Jimma University

College of Natural Science

Department of Statistics

Modelling age-at-menarche: a case study in villages of Jimma zone

By:

Biniyam Yegoraw

A thesis submitted to the Department of Statistics, College of Natural Science,
Jimma University as partial fulfillment of the requirements for the
degree of Master of Science (MSc) in Biostatistics

January, 2014

Jimma, Ethiopia

Modelling age-at-menarche: a case study in villages of Jimma zone

By:

Biniyam Yegoraw

Advisor: Dr. Yehenew Getachew

Co-advisor: Mr. Belay Birlie (PhD Scholar)

January, 2014

Jimma, Ethiopia

STATEMENT OF AUTHOR

I declare that this thesis is a result of my genuine work and all sources of materials used, for writing it, have been duly acknowledged. I have submitted this thesis to Jimma University in the partial fulfillment for the Degree of Master of Science in Biostatistics. The thesis can be deposited in the university library to be made available to borrowers for reference. I solemnly declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate.

Biniyam Yegoraw

Date: __________________________

Signature: __________________________

Jimma University, Jimma, Ethiopia

ACKNOWLEDGMENT

First of all, I would like to acknowledge my advisor Dr. Yehenew Getachew for sharing his precious time to guide me while writing this thesis. He always gave me a clear direction to follow and also taught me with his generous patience. Next, I would like to extend my appreciation to my co-advisor Mr. Belay Birlie who gives me all the necessary advices regarding this thesis development.

I am also grateful to Prof. Tefera Belachew and Jimma Longitudinal Family Survey of Youth project for allowing me to use their dataset in this thesis.

Finally, my appreciation goes to my family and all my friends had ever given me a hand.

MODELLING AGE-AT-MENARCHE: A CASE STUDY IN VILLAGES OF JIMMA ZONE

ABSTARCT

Background: Menarche, the first occurrence of menstruation in girls, is an important milestone in the development of female adolescent. Time to menarche is the duration from the birth of an individual to the occurrence of the first menstruation cycle. Often, such time to event data are clustered (correlated) based on geographic locations. In the standard survival models the covariate effect and standard errors are estimated with the assumption that event times within the same cluster are independent of each other which leads to invalid results due to the ignored correlation and or heterogeneity in the data. Hence, in this thesis we applied various clustered or multivariate survival models in the analysis of age at menarche.

Methods: In this thesis, parametric frailty models, namely exponential, Weibull, lognormal, and loglogistic baseline hazards along with gamma, inverse Gaussian, lognormal and positive stable frailty distributions were used and the selected parametric frailty model was compared with the commonly used shared gamma frailty. AIC, model adequacy and standardized variability of coefficients were used in the comparison of various clustered survival models.

Results: The median age at menarche was about 14 years. The estimated heterogeneity parameter on menarcheal age across villages found to be significant except for exponential based frailty models. Comparison output shows that loglogistic-gamma frailty model has smallest AIC and has a better fit to the age at menarche data. Mother's education level, house hold income, BMI for age and height for age are important prognostic factors of age at menarche.

Conclusions: The log logistic-gamma frailty model found to be a good time to event model that fits the data better than other frailty models used in this thesis. The estimated heterogeneity parameter found to be significant indicating there is clustering/heterogeneity in timing of menarche across villages of Jimma zone. Hence, it is appropriate to employ a multivariate survival model that take into account the clustering or heterogeneity in the data.

ACRONYMS

AFT:	Accelerated Failure Time
AIC:	Akaike Information Criteria
BAZ:	Body Mass Index for age z-score
BMI:	Body Mass Index
HR:	Hazard Ratio
HAZ:	Height for Age z -score
LR:	Likelihood Ratio
MUAC:	Mid Upper Arm Circumference
PH:	Proportional Hazard
PVF:	Power Variance Function
SE:	Standard Error
USA:	United States of America
WHO:	World Health Organization

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1. INTRODUCTION

1.1. Background

Menarche which is the first occurrence of menstruation in girls, is an important milestone in the development of female adolescent. Unlike other pubertal changes that are gradual and continuous, menarche is an event with a sudden and dramatic onset. It is considered as a distinct benchmark for sexual maturation and also an indicator of quality of life of a population since a number of biological as well as socio-economic factors influence time to menarche (Prado et al, 1995). Variation in the timing of puberty are marked between well of and under privileged population with a marked delay in menarche reported in under privileged girls (Thomas et al, 2001). Association between nutritional status and onset of menarche has been studied (Belachew et al, 2011; Chowdhury, 2000; Osteria TS, 1983). In general adolescent who are taller and heavier with a greater body fat mass tend to reach menarche at younger age (Chowdhury, 2000; Osteria, 1983).

Time to event analysis (most commonly known as survival analysis) is a set of methods applied for analysis of time to event data, where the dependent variable is the time until the occurrence of an event of interest. The event could be the onset of menarche, death, occurrence of a disease, marriage, divorce, etc.

There are various survival analysis methods available to analyze the relationship of a set of predictor variables with the time to event outcome variable. These includes, parametric, non-parametric and semi parametric approaches. Among various approaches, Cox proportional hazard (PH) model is the most widely used models in survival analysis with the strong assumption of proportional hazards. Often, this PH assumption is reasonable for short follow-up studies (Perperoglou et al,2007).

When there is at least one unaccounted predictors(univariate frailty) in the model or when there is clustering in the data, random effects survival model, also known as frailty model is used. That is, a frailty term, which is a random component designed to account for variability due to unobserved factors is considered during such situations (Kleinbaum et al, 2005).

Regardless of the unobserved factors if widespread models such as Cox PH are used without taking into account the heterogeneity/clustering in the data, then the estimates are biased and the variances of the parameters are underestimated (Jonker et al, 2009). According to Keyfitz and Littman (1979), ignoring heterogeneity or clustering in the data overestimates life expectancy based on a study on estimating life expectancy in a heterogeneous population. Lancaster (1990) also showed that, when heterogeneity is ignored in the study of unemployment rates, it resulted in underestimation of covariate effects.

Therefore, the presence of clustering or dependence in the data can easily be handled using frailty models. In handling clustering in the data, the choice of frailty distribution is very important. Depending on the type of dependence in the data, various frailty distributions has been suggested in literature. Distributions with a large right tail such as positive stable distribution lead to strong early dependence, whereas distributions with a large left tail such as gamma distributions lead to strong late dependence (Hougaard, 1984). Mainly because of its easy mathematical properties, gamma distribution is the most widely used frailty distribution (Clayton, 1978). Hougaard (1986) suggested the gamma, the degenerate and the inverse Gaussian distributions on the positive stable family of distributions for the frailty model. Oakes D. (1989) suggested the inverse Gaussian and lognormal models for the distribution of the frailty.

In this thesis, multivariate time to event models were applied with expectation that, girls with in the same village share similar unobserved factors which affect their timing of puberty and it can be handled by using frailty model which consider place of residence(village) as clustering variable.

To examine the relationship between different covariates (parental education, house hold income, work load, BMI for age, height for age and MUAC) and age at menarche various multivariate time to event models were used. The multivariate survival models used in this thesis can be categorized as parametric frailty and semiparametric frailty models. The parametric frailty models includes, exponential, weibull , log-normal, and log-logistic baseline hazards along with gamma, inverse Gaussian, lognormal and positive stable frailty distribution. And, the semiparametric multivariate survival model used is the shared gamma frailty model. For comparison of different candidate models AIC, model adequacy and standardized variability of the estimates were used.

1.2. Statement of the problem

Though menarche is a vital event in continuation of human being, only few studies are conducted to investigate the associated factors related to late or early timing of menarche using in-depth statistical analysis, such as multivariate survival models. Early timing of menarche has biomedical, emotional, and socio-cultural consequences, including predisposition to cancer and heart disease and early participation in risky behaviors, such as cigarette smoking, alcohol abuse, and sexual activity (Chodick et al., 2005). To model the age at onset of menarche, various statistical approaches including ordinary linear regression models and the classical mixed models can be applied. However, these methods are less efficient compared to time to event models, mainly because of the presence of incomplete information, i.e., censoring and non-normal nature of event times in the data.

The standard or common application of survival methods implicitly assumes a homogenous population to be studied. Consequently, the classical Cox PH (without a random effect term), which is known as the most popular model in survival analysis is applied. These models assumes the baseline hazard to be common to all the individuals in the study population and the covariates act multiplicatively on the baseline hazard, which adds additional risks based on each individual's prognostic information. However, when subgroups of population share a common trait that cannot be observed, the covariates cannot always fully account for the true differences in risk and this may be due to the heterogeneity or clustering of event times, which is not accounted in the model. Hence, not taking into account the heterogeneity/ clustering in the data may lead to poor standard error estimates and a biased estimate of covariate effect.

Therefore, adjusting for the existed correlation or clustering in the data may allow, correctly measure the standard errors and also avoid underestimation or overestimation of the parameters of interest. In general, the motivation behind this study is to address the following research questions:

- What are the factors that are associated with time to menarche?
- Which combination of baseline hazard and frailty distribution describe well the time to menarche data?

1.3. Objectives

1.3.1. General Objective

The overall objective of this study is to model age at menarche by applying appropriate multivariate time to event (frailty) model using the data obtained from Jimma longitudinal and family survey of youth Jimma zone, southwest Ethiopia.

1.3.2. Specific Objectives

The specific objectives of this study are:

1. to compare the relative performance of different parametric and semi parametric baseline hazards in modeling age at menarche;
2. to test the clustering or heterogeneity in timing of menarche across villages of Jimma zone;
3. to identify important prognostic factors in modelling age at menarche using appropriate multivariate survival model, and
4. to compare the timing of menarche for different groups of girls.

1.4. Significance of the study

The findings of this study are useful for policy makers and other organizations in creating effective policy mainly regarding reproductive health of females by identifying group of girls with lower and higher age at menarche using appropriate multivariate survival model. The outcomes of this thesis also contribute to the existing literature on the use of appropriate statistical modelling approaches in the analysis of time to event data mainly in the presence of unobserved heterogeneity or clustering in the data.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Age at menarche, anthropometric measures and physical activity

Generally, body size parameters, such as weight, body mass index (BMI) and height are strongly correlated with the age at menarche. Higher fat levels and BMI at pre-pubertal ages are associated with increased likelihood of early, less than 11 years, menarche (Freedman et al, 2002). Age at menarche is negatively related to hip and thigh circumference and positively related to waist circumference, status and biiliac breadth (Lassek et al, 2007). Garn et al (1986) found that girls who experienced menarche before the age of 11 years were 2 to 3 kg/m2 heavier than those who experienced menarche aged 14 years. Similarly, Ayatollahi et al (2002) also indicated that BMI was significantly correlated with age at menarche. They further demonstrated that, menarche age was delayed for underweight individuals. Low body weight delayed menarche by approximately 15 weeks, while high body weight and obesity induce it by 13 and 19 weeks, respectively compared to girls with normal body weight. According to the study conducted in Bangladesh, age at menarche is associated with anthropometric indices of both childhood and adolescent such as height, mid-arm circumference (MUAC), BMI, height for age (HAZ) which in turn influenced by nutritional status (Bosh et al, 2008).

Though the difference between the exercising and the non-exercising group were statistically non-significant(Bagga and Kulkarni, 2000), girls who had to do more physical work, or had a long, tiresome way to school and spent greater expenditure of calories delay the process of puberty (Serap et al, 2009). A cross sectional study performed in a group of Colombian University, women demonstrated that age at menarche was positively associated with the practice of at least two hours daily of physical activity (Chavarro, 2004). Menarche, on average, occurs later in athletes, including ballet dancers, than in the general population, with the exception of swimmers, suggesting that intense exercise delays puberty (Malina, 1982).

2.2. Age at menarche and socio economic factors

Socio-economic factors are among the determinants of age at menarche. A study conducted in India showed that in general daughters of hamals housemaids and day laborers experienced menarche later than the girls of the middle and higher economic groups. The difference was about 12 months (Bagga and Kulkarni, 2000). Contrary to the study on British teenagers which the study found out no difference on median age at menarche by social class or ethnic group, the study on Iranian School girls showed that nearly one-fourth of girls who were from poor families had higher age at menarche than girls of the middle and higher classes (Ayatollahi *et al.*,1999; Whincup, 2001). Similar study in Poland also shows that socioeconomic factors such as family income, level of parental education, also influence pubertal development. Girls from families with a high socioeconomic status experience menarche at an earlier age than girls from families with lower socioeconomic status (Wronka et al, 2005). Similar study on third World girls adopted in Western European countries revealed that the pattern of early menarche, which indicates the role of transition from an underprivileged to a privileged environment as of the determinant factor of menarche (Proos et al, 1991).

Wronka et al (2005) showed higher parental education has been associated with earlier timing of puberty. However, in a similar study at a Bangladeshi University, the mother's educational level and occupation(but not fathers) was found to have a significant influence on their daughter's age at menarche (Hossain et al., 2010). On other hand, study conducted on female University students in Portugal (Padez, 2003) found no association between student's age at menarche and their parents' educational levels and occupations which is similar to the result obtained at USA, parents' education was not a predictor of early age at menarche amongst the American population (Braithwaite et al., 2009).

2.3. Literature on survival models

2.3.1. Common survival methods

The beginning of survival analysis goes back to the time when mortality tables were introduced. Life tables are one of the oldest statistical techniques and are extensively used by medical statisticians and by actuaries. Kaplan and Meier (1958) gave a comprehensive review of earlier work and many new results. Cox (1972) extended the results of Kaplan and Meier to the comparison of life tables and more generally to the incorporation of regression-like arguments into life table analysis.

