ANALYSIS OF CLUSTERED WOMEN'S BIRTH INTERVALS IN MALAWI: APPLICATION OF PARAMETRIC AND SEMI-PARAMETRIC SURVIVAL MIXED-EFFECTS REGRESSION MODELS

MASTER OF SCIENCE (BIOSTATISTICS) THESIS

CHIKUMBUTSO WILLARD WASELA

UNIVERSITY OF MALAWI

AUGUST 2024

ANALYSIS OF CLUSTERED WOMEN'S BIRTH INTERVALS IN MALAWI: APPLICATION OF PARAMETRIC AND SEMI-PARAMETRIC SURVIVAL MIXED-EFFECTS REGRESSION MODELS

MSc.(BIOSTATISTICS) THESIS

By

CHIKUMBUTSO WILLARD WASELA

BSc (Statistics)- University of Malawi

Thesis submitted to the Department of Mathematical Science, School of Natural and Applied Science, in partial fulfillment of the requirements for the degree of Master of Science(Biostatistics)

UNIVERSITY OF MALAWI

AUGUST, 2024

Declaration

I the undersigned hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

CHIKUMBUTSO WILLARD WASELA

Full Legal Name

Signature

Date

Certificate of Approval

The undersigned certifies that this thesis represents the student's own work and effort and has been submitted with my approval.

Signature:_____ Date:_____

Tsirizani Kaombe, PhD(Senior Lecturer)

Supervisor

Dedication

Dedicated to my beloved mother, whose unwavering love, support, and sacrifices have been the guiding light throughout my academic journey. Your strength, encouragement, and belief in me have inspired me to pursue my dreams relentlessly. This thesis is a tribute to your boundless love and the countless sacrifices you've made for our family. Thank you for always being my pillar of strength.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Tsirizani Kaombe, for his invaluable guidance, support, and encouragement throughout the entire process of writing this thesis. His expertise, patience, and unwavering commitment to excellence have been instrumental in shaping the direction of my research and helping me navigate through various challenges.

I am also deeply grateful to my parents for their unconditional love, encouragement, and unwavering belief in my abilities. Their endless support and sacrifices have been the driving force behind my academic journey, and I am forever indebted to them for their unwavering support.

Lastly, I would like to extend my heartfelt appreciation to my friends for their constant encouragement, understanding, and motivation. Their companionship and camaraderie have provided me with the strength and resilience to overcome obstacles and pursue my academic goals with determination and perseverance.

Abstract

The short or long birth interval is a common public health issue that affects women's and children's health in sub-Saharan Africa. Despite a higher burden of short or long birth intervals, establishing causal relationships between birth intervals and their determinants remains challenging due to the potential for unobserved confounding factors. The objective of this research is to apply parametric and semiparametric survival mixed-effects regression, especially the shared frailty model to model the length of time to successive second births within the reproductive mothers and identify the sociodemographic and sociocultural factors that cause variation in the length of the birth interval using 2015-16 Malawi Demographic Health Survey data. Weibull mixed-effects survival model with gamma-distributed random effects provides a better fit for describing the birth interval data than Cox proportional hazard mixed-effects with normal and gamma random effects, exponential mixed-effects with normal and gamma random effect, and Weibull mixed effects with normal random effect. The findings also revealed that mothers from rural areas, the central region, and those who use traditional or modern contraceptive methods had a higher likelihood of having a second child. In contrast, women from middle or rich families, those with secondary or tertiary education, and those who had their first child alive had a lower likelihood of giving birth to a second child. This study uses gamma and normal densities for random effects distribution, suggesting future studies should consider other random effects distributions like positive stable, compound poisson, and inverse Gaussian.

Contents

List of Abbreviations

c.d.f.	cumulative density function
AIC	Akaike Informative Criterion
CI	Confidence Interval
E	expectation
EM	Expectation Maximization
MCMC	Markov Chain Monte Carlo
P(A)	probability of event A
BLUP	Best Linear Unbiased Predictor
LBW	low birth weight
SDGs	Nations Sustainable Development Goals
SGA	Small for Gestational Age
p.d.f.	probability density function

List of Symbols

f	probability density function
F	cumulative density function
Γ	Gamma function
$\mathcal{L}$	Laplace transform
h	hazard function
H	cumulative hazard function
L	likelihood function
S	survival function
V	variance
I	observed information matrix
$\mathcal{I}$	Fisher information matrix
$N(\mu, \sigma^2)$	normal distributed random variable with parameters μ, σ^2
$Exp(\lambda)$	exponential distributed random variable with parameter λ
$W(\lambda, \alpha)$	Weibull distributed random variable with parameters α, λ
$Gamma(\alpha, \lambda)$	Gamma distributed random variable with parameters α, λ

