change in the log-likelihood function is sufficiently small, or when the largest of the relative changes in the values of the parameter estimates is sufficiently small.

If the iterative procedure converges, the variance-covariance matrix of the parameter estimates can be approximated by the inverse of the information matrix, evaluated at $\hat{\beta}$, that is, $I^{-1}(\hat{\beta})$. The square roots of the diagonal elements of this matrix,

respectively, are the standard errors of the estimated values of the parameters, $\beta_1, \beta_2, \ldots, \beta_p$.

Treatment of ties

The proportional hazards model for survival data assumes that the hazard function is continuous, and under this assumption, tied survival times are not possible. However, the survival times are usually recorded to the nearest day, month, or year (in this study to the nearest day), and so tied survival times can arise as a result of this rounding process. In order to accomadate tied observations, the partial likelihood function in equation (11) has to be modified.

Let s_j be the vector of sums of each of the p explanatory variables for those individulas who die at the j^{th} time, $t_{(j)}$, j = 1, 2, ..., r. If there are dj deaths at t(j), the hth element of s_j is

$s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, where x_{hjk} is the value of the h^{th} explanatory variable, h = 1, 2, ... p, for k^{th} of d_j individuals, k = 1, 2, ..., d_j who die at the j^{th} death time, j = 1, 2, ..., r. The partial likelihood function which incorporates tied deaths is given by

$$L(\beta) = \prod_{j=1}^{n} \left[\frac{\exp(\beta' s_j)}{\left[\sum_{l \in R(t_{(j)})} \exp(\beta' \mathbf{x}_l) \right]^{d_j}} \right]^{\delta_j}.$$

1.7.4 INTERPRETATION OF PARAMETER ESTIMATES

The coefficients of the explanatory variables in the model can be interpreted as logarithms of the ratio of the hazard of death to the baseline hazard.

Models with a variate

Suppose that a proportional hazards model contains a single continuous variable X, so that the hazard function for the i^{th} of n individuals for whom X takes the value x_i; is

$$h_i(t) = h_o(t)\exp(\beta x_i)$$

Now consider the ratio of the hazard of death for an individual for whom the value x+1 is recorded on X, relative to one for whom the value x is obtained, i.e.,

$$\frac{\exp\{\beta(x+1)\}}{\exp\{\beta x\}} = \exp\{\beta\}$$

Thus, $\hat{\beta}$ in the fitted proportional hazards model is the estimated change in the logarithim of the hazard ratio when the values of X is increased by one unit. Therefore, the coefficient of X can be interpreted as the logarithim of a hazard ratio.

Models with a factor

When individuals fall into one of g groups, g≥2, which correspond to categories of an explanatory variable, the groups can be indexed by the levels of a factor. The hazard function for an individual in the j^{th} group, j=1,2,..., g is given by

$$h_j(t) = \exp\{\gamma_j\}h_0(t)$$

where γ_j is the effect due to the j^{th} level of the factor, and, $h_0(t)$ is the baseline hazard function.

Usually, we take $\gamma_1 = 0$ so that the baseline hazard function corresponds to the hazard of death at time t for an individual in the first group. *The ratio of the hazards at*

time t for an individual in the j^{th} group, $j \geq 2$, relative to an individual in the first group (reference category) is then $\exp(\gamma_j)$. Consequently, the parameter γ_j is the logarithim of this relative hazard. that is,

$$\gamma_j = \log\left(h_j(t)/h_0(t)\right)$$

A model which contains the terms γ_j , j = 1,2,..., g, with $\gamma_1 = 0$. can be fitted by defining g-1 indicator variables, $X_2, X_3, \ldots, X_g$. Fitting this model leads to estimates $\hat{\gamma}_2, \hat{\gamma}_3, \ldots, \hat{\gamma}_g$, and their standard errors. The estimated logarithim of the relative hazard for an individual in group j, relative to an individual in group 1 is $\hat{\gamma}_j$.

Sometimes, the hazard ratio relative to the level of a factor other than the first (reference category) level may be required. The hazard functions for individuals at level j and k of the factor are respectively $\exp(\alpha_j)h_0(t)$ and $\exp(\alpha_k)h_0(t)$, and so the hazard ratio for an individual at level j, relative to one at level k, is $\exp(\alpha_j - \alpha_k)$. The log hazard ratio is then $\alpha_j - \alpha_k$ which is estimated by $\hat{\alpha}_j - \hat{\alpha}_k$.

Models with combinations of terms

If a fixed model contains terms corresponding to a number of variates, factor, or combination of the two, the parameter estimates can again be interpreted as logarithims of hazard ratios. In such cases, the parameter estimate associated with a particular effect is said to be adjusted for the other variables in the model, and so the estimates are log-hazard ratios, adjusted for the other terms in the model.

1.7.5 STATISTICAL TESTS ON THE COEFFICIENTS

The statistical inference on $\beta = (\beta_1, \beta_2 ..... \beta_p)'$ relies on a test based on the properties of the likelihood function. Suppose particular values for g of the p parameters are postulated. Thus the hypothesis to be tested is

$$H_0 : \beta_1 = \beta_{10}, ..., \beta_g = \beta_{g0}$$

where $\beta_{10}, \beta_{20}, ..., \beta_{g0}$ are specified values. Let the parameters in H_0 to be the first g components of the vector of the parameters ($g \leq p$). If we split β in to two vectors of parameters (β_g, β_r), the above null hypothesis becomes

$$H_0 : \beta_g = \beta_{g0}.$$

The test imposes g restrictions on the parameter to be estimated. The parameter β_r indicates the set of the remaining r = p-g parameters which are left unspecified by the hypothesis. The vector β_g may contain just one or more parameters. One of the following three tests, all leading to the same result, could be applied on the hypothesis.

The **Likelihood ratio (LR) test**:

This test procedure needs to fit both the unrestricted and the restricted models. Obtain the value of the log-likelihood function $LL(\hat{\beta}_g, \hat{\beta}_r)$ where $LL(\hat{\beta}_g, \hat{\beta}_r)$ is the joint maximum likelihood estimate of β_g and β_r in the unrestricted model, and of $LL(\beta_{g0}, \tilde{\beta}_r)$ where $\tilde{\beta}_r$ is the ML estimate of β_r when the model imposes the g restrictions in H_0. The test statistic for H_0 is based on the ratio of these log-likelihood values. Thus under H_0 the statistic

$$Q_{LR} = 2\left[LL(\hat{\beta}_g, \hat{\beta}_r) - LL(\beta_{g0}, \tilde{\beta}_r)\right]$$

is asymptotically distributed as central χ^2 with g degrees of freedom equal to the number of restrictions imposed by the null hypothesis, g.

The **Wald test:**

Requires fitting the unrestricted model, and is based on the ML estimator $\hat{\beta}_g$. The test statistic is

$$Q_W = (\hat{\beta}_g - \beta_{g0})' \left[I^{-1}_{g,g}(\hat{\beta}_g, \hat{\beta}_r) \right]^{-1} (\hat{\beta}_g - \beta_{g0})$$

where $I^{-1}_{g,g}(\hat{\beta}_r, \hat{\beta}_r)$ is a sub-matrix of dimension (gxg) of the entire variance-covariance matrix of $(\hat{\beta}_g, \hat{\beta}_r)$ estimated under the unrestricted model.

This test statistic is approximatelly distributed as central χ^2 with g degrees of freedom under H_0. If the null hypothesis imposes one restriction only on the parameter β_k, i.e. g=1, then the statistics reduces to

$$Q_W = \frac{(\hat{\beta}_k - \beta_{k0})^2}{\text{var}(\hat{\beta}_k)} \sim \chi^2_{(1)}.$$

Rao's score test:

It requires only fitting the restricted model. It is based on the $LL(\beta_{g0}, \tilde{\beta}_r)$. The test statistics is

$$Q_R = U' \left[I^{-1}_{g,g}(B_{g0}, \tilde{B}_r) \right]^{-1} U$$

where $u = \hat{\beta}_g - \beta_{g0}$.

Q_R has approximately a central χ^2 distribution with g degrees of freedom, under H_o.

The three tests often lead to identical conclusions on the parameters.

1.7.6 VARIABLE SELECTION

In survival data analysis, one of our goals is to identify variables related to survival and build a model that excludes variables that do not appear to be good predictors. So, the aim in variable selection is to determine which of the variables has/have an effect on the hazard function. The usual methods for variable selection in statistical modelling can be used to build Cox's regression model. In this study the Rao's score statistic is used for entering variables into a model, and likelihood ratio statistic based on conditional parameter estimates is used for variable removal. The Forward stepwise (conditional LR) automatic routine is used. For the steps of this routine see References 2 and 12.

1.8 ESTIMATING THE HAZARD AND SURVIVOR FUNCTIONS IN THE PRESENCE OF COVARIATES

So far we have only considered the estimation of the β parameters in the linear component of a proportional hazards model. This estimation is required in order to draw inferences about the effect of explanatory variables on the hazard function. Once a suitable model for a set of survival data has been identified, the hazard function and the corresponding survivor function, can be estimated. Suppose there are p explanatory variables and the regression coefficients are estimated. The estimated hazard function for the i^{th} individual in the study is given by

$$\hat{h}_i(t) = e^{\hat{\beta}' x_i} \hat{h}_0(t) \quad , i=1,2, \ldots, n.$$

Using this equation, the hazard function for an individual can be estimated once an estimate $\hat{h}_0(t)$, baseline hazard function, has been found. The relationship between the hazards, cumulative hazard and survivor functions (Section 1.4.3) can then be used to give estimates of the cumulative hazard function and the survivor function.