As the application of survival models became popular, parametric models gave way to non-parametric and semi-parametric approaches for their demand in dealing with the growing field of clinical trials in medical study. Survival models have the capability of handling incomplete (censored) information in the data. Kalbflesch and Prentice (1963), Cox (1972), Cox and Oakes (1984), Miller (1981) used survival analysis in modeling human lifetimes. Allison (1984), and Bloessfeld *et al* (1989), Tuma and Hannan (1984) shows the application of survival analysis in social science. Fergusson et al (1984) used hazard functions to study the time to marital breakdown after the birth of chil. Hazard functions had been also applied in studies of time to shift in attentions in classroom (Felmlee et al., 1985) time to change decision in the face of irrelevant information from a low-status partner (Hembroff and Myers, 1984), and in the study of relapse of mental illness (Lavori et al., 1984).

PH modeling is the most commonly used type of the survival models in many research areas. It has been applied in diversified fields of specializations, such as, to topics like smoking relapse Stevens and Hollis (1989), affective disorders Shapiro *et al* (1989), childhood family breakdown Fergusson *et al* (1985), interruptions in conversation Dress (1986), and in medical areas for identification of important covariates that have as significant impact on the response of the interested variables.

A distinguishing feature of survival and event history models is that they take censoring into account. A simple definition of censoring is that we have information about an individual's survival time, but do not know the exact survival time (Kleinbaum and Klein, 2005). Various types of censoring can occur, with the most common type being right-censoring, which will also be the primary focus in many studies. In most cases, truncation refers to the complete lack of information about the occurrence of the event. There is often some confusion as to whether observations are censored or truncated. Strictly speaking, truncation refers to the cases where subjects do not appear in the data because they are not observed. Censoring refers to cases when subjects are known to fail within a particular episode, but the exact failure time is unknown (Allison, 1984; Tuma and Hannan, 1984; Yamaguchi ,1992).

2.3.2. Frailty models

Frailty is an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals. To address the problem of unobserved heterogeneity in event times resulting from, first Beard (1959) and later Vaupel et al. (1979) and Lancaster (1979) independently suggested a random effects model for durations. Then after, investigators have recognized that ignoring individual heterogeneity or clustering in the data may lead to inaccurate conclusions. Models for heterogeneity have been proposed by Vaupel et al. (1979), who introduced frailty as an unobserved quantity in population mortality. Oakes (1989) proposed frailty models for bivariate survival times and introduced several possible frailty models. Flinn and Heckman (1982) also introduced heterogeneity into their model for analyzing individual event histories. They showed that improper modeling of heterogeneity will result in biased estimates, since the covariates in the model fail to explain the true effect of the covariates on a response variable. Keyfitz and Littman (1979) showed that ignoring heterogeneity will lead to an incorrect calculation of the life expectancy from known death rates. A similar conclusion was reached by Vaupel et al. (1979) using a continuous mixture model in which an unobserved non-negative random frailty represents all individual differences in endowment for longevity.

2.3.2.1. Frailty Distributions

One of the practical problems in the area of frailty modeling is the choice of the frailty distribution. The frailty distributions most commonly used in practice are the gamma distribution (Clayton, 1978; Vaupel *et al.*, 1979), the positive stable distribution (Hougaard 1986b), a three-parameter distribution (PVF) (Hougaard 1986a), the compound Poisson distribution (Aalen 1988, 1992) and the log-normal distribution (McGilchrist and Aisbett, 1991). The frailty distributions that have been studied most belong to the power variance function family, a particular family of distributions introduced first by Tweedy (1984) and later independently studied by Hougaard (1986b). For reasons of convenience, analysts often choose parametric representations of frailty models that are mathematically tractable. Hougaard (1986a, 1986b) used several distributions for frailty including gamma, inverse Gaussian, positive stable distributions and claimed that gamma and inverse Gaussian distributions are relevant and mathematically tractable as a frailty distribution for heterogeneous populations. Flinn and Heckman (1982) used a lognormal distribution for frailty, whereas Vaupel et al. (1979) assumed that frailty is distributed across individuals as a gamma distribution.

CHAPTER THREE

3. DATA AND METHODOLOGY

3.1. The Dataset

The data for this thesis was taken from a longitudinal study of adolescents in the Jimma zone, southwest Ethiopia. The data was collected for the purpose of assessing the life events of adolescents as they transit to adulthood in Jimma zone, southwestern part Ethiopia. The first three rounds of the longitudinal data has been taken from 18 villages (also locally known as kebeles) selected from *Jimma city* and three rural districts, namely *Kersa*, *Dedo* and *Manna*.

To select the target sample of adolescents a two-stage sampling plan was used. At the first stage, households were randomly sampled with the sample size in each “kebele” determined by the relative proportion of the study population in the “kebele” and the overall target sample size. In the second stage, one adolescent (a boy or a girl) was randomly selected from each household. Using this sampling strategy a total of 1059 boys and 1025 girls were interviewed in the first round. But, this study is based on 924 female adolescents, since 101 girls were not interviewed for the response variable at any of consequent rounds. The questionnaires were interviewer-administered and the girls’ timing of menarche was obtained by female interviewers asking girls whether they had experienced their menarche and the timing at which the event had happened. The studied girls (924 girls) were followed for three rounds that are one year apart spanning starting at 2005.

3.2. Variable Description

3.2.1. Dependent variable

The dataset for this thesis is right censored survival data; accordingly, the response variable has two components. The first component is the observed time and the second component is a status indicator which indicates whether the observed time is event time or censoring time. In modelling age at menarche survival time is defined from birth to the age at onset of menarche where the length of time is measured in years. Here, the status variable shows whether study subjects (girls) experience menarche or not at provided age.

3.2.2. Independent Variables

To model age at menarche, two continuous and five categorical candidate covariates of interest were considered. Throughout this thesis village is considered as a clustering variable. These covariates are described together with their values or codes in Table 1.

Table 1: Description of covariates used in modeling age at menarche

Characteristics	Categories/description
Father's education level	(0) Secondary and higher (1) Primary (2) No education
Mother's education level	(0) Secondary and higher (1) Primary (2) No education
House hold income category	(0) Low (1) Middle (2) High
BMI for age (Z-score)	(0) Normal (1) Under weight (2) Over weight
Height for age (Z-score)	(0) Not-stunted (1) Stunted
Workload	Continuous (index)
MUAC	Continuous (measured in centimeters)
Place of residence	A total of 18 villages (clustering variable)

As described in the above table, baseline MUAC and workload index are continuous variables. Whereas, for the rest of baseline covariates, a value code has been assigned so that the covariates can enter into the statistical analysis and interpreted in relative to a single reference category. Parental education level, the first and second categorical predictors were classified as secondary and above, Primary and no-education. Household income is classified as low, middle and high based on computed tertile values. BMI for age and height for age were categorized based on WHO standards (using WHO Anthro-Plus software v3.2.2). BMI for age has three levels: *underweight* (< -2SD), *normal* (between -2SD and +1SD), and *over weight* (>+1SD). Similarly height for age is classified as *stunted* (< -2SD) and *non-stunted* (≥ -2SD).

3.3. The Survival Methods

In this thesis, we have used non parametric tests, such as log-rank and peto-peto tests. Among frailty models, shared gamma frailty (with unspecified baseline hazard) and frailty models (with parametric baseline hazard, such as exponential, Weibull, lognormal, and log logistic) with gamma, inverse Gaussian, lognormal and positive stable frailty distribution have been applied. R (*surval* and *parfm* packages) and STATA version 11 statistical packages were used in the analysis.

3.3.1. Basic Survival functions

Let T be a non-negative random variable representing the time until the occurrence of an event, such as menarche. For simplicity we will adopt the terminology of survival analysis, referring to the event of interest as ‘death’ and to the waiting time as ‘survival’ time, but the techniques to be studied have much wider applicability. They can be used, for example, to study age at menarche, age at marriage, intervals between successive births to women, the duration of stay in a city (or in a job) and the length of life.

3.3.1.1. The survival functions

Suppose T is a continuous random variable with probability density function f(t) and cumulative distribution function (c.d.f).

$F(t) = \Pr(T \leq t)$ is the probability that the event has occurred by duration t.

Often, it is convenient to work with complement of the c.d.f, the survival function

$$S(t) = \Pr(T > t) = \int_t^\infty f(t)\, d(x) \qquad (3.1)$$

Which gives the probability of being alive(not experience the event) at duration t. More generally, the probability that the event of interest has not occurred by duration t.

3.3.1.2 The Hazard Function

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of event, defined as

$$\lambda(\mathrm{t}) = \lim_{dt\to 0} \frac{\Pr\,(t<T\leq t+dt|T>t)}{dt} \qquad (3.2)$$

The numerator of this expression is the conditional probability that the event will occur in the interval $(t, t + dt)$ given that it has not occurred before, and the denominator is the width of the interval. Hence, dividing one by the other will result in a *rate* of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, one can obtain an instantaneous rate of occurrence. The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $(t, t + td)$ and T > t (which is, the same as the probability that t is in the interval), to the probability of the condition T > t. The former can be written as $f(t)\mathrm{dt}$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result

$$\lambda(\mathrm{t}) = \frac{f(t)}{S(t)} \qquad (3.3)$$

which some authors give as a definition of the hazard function. In words, the rate of occurrence of the event at duration *t* equals the density of events at *t*, divided by the probability of surviving to the duration without experiencing the event.

From equation (3.1) that $-f(t)$ is the derivative of $S(t)$. This suggests rewriting (3.3) as

$$\lambda(\mathrm{t}) = -\frac{d}{dt} logs(t)$$

If we now integrated from 0 to t and introduce the boundary condition S(0) = 1

The above expression to obtain a formula for the probability of surviving to duration t as a function of the hazard at all duration up to t:

$$S(t) = \exp\{-\int_0^t \lambda(t)dx\} \qquad (3.4)$$

The integral in curly brackets in this equation is called Cumulative hazard (or cumulative risk) and is obtained from the hazard function,

$$\Lambda(\mathrm{t}) = \int_0^t \lambda(x)dx \qquad (3.5)$$

Λ(t) can be taken as the sum of the risks one face going from duration 0 to t.

3.3.2. Comparison of Survivorship Functions

In survival analysis, it is always a good practice to present numerical summaries of the survival times for the individuals using non parametric methods. The methods are said to be non-parametric methods, since they don't require any assumption about the distribution of survival time.

$$Q = \frac{\left[\sum_{i=1}^{m} wi(d_{1i}-\hat{e}_{1i})\right]^2}{\sum_{i=1}^{m} w_i^2 \hat{v}_{1i}} \qquad (3.6)$$

Where, $\hat{e}_{1i}= \frac{n_{1i}d_i}{n_i}$ and $\hat{v}_{1i}=\frac{n_{1i}n_{0i}d_i(n_{1i}-d_i)}{n_i^2(n_i-1)}$

n_{0i} is the number at risk at observed survival time $t_{(i)}$ in group 0;

n_{1i} is the number at risk at observed survival time $t_{(i)}$ in group 1;

d_{1i} is the number of observed event in group 1;

n_i is the total number of individuals or risk before time $t_{(i);}$ and

d_i is the total number of event at $t_{(i)}$.

The contribution to the test statistic depends on which of the various tests is used, but each may be expressed in the form of a ratio of weighted sums over the observed survival time points. Under the null hypothesis, the two survivorship functions are the same, and assuming the censoring experience is independent of group, and that the total number of observed events and the sum of the expected number of events is large, test statistic (Q) follows a chi-square distribution with one degree of freedom. We can also use the above test to compare k-groups. In this study we used log-rank and generalized Wilcoxon tests to compare survival functions of different groups.

3.3.2.1. The Cochran-Mantel-Haenszel log-rank test

The log rank test, sometimes called the Cox-Mantel test, is the most well-known and widely used test in various areas of applied statistics. This test is based on weights equal to one (i.e. w_i=1). The log rank test statistic is given as

$$Q_{LR} = \frac{\left[\sum_{i=1}^{m}(d_{1i}-\hat{e}_{1i})\right]^2}{\sum_{i=1}^{m}\hat{v}_{1i}} \qquad (3.7)$$

3.3.2.2. The Generalized Wilcoxon test

Gehan (1965) and Breslow (1970) generalized the Wilcoxon rank sum test to allow for censored data. This test uses weights equal to the number of subjects at risk at each survival time i.e. w_i=n_i and is called Wilcoxon or generalized Wilcoxon. The generalized wilcoxon is given as

$$Q_{GWt} = \frac{[\sum_{i=1}^{m} ni(d_{1i} - \hat{e}_{1i})]^2}{\sum_{i=1}^{m} n_i^2 \hat{v}_{1i}} \qquad (3.8)$$

3.3.3. Basic Survival models

3.3.3.1. Cox PH models

In survival analysis, a popular approach is to model the hazard function rather than the mean of the survival times as in classical regression models. Since a hazard function may be complicated, we can avoid a parametric assumption and allow the hazard function to be nonparametric. Then, one may link the hazard function to covariates xi through the usual (parametric) linear predictor $\mathbf{X}_i'$, which is sometimes called a risk score or prognostic index in survival analysis. This leads to a semiparametric regression model.

A widely used semiparametric survival regression model is the following Cox proportional hazards model (Cox 1972)

$h_i(t)$= $h_0(t)$exp($\mathbf{X}_i'\boldsymbol{\beta}$), or $S_i(t)$=$S_0(t)^{\exp(\mathbf{X}'_i\boldsymbol{\beta})}$, (3.9)

Where $h_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{\beta}$ is a vector of unknown regression parameters, and $\boldsymbol{X}_i$ = $(x_{i1}, \dots, x_{ip})'$ is a vector of covariates, and $i = 1,2, \dots .n$ .In the Cox proportional hazard model (3.9) , no distributional assumption is made for for the survival data, so it is very flexible. The assumption in the model is that the hazards ratio $\frac{h_i(t)}{h_0(t)}$ does not change over time (i.e., proportional hazards), which should be checked in a particular application.