Breslow [13] derived a maximum likelihood estimator for $H_0(t)$, the cumulative hazard for baseline after assuming the failure time distribution has a hazard function which is constant between each pair of successive observed failure times [13]. The estimate of $h_0(t)$ in the interval $(t_{(j-1)}, t_{(j)})$ then becomes

$$\hat{h}_j = \frac{d_j}{(t_{(j)} - t_{(j-1)}) \sum_{i \in R(t_{(j)})} \exp(\beta' \mathbf{x}_i)}.$$

Note that when $\beta = 0$ this equation reduces to 9 in Section 1.5.3.

Here a rough estimate of $H_0(t_{(j)})$-$H_0(t_{(j-1)})$ is $\hat{h}_0(t_{(j)})(t_{(j)} - t_{(j-1)})$. Summing such terms overall $t_{(j)} \leq t$ gives the cumulative failure rate at time t

$$\hat{H}_0(t) = \sum_{j | t_{(j)} \leq t} \frac{d_j}{\sum_{i \in R(t_{(j)})} \exp(\beta' \mathbf{x}_i)} \qquad \text{for } t \in (t_{(j)}, t_{(j-1)}]$$

which is a step function.

Thus, using the relationship (3) the baseline survivor function can be estimated by

$$\hat{S}_0(t) = \prod_{t_{(j)} \leq t} \left[\exp\left(\frac{-d_j}{\sum_{i \in R(t_{(j)})} \exp(\hat{\beta} \mathbf{x}_i)} \right) \right].$$

For individuals with a certain covariate vector x, not necessarily in baseline, the estimated cumulative hazard and survivor functions are

$$\hat{H}(t, \mathbf{x}) = \hat{H}_0(t) \exp(\hat{\beta}' \mathbf{x}) \qquad \text{and}$$

$$\hat{S}(t, \mathbf{x}) = [\hat{S}_0(t)]^{\exp(\hat{\beta}' \mathbf{x})}$$

1.9 VALIDATION OF THE PROPORTIONAL HAZARDS MODEL

After a model has been fitted to an observed set of survival data, the adequacy of the fitted model needs to be assessed. Here in the Cox-regression model the proportionality of hazards is the fundamental assumption, and it is worth checking as thoroughly as possible. Also regression diagonstics for assessing goodness-of-fit will be done.

A.. GRAPHICAL METHOD FOR CHECKING PROPORTIONALITY OF HAZRADS

Under Proportionality Hazard (PH) assumption the model satisfies the relationship

$$H\ (t, \mathbf{x}) = H_0(t) \exp\ (\beta' \mathbf{x})$$

it holds that $$-\log S(t, \mathbf{x}) = \exp(\beta' \mathbf{x})[-\log(S_0(t))].$$

The logarithim of this last function exhibits a constant distance $\beta' \mathbf{x}$, under proportional hazard assumption

$$\log[-\log S(t, \mathbf{x})] - \log[-\log S_0(t)] = \beta' \mathbf{x}$$

Consider two different covariate patterns, that is, two sets of values $\underline{X}_1$ and $\underline{X}_2$ for the vector $\underline{X}$. Thus, under proportional hazard assumption, the functions $\log[-\log S(t, \mathbf{x}_1)]$ and $\log[-\log S(t, \mathbf{x}_2)]$ would exhbit a constant distance from the reference cummulative hazard $\log[-\log S_0(t)]$, and would therefore be parallel.(see References 1 and 2). This property can be used for a graphical check of the proportional hazard assumption in the presence of the variables. We use the stratified Cox-model, for the p variables under consideration, each one, say x_k, be checked for the validity of the assumption as follows. Let the vector $\underline{X}$ be partitioned as $X=(x_k, X^-)$ where $X^- =(X_1, X_2, \ldots, X_{K-1}, X_{K+1}, \ldots, X_P)$ is the vector of p-1 variables after , the k^{th} component is excluded. For example, if x_k is a binary with values 0 or 1; the basic Cox model (10)

would assume the form

$$h(t,\mathbf{x}) = \begin{cases} h_0(t)\exp(\beta'^{-}\mathbf{x}^{-}) \ , & \text{for } x_k = 0 \\ h_0(t)\exp\beta_k \exp(\beta'^{-}\mathbf{x}^{-}) \ , & \text{for } x_k = 1 \end{cases}$$

That is, hazards of two individuals with the same X^- but different values of x_k are proportional, with ratio $\exp(\beta_k)$. So, to verify this assumption on x_k let us consider two strata identified by the values of x_k . The first stratum for those with value x_k=0 and the other stratum for those with value of x_k=1. Applying the Cox's basic model on each strata results

$$h(t,\mathbf{x}) = \begin{cases} h_{01}(t)\exp(\beta'^{-}\mathbf{x}^{-}) & , \text{for } 1^{st} \textit{ stratum} \\ h_{02}(t)\exp(\beta'^{-}\mathbf{x}^{-}) & , \text{for } 2^{nd} \textit{ stratum} \end{cases}$$

where the baseline functions, $h_{01}(t)$ and $h_{02}(t)$ in the two strata are left arbitrary and unrelated.

Here the covariates X^- are assumed to satisfy proportional hazard assumption approximately, and the estimates of the coefficients, $\hat{\beta}^-$ are obtained by maximizing the partial likelihood over the entire sample.

B. IDENTIFICATION OF INFLUENTIAL OBSERVATIONS

The aim is to determine whether any particular observation has an impact on inferences made on the basis of a model fitted. As most of the inferences from the model are made through the estimates of the parameter coefficients, we try to identify those observations which may significantly affect the estimates of the coefficients. Suppose that we wish to determine whether any particular observation has an effect on $\hat{\beta}_j$, the j^{th} parameter estimate, j=1,2,...,p, in the fitted Cox regression model. A statistic that estimates the change in the j^{th} coefficient with and without a case called Dfbeta is computed for each

case. Then cases with outlying values for Dfbeta will be identified as influential observations and examined. An easy way of examination is using the graphical plot of the Dfebta against the case identification

number [2,12] .

C. RESIDUALS FOR IDENTIFICATION OF OUTLIERS

The Cox-Snell residual for the i^{th} individual is given by

$r_{Ci} = \exp(\hat{\beta}' \mathbf{x}_i)\hat{H}_0(t_i)$, i = 1, 2, ..., n.

The censored observations lead to residuals that can not be regarded on the same footing as residuals from uncensored observations. Therefore, the Cox-Snell residuals need modification so that explicit account can be taken for censoring. Thus the modified Cox-Snell residual for the i^{th} individual is

$r_{Mi} = \delta_i - r_{Ci}$, i =1, 2, ..., n.

These residuals are used in the analysis of survival data, and are called Martingale residuals [1, 2, 12].

Plotting Martingale residuals against the linear predictor score, which is the sum of the products of the coefficients and the variables for each case, would facilitate checking the presence of an outlier observation. We expect the values appear to be randomly distributed in a band arround 0.

1.10 OBJECTIVES OF THIS STUDY

- General objectives:

  To determine the survival pattern up to the first year of life and investigate the possible risk factors that contribute most to early mortality.

- Specific objectives:
  - To see the pattern of survival at first year of life,
  - To identify the risk groups with respect to variables such as the socioeconomic, demographic, environmental, traditional practices and health service usage.
  - To contribute findings to the Jimma Infant Survival and Differentials Project (JISDP), which is planning strategies for the next phase, and
  - To provide detail information on infant survival for concerned policy makers and the scientific community for further research.

CHAPTER 2

DETERMINANTS OF INFANT SURVIVAL FROM RELATED STUDIES

It is obvious that the burden of diseases and death is high in developing countries. This is seen mainly in vulnerable groups such as infant, children and women in the reproductive phase of life. The hostile physical environment in the developing world also manifests itself in a high mortality rate. The indices which (some times taken twice, also indirectly as life expectancy) would help to assess performances of health programs and also indicates the socioeconomic status of the community [14]. Although neonatal and post neonatal (or jointly infant) mortality rates are estimated to be high in many developing countries very few studies were conducted in this area, and most of the studies were either retrospective or cross-sectional studies. Specifically, there is relatively little knowledge regarding mortality differentials in Sub-Saharan Africa. The cause of the problem may be absence of data, and another cause with equal importance may be failure to exploit data that already exist. Census and survey organizations that are responsible for data gathering activities often lack the capacity to convert their data into information, and the data never reach those who could use them. There may be surveys on mortality that have never been coded, coded but never analyzed, and analyzed but never published. Donors often find it easier to justify launching a new survey than providing financial aid to complete an unfinished project [15].

In Ethiopia, according to the 1994 population and housing census result the infant mortality rate (IMR) is 116 per 1000 live births. For male and female it is 125 and 108 per 1000 live births, respectively. The IMR is 98 and 121 per 1000 live births for urban

and rural, respectively [16]. For Oromiya region, where the majority of cases in this study came from, regional infant mortality rate is 118 per 1000 live births (for male 128, for female 108, for urban 93 and for rural 121 per 1000 live births) [17]. Though not reported the result of the national census, neonatal and post neonatal mortality rates are undoubtedly very high as compared to other countries. In study based on two woredas of Addis Ababa, early neonatal mortality (Death at first week of life) was 50.9 per 1000 live births and late neonatal (Death at 8^{th} to 28^{th} day of life) was 20.9 per 1000 live births. The overall neonatal mortality was 71.9 per 1000 live births [18]. A study on Jimma infant's birth cohort reported that IMR is 115/1000 [4]. This estimate is substantially higher than the 1994 census estimate for Jimma, 93/1000 live births [17].

To allocate limited resources appropriately and to evaluate the effects it is very important to know the major causes of mortality and how these causes will contribute to affect survival in due courses of life. A "determinant of infant survival" can be defined as a variable that would change an infant's survival level if its own value altered. Many characteristics of the mother play a vital role in the survival chance of the infant. Many of past studies have attempted to identify the most important determinants of infants' survival in different geographical areas, but nothing is done in Ethiopia. The relative importance of socioeconomic, environmental and demographic factors as determinants of infants' chance of survival varies with the level of economic development of a society. Studies identified the type of variables that affect the infant's survival, but the effect of these variables may vary from population to population. Though the suspected list of variables that may affect survival chances of infants is long, most of the studies concentrated on few variables. An assessment of literature on infant survival show that

those variables that have different effects on the risk of mortality can be classified in to two broad categories. namely

- Socioeconomic and environmental factors. and
- Demographic and factors related to reproductive or maternal characteristics

These measures are indicative of the status of the environment and women, with effects on the survival chances of infants. Though it is hypothesized that in traditional society the latter set of factors affect infants' chance of survival than the former set [19, 20], it is advantageous to concentrate jointly on both of them. As almost all analyses on infant survival include variables from one (or both) of these sets, brief review on each set is given.

a. Socioeconomic and environmental factors

These are generally described as characteristics of the surrounding environment and the status of the family or household into which the infant is born. Variables commonly included in infant survival are mother's and husband's education, place of residence, family income, sanitary facilities such as toilet, water source and waste disposal, marital status, mother's occupation, cultural and breast feeding practices, public health factors such as access and usage of medical cares for mother and infants [15,18-31].

Though studies have shown that these socioeconomic and environmental factors should be considered in the analysis, little is known about the relationship between each factor and the survival chances of infants. Naturally, one would expect that infants from poor environment and low socioeconomic classes face higher risk of

dying, but this is not theoretically well established and is intensively debated among researchers [30]. Also some past studies show that the influence of these variables differ from country to country [23] as observed from the following brief discussion on determinants.

i. Mother's education

Almost all studies on infant and child survival analyses focus on this variable, and some studies have got a surprising results. Naturally, it is expected that a higher level of mother's education would be associated with lower risk of mortality and higher chance of survival, because it is a proxy for increased command over resources, resulting in higher quality of life, clothing, shelter, nutrition, medical care, sanitary facilities and water supply. Thus this expected positive association between education and infant survival has been observed in different population [15, 20-21, 23 -26]. On the contrary, in some populations it has been found that the effect of education on infant mortality is weak [30,32].

ii Place of residence

Place of residence of the infant is one of the most important determinant in survival analyses. It could be used as a proxy for living condition, and both public and medical health provisions are expected to be associated with this variable. Also it could be used as proxy for inequality in the provision of services such as water supply, sewage and waste disposal. In many studies differentials by urban and rural residence have been observed, with urban areas having the advantage [23-26]. However, this urban/rural differential may be substantially reduced if the effects of other variables are controlled.

iii Water sources and sanitary facilities

These characteristics may be directly associated with infant's survival chance. It is expected that infants from a household with safe water source and better sanitary facilities would have a better chance of survival. One study found that. the mortality of infants born to households with pipe water supplies was significantly lower than that of infants in houses where water came from other source, e.g. wells, rivers or canals. The risk is, especially, higher during the first month and last six months of the first year of life. In the same study babies born to households with toilet facility were significantly less likely to die within their first year, except during the last three weeks of the first month [33].

iv Access to Health services and usage

It is expected that a good health care system with a quality service for mothers and infants would lead to a better chance of survival for a baby. Mothers who have properly attended antenatal care services would have a less chance of delivery complication and infant deaths. Vaccination to newly born babies is very vital for their survival. The objective of immunization is to reduce mortality and morbidity in children from the Extended Program of Immunization (EPI) target diseases, while they are under the age of one [34]. Multiple antigens can be administered simultaneously. The standard immunization schedule shows that Bacille Calmette Guerin (BCG) and Polio0 would be given at birth, measles at 9 months and the remaining others within the range of 6 weeks and 10 weeks after birth.

v Breast feeding

Various studies have indicated that breast-fed infants experience a lower mortality risk than do artificially fed infants [29,35]. The advantages of breast-feeding comes from the fact that it is naturally ideal assuring that virtually all the infant's nutritional requirements will be met, and that, at least for the initial period of infancy, it provides some immunity against respiratory diseases [34]. Breast-feeding practice is nearly universally accepted in rural Ethiopia. It is part of the strong bonding mechanism that provides security to the mother and the infant. Most of the time estimates of the relationship between breast-feeding and infant's risk of mortality can be biased by methodological problems. Since infants' death curtails the duration of breast feeding, the average duration of breast feeding among those who survive beyond the age interval is more than among those who die before the end of the interval. The resulting truncation of breast-feeding artificially inflates the association between survival status and breast-feeding duration when in fact it is death that leads to shorter duration rather than the other way round. To avoid this bias it is suggested to consider only the duration to the beginning of the period at risk or alternatively, by disregarding the length of breast-feeding and consider only whether or not the infant was breast-fed at all up to the end of the period at risk.

b. Demographic and factors related to reproductive and maternal characteristics

These characteristics directly focus on the demographic characteristics of the mother and the infant and health-related experience of the mother rather than the surrounding environmental conditions. Factors included under this category are maternal age,

ethnicity, religion, sex of infant, birth weight, parity, survival of the preceding child, previous birth interval, birth order, previous abortions, usage of contraceptives etc ... [18, 19-20, 26, 36]. The major determinants are discussed in brief.

i Sex of an infant

Sex is almost universally recognized determinant of survival that, with equal care and feeding, females experience lower mortality at early ages and gradually get higher at childhood ages [25-26,32,36]. Specially, a study based on World Fertility Survey data of 39 countries found that the mortality of girls after first month of life seems to be somewhat higher than that of boys in Egypt, Pakistan and less clearly in Bangladesh and Yemen [36]. Using similar data from Korea another study [32] also found that mortality for males is higher around early infancy and higher for females at latter ages.

An interesting argument given for such pattern is, since biological strength of infants is the most important factor at early ages and environmental factors at later ages, and as males are weaker biologically they are exposed to risk at early ages where biological factors play a prominent role. If once they pass this stage, where the biological strength is required, due to male children preference by most societies they would be cared for better, that is, male child gets more attention and resources at stage where the environmental factors are more important [32].

ii. Maternal age

The age of the mother at the time of birth has a significant effect on the survival chance of an infant. Teenage pregnancy carries a number of risks to both mother and baby which include the risk of low birth weight and increased risk of mortality for the

infant. Also the maternal depletion syndrome and weakness of maternal reproductive tissue may contribute to the higher risk of birth outcomes from older mothers [28,37]. So one would expect that the risk of dying of infants born to older mothers as well as younger mothers would be greater than the middle-aged mothers. And most studies confirmed this U shape pattern of infant's risk with mother's age [20, 24, 35, 37]. But one study found unexpected pattern for Indonesia and Pakistan whereas the usual pattern was observed for Philippines [25].

CHAPTER 3

METHODOLOGY

3.1 BACKGROUND ABOUT THE SOURCE OF THE DATA

The survival data for this analysis were obtained from Jimma Infant Survival and Differentials Project (JISDP) which is centered at Jimma Institute of Health Sciences, my host institution. The project was designed to generate reliable data through a longitudinal follow-up study of live births for one year of life. This community-based follow-up study incorporates 46 urban and 65 rural Kebeles within an estimated population of 300,200 in the Zones of Jimma, Illubabour and Keffa-Sheka, South West Ethiopia in 1992-1994. Details of the study design have been given elsewhere [4,38]. The main aim of the project was to describe infant growth and investigate some of its determinants [4]. Seven experts from different professions, comprised of epidemiologist, public health experts, anthropologist and statisticians, were involved in the project for its co-ordination, management and interpretation. Traditional birth attendants (TBA) in the study areas collaborated in the study. TBAs are women residents of the Kebele they serve and by tradition they visit and assist women during pregnancy and delivery. The TBAs had easily access to women in the fertile age group and were able to assess their pregnancy status. TBAs went house to house regularly to locate pregnant women in their second trimester, and reported daily in person to the interviewer responsible for her Kebele. After the interviewer registered the address of the expectant mother both the TBA and the interviewer monitor the mother so as to reach her on time soon after delivery. Questionnaires were initially prepared in English, and then translated into Amharic and

back translated. Interviewers scheduled seven visits to the index infant's family soon after delivery, regularly at 2,4,6,8,10 and 12 months of age or up to his/her death. The interviewers handed in their work to their supervisors daily for early edition of the data. The data collection started on September 11, 1992 in Jimma town and extended its coverage to other urban and rural settings and ended in October 1994. A total more than 8000 live births were identified and followed until age one year or until death or drop out from the study. In this analysis only singleton births from all area are considered. Births from Jimma town were thoroughly studied for infant growth by previous two studies [4, 39] and a brief survivorship analysis focussing on multiple births was also considered by another previous paper [38]. In this study I focus on survivorship analysis in detail based on the methodologies given under Chapter 1.