3.3.3.2. Accelerated failure time (AFT) models

Although the semiparametric Cox proportional hazards model (3.9) is widely used in the analysis of survival data, parametric regression models for survival data have also been developed. These parametric models assume that the survival data follow some parametric distributions, and they may be preferred if the distributional assumptions hold. A major advantage of parametric regression models is that, if the parametric distributional assumption holds for the survival data, statistical inference based on the parametric model will be more efficient than a semiparametric model which makes no distributional assumptions. On the other hand, a major advantage of semiparametric models such as the Cox proportional hazards models is that they are robust against distributional assumptions.

For the proportional hazard models, we assume that the hazards ratio $\frac{h_i(t)}{h_0(t)}$ is constant over time. In practice, however, the proportional hazards assumption may not be valid, so the PH models may not be appropriate in these situations. An alternative and popular survival regression model is the so-called accelerated failure time (AFT) model, which does not require the proportional hazards assumption.

For AFT models, the following log-linear representation is widely used

$$\log(T_i) = \mathbf{X}_i'\boldsymbol{\beta} + \omega\varepsilon_i, \quad i=1,2, \ldots, n, \tag{3.10}$$

where ω is scale parameter and ε_i's are random errors. If we assume that ε_i follows a parametric distribution, we have a parametric AFT model. Different choices of the distributions for ε_i lead to different AFT models. The following are three commonly used parametric AFT models.

If ε_i follows the Gumbel distribution, the survival time Ti follows a Weibull distribution. Thus, the Weibull distribution has both the proportional hazards property and the accelerated failure time property, which is very appealing and makes the Weibull survival model very attractive. Additionally if $\omega = 1$, then the weibull model will be reduced to an Exponential model. Another common choice for ε_i is the distribution is the standard normal distribution N (0,1). If ε_i follows N (0,1), the survival time T_i follows a log-normal distribution. And, the third common choice for the distribution of ε_i is the logistic distribution, if ε_i follows a logistic distribution, which leads to loglogistically distributed the survival time T_i follows a log-logistic distribution. Inference for parametric AFT model (3.10) can be based on the likelihood method.

3.3.4. Frailty models

3.3.4.1. Univariate frailty models

The standard situation of the application of survival methods in scientific research projects assumes that the population being investigated is homogenous. Consequently, appropriate survival model assumes that the survival data of the different subjects are independent form each other and that each subject's individual survival time distribution is the same (independent and identically distributed failure times). However, the effect of treatment or the influence of various explanatory variables may differ greatly between subgroups of subjects. To account for such unobserved heterogeneity in the study population Vaupel et al. (1979) introduced univariate frailty models into survival analysis.

In frailty models, the variability of survival times can be divided into two parts. One part is observed risk factors, known as covariates, and the other part is unobserved risk factors, known as *frailty*. In the univariate frailty model, the population under study is considered as a mixture in which baseline hazard is common to all individuals but each individual has its own frailty term.

Suppose that there are samples of *n* observations in a study; with $\mathbf{x}_i$ the observable covariate vector for the i^{th} individual. The hazard function of the i^{th} individual at time t is

$$h_i(t \,/\, x_i, u_i) = u_i h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x_i}) \qquad , \ i=1,\ldots,n \qquad (3.11)$$

This shows that the hazard of an individual also depends on an unobservable random variable, u_i which acts multiplicatively on the hazard rate. In the univariate case, frailty models are used to make adjustments for over dispersion. When unobserved or unmeasured effects are ignored, the estimates of survival may be misleading. Therefore, a correction for this over dispersion is needed in order to allow for adjustments for those important frailties.

3.3.4.2. Shared (Multivariate) Frailty Models

Multivariate frailty model is a conditional independence model in which frailty is common to all subjects in a cluster. It is also known as a mixture model because the frailties in each cluster are assumed to be random. It assumes that, given the frailty, all event times in a cluster are independent. Shared frailty model was introduced by Clayton (1978) without using the notion frailty and extensively studied in Hougaard (2000), Therneau and Grambsch (2000), Duchateau et al. (2002), and Duchateau and Janssen (2008).

Generally, multivariate frailty model can be taken as extension of univariate frailty(3.3.4.1) model which allows the individuals in the same cluster to share the same frailty value. When frailty is shared, dependence between individuals who share frailties is generated. However, when conditioning on the frailty, the individuals are independent of each other. Shared frailty models are very important in analysing multivariate or clustered survival data. Shared frailty model assumes that individuals in a subgroup share the same frailty u, but frailty from group to group may differ.

Suppose there are G groups with ni individuals in the i^th group; $\mathbf{x}_{ij}$ is the observable covariate vector for the j^th individual in the i^th group. The hazard function of the j^th individual in the i^th group is

$$h_{ij}(t / x_{ij}, u_i) = u_i h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_{ij}) \qquad i = 1, \dots, G \;\; j = 1, \dots, n_i \qquad (3.12)$$

where u_i is a random variable assumed to have a one-dimensional distribution. This model is a random effect model with two sources of variation. There is a group variation, described by the random variable, u. Secondly, there is the individual variation described by the hazard function $h_0(t)\exp(\beta' x_{ij})$ .In multivariate frailty model, groups with a large value of the frailty will experience the failure at earlier times than groups with small values of the frailty.

3.3.4.2.1. The shared gamma frailty model

3.3.4.2.1.1. The model

Here we consider models in which the hazard function partly depends on an unobservable random variable thought to act multiplicatively on the hazard, so that large values of the variable increase the hazard. In gamma shared frailty model the baseline hazard (h_0) does not assumed to be a certain parametric distribution i.e. $h_0\ (t)$ left unspecified.

Let j^{th} ($j = 1,2,...,n_i$) individual of the i^{th} group ($i=1,2,\dots,G$), and let T_{ij} denote the survival times under study and let C_{ij} be the corresponding right censoring times. The observations are $Y_{ij} = \min(T_{ij},C_{ij})$and the censoring indicators are $\delta_{ij} = I_{\{Tij \leq Cij\}}$. The frailty model specifies that the hazard function conditional on the frailty is:

$$h_{ij}(t / x_{ij}, u_i) = u_i\, h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}_{ij})$$

Where $h_0(t)$ is the baseline hazard ; $X_{ij}=(X_{1ij}, \dots, X_{pij})$' denotes the covariate vector for the j^{th} individual of group i, and **β** is the corresponding vector of regression parameters. Conditionally on frailty u_i's are independently and identically distributed with mean 1 and unknown variance θ; the probability density function for gamma frailty is given by:

$$f(u) = \frac{u^{(\frac{1}{\theta}-1)}\exp\{-\frac{u}{\theta}\}}{\Gamma\left(\frac{1}{\theta}\right)\theta^{\frac{1}{\theta}}}\ ;\ \ \theta > 0 \qquad (3.13)$$

Large values of θ signify a closer positive relationship between the subjects of the same group and greater heterogeneity among the groups. As discussed by Nielsen et al. (1992), we assume the censoring times C_{ij} to be independent of the event times and of the frailties u_i.

3.3.4.2.1.2. Penalized partial likelihood for Frailty models

The addition of frailties to the Cox model leads to unobserved entities in the model which also prevail in the partial likelihood. It is however assumed that these frailties come from a gamma density with mean equal to 1 and unknown heterogeneity parameter θ. Therefore, a penalty is added to the partial likelihood that decreases with the distance of the frailty from one, the mean of the frailty density.

The penalty term on the log scale in the case of the gamma density is given by $-\sum_{i=1}^{n} \log\{f_u(u_i)\}$. The penalized partial likelihood for the frailty model is then given (McGilchrist, 1993) by

$$l_{ppl}^{F}(\beta_0, u, \theta) = log\left[\prod_{i=1}^{n}\prod_{j=1}^{ni}\left[\frac{u_i\exp\{\beta_0 x_{ij}(t_{ij})\}}{\sum_{R(t_{ij})} u_q \exp\{\beta_0 x_{qs}(t_{ij})\}}\right]^{\delta_{ij}}\right] + \sum_{i=1}^{n}\log\{f_u(u_i)\} \qquad (3.14)$$

For fixed value of the heterogeneity parameter θ, maximization of the penalized partial likelihood criterion leads to the same parameter estimates for the fixed effects β and the frailties u_i as the EM algorithm (Therneau et al., 2003). For a particular value of θ, estimates for the fixed effects, frailties and baseline hazards can thus be obtained by maximizing the penalized partial likelihood.

To make it more clear, let's keep θ fixed in $l^{F}_{ppl}(\beta_0, u, \theta)$ (model 3.14) and write as $l^{F}_{ppl}(\beta_0, u|\theta)$. We further use $\hat{\beta}^{\theta}_{0}$ and $\hat{u}^{\theta}$ to denote the value of β_0 and u, for a given value of θ, and get $l^{F}_{ppl}(\beta_0, u|\theta)$. We now consider the profile penalized partial likelihood $l^{F}_{ppl}(\hat{\beta}^{\theta}_{0}, \hat{u}^{\theta}|\theta)$ as a function of θ. Note that, the profile penalized partial likelihood is increasing with increasing values of θ.

A way to obtain an estimate for θ that corresponds with the EM estimate is to replace the profile penalized partial likelihood by the profile marginal likelihood of θ and to estimate θ as the argument that maximizes this profile marginal likelihood. The marginal likelihood is obtained by integrating out the frailties from the joint density of the observed event/censoring time and the frailties (Klein, 1992) and is of form

$$\mathrm{l}^{\mathrm{F}}_{\mathrm{marg}}(\theta, \mathrm{h}_0(.), \beta_0) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \Big[\mathrm{D_i}\log\theta - \log\Gamma\left(\frac{1}{\theta}\right) + \log\Gamma\left(\frac{1}{\theta} + \mathrm{D_i}\right) - \left(\frac{1}{\theta} + \mathrm{D_i}\right)\log\,[1 + \theta\sum_{\mathrm{j}=1}^{\mathrm{ni}} \mathrm{H_{ij}}(\mathrm{t_{ij}})] + \sum_{\mathrm{j}=1}^{\mathrm{ni}} \delta_{\mathrm{ij}}\{\beta_0\,\mathrm{x_{ij}}(\mathrm{t_{ij}}) + \log\mathrm{h_0}(\mathrm{t_{ij}})\}\Big] \quad (3.15)$$

With D_i the number of events at i$^{\text{th}}$ cluster and $H_{ij}(.)$ is the cumulative hazard for j$^{\text{th}}$ subject in cluster i.

To arrive at the profile marginal likelihood, we replace β_0, $h_0(.)$, and $H_{ij}(.)$ in this general expression for the marginal likelihood by their respective estimates $\hat{\beta}^{\theta}_{0}$, $\hat{h}^{\theta}_{0}(.)$, and $\widehat{H}^{\theta}_{ij}(.)$. In terms of the estimates $\hat{\beta}^{\theta}_{0}$ and $\hat{u}^{\theta}$ we can give explicit expression for the estimated baseline hazard and cumulative baseline hazard. With the total number of ordered distinct event times t(1) <⋯<t(e) and with d(k) the number of events at time t(k), k=1,...,e, define (Duchateau et al., 2002)

$$\hat{h}^{\theta}_{0}(t_{(k)}) = \frac{d_{\{k\}}}{\sum_{R(t_{(k)})} \hat{u}^{\theta}_{q} \exp\{\beta^{\theta}_{0} x_{qs}(t_{(k)})\}} \quad (3.16)$$

An estimate for the cumulative hazard

$$\widehat{H}^{\theta}_{ij}(t_{(ij)}) = \sum_{t_{(l)} \le t_{(ij)}} \hat{h}^{\theta}_{0}(t_{(l)}) \exp\{\beta_0 x_{ij}(t_{(l)})\} \quad (3.17)$$

3.3.4.3. Baseline hazards for parametric frailty models

Under the parametric frailty approach the baseline hazard is defined as a parametric function and the vector of its parameters is estimated together with the regression coefficients and the frailty parameters. A number of possibilities for the baseline hazard are considered in the literature; but in this thesis our focus will be on exponential, Weibull, lognormal, and log logistic distributions. Table 2, presents the hazard and cumulative hazard functions for each of these distributions.

Table 2: Baseline hazards for parametric frailty models

Distribution	ho(t)	Ho(t)=$\int_0^t ho(s)ds$	Parameters space
Exponential	λ	λt	$\lambda > 0$
Weibull	$\lambda \rho t^{\gamma-1}$	λt^{ρ}	$\lambda, \rho > 0$
Log normal	$\frac{\phi(\frac{\log(t)-\mu}{\sigma})}{\sigma t[1-\Phi\left(\frac{\log(t)-\mu}{\sigma}\right)]}$	-log[1- $\Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$]	$\mu \in \mathcal{R}, \sigma > 0$
Log logistic	$\frac{exp^{\lambda}\gamma t^{\gamma-1}}{1+exp^{\lambda}t^{\gamma}}$	Log(1+$exp^{\lambda}t^{\gamma}$)	$\lambda \in \mathcal{R}, \gamma > 0$

Where, $\phi(.)$ and $\Phi(.)$ respectively represent the probability density and the cumulative distribution functions of a standard normal random variable.

3.3.4.4. Frailty distributions

Various frailty distributions have been proposed in the literature (Duchateau and Janssen, 2008). Here in this thesis, we have considered gamma, inverse Gaussian, lognormal and positive stable frailty distributions. In all cases, a single heterogeneity parameter indexes the degree of dependence.

3.3.4.4.1. Gamma frailty

The gamma frailty distribution has been widely used to model intra-cluster dependency because of its simple interpretation and mathematical tractability (Greenwood and Yule, 1920; Vaupel et al., 1979 ; Hougaard, 2000). For this frailty distribution, it is easy to derive the closed form expressions of unconditional survival, cumulative density and hazard function because of the simplicity of the Laplace transformation. Despite these advantages there is no biological reason which makes the gamma distribution more preferable than other frailty distributions.

The probability density function of gamma frailty distribution (which is similar to equation 3.13) is given by

$$f(u) = \frac{u^{(\frac{1}{\theta}-1)}\exp\{-\frac{u}{\theta}\}}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}} \ ; \ \theta > 0 \qquad (3.18)$$

where г(˙) is the gamma function. It corresponds to a gamma distribution Gam (μ;θ) with μ fixed to 1 for identifiably. Its variance is then θ. The associated Laplace transform is

$$L(s) = (1 + \theta s)^{-\frac{1}{\theta}}, \ s \geq 0$$

For the gamma distribution, the Kendall's tau which measures the association between any two event times from the same cluster in the multivariate case, can be computed as

$$\tau = \frac{\theta}{\theta+1} \in (0,1) \qquad (3.19)$$

3.3.4.4.2. Inverse Gaussian frailty

The inverse Gaussian (inverse normal) distribution was introduced as an alternative to the gamma distribution by Hougaard (1984) and has been used by Keiding et al. (1997) and Price and Manatunga (2001). It has uni-modal density and is the member of exponential family. While its shape resembles the other skewed density functions, such as log-normal and gamma.

The probability density function of an inverse normal distributed random variable with mean one and variance θ given by

$$f(u) = \frac{1}{\sqrt{2\pi\theta}} u^{-\frac{3}{2}} \exp\left(-\frac{(u-1)^2}{2\theta u}\right), \theta > 0 \qquad (3.20)$$

Consequently, the Laplace transform is given by

$$L(s) = \exp(\frac{1}{\theta}(1 - \sqrt{1+2\theta s}), \quad s \geq 0$$

With multivariate data, inverse Gaussian distributed frailty yields a Kendall's tau is

$$\tau = \frac{1}{2} - \frac{1}{\theta} + \frac{\exp\left(\frac{2}{\theta}\right)}{\theta^2} \int_{\frac{2}{\theta}}^{\infty} \frac{\exp(-u)}{u} \quad du \in (0, \frac{1}{2}) \qquad (3.21)$$

3.3.4.4.3. Log-normal frailty

Mc Gilchrist(1993) develop methodology for fitting frailty model that parallels the classical mixed model which is not a member of power variance function family. Log-normal frailty models are frequently used in modelling dependence structures in multivariate frailty models, McGilchrist and Aisbett (1991), McGilchrist (1993). Unfortunately for this model, there is no explicit form of the unconditional likelihood. Consequently, estimation strategies based on numerical integration in the maximum likelihood approach are required.

The probability density function of log normal frailty distribution is given by

$$f(u)=\frac{1}{u\sqrt{2\pi\theta}}\exp(-\frac{\log(u)^2}{2\theta}) \text{ with } \theta>0 \qquad (3.22)$$

For lognormal frailty distribution the Laplace transformation does not take a simple form and, no explicit formula exist for Kendall's τ.

3.3.4.4.4. Positive Stable frailty

Hougaard(2000) introduces the positive stable distributions as a family with two parameters: a scale $\delta > 0$ and the so called index $\alpha < 1$. Imposing $\delta = \alpha$ the positive stable frailty distribution PS*(ν) is obtained, with ν =1- α .

The associated probability density function is then

$$f(u)= -\frac{1}{\pi u}\sum_{k=1}^{\infty}\frac{\Gamma(k(1-\nu)+1)}{k!}(-u^{\nu-1})^k\sin((1-\nu)k\pi),\ \nu \in (0,1) \qquad (3.23)$$

The mean and variance are both undefined. Therefore, the heterogeneity parameter ν does not correspond to the variance of the frailty term. Because of that, we intentionally call it ν instead of θ to avoid misinterpretation.

In contrast to the probability density function, the associated Laplace transform takes a very simple form,

$$L(s)=\exp(-s^{1-\nu}) \quad , \quad s\geq 0$$

With clustered data, the Kendall's tau for positive stable distributed frailties is

$$\tau = \nu \ \epsilon(0,1) \qquad (3.24)$$

3.3.5.5. Parameter estimation in parametric frailty models

For right-censored clustered survival data, the observation for subject j∈Ji ={1,...,n_i} from cluster i ∈ I={1,…,s} is the couple (y_{ij}, δ_{ij}) , where y_{ij}=min(t_{ij}, c_{ij}) is the minimum between the survival time t_{ij} and the censoring time c_{ij} , and where δ_{ij} = I($t_{ij} \leq c_{ij}$) is the event indicator. Covariate information may also have been collected; in this case, z_{ij}= (y_{ij}, $\delta_{ij,}$ $\mathbf{x}_{ij}$), where x_{ij} denote the vector of covariates for the (ij)th observation. In the parametric setting, estimation is based on the marginal likelihood in which the frailties have been integrated out by averaging the conditional likelihood with respect to the frailty distribution. Under assumptions of non-informative right-censoring and of independence between the censoring time and the survival time random variables, given the covariate information, the marginal log-likelihood of the observed data $z = \{z_{ij}; i \in I, j \in J_j\}$, can be written as

$$\mathrm{lmarg}(\psi, \beta, \xi; z, X) =$$
$$\sum_{i=1}^{g}\Big\{\Big[\sum_{j=1}^{ni} \delta_{ij}\big(\log\big(h_0(yij)\big) + \mathbf{x}_{ij}^{T}\beta\big)\Big] + \log\Big[(-1)^{di} L^{(di)}\Big(\sum_{j=1}^{ni} H_0\big(y_{ij}\big)\exp\big(x_{ij}^{T}\beta\big)\Big)\Big] -$$
$$\log\,[L(\sum_{j=1}^{ni} H_0\big(\tau_{ij}\big)\exp\big(x_{ij}^{T}\beta\big)\ ]\ \} \tag{3.25}$$

where $d_i = \sum_{j=1}^{ni} \delta_{ij}$ is the number of events in the ith cluster, and $L^{(q)}$(.) the qth derivative of the Laplace transform of the frailty distribution defined as

$$L(s)= E[\exp(-Us)] = \int_0^{\infty} \exp(-uis)\, f(ui)dui\ ,\ s \geq 0$$

ξ is used as generic notation to denote either θ or v (for positive stable frailty model). Estimation of ψ, β, and ξ are obtained by maximizing the marginal log likelihood. This can be done if one is able to compute higher order derivatives $L^{(q)}$(.) of the Laplace transform up to q=max{d_1,…,d_G}.

3.3.5. Model diagnosis

3.3.5.1. Cox-Snell residuals

The residual that is most widely used in the analysis of survival data is the Cox-Snell residual, so called because it is a particular example of the general definition of residuals given by Cox and Snell (1968). For observation j at timet_j, the Cox snell residual can be defined as

$\widehat{H}_j(t_j)$ = - log$\hat{S}_j(t_j)$,

Where $\widehat{H}_j\left(t_j\right)$ and $\hat{S}_j\left(t_j\right)$ are the estimated cumulative hazard and survivor functions, respectively, for the j$^{\text{th}}$ individual at the censored survival time.

The estimated cumulative hazard function ($\widehat{H}_j\left(t_j\right)$) is obtained from the fitted model (Collett, 2003). Cox and Snell argued that if the correct model has been fit to the data, these residuals are n observations from an exponential distribution with unit mean. Thus a plot of the cumulative hazard rate of the residuals against the residuals themselves should result in a straight line of slope 1. Cox–Snell residuals can never be negative and therefore are not symmetric about zero

3.3.5.2. Deviance residuals

The deviance residual help in identifying poorly fitted subjects, which is defined in (Therneau, T. M. et al., 1990) as

D_i=sign$(\widehat{M_i})\sqrt{-2(\widehat{M}_i + di) + \log{(di - \widehat{M}_i)}}$).

Where, $(\widehat{M_i})$ is martingale residual and di is censoring indicator.

The function sign (.) is the sign function which takes the value 1 if $\widehat{M}_i$ is positive and -1 if $\widehat{M}_i$ negative. The martingale residuals take values between negative infinity and unity. They have a skewed distribution with mean zero (Therneau et al., 1990). The deviance residuals are, however, a normalized transform of the martingale residuals. They also have a mean of zero but are approximately symmetrically distributed about zero when the fitted model is appropriate. Deviance residual can also be used like residuals from linear regression. The plot of the deviance residuals against the covariates can be obtained. Any unusual patterns may suggest features of the data that have not been adequately fitted for the model. Very large or very small values suggest that the observation may be an outlier in need of special attention. A plot of the deviance residuals versus the risk score(linear predictor) is also helpful diagnostic to assess a given individual on the model. Potential outliers will have deviance residuals whose absolute values are very large.

3.3.6. Models comparison criteria

After fitting a number of different models for a given dataset, it will be wise idea to compare them using some standard methods of model comparison criteria. Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model.

In this thesis too, we have used AIC as a means to compare different candidate models. To apply AIC in practice, one need to start with a set of candidate models, and then search for a model with minimum AIC values. There will always be information lost due to using one of the candidate models to represent the "true" model. Hence, the goal to select among M candidate models, the model that minimizes this information loss. Suppose, the AIC values of the candidate models are AIC_1, AIC_2, AIC_3, …, AIC_M, then a model with minimum AIC value will be considered as the ‘best’ model out of these M-candidate models.

In addition to AIC, model adequacy and the standardized measure of variability for coefficients were used to identify a ‘best’ model among selected parametric frailty model and shared gamma frailty model. Standardized measure of variability is defined by the ratio of standard error to the corresponding parameter estimate ($SVC=\frac{SE(\beta)}{|\beta|}$) in which the smaller is preferred.

CHAPTER FOUR

4. STATISTICAL ANALYSIS AND RESULTS

4.1. Descriptive summaries

Out of 924 girls included in the study 87.77% (811) of them experience the event of interest, i.e. menarche. The remaining 113 (12.23%) were censored. The median age at menarche found to be 14 years. For all covariates the median age at menarche appeared to vary between 14 to 15 years for different levels of categorical predictors.

Table 3: Descriptive summaries for baseline characteristics and survival status

Characteristics	%n; Mean($\pm SD$)	% Event	% Censored	Median age (years)
Father's education level				
Secondary & higher	29.87	92.39	7.61	14
Primary	41.77	86.01	13.99	14
No education	28.36	85.49	14.51	14.5
Mother's education level				
Secondary & higher	13.42	89.52	10.48	14
Primary	34.52	87.77	12.23	14
No education	52.06	87.32	12.68	15
Household income				
Low	32.57	86.71	13.29	15
Middle	34.31	84.23	15.77	14
High	33.12	92.48	7.52	14
BMI for age				
Normal	86.90	88.04	11.96	14
Under weight	6.93	92.19	7.81	15
Over weight	6.17	78.95	21.05	14
Height for age				
Non-stunted	88.96	88.81	11.19	14
Stunted	11.04	79.41	20.59	15
Workload index	35.14(±16.97)	-	-	-
MUAC	22.72(±2.92)	-	-	-

Source: Jimma Longitudinal Family Survey of Youth; Round 1-3

According to Table 3, the median ages at menarche for girls with parental education level of illiterate were higher (14.5 and 15 years for illiterate fathers and mothers respectively) compared to girls with parental education of primary and, secondary & higher. On the other hand the median age at menarche is delayed for girls from low income family as compared to girls from middle and high income households. Similarly, median timing of menarche for stunted and underweight girls were delayed by one year as compared to girls found in normal range of BAZ and HAZ.

Table 4: log-rank and Peto and Peto tests for equality of survival function of categorical covariates

Categorical Predictors	DF	Logrank Test		Peto-Peto Test	
		Test Statistics	P-value	Test Statistics	P-value
Father’s Education	2	1.42	0.491	2.45	0.300
Mother’s Education	2	59.79	<0.001	57.99	<0.001
House hold income	2	11.35	0.004	15.11	0.001
BMI for age	2	21.90	<0.001	19.49	<0.001
Stunting	1	46.18	<0.001	30.37	<0.001

Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, DF=degree of freedom

According to table 4, the hypothesis on equality of survival functions for different categories of father’s education level is not rejected(in both logrank and peto-peto tests). However, for the rest categorical predictors (mother’s education level, house hold income, BMI for age, and stunting), we have evidence that at least one category is significantly different from the rest.

4.3. The Shared gamma frailty model

The first multivariate survival model used to analyze age at menarche is shared gamma frailty model. Gamma shared frailty model is a model with non-parametric baseline hazard and with gamma distributed random effects or frailty terms. In this thesis, variable selections for gamma shared frailty model is performed first by considering a univariable analysis (Annex 1) and then by including variables with significant association with the dependent variable (p-value < 0.10) in multivariable analysis. Unlike standard PH models, in frailty models the event times are not expected to be independent with in cluster (villages) but do so across villages.