3.2. VARIABLES CONSIDERED

Variables on survival time of infants and censoring indicator were considered. Based on the availability of data in the source project and literature the following variables were suspected to be related with infant survival and included in the study.

A) Demographic variables

- Maternal age at birth
- Mother's ethnic group
- Religion
- Family size.
- Deaths of previous children
- Number of previous abortions
- Number of still births

- Total number of live births
- Sex of an infant
- Birth weight
- Season of birth

B) Socioeconomic variables

- Education in grades completed
- Occupational status
- Place of residence
- Monthly income
- Marital status

C) Environmental health indicator variables

- Sources of water
- Availability of latrine
- Type of the roof of the house
- Type of the floor of the house

D) Health service usage indicator variables

- Frequency of visit to mothers clinic
- Place of delivery
- Attendant of birth
- Vaccination status of the infant at birth

E) Practices on infants

- Swallowing butter after birth
- Breast-feeding

3.3 STATISTICAL PROCEDURES EMPLOYED

After the data were coded using standard techniques and edited for each variable the following methods of analysis were applied.

A Uni-variate analysis based on frequency distribution were considered and a given variable is excluded from the analysis if almost all mothers have similar characteristic on that variable (homogeneous).

B Overall survival pattern (curve)

The survival function, survivorship or its graphical presentation, the survival curve obtained using the Product-Limit (PL) method of Kaplan–Meier discussed in Section 1.5.1.

C Identification of risk groups

To identify high and low risk groups for the risk variables:

i) Examination of the individual variables- Bivariate analysis

A preliminary examination of the data is done before a model is built. Examination of each variable's relationship to the length of survival was tested using the log-rank test, which discussed in Section 1.6.

ii) Simultaneous examination of the factors

The Cox's Regression model, which is a semi-parametric model appropriate for the analysis of survival data with censoring is used [1, 2]. Its methodological details are given in Section 1.7

D Diagnostic checking

Application of graphical methods and residuals discussed in Section 1.9 to see:

-The validity of the proportional hazard assumption for the Cox's model,

-The presence of outliers and influential points.

For all of the above mentioned statistical procedures the Statistical Package for Social Scientists (SPSS) was used.

CHAPTER 4

RESULTS AND DISCUSSION

Results of univariate analysis

A total of 8050 singleton live birth and a data on these were obtained from the source for all study areas of the project. Data on ten infants were found to have missing values on major variables and were excluded from the analysis. Five observations were found to be outliers and discarded from the analysis. Due to various reasons the data on 704 (8.8%) infants were withdrawn from the study before their first year of life. Hence 7331(91.2%) were successfully followed until first year of life or death before first year. The descriptive analysis of the data showed that neonatal mortality (death in the first 28 days) rate is found to be 26.6 per 1000 live births. Post natal mortality (death after day 28) rate is 73 .1 per 1000 live births. The overall Infant mortality rate is found to be 101.9 per 1000 live births.

Estimates of cumulative survival functions based on Kaplan-Meier method is 0.9855 (s.e.=0.0013), 0.9736 (s.e.=0.0018), 0.9039 (s.e.=0.0033) for days 7, 28 and 360, respectively. The corresponding graphical presentation of the survival function, the survival curve is given in Figure 1. Also the graphical presentation of the hazard function is given in Figure 2.

The distribution of the variables under study is given in Table 1. The third column shows the percentage of the event death before age 1 for the corresponding categories.

Fig.1 The overall Survival Function

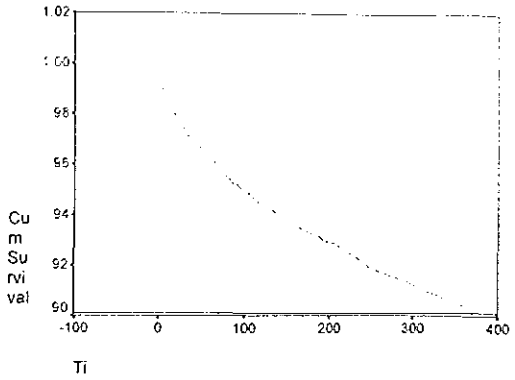


Fig2. The Hazard Function

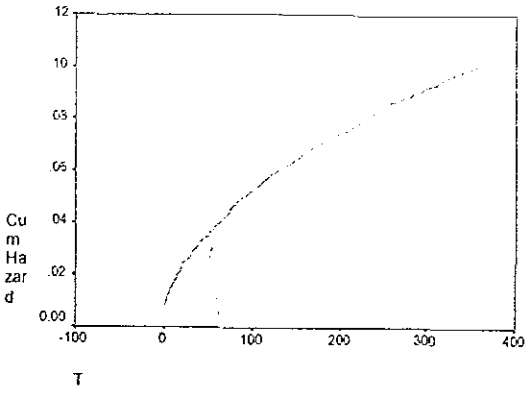


Table 1: Levels and frequency distribution of the factors under study.

Characteristics	Number	% died
Maternal age group		
<20	959	11.0
20-34	6536	9.1
>34	540	8.1
Ethnicity		
Oromo	5389	9.8
Amhara	651	7.2
Tigrie	110	6.4
Dawaro	494	10.9
Kefa	601	9.7
Gurage	412	5.8
Yeim	289	8.3
Others	89	3.4

Religion		
Christian	2479	8.2
Muslim	5556	9.8
Marital status		
Married	7468	8.9
Others	567	14.5
Grades completed		
>8	790	5.8
1-8	2382	8.1
Illiterate	4863	10.5
Occupation		
House wife	6524	9.3
Others	1511	9.1
Monthly household income		
< 150.00 birr	6051	9.5
≥150.00 birr	1984	8.5
Place of residence		
Urban	3804	8.1
Rural	4231	10.3
Family size		
1-3	1539	12.4
≥4	6496	8.5

Sex of infant		
Female	3937	8.5
Male	4098	10.0
Weight at birth		
≤2500 grams	1032	15.4
>2500 grams	7003	8.4
Months of birth		
October-march	4105	9.2
April- September	3930	9.4
Deaths from previous children		
None	5419	8.0
One	1536	11.6
Two or more	1080	12.5
Previous still births		
None	7867	9.3
Two or more	168	10.7
Previous abortions		
None	7423	9.3
One or more	612	9.2
Total live births		
1-3	4589	9.6
4-6	2372	8.8
≥ 7	1074	9.3

Sources of water		
Safe	4072	8.6
Unsafe	3963	10.0
Latrine availability		
Available	2941	7.2
Not available	5094	10.5
Type of roof		
Iron sheet	3683	8.3
Tatched	4352	10.1
Type of floor of the house		
Cement/Timber	812	6.5
Soil/Muddy	7223	9.6
Frequency of ANC visit		
None	3890	11.3
1-3	2621	8.2
≥ 4	1524	6.0
Place of delivery		
Health institution	1312	7.6
Home	6723	9.6
Birth attended by		
Health oriented person	1545	8.7
Non-health person	6490	9.4

Vaccinated at birth (BCG and Polio 0)		
Yes	1754	5.9
No	6281	10.2
Swallowing butter at birth		
No	2614	9.0
Yes	5421	9.4
Stopped breast-feeding before 4 months		
Yes	132	23.5
No	7081	3.6

Results from bivariate data analysis

For the suspected variables survival functions are computed for each category, and in Table 2 the cumulative survival functions at time equals 1 year is given. To see whether a given variable has an effect on survival a log-rank test described in Section 1.6 is done and the test statistic and significance level are given in columns 3 and 4 of Table 2, respectively. Among the demographic variables ethnicity which shows cultural difference, family size, mother's age, deaths of previous children, birth weight and sex of an infant are associated with survival at 5% level of significance. Religion, the month of birth, number of live births, experience of previous abortions and number of previous still births are not significantly related to the survival chance of an infant.

From the socioeconomic factors, education, residence and marital status show a significant association with survival at 5% level of significance. Occupation of the mother and household income do not show a significant association.

Among the environmental variables latrine status, floor and roof type of the house are significantly associated with survival, while the source of water is not.

From the health service usage variables, frequency of Antenatal Care (ANC) visit and the vaccination status (BCG and POLIO 0) of the infant at birth show a highly significant association. Place of delivery has a weak association, and type of birth attendant does not show an association with infants survival.

Swallowing butter after birth does not statistically show a decrease in infants' chance of survival.

Concerning the breast-feeding variable, it is considered only for those alive and under follow up at the age of 4 months. Its effect on survival from the age of 4 months up to 1 year of life was analyzed. It shows a high significance, and those who did not stop breast-feeding before 4 months had more chance of survival / to survive than those who stopped breast-feeding before the age of 4 months.