Table 5: Multivariable analysis using shared gamma frailty model

Covariates	Coefficient	SE	HR	95% CI
Mother's education level				
Secondary & higher	Ref			
Primary	-0.009	0.115	0.991	(0.792, 1.241)
No-education	-0.433	0.112	0.649	(0.521,0.808)*
Work load index	-0.003	0.003	0.997	(0.992, 1.002)
House hold income				
Low	Ref			
Middle	0.066	0.089	1.068	(0.897, 1.272)
High	0.173	0.095	1.189	(0.987, 1.431)
BMI for age				
Normal	Ref			
Under weight	-0.344	0.139	0.709	(0.539, 0.932)*
Over weight	0.097	0.156	1.102	(0.812, 1.495)
Height for age				
Non-stunted	Ref			
Stunted	-0.582	0.123	0.558	(0.439,0.711)*
θ= 0.048(SE=0.027)*		τ= 0.023		AIC= 9889.789

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.05, Ref=Reference group, SE=Standard Error, HR=Hazard Ratio, AIC=Akaike Information Criteria*

According to the result obtained using gamma shared frailty model, given the frailty the hazard of having menarche among girls with maternal education level of illiterate is lower (HR=0.649 , 95% CI of 0.521,0.808) compared to girls with maternal education level of secondary and above. However, the estimated conditional HR for girls with maternal education level of primary is not statistically significant (HR=0.991, 95%CI: 0.792, 1.241). It is also possible to check the significance of the hazard ratio by looking the confidence intervals i.e. if one is included in the two estimates (the upper and lower confidence estimates) , it is evidence that the hazard ratio of the two groups is not significantly different from one.

The hazard of menarche for stunted girls is lower (HR=0.558, 95% CI: 0.439, 0.711) compared to non-stunted girls. The hazard of menarche for underweight girls is also lower (HR=0.709, 95% CI: 0.539, 0.932). However, the estimated conditional hazard ratio for being overweight is not statistically significant (HR=1.102, 95% CI: 0.812, 1.495).

The estimated regression coefficients for work load index and household income are not statistically significant in shared gamma multivariable model (unlike the univariable case).

The estimated variance of random effect under shared gamma frailty model is 0.048 (SE=0.027). On the other hand, the correlation between menarcheal ages for any two girls live within the same village estimated to be 0.023. The hypothesis on the significance of θ is performed by considering ordinary Cox PH model as reduced model where the critical value obtained from mixture of chi square distribution with zero and one degree of freedom.

4.3.1. Diagnostic for Shared gamma frailty model

4.3.1.1. Plot of Cox-Snell residuals

Figure 1, which is plot of Cox snell residuals versus cumulative hazard of residuals, indicates the fitted line was better at the beggning(at smaller values of cox snell residuals) but deviate heighly from aline with zero intercept and unit slope as the magnitude of cox snell residual increases. Accordingly, the plotted line does not provide positive evidence in favor of the fitted shared gamma frailty model.

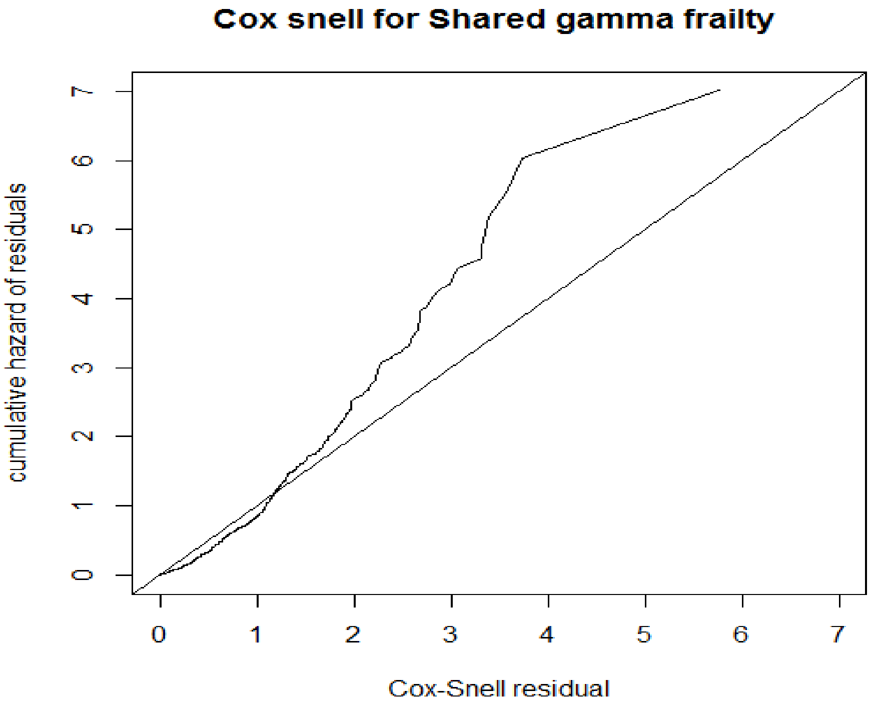

Figure 1: Plot of cumulative hazard vs Cox Snell residuals

4.3.1.2. Plot of deviance residuals

According to plot of deviance residuals against the risk score (figure 2), the distribution of residuals is approximately symmetric about zero with random pattern and there exists no outlying observation.

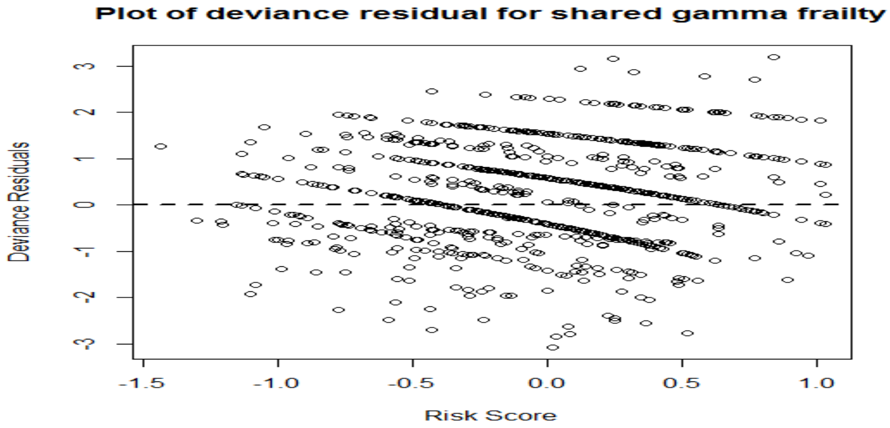

Figure 2: plot of deviance residuals with risk score

4.4. Parametric frailty models

A popular approach to model clustered survival data is to use parametric frailty models. In parametric frailty models the baseline hazard assumed to follow some parametric distribution. Before performing a univariable analysis(one predictor at a time) using candidate frailty models, it is better to check if the estimated heterogeneity parameter is significant using likelihood ratio test.

Table 6: Summary of Estimated heterogeneity parameters with LR test

	Frailty Distribution							
Baseline hazard	Gamma		Inverse Gaussian		Log normal		Positive stable	
	θ	SE(θ)	θ	SE(θ)	θ	SE(θ)	ν	SE(ν)
Exponential	5.3e-72	3.3e-69	6.9e-09	6.2e-06	7.8e-09	7.1e-05	4.9e-07	3.1e-04
Weibull	0.093*	0.041	0.102*	0.048	0.097*	0.044	0.115*	0.042
Log normal	0.096*	0.044	0.099*	0.045	0.105*	0.049	0.105*	0.037
Log logistic	0.136*	0.062	0.133*	0.059	0.149*	0.068	0.118*	0.038

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3; SE=standard error,*p<0.05*

As displayed in Table 6, the estimated heterogeneity parameters are significant for weibull, lognormal and loglogistic based frailty models but not for exponential based models. Consequently, it is possible to ignore exponential based frailty models from our list of candidate frailty models. Even if exponential based frailty models are ignored, it is possible to see whether these models are appropriate based on p-values for estimate of log(shape) or log(ρ) in Weibull based models. That is, if log(ρ) is not significantly different from zero exponential based models are better than Weibull based models. The statistical significance of heterogeneity parameters can also be seen by constructing CI for thetas(ν for positive stable frailty) i.e. if the estimated lower confidence limit is zero, it is evidence that the estimated heterogeneity can be ignored.

4.4.1. Variable Selection and Comparison of Parametric frailty models

To identify important covariates we use a univariable analysis (one predictor at a time) which is similar to method used in section 4.3 with similar level of significance. The results of univariable and multivariable analysis for gamma and inverse Gaussian frailty are displayed in Annex1 & Annex 2 of this thesis, respectively.

After performing multivariable analysis for each model, the next step is to identify a model which is 'best' for describing the association of age at menarche with different potential covariates. The following table displays the AIC values for each combination of baseline hazard and frailty distribution.

Table 7: AIC values for candidate parametric frailty models

Baseline hazard	Frailty Distribution			
	Gamma	Inverse Gaussian	Log normal	Positive stable
Weibull	2714.140	2714.103	2713.81	2720.814
Log normal	2525.285	2526.327	2525.805	2534.577
Log logistic	**2479.017**	2481.278	2480.795	2488.393

Source: Jimma Longitudinal Family Survey of Youth; Round 1-3; AIC=Akaike information criteria

According to Table 7, AIC values of all weibull based frailty models are larger indicating worst fit of models compared to the rest candidate models. On the other hand, log logistic-gamma frailty model has the smallest AIC(2479.017) demonstrating better fit as compared to the rest candidate models.

4.4.2. Log logistic gamma frailty model

As described in section (4.4.1), the AIC value of log logistic gamma frailty model is the smallest as compared to the rest. Accordingly, it has been selected as 'best' parametric frailty model for modeling age at menarche.

Table 8: multivariable analysis of logistic gamma frailty models

Covariates	Coefficient	SE	φ	95% CI
Mother's education level				
Secondary & higher	Ref			
Primary	0.005	0.008	1.005	(0.988, 1.022)
No-education	0.046	0.008	1.048	(1.031, 1.065)*
Workload index	0.0002	0.0002	1.0002	(0.999, 1.001)
House hold income				
Low	Ref			
Middle	-0.007	0.006	0.993	(0.982, 1.005)
High	-0.022	0.007	0.979	(0.966,0.992)*
BMI for age				
Normal	Ref			
Under weight	0.032	0.010	1.033	(1.013, 1.053)*
Over weight	-0.015	0.011	0.985	(0.964, 1.007)
Height for age				
Non stunted	Ref			
Stunted	0.060	0.008	1.062	(1.044,1.079)*

Cons= 2.632 (SE=0.012)* γ=0.043 (SE=0.002)*

θ=0.090(SE=0.049,95%CL[0.030,0.267])* τ = 0.043 AIC= 2479.017

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.05, Ref=Reference group, SE= Standard Error, φ=Acceleration factor, AIC=Akaike Information Criteria, CI= confidence interval*

According to the result obtained using loglogistic gamma frailty model, age at menarche for girls with maternal education level of illiterate is delayed with a factor of φ=1.048 (95% CI: 1.031, 1.065) compared to girls with maternal education level of high school and higher. However, the acceleration factor for maternal education level of primary is not statistically significant (φ=1.005, 95% CI: 0.988, 1.022).

Girls in the households with the highest income experience menarche earlier (φ=0.979, 95% CI: 0.966, 0.992) compared to those in low income category. But, the estimated acceleration factor for girls in the households with middle income tertile is not statistically significant (φ=0.993, 95% CI: 0.982, 1.005).

The timing of menarche for stunted girls is delayed by a factors of 1.062 (95% CI: 1.044, 1.079) compared to non-stunted girls. The menarcheal age for underweight girls also delayed by a factor of φ=1.033 (95% CI: 1.013, 1.053) compared to girls with normal BAZ range. However, the age at menarche for overweight girls is not significantly different from BAZ-normal girls (φ=0.985, 95% CI: 0.964, 1.007).

The estimated acceleration factor for work load is not statistically significant at multivariable loglogistic gamma frailty model (unlike the univariate case)

In log logistic gamma frailty model (Table 8) the estimated variance of random effect is 0.090 (95% CI: 0.030, 0.267) and the lower confidence interval estimate for θ is 0.030. The significance of heterogeneity parameter can be also confirmed using likelihood ratio test p-value, where the reduced model is log logistic survival model and the full model includes one additional parameter θ. So, the achievement in loglikelihood obtained by including heterogeneity parameter θ will be illustrated by p-value in our case, the inclusion of θ is supported. In frailty models Kendell's tau (τ) represent the correlation between any two failure times within the same cluster. Based on the final parametric frailty model Kendell's tau is estimated to be 0.043. Here the estimated shape parameter is less than one (0.043) indicating the hazard of menarche decreases to some points then rises.

4.4.3. Model Diagnostic

4.4.3.1. Checking adequacy of Log-logistic distribution

Graphical evaluation of the log logistic baseline distribution can be assessed by plotting the log failure of odds i.e. $log(\frac{1-\hat{S}_t(t)}{(\hat{S}_t)})$, which is a linear function of log(t).And it is expected to be a straight line, if it is plotted against log(t). If so, it is evidence that the assumed log logistic baseline distribution holds. Where $(\hat{S}_t)$ is a KM survival estimate. From figure 3, it can be seen that the plotted points fall in a linear fashion. This is evidence that the assumed log-logistic baseline distribution is appropriate for modelling age at menarche.