Table 2: Summary results of bivariate data analysis

Characteristics	Cumulative survival to 1 year(S.E.)	Log-rank statistic	Significance level	Relative Risk
Mothers age at birth		5.73	0.05	
(20-34)	0.9057(0.0037)			
35^{+}	0.9176(0.0119)		0.35	0.8650
≤ 19	0.8845(0.0106)		0.04	1.2468

Ethnicity (Oromo)	0.8993(0.0041)	15.72	0.02	
Amhara	0.9242(0.0106)		0.56	0.7476
Tigray	0.9337(0.0242)		0.27	0.6571
Dawaro	0.8864(0.0146)		0.44	1.1175
Kefa	0.8953(0.0131)		0.93	1.0116
Gurage	0.9380(0.0123)		0.01	0.5938
Yeim	0.9127(0.0171)		0.44	0.8512
Others	0.9632(0.0209)		0.07	0.3479
Religion(Christian)	0.9124(0.0059)	3.57	0.06	
Muslim	0.9004(0.0041)			1.1679
Family size (≥4)	0.9117(0.0036)	26.38	0.00	
1-3	0.8702(0.0088)		0.00	1.5331
Birth weight (>2500)	0.9130(0.0034)	56.58	0.00	
≤ 2500 gram	0.8417(0.0115)			1.9348
Sex (Female)	0.9120(0.0046)	5.83	0.02	
Male	0.8960(0.0049)			1.1939
Month of birth		0.006	0.93	
(October-march)	0.9050(0.0046)			
April-September	0.9027(0.0048)			1.0059
Deaths of children		31.99	0.00	
(none)	0.9164(0.0038)			
One	0.8818(0.0083)		0.00	1.4710
Two or more	0.8728(0.0102)		0.00	1.5699

Live births (4-6)	0.9102(0.0059)	2.02	0.36	
1-3	0.9000(0.0045)			1.124
7^{+}	0.9060(0.0089)			1.0517
Previous abortions		0.02	0.88	
(none)	0.9037(0.0035)			
One or more	0.9059(0.0120)			0.9785
Still births (none)	0.9041(0.0034)	0.46	0.49	
One or more	0.8904(0.0244)			1.1745
Marital status		26.00	0.00	
(married)	0.9083(0.0034)			
Single	0.8413(0.0160)			1.7956
Grades completed		20.34	0.00	
(9^{+})	0.9365(0.0084)			
1-8	0.9158(0.0058)		0.07	1.3538
Illiterate	0.8929(0.0045)		0.00	1.7579
Residence (Urban)	0.9136(0.0047)	8.22	0.00	
Rural	0.8952(0.0047)			1.2368
Monthly income		1.31	0.25	
(>150)	0.9096(0.0066)			
≤ 150 birr	0.9019(0.0039)			1.1053
Occupation		0.00	0.98	
(house wife)	0.9039(0.0037)			
Some time Out of house	0.9036(0.0078)			1.0018

Water source (safe)	0.9095(0.0046)	3.12	0.08	
Unsafe	0.8981(0.0049)			1.1381
Latrine available		20.32	0.00	
(yes)	0.9234(0.0051)			
No	0.8928(0.0044)			1.4386
Roof made of		6.30	0.01	
(iron sheet)	0.9123(0.0048)			
Thatched	0.8967(0.0047)			1.2048
Floor made of		6.78	0.01	
(cement/tumbler)	0.9300(0.0093)			
Soil	0.9010(0.0036)			1.4459
ANC visit (4^{+})	0.9362(0.0064)	39.92	0.00	
1-3	0.9149(0.0056)		0.02	1.3608
none	0.8839(0.0052)		0.00	1.900
Delivery place		3.95	0.05	
(Health institution)	0.9189(0.0078)			
Home	0.9010(0.0037)			1.2375
Birth attendant		0.44	0.51	
(health personnel)	0.9077(0.0076)			
Non-health				
personnel	0.9029(0.0037)			1.0651
Vaccination at birth		29.77	0.00	
(yes)	0.9370(0.0060)			
No	0.8946(0.0039)			1.7658

Swallowing butter at birth (no) Yes	0.9078(0.0057) 0.9020(0.0041)	0.92	0.34	1.0782
Breast feeding at 4 month (not stopped) Stopped	0.9626(0.0023) 0.7533(0.0386)	167.00	0.00	7.9493

RESULTS OF MULTIVARIATE ANALYSIS

All the suspected variables seen in bivariate analysis were included in the Cox's regression model to see the adjusted effect of the variables. The best possible combination of variables including the joint effect of terms that could describe the relationship between time to an event (death) and the variables has been obtained, and the results which are significantly related with infant survival are given in Table 3. Column 8 shows the hazard ratio or commonly known as the relative risk of infants within a given category of the variable with respect to those in the reference category of that variable, after adjusting for the effect of all other variables. In this final model almost all of the variables are demographic and health service usage variables.

From the table an infant from low family size (1-3) had 1.42 times higher risk of dying as compared to an infant family size greater than 3. Male infants had 1.52 times higher risk of dying than female infants at any time before they reach their first year. But male infants had more survival chance in rural areas and had 0.74 times lower risk of death as compared to other infants. Non-vaccinated and low birth weight infants had almost 2 times higher risk of dying as compared to other infants, adjusted for other variables. Previous death of children to a mother had a significant association with the chance of her infant's survival. If a mother had lost one previous child then her current infant had 1.32 times higher risk of dying relative to an infant whose mother had no experience of

child death. Similarly if a mother had experienced at least two child deaths, then her current infant had 1.57 times higher risk of dying as compared to an infant whose mother had no previous child death.

Infants from mothers who had never visited mother's clinic had 1.69 times higher risk of dying relative to those infants whose mother had visited the clinic at least 4 times. Similarly, mothers who had low visit (1-3 times) had 1.27 times higher risk of their infants' death as compared to mothers with highest frequency of ANC visit. Non-vaccinated infants at birth had 2 times higher risk of death as compared to those vaccinated at birth. Infants delivered at home had a risk of dying which is 0.60 times that of infants delivered at health institutions.

Infants who swallow butter had 1.29 times higher risk of death than those who don't. Unmarried women had 1.78 times higher chance of losing their infants as compared to others. Those house-holds with no latrine facility had 1.30 times higher risk of losing their infant. Teenage mothers with any previous experience of child death had 3.49 times higher risk of losing their infants as compared to other mothers. Families with any previous child death and with low family size (<4) currently had 1.43 times higher risk of infant death as compared to others.

Breast-feeding status was observed only for those who had survived to the age of four months, and the result showed that those infants who had reached the age of 4 months but stopped breast-feeding before the age of 4 months had 9.69 times higher risk of dying.

The overall survival curves for the first year of life for factors in the final model, adjusted for other factors are given in Figure 3.

Table 3: Summary results for the significant factors in the final Cox's regression model for the overall survival to first birth day

CHARACTERISTICS	COEFFICIENTS	S.E.	WALD STATISTICS	D.F	SIGNIFICANCE	R	RELATIVE RISK
Family size (>=4)							
1-3	0.3493	0.1059	10.8879	1	0.0010	0.0258	1.4181
Sex (Female)							
Male	0.4194	0.0993	17.8293	1	0.0000	0.0345	1.5211
Child deaths (None)			19.6358	2	0.0001	0.0343	
One	0.2739	0.1041	6.9241	1	0.0085	0.0192	1.3151
Two or more	0.4515	0.1056	18.2678	1	0.0000	0.0350	1.5707
Vaccinated at birth (Yes)							
No	0.6256	0.1432	19.0940	1	0.0000	0.0358	1.8694
Swallowing butter (No)							
Yes	0.2596	0.0826	9.8717	1	0.0017	0.0243	1.2965