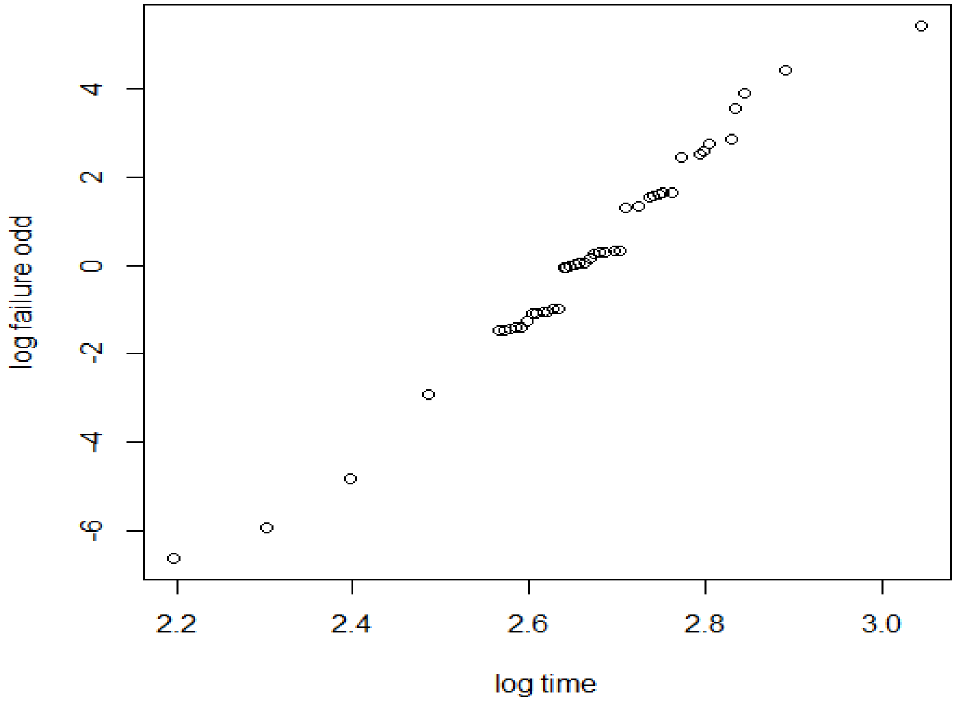

Figure 3: plot of log time vs log failure odd

4.4.3.2. Plot of Cox-Snell residuals

The Cox-Snell residuals obtained from Weibull, lognormal and log-logistic gamma frailty models were plotted with cumulative hazard of residuals to assess goodness of the respective fitted models. Among the three plots (Figure 4) the line fitted by log-logistic gamma frailty model lies in a better fashion on a line which is with unit slope and zero intercept. So based on plot of cox Snell residuals log-logistic-gamma model fit the data well compared with others.

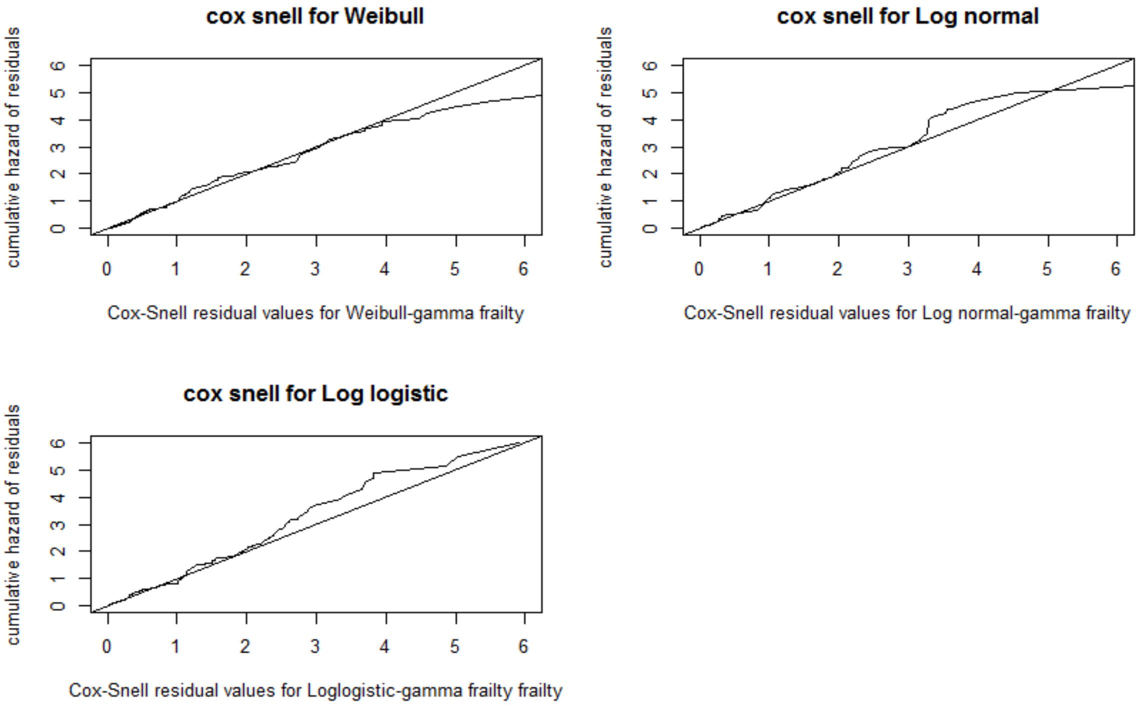

Figure 4: Cumulative hazard plot of the Cox-Snell residual for frailty models.

4.4.3.3. Plot of Deviance residual

The plot of deviance residual against the risk score (figure 5) shows that the deviance residuals seem to be approximately symmetric about zero with random pattern and there exists no outlying observation. Therefore, we have almost no concern about the adequacy of the fitted log-logistic gamma frailty model.

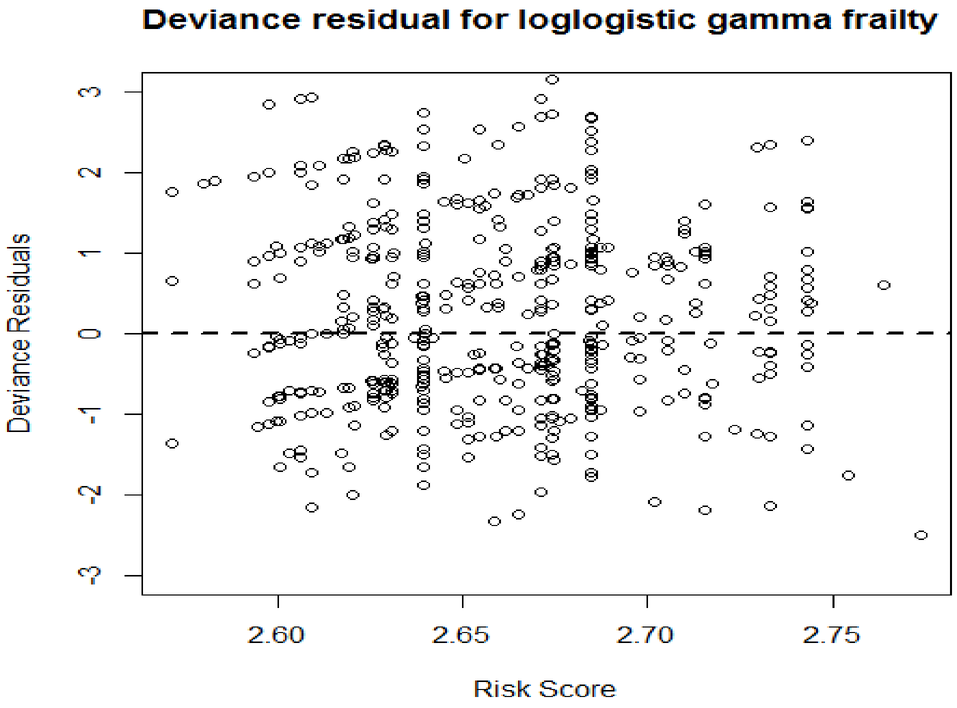

Figure 5: plot of deviance residual vs risk score

4.5. Comparison of Shared gamma and log logistic–gamma frailty models

The AIC value for shared gamma frailty model which is based on penalized partial likelihood approach is 9889.789 which is higher than the AIC value for log-logistic gamma frailty model (2479.017) which may give us some insight in selecting our appropriate model. In addition to AIC values of the two models we used the standardized variability of coefficients and their adequacy using plot of Cox-Snell residuals.

Table 9: SVC for shared gamma and log logistic gamma frailty models

	Shared gamma frailty			Log logistic–gamma frailty		
Covariate	Estimate	SE	SVC	Estimate	SE	SVC
Mother's education level						
Secondary & higher	Ref			Ref		
Primary	-0.009	0.115	12.778	0.005	0.008	1.6
No-education	-0.433	0.112	0.259	0.046	0.008	0.174
Workload index	-0.003	0.003	1	0.0002	0.0002	1
House hold income						
Low	Ref			Ref		
Middle	0.066	0.089	1.348	-0.007	0.006	0.857
High	0.173	0.095	0.549	-0.022	0.007	0.318
BMI for age						
Normal	Ref			Ref		
Underweight	-0.344	0.139	0.404	0.032	0.010	0.312
Over weight	0.097	0.156	1.608	-0.015	0.011	0.733
Height for age						
Non stunted	Ref			Ref		
Stunted	-0.582	0.123	0.211	0.060	0.008	0.133

Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, Ref=Reference group, SE= Standard Error, SVC=Standardized Variability of Coefficients

As shown in Figure 1, the Cox Snell residual plot for shared gamma frailty model indicates the plotted line lie far from a line which has zero intercept and unit slope. Since plot of Cox-Snell residuals indicate the overall goodness of fit of used model to the dataset, we lost evidence to accept shared gamma frailty model as appropriate model in modeling age at menarche. On the other hand, the plotted line of Cox-Snell residual for log logistic- gamma frailty model (Figure 4) shows good fit of the model to the dataset. Similarly, the computed SVC values of all estimate of covariate effect using log logistic-gamma frailty model is less than the corresponding values for shared gamma frailty model. To model age at menarche, we found appropriate to use log-logistic gamma frailty model, mainly because of its smallest AIC value, adequacy of model fitness in modeling age at menarche and due to its smallest SVCs compared to shared gamma frailty model.

CHAPTER FIVE

5. DISCUSSION AND CONCLUSION

5.1 Discussion

In this thesis, we modeled age at menarche taking in to account the clustering within a village in the dataset. In survival analysis if the assumption of independence between the event times is questionable, then the classical or standard survival models will result in incorrect inference. During such situation, where there is clustering or dependence in the event times, then random effect survival models (also known as frailty models) are appropriate to consider. In this thesis, time to menarche data was analyzed using gamma, inverse Gaussian, lognormal and positive stable frailty models with different specification of baseline hazards.

Among various frailty models (with parametric and semi-parametric baseline hazards), log logistic-gamma frailty model was selected due to its lowest AIC value, fit the data well, and has smallest standardized variability of coefficients as compared to shared gamma frailty model. Overall, models with flexible hazard fit time to menarche data well as compared to models with monotone and constant hazard based frailty models (Table 6 & 7). In modeling clustered survival data using frailty models, often it is better to test whether the heterogeneity parameter is significantly different from zero or not. Such a test is based on 50:50 mixture of chi-square distribution with zero and one degrees of freedom (Duchateau and Janssen, 2008). In modeling age at menarche using frailty models the estimated heterogeneity parameters were significant except for exponential based models.

The estimated heterogeneity parameter was estimated to be 0.090 (95% CI: 0.030, 0.267). The correlation of age at menarche for any two girls lives within the same village (Kebeles) was estimated to be 0.043. From Annex-1 and 2, the estimated log(shape) parameters in weibull based frailty models are significant indicating Weibull based models are significantly better than exponential based frailty models in modeling age at menarche. It can be also seen that among estimated heterogeneity parameters the variance of random effect estimated by log-logistic-gamma frailty model is the maximum.

In modeling cluster survival data the shape of the frailty distribution also plays an important role. Accordingly, a gamma distribution which has large left tail leads to strong late dependence (Hougaard, 2000). There are many applications of the gamma frailty model. Lancaster (1979) suggested this model for the duration of unemployment. Aalen (1987) studied the expulsion of intrauterine contraceptive devices. Ellermann et al. (1992) studied recidivism among criminals using gamma-Weibull model. Andersen et al. (1993) used the gamma frailty model to check the proportional hazards assumptions in his study of malignant melanoma. Vaupel et al. (1979) used the gamma distribution in their studies on population mortality data from Sweden.

Generally, the gamma distribution has two advantages as a frailty distribution beside its mathematical simplicity. The frailty distribution of the survivors at any given age is again a gamma distribution, with the same parameter and a different scale parameter. The second advantage is that the frailty distribution among the subjects experiencing event at any time is also a gamma distribution, with the same shape parameter plus one, and a scale parameter as a function of the event time (Abdulkarimova, 2013).

This study also closely examine the effect of parental education level, house hold income category, work load, stunting, body mass index for age, and MUAC on the age at onset of menarche. Each predictor effect was assessed with univariable analysis and those predictors significant at level of 0.10 were analyzed together in multivariable model. Consequently, mother’s education level, household income, height for age and BMI for age were identified as important prognostic factors in modelling age at menarche. Results of loglogistic-gamma frailty model indicates significant effect of maternal education level on timing of menarche which is similar result to study done in Bahrain (Al-Sayyad et al,1991). Similarly, study at a Bangladeshi University also shows the mother’s educational level and occupation (but not fathers) was found to have a significant influence on their daughter’s age at menarche (Hossainet al., 2010).

Study done in Iran also showed similar result that BMI was significantly associated with age at menarche and being underweight delayed menarche (Ayatollahi et al, 1999). Though the work load was not statistically significant similar to Bagga and Kulkarni(2000) at our multivariable result, there are studies showing girls who had to do more physical work, or had a long, tiresome way to school and spent greater expenditure of calories delay the process of puberty (Serap *et al*, 2009) supporting the result of univariable analysis on work load.

The result of our selected model also showed that stunted girls have delayed age at menarche which is consistent with the result of studies in Bangladesh (Bosh et al, 2008). Other studies have also declared that the stunted girls experienced a significant delay in age at menarche of as compared to the tallest girls (Simondom et al, 1998). Similarly, girls who were from poor families had higher age at menarche than girls of the higher classes (Ayatollahi et al, 1999) which consolidate with our result.

5.2. Conclusions

This study was based on a dataset of time to menarche obtained from Jimma longitudinal and family Survey of youths. The aim of this study was to model time to menarche data using appropriate multivariate survival models. Hence, the data was modeled using various frailty models. The modeling for each candidate models was done first by performing univariable analysis to identify important prognostic factors for time to menarche data. Since the analysis is based on correlated event time analysis, for frailty models we first observe whether the failure times are really correlated or not. The result shows that exponential based models produce insignificant variance of random effect (ν in positive stable case which is not the variance of random effect). Consequently, exponential-based frailty models were not compared with other parametric frailty models, but these models found to be inappropriate compared to Weibull based frailty models.

To identify a best distribution AIC values were used. And, the result shows that logistic – gamma frailty models fit the data well. The adequacy of the selected model also checked using different diagnostic mechanisms. Plot of Cox Snell residual for log-logistic-gamma frailty model is well straight and close to a line which has zero intercept and unit slope.

Similarly, adequacy of log logistic model (log-logistic baseline distribution for frailty model) was assessed using plot of log(t) versus log(failure odd). And the plot gives positive evidence on adequacy of log logistic distribution for our selected model. Finally, the assessment for outliers was performed by plotting deviance residual with risk score. And, the plot does not identify any clear outlier for our selected model. It has been also founded that mother's education, house hold income category, BMI for age and height for age (stunting) were important prognostic factors in modeling age at menarche.

REFERENCES

Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine*. **7:** 1121-1137.

Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*. **4**: 951 – 972

Abdulkarimova, U.(2013).Frailty Models for Modelling Heterogeneity. *Open Access Dissertations and Theses*. Paper 7757.

Allison, P. D. (1984). Event history analysis. Thousand Oaks, CA: Sage.

Al-Sayyad, A., Al-Nooh, A., Al-Sayed, F.(1991). Demographic and health survey in Bahrain. Bahrain: Arabian Gulf University.

Ayatollahi, S.M., Dawlatabadi, E., Ayatollahi, S.A. (1999). Age at menarche and its correlates in Shiraz, Southern Iran. *Irn J Med* **24** (1&2):20-25

Ayatollahi, S.M., Dawlatabadi, E., Ayatollahi, S.A.(2002). Age at menarche in Iran. *Ann Hum Biol*. **29**:355-362.

Bagga, A., Kulkarni, S.(2000). Age at menarche and secular trend in Maharashtra Indian girls. *Act bio Szeged* .**44**(1-4): 53-57

Beard, R.E. (1959). Note on some mathematical mortality models. Ciba Foundation Colloquium on Ageing, Little, Brown, Boston, 302–311.

Belachew, T., Hadley, C., Lindstrom, D., Getachew, Y., Duchateau, L., Kolstren, P. (2011). Food insecurity and age at menarche among adolescent girls in Jimma Zone Southwest Ethiopia. *Reproductive Biology and Endocrinology*. **9**:125

Bloessfeld, H.P., Hamerle, A. and Mayer, K.U. (1989). Event history analysis: Statistical theory and application in the social sciences. Hillsdale, NJ: Erlbaum.

Bosch, A.M., Willekens, F.J., Baqui, A.H., Van Ginneken, J.K., Hutter, I.(2008). Association between age at menarche and early-life nutritional status in rural Bangladesh. *J Biosoc Sci*. **40**(2):223–237

Braithwaite, D., Moore, D. H., Lustig, R. H., Epel, E. S., Ong, K. K., Rehkopf, D. H.,et al. (2009). Socioeconomic status in relation to early menarche among black and white girls. *Cancer Causes & Control*. **20**(5), 713–720.

Breslow, N. E.(1970). A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* .**57**: 579–594.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrika* . **30**, 89-100.

Chavarro, J., Villamor, E., Narvaez, J., Hoyos, A.(2004). Socio-demographic predictors of age at menarche in a group of Colombian university women. *Ann Hum Biol.* **31:**245-257.

Chodick, Gabriel, Alfred Rademaker, Michael Huerta, Rn D. Balicer, Nadav Davidovitch, and Itamar Grotto.(2005). Secular trends in age at menarche, smoking, and oral contraceptive use among Israeli girls. *Preventing Chronic Disease*. **2**(2): A12.

Chowdhury, S., Shahabuddin, A.K., Seal, A.J., Talukder, K.K., Hassan, Q., Begum, R.A. (2000). Nutritional status and age at menarche in a rural area of Bangladesh. *Ann Hum Biol*. **27**: 249-56

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika,* **65**:141-151.

Clayton, D., Cuzick, J. (1985).The semi-parametric Pareto model for regression analysis of survival times. Proceedings of the Centenary Session of the International Statistical Institute, Amsterdam

Collett, D.(2003). Modelling Survival Data in Medical Research, 2nd edition. London, UK: Chapman & Hall/CRC

Cox, D.R. and Snell, E. J.(1968). A general definition of residuals with discussion . *Journal of the Royal Statistical Society*. Series B **30**, 248-275.

Cox, D.R. (1972). Regression models and life tables. Journal of the Royal Statistical Society. *Methodologica*., **34** : 187-220.

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

Dress, K.A. (1986). The effect of gender identify on conversation. *Social Psychology Quarterly*. **49**: 294-301.

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R. and Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics & Data Analysis*. **40**:603-630.

Duchateau, L., Paul. J. (2008). The Frailty Model. New York, Springer

Fergusson, D.M., Horwood, L.J. and Diamond, M.E. (1985). A survival analysis of childhood family history. *Journal of Marriage and the Family.* **47**, 287-295.

Femalee, D., Eder, D. and Tsui, W.Y. (1985). Peer influence on classroom attention. Social Psychology Quarterly.**48**: 215-226.

Flinn, C.J., and Heckman, J.J. (1982). New methods for analyzing individual event histories. *Socialogical Methodology*, 99-140.

Freedman, D.S., Khan, L.K., Serdula, M.K., Dietz, W.H., Srinivasan, S.R., Berenson, G.S. (2002). Relation to age at menarche to race, time period, and anthropometric dimensions: the Bogalusa Heart Study. *Pediatrics* . **110**(4): e43.

Garn, S.M., LaVelle, M., Rosenberg, K.R., Hawthorne, V.M. (1986). Maturational timing as a factor in female fatness and obesity. *Am J Clin Nutr*. **43**:879-883.

Gehan, E. A. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. **52**: 203–223.

Greenwood, M., Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*.**83**,*255-279*

Gross, A.J. and Clark, V.A. (1975). Survival distributios. John Wiley and Sons Inc., New York.

Hembroff, L.A. and Myers, D.E. (1984). Status characteristics: Degrees of task relevance and decision processes. *Social Psychology Quarterly*. **47**: 337-346.

Hossain, M. G., Islam, S., Aik, S., Zaman, T. K. and Lestrel, P. E.(2010). Age at menarche of university students in Bangladesh: secular trends and association with adult anthropometric measures and socio-demographic factors. *Journal of Biosocial Science*. **42**(5), 677–687.

Hougaard, P. (1984). Life table methods for heterogeneous populations. *Biometrika.* **71**, 75 – 83

Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika.* **73**, 387 – 396

Hougaard, P. (1986b). A class of multivariate failure time distributions. *Biometrika*. **73**, 671 – 678

Jonker, M., Bhulai, S., Boomsma, D.I., Ligthart, R.S., Posthuma, D., Van der Vaart, A.W. (2009). Gamma frailty model for linkage analysis with application to interval-censored migraine data. *Biostatistics*. **10**(1):187-200.

Kalbfleisch, J. D., and Prentice, R. L. (1963). Marginal likelihoods based on Cox's regression and life model. *Biometrika*. **60** : 267-278.

Kaplan, E.L. & Meier, P. (1958).Nonparametric estimation from incomplete observations. *Journal of the American Statistical Associatio.* **53**:457-481

Keiding, N., Andersen, P., Klein, J. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*. **16**, 215 – 224

Keyfitz, N. and Littman, G. (1979). Mortality in a heterogeneous population. *Population Studies*, **33**, 333-342.

Kleinbaum, D.G., Klein, M. (2005). Survival analysis: a self-learning text. 2nd ed. New York: Springer.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*. **47**, 939–956.

Lassek, W.D., Gaulin, S.J.(2007)**.** Menarche related to fat distribution. *Am J Phys Anthropol.* **133:**1147-1151

Lavori, P.W., Keller, M.B. and Klerman, G.L. (1984). Relapses in affective disorders: A reanalysis of the literature using life-table methods. *Journal of Psychiatric Research*. **18**,13-25.

Malina, R.M.(1983): Menarche in athletes: a synthesis and hypothesis**.** *Ann Hum Biol.* , **10:**1-24.

McGilchrist, C.A., Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics*. **47**, 461 – 466

Miller, R.G. (1981). Survival analysis. John Wiley and Sons, New York.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journalof the American Statistical Association*. **84**, 487-93.

Osteria, T.S.(1983). Nutritional status and menarche in a rural community in the Philippines. *Philipp J Nutr*. **36**(4): 150-156.

Padez, C.(2003). Social background and age at menarche in Portuguese university students: a note on the secular changes in Portugal. *American Journal of Human Biology*. **15**(3), 415– 427.

Perperoglou, A., Keramopoullos, A., van Houwelingen, H.C. (2007). Approaches in modelling long-term survival: an application to breast cancer. *Stat Med*. **26**(13): 2666-2685.

Prado, C., Martinez, R., Perez-de landozabal, E. (1995). Menarcheal age as an indicator of socio-economic level in emigrants. *J. Hum. Ecol.*, **4**: 157-171

Price, D.L., Manatunga, A.K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine.* **20**, 1515 – 1527

Proos, L.A., Hofvander, Y., Tunevo, T.(1991). Menarcheal and growth pattern in Indian girls adopted in Sweden. I. Menarcheal age. Acta Paediatr Scand. **80**:852–858.

Rodgers, J.L., Doughty, D. (2000). Genetic Influences on Human Fertility and Sexuality: Behavior Genetic Modeling of Menarche in US Females. Boston: Kluwer Academic Publishers.

Serap S, Funda K, Devrim T K, Mehmet Z, Özgür S.(2009) Factors affecting onset of puberty in Denizli province in Turkey. *The Turkish Journal of Pediatrics.* **51**(1) : 49-55

Shapiro, D.R., Quitkin, F.M. and Fliess, J.L. (1989). Response to maintenance therapy in bipolar illness: Effects of index episode. *Archives of General Psychiatry*. **46**, 401-405.

Stevens, V.J. and Hollis, J.F. (1989). Preventing smoking relapse using an individually tailed skills training technique. *Journal of Consulting and Clinical Psychology.* **57**, 420-424.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990).Martingale-based residuals for survival models. *Biometrika* .**77**: 147-160.

Therneau, T. and Grambsch, P. (2000). Modelling Survival Data: Extending the Cox Model. Springer-Verlag, New York.

Thomas, F., Renaud, F., Benefice, E., de Meeüs, T., Guegan, J.F.(2001) . International variability of ages at menarche and menopause: patterns and main determinants. *Hum Biol*. **73**(2):271-290

Tuma, N.B. and Hannan, M.T. (1984). Social dynamics: Models and Methods. San Diego, C.A. Academic Press

Tweedie, M.C.K. (1984) .An index which distinguishes between some important exponential families. In: Statistics: Applications and New Directions. Proc. of the Indian Statistical Institute Golden Jubelee International Conference. (eds. J.K. Ghosh and J. Roy). 579-604.

Vaupel, J.W., Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity on individual frailty on the dynamic of mortality. *Demography*.**16**(3), 439-454.

Whincup, P.H., Gilg, J.A., Odaki, K., Taylor, S.J.C., Cook, D.J. (2001). Age at menarche in contemporary British teenagers: survey of girls born between 1982 and 1986. *BMJ*; **323**:232-237.

Wronka, I., Pawlinska, R. Menarcheal age and socioeconomic factors in Poland.(2005). *Ann Hum Biol*. **32**:630–638.

Yamaguchi, Kazuo. (1992).Accelerated Failure-Time Regression Models with a Regression Model of Surviving Fraction: An Application to the Analysis of 'Permanent Employment' in Japan. *Journal of the American Statistical Association* **87**: 284-292.