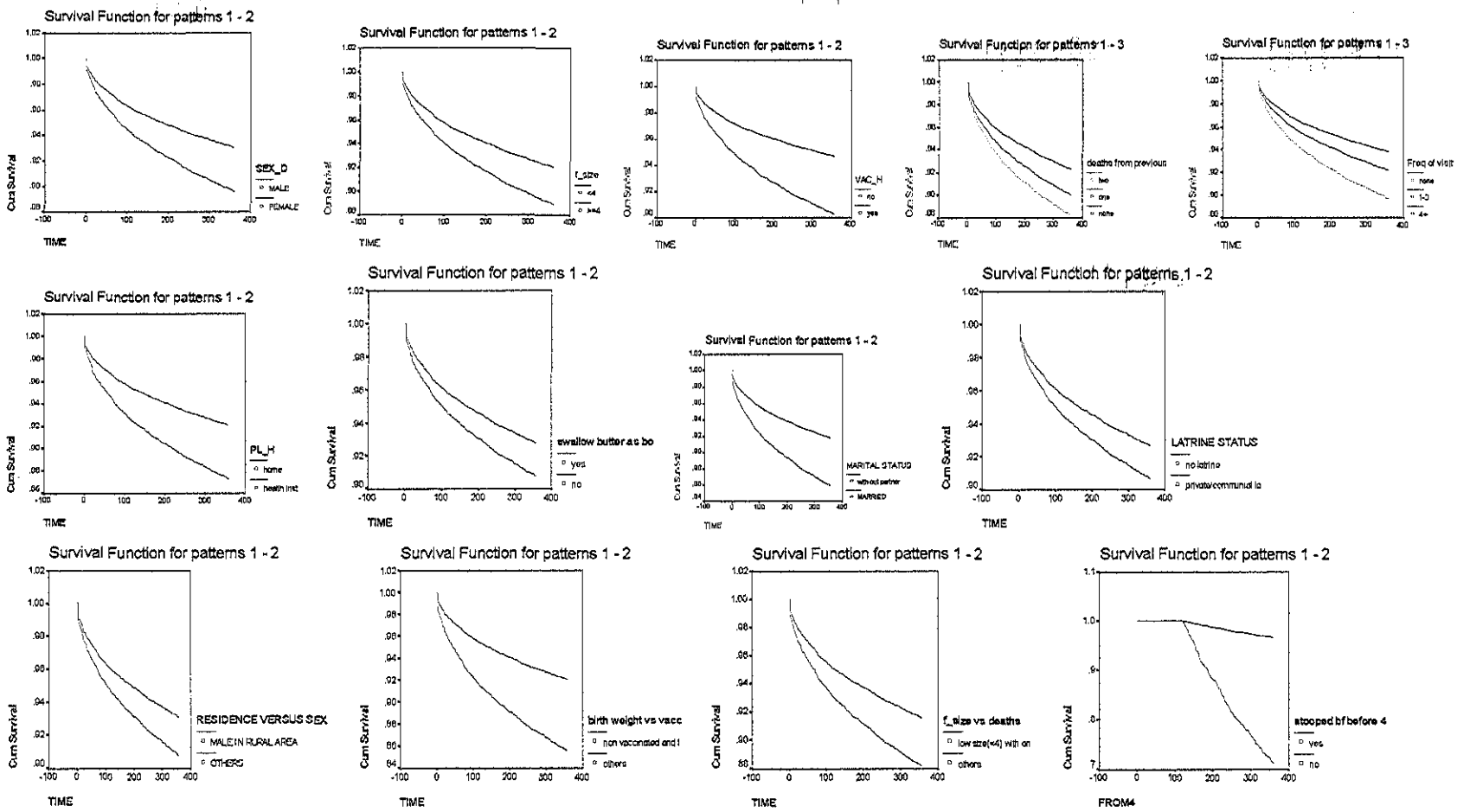
ANC Visit (at least 4)			21.6337	2	0.0000	0.0364	
1-3	0.2388	0.1278	3.4924	1	0.0617	0.0106	1.2698
None	0.5233	0.1256	17.3517	1	0.0000	0.0340	1.6876
Delivery place (Health In.)							
Home	-0.5019	0.1474	11.5973	1	0.0007	-0.0269	0.6054
Marital status (with partner)							
Single	0.5750	0.1194	23.1748	1	0.0000	0.0399	1.7771
Latrine availability (yes)							
No	0.2550	0.0961	7.0372	1	0.0080	0.0195	1.2905
Teenage with **child death**	1.2496	0.2674	21.8351	1	0.0000	0.0386	3.4890
Male infant in rural area	-0.3008	0.1143	6.9294	1	0.0085	-0.0192	0.7402
Under-Weight non vaccinated	0.6309	0.0962	42.0100	1	0.0000	0.0555	1.8793

Low family size with experience of child death	0.3584	0.1821	3.8739	1	0.0490	0.119	1.4311
Stopped breast-feeding before* 4 months (No) Yes	2.2710	0.1968	133.2278	1	0.0000	0.1609	9.6891
OVERALL CHI-SQUARE STATISTIC (with out the breast-feeding factor) = 274.105				15	0.0000		

Levels in parentheses are reference categories.

*The breast-feeding factor is considered only for those who survived and under follow-up to 4 months, and the model is fitted separately by considering this time constraint.

Fig 3. SURVIVAL CURVES FOR THE FACTORS IN THE FINAL MODEL



Results of model adequacy measures

Models with different combinations of covariates had been identified and the model with combination of variables discussed above (given in Table 3 for the overall survival to first year) has been selected based on Akaike's information criterion (AIC). The AIC criterion states that for different competing models AIC is calculated as AIC=-2log likelihood + α p, where α is any constant between 2 and 6, and p is the number of estimated parameters in the model. Then the one with minimum AIC would be selected as the best model (for further information see Reference 2).

Also based on the usual likelihood ratio test statistic, this model has a chi-squared value of 274.105, which is significant. Thus there is no statistical reason to conclude that this model does not show the best combination of the covariates. The same is true for other models given in Tables 4-6.

For checking the assumption of the proportionality in Cox's regression model the log minus log plot for each variable in the final model is shown in Figure 4. From the plots the lines for each case are almost parallel, thus the assumption is almost satisfied.

The plot of the martingale based residuals versus the linear predictor (X'Beta) is given in Figure 5. Due to space limitation the plots of Dfbetas versus case numbers are not shown.

Fig 4: The log-minus log plot for factors in the final model to check the assumption of the proportional hazard

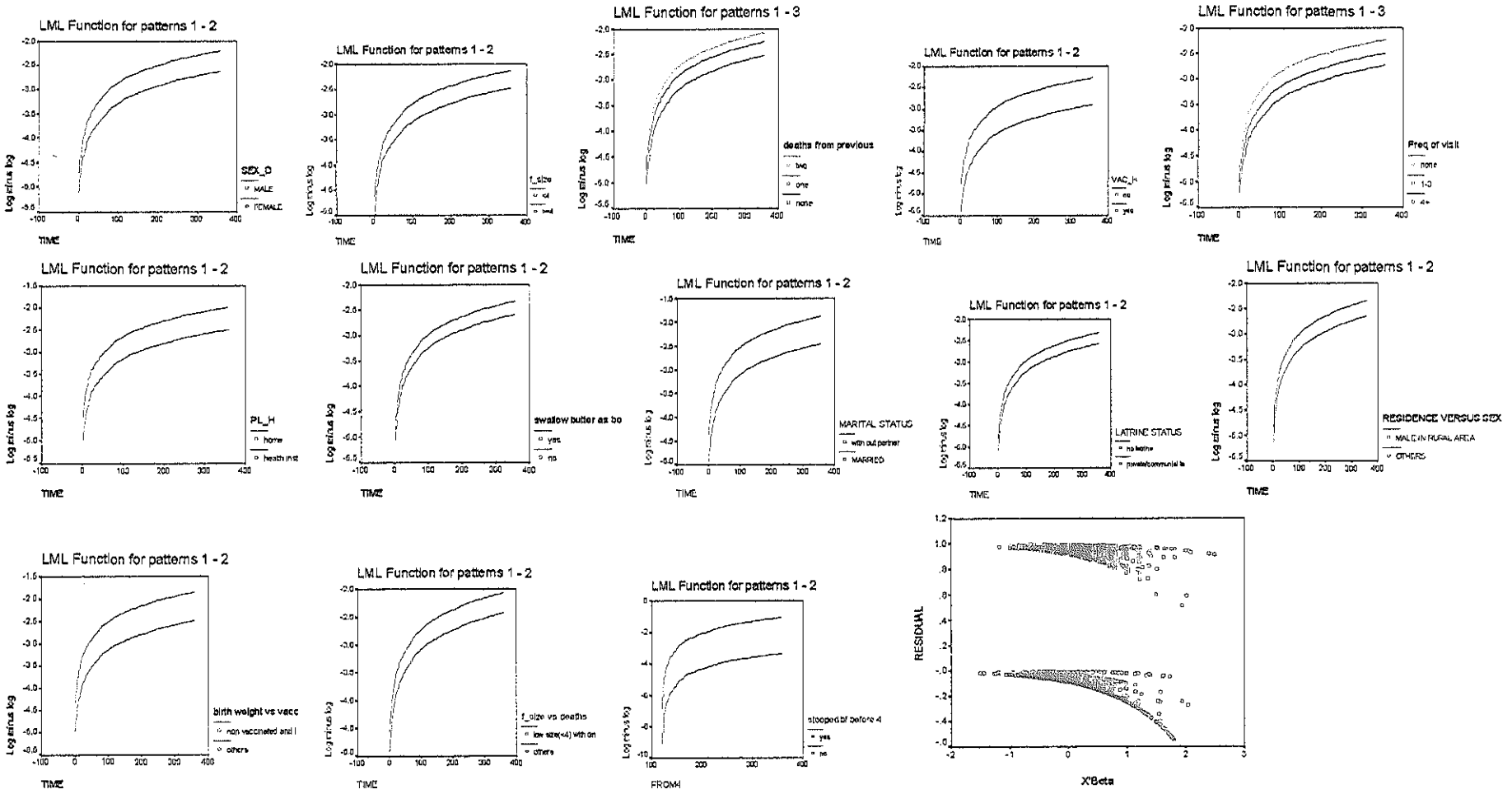


Fig. 5 The residuals versus linear predictor plot for checking outliers

DSCUSSION

The interpretation of the differences among factors in infant survival analysis is never easy. Probably the factors considered here are more or less remote on the causal chain leading to avoidable deaths in early life. Thus a detailed understanding of how the differences arise is beyond the objective and scope of this analysis. Yet the differences observed among the factors in the outcome of the first year of life are sufficient to suggest that the analysis has brought out important aspects of the problems associated with survival. Information on provision of health facilities or the availability of medical care services are not available. Also due to methodological complexity and lack of high capacity computer the information on traditional practices is not fully used. These are probably the most important factors that determine infant survival in our society. Without ignoring these limitations one can deduce the following.