ANNEX 1 : Univariable analysis

Univariable Shared gamma frailty model

Covariate	Coefficient	SE	HR	90% CI
Father's education level				
Secondary & higher	Ref.			
Primary	-0.035	0.085	0.966	(0.840,1.111)
No-education	-0.083	0.096	0.921	(0.787,1.077)
Mother's education level				
Secondary & higher	Ref.			
Primary	-0.020	0.114	0.983	(0.815,1.185)
No-education	-0.489	0.110	0.613	(0.512,0.733)*
Work load index	-0.005	0.002	0.995	(0.991,0.999)*
HH income category				
Low	Ref			
Middle	0.051	0.088	1.052	(0.909,1.217)
High	0.294	0.091	1.341	(1.156,1.557)*
BMI for age				
Normal	Ref			
Underweight	-0.394	0.139	0.675	(0.536,0.849)*
Over-weight	0.075	0.155	1.078	(0.835,1.393)
Height for age				
Non stunted	Ref			
Stunted	-0.655	0.122	0.520	(0.425,0.636)*
MUAC	-0.003	0.004	0.997	(0.990,1.004)

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10,*

Ref=Reference group, SE=Standard Error, HR =Hazard ratio, CI=confidence interval

Univariable :Weibull –gamma frailty

Covariate	Coefficient	SE	φ	90% CI
Father's education level				
Secondary & higher	Ref.			
Primary	0.002	0.008	1.002	(0.989,1.015)
No-education	0.003	0.009	1.003	(0.989,1.017)
log(ρ)=2.42(SE=0.024)*				
Mother's education				
Secondary & higher	Ref.			
Primary	0.004	0.009	1.004	(0.988,1.019)
No-education	0.067	0.009	1.069	(1.053,1.085)*
log(ρ) =2.481(SE=0.023)*				
Workload index	0.001	0.0002	1.001	(1.0006,1.001
log(ρ) =2.433(SE=0.024)*				
House hold income				
Low	Ref			
Middle	-0.023	0.008	0.977	(0.965,0.990)*
High	-0.034	0.008	0.966	(0.953,0.979)*
log(ρ)=2.425(SE=0.024)*				
BMI for age				
Normal	Ref			
Underweight	0.084	0.013	1.088	(1.065,1.111)*
Over-weight	-0.014	0.013	0.986	(0.964,1.008)
log(ρ) =2.459(SE=0.024)*				
Height for age				
Non stunted	Ref			
Stunted	0.089	0.011	1.094	(1.074,1.113)*
log(ρ) =2.469(SE=0.024)*				
MUAC	0.0005	0.0003	1.0005	(0.999,1.001)
log(ρ) =2.420(SE=0.024)*				

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10, Ref=Reference group, SE= Standard Error, φ =Acceleration factor, CI=confidence interval*

Univariable lognormal –gamma frailty

Covariate	Coefficient	SE	φ	90% CI
Father education level				
Secondary & higher	Ref.			
Primary	0.007	0.008	1.006	(0.994,1.021)
No-education	0.010	0.007	1.010	(0.998,1.022)
Mother education level				
Secondary & higher	Ref.			
Primary	0.015	0.009	1.015	(1.001,1.029)*
No-education	0.065	0.008	1.067	(1.052,1.082)*
Workload index	0.0006	0.0001	1.0006	(1.0003,1.001)*
Household income				
Low	Ref			
Middle	-0.007	0.007	0.993	(0.982,1.005)
High	-0.040	0.007	0.961	(0.949,0.972)*
BMI for age				
Normal	Ref			
Underweight	0.039	0.011	1.040	(1.021,1.059)*
Over-weight	-0.014	0.012	0.986	(0.966,1.006)
Height for age				
Non stunted	Ref			
Stunted	0.070	0.010	1.072	(1.056,1.089)*
MUAC	0.0003	0.0005	1.0003	(0.999,1.001)

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10, Ref=Reference group, SE= Standard Error, φ =Acceleration factor, CI=confidence interval*

Univariable Loglogistic–gamma frailty

Covariate	Coefficient	SE	*φ*	90% CI
Father's education level				
Secondary & higher	Ref.			
Primary	0.003	0.006	1.003	(0.993,1.014)
No-education	0.012	0.007	1.012	(0.999,1.024)
Mother's education level				
Secondary & higher	Ref.			
Primary	0.011	0.009	1.011	(0.996,1.025)
No-education	0.058	0.008	1.060	(1.046,1.075)*
Work load index	0.001	0.0002	1.001	(1.0003,1.001)*
House hold income				
Low	Ref			
Middle	-0.004	0.007	0.996	(0.985,1.006)
High	-0.037	0.007	0.964	(0.953,0.975)*
BMI for age				
Normal	Ref			
Underweight	0.033	0.011	1.033	(1.011,1.055)*
Over-weight	-0.013	0.012	0.987	(0.964,1.010)
Height for age				
Non stunted	Ref			
Stunted	0.071	0.009	1.073	(1.058,1.089)*
MUAC	0.001	0.001	1.001	(0.9997,1.0014)

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10, Ref=Reference group, SE= Standard Error, φ =Acceleration factor, CI=confidence interval*

Univariable Analysis weibull –Inverse gaussian frailty

Covariate	Coefficient	SE	φ	90% CI
Father's education level				
Secondary & higher	Ref.			
Primary	0.002	0.008	1.002	(0.987,1.018)
No-education	0.003	0.009	1.003	(0.986,1.021)
log(ρ)=2.42(SE=0.024)*				
Mother's education level				
Secondary & higher	Ref.			
Primary	0.003	0.010	1.003	(0.985,1.022)
No-education	0.067	0.009	1.069	(1.050,1.088)*
log(ρ)=11.951(SE=0.272)*				
Workload index	0.001	0.0002	1.001	(1.001,1.0014)*
log(ρ)=2.43(SE=0.024)*				
House hold income				
Low	Ref			
Middle	-0.023	0.008	0.977	(0.962,0.992)*
High	-0.034	0.008	0.966	(0.950,0.982)*
log(ρ)=2.425(SE=0.024)*				
BMI for age				
Normal	Ref			
Underweight	0.084	0.013	1.087	(1.060,1.116)*
Over-weight	-0.014	0.013	0.986	(0.960,1.012)
log(ρ) =2.458(SE=0.024)*				
Height for age				
Non stunted	Ref			
Stunted	0.089	0.011	1.094	(1.071,1.117)*
log(ρ)=2.47(SE=0.02)*				
MUAC	0.0005	0.0003	1.0005	(0.999,1.001)
log(ρ) =2.419(SE=0.024)*				

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10, Ref=Reference group, SE= Standard Error, φ =Acceleration factor, CI=confidence interval*

Univariable lognormal –Inverse gaussian frailty

Covariate	Coefficient	SE	φ	90%CI
Father's education level				
Secondary & higher	Ref.			
Primary	0.006	0.007	1.006	(0.992,1.019)
No-education	0.015	0.007	1.015	(0.999,1.031)
Mother's education level				
Secondary & higher	Ref.			
Primary	0.015	0.009	1.015	(0.998,1.033)
No-education	0.065	0.008	1.067	(1.050,1.085)*
Workload index	0.001	0.0002	1.001	(1.0003,1.001)*
House hold income				
Low	Ref			
Middle	-0.007	0.007	0.993	(0.979,1.007)
High	-0.040	0.007	0.961	(0.947,0.974)*
BMI for age				
Normal	Ref			
Underweight	0.039	0.011	1.040	(1.017,1.063)*
Over-weight	-0.014	0.012	0.986	(0.962,1.010)
Height for age				
Non stunted	Ref			
Stunted	0.070	0.009	1.073	(1.053,1.093)*
MUAC	0.0003	0.0004	1.0003	(0.999,1.001)

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10, Ref=Reference group, SE= Standard Error, φ =Acceleration factor, CI=confidence interval*

Univariable loglogistic–Inverse gaussian frailty

Covariate	Coefficient	SE	φ	90% CI
Father's education level				
Secondary & higher	Ref.			
Primary	0.003	0.006	1.003	(0.990,1.016)
No-education	0.012	0.007	1.012	(0.997,1.027)
Mother's education level				
Secondary & higher	Ref.			
Primary	0.010	0.009	1.010	(0.993,1.028)
No-education	0.058	0.008	1.060	(1.043,1.077)*
Workload index	0.001	0.0002	1.001	(1.0003,1.001)*
House hold income				
Low	Ref			
Middle	-0.004	0.007	0.996	(0.983,1.008)
High	-0.037	0.007	0.964	(0.951,0.977)*
BMI for age				
Normal	Ref			
Underweight	0.033	0.011	1.033	(1.012,1.055)*
Over-weight	-0.013	0.012	0.987	(0.964,1.010)
Height for age				
Non stunted	Ref			
Stunted	0.071	0.009	1.073	(1.055,1.092)*
MUAC	0.001	0.0005	1.001	(0.999,1.002)

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.10, Ref=Reference group, SE= Standard Error, φ =Acceleration factor, CI=confidence interval*

Annex 2: Multivariable Analysis

Weibull - gamma frailty Model

Covariates	Coefficient	SE	φ	95% CI
Mother’s education level				
Secondary & higher	Ref			
Primary	0.003	0.008	1.003	(0.986, 1.021)
No-education	0.053	0.009	1.054	(1.037, 1.072)*
Work load index	0.001	0.0002	1.001	(1.0004,1.0011)*
House hold income				
Low				
Middle	-0.006	0.008	0.994	(0.979,1.009)
High	-0.017	0.007	0.983	(0.970, 0.996)*
BMI for age				
Normal	Ref			
Under weight	0.054	0.011	1.056	(1.032, 1.079)*
Over weight	- 0.015	0.012	0.985	(0.962, 1.008)
Height for age				
Non stunted	Ref			
Stunted	0.069	0.009	1.072	(1.052,1.092)*

Cons= 2.629(SE=0.012)* ρ=13.087 (SE=0.303)* log(ρ)=2.571(SE=0.023)

θ=0.083(SE=037,95%CL[0.034,0.199]) τ = 0.039

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.05, Ref=Reference group, SE= Standard Error, HR=Hazard Ratio*

Multivariable

Lognormal gamma frailty

Covariates	Coefficient	SE	φ	95% CI
Mother’s education level				
Secondary & higher	Ref			
Primary	0.012	0.008	1.012	(0.995,1.029)
No-education	0.055	0.008	1.056	(1.039,1.074)*
Work load index	0.0002	0.0002	1.0002	(0.999,1.001)
House hold income				
Low	Ref			
Middle	-0.007	0.006	0.993	(0.980, 1.005)
High	-0.023	0.007	0.977	(0.964,0.991)*
BMI for age				
Normal	Ref			
Under weight	0.038	0.010	1.039	(1.018, 1.059)*
Over weight	-0.015	0.011	0.985	(0.963, 1.007)
Height for age				
Non stunted	Ref			
Stunted	0.058	0.009	1.060	(1.042,1.079)*

Cons= 2.617 (SE=0.011)* σ =0.078 (SE=0.002)*

θ=0.058(SE=0.032,95%CL[0.019,0.172])* τ = 0.029

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = $p < 0.05$, Ref=Reference group, SE= Standard Error, φ=Acceleration factor*

Multivariable

Weibull - Inverse gaussian frailty

Covariates	Coefficient	SE	φ	95% CI
Mother's education level				
Secondary & higher	Ref			
Primary	0.003	0.009	1.003	(0.986, 1.021)
No-education	0.053	0.009	1.054	(1.036, 1.072)*
Work load index	0.001	0.0002	1.001	(1.0004,1.001)*
HH income cat				
Low	Ref			
Middle	-0.006	0.007	0.994	(0.979, 1.009)
High	-0.017	0.007	0.983	(0.969, 0.996)*
BMI for age				
Normal	Ref			
Under weight	0.054	0.011	1.055	(1.032, 1.079)*
Over weight	-0.015	0.012	0.984	(0.962, 1.008)
Height for age				
Non stunted	Ref			
Stunted	0.069	0.010	1.072	(1.052,1.093)*

Cons=2.629(SE=0.012)* ρ =13.081 (SE=0.302)* log(ρ)=2.571(SE=0.023)*

θ=0.089(SE=0.042,95%CL[0.036,0.224])

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.05, Ref=Reference group, SE=Standard Error, HR=Hazard Ratio*

Multivariable

Lognormal Inverse Gaussian frailty

Covariates	Coefficient	SE	φ	95% CI
Mother's education level				
Secondary & higher	Ref			
Primary	0.012	0.008	1.012	(0.996,1.029)
No-education	0.054	0.008	1.056	(1.039,1.074)*
Work load index	0.0002	0.0002	1.0002	(0.999,1.001)
House hold income				
Low	Ref			
Middle	-0.008	0.006	0.992	(0.979, 1.005)
High	-0.023	0.007	0.977	(0.964,0.991)*
BMI for age				
Normal	Ref			
Under weight	0.038	0.010	1.039	(1.018, 1.060)*
Over weight	-0.016	0.011	0.985	(0.963, 1.006)
Height for age				
Non stunted	Ref			
Stunted	0.058	0.009	1.060	(1.042,1.079)*

Cons= 2.617 (SE=0.011)* σ =0.078 (SE=0.002)*

θ=0.058(SE=0.032, 95% CL[0.019,0.169])*

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.05, Ref=Reference group, SE= Standard Error, φ=Acceleration factor*

Multivariable

Loglogistic Inverse gaussian frailty

Covariates	Coefficient	SE	φ	95% CI
Mother's education level				
Secondary & higher	Ref			
Primary	0.005	0.008	1.005	(0.988,1.022)
No-education	0.046	0.008	1.048	(1.030,1.065)*
Work load index	0.0002	0.0002	1.0002	(0.999,1.001)
House hold income				
Low	Ref			
Middle	-0.007	0.006	0.993	(0.981, 1.005)
High	-0.022	0.007	0.978	(0.966,0.992)*
BMI for age				
Normal	Ref			
Under weight	0.032	0.010	1.033	(1.013, 1.053)*
Over weight	-0.015	0.011	0.985	(0.963, 1.006)
Height for age				
Non stunted	Ref			
Stunted	0.060	0.008	1.062	(1.044,1.080)*
Cons= 2.631 (SE=0.011)*			γ=0.043 (SE=0.001)*	
θ=0.085(SE=0.045,95% CL[0.030,0.239])*				

*Source: Jimma Longitudinal Family Survey of Youth; Round 1-3, * = p < 0.05, Ref=Reference group, SE= Standard Error, φ=Acceleration factor*