Despite the finding that infant mortality in the area is not higher than in most regions of the country or the national value, mortality during the neonatal period is extremely high by virtually any standards.

Among the demographic variables ethnicity has a significant association with infant's survival in bivariate analysis, but its effect couldn't significantly show up when other variables are controlled. This shows that the difference observed at bivariate analysis is not due to ethnicity, but due to other confounding factors controlled at multivariate analysis. Some of these confounding factors may be traditional practices, place of residence and health service usage. The maternal age, jointly with number of previous child death, affects survival. Infants from teenage mothers with any previous experience of child death are more affected. This may be due to lack of experience in child care

among teenage mothers. The other variables, namely family size, death of previous child, sex of an infant remained consistent even after other factors are controlled. Birth weight affects survival jointly with the vaccination status of the infant. If the infant had low birth weight and not vaccinated then the risk of death is higher. It is obvious that under-weight infants are biologically weak and malnourished, and therefore, are vulnerable to any death causing hazard. In addition to this if they do not get vaccine, they may not to develop immunity and the risk will be magnified. As family size increases there is a tendency of sharing activities, so that the mother would give more time and attention for her infant care. Especially this factor is important in rural areas where women involvement in farming activity is high and it is less important in urban areas. That is why this factor is not captured in the final model for urban areas (Table 6). If a mother had experienced any previous child death, it shows that there is something unfavorable in that family for children. That could be either biological or socio demographic factor. This indicates a clustering of deaths within those households with previous child death. Therefore, a mother with previous child death is more likely to lose her current infant compared to mothers with no previous child death. If the household with a low family size has experienced any child death then it has a higher risk. Due to their biological weakness at birth male infants have a higher risk of death at early ages. This finding agrees with other studies. However, in this study, male infants from rural areas had lower risk of death. This may be due to male preference in a rural areas perhaps because a male would contribute to the agricultural activity later on, and might be given a better attention and care than females.

Next to the demographic factors, those factors related to health service, were found to be important in this study. Since mothers get health education on how to manage their

pregnancy and give care to the infant during visit to maternal clinic, an increased number of visit would result in having an appropriate infant care and end up with a better chance of infant survival. Vaccination would help to develop immunity to the infant and strengthen their biological make up to fight diseases. That is why infants vaccinated at birth had a better chance of survival. In the bivariate data analysis, infants born at health institutions had low risk of death than infants born at home. But when other factors are controlled the situation is reversed and the difference is significant. One of the possible reasons for this pattern might be due to the fact that mothers seek health institutions for delivery only after the condition gets complicated, which has a negative effect on future survival of the infant, and even on mothers.

Giving butter to an infant just after birth is the common practice in the South-West part of our country. According to this study it has a significant negative effect on infants' survival. It affects infants' health in introducing microorganisms due to the poor hygiene during swallowing and causing a problem in infants' intestine which is not ready for other foods at early life.

Breast-milk contains antibodies and other protective substances for growth and health of the infant. It is the most appropriate food for the baby's developing systems [29,34]. So the breast milk is the best natural food for babies and protects the baby from diseases. Experts highly recommend that a baby should be exclusively breast-fed during the first 4-6 months [29]. It is found out that infants who are not exclusively breast-fed up to at least 4 months had a significantly less chance of survival. The substitute foods, bottle milk and cows milk provide no protection against diseases like diarrhea, even cause it. Therefore, it is naturally the expected result in this study.

Among the socioeconomic variables marital status is independently important factor in the final model. This factor could be used as an indicator of economic and psychological factor. Those unmarried mothers, by and large, have low household income. If a women had a baby without a partner with her, there is a chance of being neglected or isolated from the society. This has a psychological influence on the mother so that she would not be encouraged to give good care to her baby. Educational status, which is important in many studies, is not independently important factor in this study for the overall survival in the first year. This shows that education could not make differences in this society with regard to infant's chance of survival. But it is to be noted that at post neonatal period education affects survival jointly with the status of maternal clinic visit. If a mother is illiterate and her frequency of maternal clinic follow-up is low (<4), the risk of infant loss at post neonatal period is higher as compared to other mothers (see Table 5). This risk is not shown at neonatal period and to the overall survival in the first year of life. This is because the neonatal period is completely dependent on infant's biological strength and demographic factors, not by socioeconomic factors. Also we found an interesting result from further analysis by age and residence. The analysis by age component shows that, at neonatal period, none of the socioeconomic and environmental factors are important (Table 4). The socioeconomic and environmental factors, marital status and latrine facility began to emerge as significant at the post neonatal period (Table 6). As rural and urban areas are different in development status a separate analysis shows that the socioeconomic and environmental factors, marital status and latrine facility, are not important factors in rural areas but they are important in urban areas.

Table 4: Summary result of multivariate analysis for Neonatal period

CHARACTERISTICS	COEFFICENTS	S.E	WALD STATISTICS	D.F	SIGNIFICANCE	R	RELATIVE RISK
Sex (Female)							
Male	0.2304	0.0741	9.6749	1	0.0019	0.0239	1.2591
Family size (>=4)							
1-3	0.3268	0.1057	9.5658	1	0.0020	0.0238	1.3865
Child death (None)			19.2149	2	0.0001	0.0337	
One	0.2848	0.1036	7.5536	1	0.0060	0.0204	1.3295
Two or more	0.4407	0.1035	17.4436	1	0.0000	0.0340	1.5538
Swallowing butter (No)							
Yes	0.2530	0.0816	9.6110	1	0.0019	0.0238	1.2879
Vaccinated at birth (Yes)							
No	0.6082	0.1417	18.4285	1	0.0000	0.0350	1.8371

ANC visit (at least 4)			22.8688	2	0.0000	0.0375	
1-3	0.2400	0.1270	3.5716	1	0.0588	0.0108	1.2713
None	0.5261	0.1231	18.2655	1	0.0000	0.0349	1.6923
Delivery place (Health institution) Home	-0.4682	0.1446	10.4882	1	0.0012	-0.0252	0.6262
Teenage with child death	1.2035	0.2673	20.2629	1	0.0000	0.0369	3.3316
Low weight with non vaccination	0.6135	0.0958	41.0321	1	0.0000	0.0540	1.8469
Unmarried with low income	0.6200	0.1322	21.9848	1	0.0000	0.0386	1.8589
Over all chi square statistic =249.603				12	0.0000		

Table 5. Summary result of multivariate analysis for Post Neonatal period

CHARACTERISTICS	COEFFICENTS	S.E	WALD STATISTICS	D.F	SIGNIFICANCE	R	RELATIVE RISK
Sex (Female)							
Male	0.4247	0.1156	13.4853	1	0.0002	0.0347	1.5291
Family size (>=4)							
1-3	0.2524	0.1045	5.8336	1	0.0157	0.0201	1.2871
ANC visit (at least 4)			6.1475	2	0.0447	0.0150	
1-3	0.2161	0.1592	1.8426	1	0.1746	0.0000	1.2412
None	0.3743	0.1612	5.3898	1	0.0203	0.0189	1.4540
Marital status (with Partner)							
Single	0.5651	0.1445	15.3012	1	0.0001	0.0374	1.7597

Availability of Latrine (Yes) No	0.2875	0.1115	6.6481	1	0.0099	0.0221	1.3331
Low weight with previous child death	0.6784	0.1650	16.9104	1	0.0000	0.0396	1.9706
Illiterate and low ANC visit	0.2249	0.1136	3.9180	1	0.0478	0.0142	1.2522
Male infant in rural area	-0.3851	0.1348	8.1605	1	0.0043	-0.0254	0.6804
Over all chi-square statistic = 80.407				9	0.0000		

Table 6. Summary result of multivariate analysis for Urban areas

CHARACTERISTICS	COEFFICENTS	S.E	WALD STATISTICS	D.F	SIGNIFICANCE	R	HAZARD RATIO
Sex (Female)							
Male	0.4489	0.1174	14.6101	1	0.0001	0.0500	1.5665
Birth weight (non-low)							
Low	0.5408	0.1719	9.9002	1	0.0017	0.0396	1.7175
Vaccinated (Yes)							
No	0.9498	0.1751	29.4266	1	0.0000	0.0737	2.5853
Delivery place (Health institution)							
Home	-0.5110	0.1601	10.1910	1	0.0014	-0.0403	0.5999

Maternal age (20-35)			11.6634	2	0.0013	0.0390	
<20	0.3706	0.1629	5.1734	1	0.0229	0.0251	1.4485
>35	-0.8353	0.3430	5.9303	1	0.0149	-0.0279	0.4337
Marital Status (with partner) Single	1.0371	0.2384	18.9213	1	0.0000	0.0579	2.8209
Availability of Latrine status (yes) No	0.4377	0.1178	13.8034	1	0.0002	0.0484	1.5492
Teenage with child death	0.7573	0.3847	3.8751	1	0.0490	0.0193	2.1826
Low family size with previous child death	0.8019	0.2265	12.5383	1	0.0004	0.0457	2.2298
Low ANC visit with previous child death	0.4756	0.1407	11.4160	1	0.0007	0.0432	1.6089

Stopped breast-feeding before 4 month (No) Yes	1.9560	0.2448	63.8617	1	0.0000	0.1609	7.0713
Over all chi square statistic with out B.F factor = 168.214				11	0.0000		

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

The findings of this study show that demographic, health service usage and traditional practice indicators are the most important factors. These factors are more or less related to the strength of the infant's biological make up. This result agrees with previous studies in other countries [19, 36]. The socioeconomic and environmental health indicators are not that much important in affecting infants' survival in the study area, and these findings agree with other studies. Thus the overall findings support the hypothesis that in economically poor society, demographic and biological factors affect infant survival than socioeconomic and environmental factors. But at the early stages of economic development the effects of the latter factors begin to emerge and completely dominate at advanced development stages [19]. Due to the reason that the population considered here is highly rural and poor in socioeconomic development, the socioeconomic and environmental factors are not very pronounced to bring about differences in infant survival. (e.g more than 60% of the population is illiterate, access to safe water is 27% at national level and health service coverage is 48.5%. in 1996) [7].

However, it can be hypothesized that socioeconomic and environmental factors are important determinants for infant survival in urban areas, which are believed to be in a better status. This hypothesis is supported by this study, that is, marital status and latrine facility are significant factors only in urban areas among the socioeconomic and environmental factors (Table 6). These factors do not show a significant difference in infant survival in rural areas.

The fact that demographic and biological factors play a major part in early ages of life while socioeconomic and environmental factors are important and influential at later ages of life has been confirmed in this study. The marital status and latrine facility factors become significant only at post neonatal period, but not at neonatal period (Tables 4 and 5). Though, this study could not substantiated, due to data limitation, most studies on child survival have shown that survival beyond infancy is more likely to be influenced by socioeconomic and environmental factors than demographic and biological factors. This orientation of effects of factors is true because, infants at early age do not share the resources available around and depend on his/her biological strength and other related factors like vaccination, breast-feeding and traditional factors. But at later ages of life children begin to share resources and the status of the household and surrounding environment would determine the quality and quantity of the share that they would receive, and influence their chance of survival.

On the basis of the findings the policy implications would be to facilitate socioeconomic developments in rural and urban areas both in quality and equitable distribution. Increasing the provision of maternal clinics and vaccination at birth is highly recommended. Though a further study on knowledge of mothers on certain issues is required, health education aiming at the following would be important.

- Emphasizing the importance of exclusive breast-feeding during the first 4-6 months.
- Avoiding harmful traditional practices.
- Use of maternal clinics so that teenage pregnancy and child deaths would be avoided.

To reduce infant and child deaths in those households where death is clustered, giving special attention in family planning and health services to the high risk households is

recommended.

There is a male infant preference in rural areas. Therefore, attention should be given in educating rural people to give similar considerations to female infants.

Finally, a further analysis of the data including the traditional practices with the aid of high capacity computers is highly recommended. Also a study incorporating children beyond infancy, up to age 5, would help to get a clear insight into how the factors influence child survival in the area.

BIBLIOGRAPHY

1. Marubini, E. and Valsecchi M. G.: Analysing survival data from clinical trials and observational studies, Chichester; John Wiley & Sons. (1995).

2. Collet, D.: Modelling Survival data in medical research, London; Chapman & Hall (1994).

3. Shamebo, D., SandstrÖm A. and Wall S.: The Butajira rural health project in Ethiopia: Epidemilolgical surveillance for research and intervention in primary health care, *Scand J prim Health care*, **10**: 198-205.

4. Asefa, M., Drewett R. and Hewison J.: An Ethiopian birth cohort study, *Paediatric and perinatal Epidemiology*, **10**(4), 443-462 (1996).

5. World Bank, World development report, Oxford university press, Inc. (1999). (http///www.worldbank.org/).

6. World population data sheet, Population Reference bureau, Inc. (1997). (http://www.prob.org/prob/.

7. Health information processing & documentation team, Planning & project department, Health and health related indicators; Ministry of health, Addis Ababa, Ethiopia.

8. Harris, E.K. and Albert A.: Survivorship analysis for clinical studies, New york, Marcel Dekker. Inc. (1991).

9. Lee, E.T.: Statistical methods for survival data analysis, Belmont; Lifetime learning publications (1980).

10. Cox, D.R.: Regression models and life-tables (with discussion), *J. Roy. Stat.soc., ser. B* **34** :187-220 (1972).

11. Cox, D. R. and Oakes, D.: Analysis of survival data, London; Chapman & Hall (1984).

12. SPSS, SPSS Inc., Chicago, USA (1991).

13. Breslow, N.: Covariance analysis of censored survival data, *Biometrics* **30**: 89-99 (1974).

14. Dale, L.P., David R.B. and Betty T.: '28 day survival rates of 6676 neonates with birth weight of 1250 grams or less, *peadiatr* **87**(1): 7-5 (1991).

15. Farah, A. and Prerston S.H.: Data and perspectives. Child mortality differentials in Sudan. *Population and development review*, **8**(2): 365-383 (1982).

16. CSA: The 1994 population and housing census of Ethiopia Results at country level, Volume I part I. (1998).

17. CSA: The 1994 population and housing census of Ethiopia, Results for Oromiya region, Volume I Part I (1996).

18. Yodit, S.: Neonatal survival in Addis Ababa, A thesis presented to Graduate School, Addis Ababa University (1995).

19. Gubhaju, B., Streatfield K. and Majumder A. K.: Socioeconomic, Demographic and Environmental determinants of infant mortality in Nepal, *J.biosoc. sci.* **23**:425-435 (1991).

20. Kost, K. and Amin S.: Reproductive and socioeconomic determinants of child survival: confounded, Interactive, and age-dependent effects, *Social Biology* **39**(1-2):139-150 (1992).

21. Frenzen, P.D. and Hogan D. P.: The impact of class, education and health care on infant mortality in a developing society: The case of rural Thailand, *Demography* **19**(3):391-408 (1982).

22. Meegama, S.A.: Socioeconomic determinants of infant and child mortality in Sri Lanka. An analysis of post war experience, *World Fertility Survey Scientific reports*, **8**:International Statistical Institute (1980).

23. Hobcraft, J.N. McDonald J.W. and Rustein S.O.: Socio-economic factors in infant and child mortality. A cross-national comparison, *Population studies* **38**:193-223 (1984).

24. Trussell, J. and Hammerslough C.: A hazards-model analysis of the covariates of infant and child mortality in Sri Lanka, *Demography* **20**(1): 1-26 (1983).

25. Martin, L.G. Trussell J. Salvail F.R. and Saha N.M.: Covariates of child mortality in the Philippines, Indonesia, and Pakistan: An analysis based on hazard models, *Population Studies* **37**:417-432 (1983).

26. Boulier, B.L. and Paques V.B.: On the theory and measurement of the determinants of mortality, *Demography* **25**(2): 249-263 (1988).

27. O'toole, J. and Wright R.E.: Parental education and child mortality in Burundi, *J. Biosoc. Sci.* **23**:255-262 (1991).

28. Ahmad, O.B., Eberstein I.W. and Sly D.F.: Proximate determinants of child mortality in Liberia, *J.Biosoc. Sci.* **23**:313-326 (1991).

29. Holland, B.: Breast-feeding, social variables, and infant mortality: A hazards model analysis of the case of Malaysia, *Social Biology* **34**(1-2): 78-93 (1987).

30. Adetunji, J.A.: Infant mortality and mother's education in Ondo state, Nigeria, *Soc.Sci. Med.* **40**(2):253-263 (1995).

31. Smucker, C.M., Simmons G.B. Bernstein S. and Misra B.D. Neo-natal mortality in South Asia: The special role of Tetanus, *Population Studies* **34**: 321-335 (1980).

32. Choe, M.K.: Sex differentials in infant and child mortality in Korea, *Social Biology* **39**(1-2): 139-150 (1992).

33. Davanzo, J., Butz W.P. and Hobicht J.P.: How biological and behavioral influences on mortality in Malaysia vary during the first year of life, *Population Studies* **37**: 381-402 (1983).

34. Manual on Maternal and Child health care, 3rd ed., Ministry of Health, Addis Ababa, Ethiopia (1995).

35. Eberstein, I.W., Nam C.B. and Hummer R.A.: Infant mortality by cause of death: Main and interaction effects, *Demography* **27**(3) :413-430 (1990).

36. Hobcraft, J.N., McDonald J.W. and Rustein S.O.: Demographic determinants of infant and early child mortality: A comparative analysis, *Population Studies* **30**: 363-385 (1985).

37. Pebley, A, R. and Stupp P.W.: Reproductive patterns and child mortality in Guatemala, *Demography* **24**(1):43-60 (1987).

38. Asefa, M. and Tessema F.: Infant survivorship and occurance of multiple births: A longitudinal community-based study, South west Ethiopia, *Ethiop. J. Health Dev.* **11**(3): 283-288.

39. Lesaffre, E., Asefa M. and Verbeke G.: Assessing the goodness of fit of the Laird and Ware model-An example: The Jimma infant survival differential longitudinal study, *Statist. Med*, **18**:835-854 (1999).