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Nowadays, the most concerning issue in world is overpopulation. Although rates of population increase are now much lower in the developed world than in the developing countries, the world's population has been growing rapidly in recent decades (Kopp et al., 2018; Gu et al., 2021; Seyedtabib et al., 2020). Much of this increase comes from high-fertility African countries (Kaba, 2020; Gu et al., 2021). Fertility rate is one of the key factors that determines population growth; it is the primary factor responsible for changing the size of the population (Gu et al., 2021). The decisions made by the couple, the length of the effective reproductive period, the length of the birth intervals, and socioeconomic, health-related, and emotional variables all influence fertility rate (Ranjbar et al., 2023). The birth interval is defined as the number of months between two successive live births (Pimentel et al., 2020; Aleye et al., 2024). Due to its implications for fertility rate, maternal, and child health, it has drawn particular attention in the fields of family planning and public health.

The World Health Organization (WHO) currently recommends an interval between the last live birth and the next pregnancy of at least 24 months, a birth interval of 33 months Bongaarts and Hodgson (2022). However, there are some inconsistencies in the literature

with cutoff points for optimal or ideal birth intervals. Tesema et al. (2021) suggest that the ideal time between successive live births is between 18 and 23 months while another study by Erfani et al. (2018) suggests that the ideal time is between 18 - 35 months to guarantee survival throughout childhood.

Numerous studies have provided empirical data indicating a correlation between the birth interval and the health of mothers, infants, and children Shifti et al. (2023); Tesema et al. (2023); Kozuki et al. (2013); Ball et al. (2014). Shifti et al. (2023) showed that both short and long intervals were linked to poor results such as preterm birth, low birth weight(LBW), and Small for Gestational Age(SGA) in high-income and low-income nations based on their meta-analysis in Asia-Pacific region. Short birth intervals were also linked to even more severe outcomes, like perinatal mortality, in low-income nations (Ball et al., 2014). Tesema et al. (2023) concluded that births to mothers with short birth intervals had a higher chance of dying in the first five years of life. (Kozuki et al., 2013) shows that birth intervals shorter than 18 months are significantly associated with preterm birth and death in the first year of life of the child.

Conde-Agudelo et al. (2007) conducted a second meta-analysis which revealed that there was a correlation between maternal health risks and both short and long birth intervals. Long birth intervals, whether intended or unintended, was associated with adverse maternal outcomes such as pre-eclampsia (Conde-Agudelo et al., 2007; Shifti et al., 2023). The long interval may also be correlated with negative outcomes if it is not a result of conscious family planning; for instance, mothers may be struggling with secondary infertility (Rutstein and Winter, 2014).

In Malawi, the births that happened less than 15 months following the previous birth have the greatest perinatal mortality rate (55 deaths per 1000 pregnancies) (Kopp et al., 2018). Contrarily, women who have regular birth intervals are better able to recuperate from the macro- and micronutrient depletion that happens throughout pregnancy and lactation (Ranjbar et al., 2023; Tesema et al., 2023). Thus, this enhances the health of children and future pregnancies. According to Rutstein and Winter (2014), an estimated 1.6 million deaths of children under five would be prevented each year if all birth-to-pregnancy intervals were separated by at least three years.

The majority of the literature that is relevant to understanding the factors influencing the duration of birth intervals can be categorized as research on the fertility transition, birth spacing, fertility intentions and preferences, and postpartum contraceptive use (Seyedtabib et al., 2020). Numerous demographers have written about the idea of "natural fertility" and the African fertility transition, attempting to illustrate how African countries differ or are similar to European ones in their move to lower birth rates (Muza and Mangombe, 2021; Bongaarts and Hodgson, 2022; Shapiro and Hinde, 2020). Other researchers have studied reproductive goals or preferences, using a model of planned, logical reasoning to space out births and/or reduce the size of the family (Greenhalgh, 2023; Blake, 2022).

Since short and long birth intervals are a potentially modifiable problem, better knowledge and understanding of its determinants is imperative and essential to improve maternal health by designing and applying specifically targeted interventions thereby decreasing catastrophic pregnancy outcomes. A review of determinants of birth interval could encompass a range of factors that influence the timing between successive pregnancies. Consequently,

in recent years, many studies have examined the interval between marriage and first birth, and inter-birth intervals (Seyedtabib et al., 2020; Bagheri and Saadati, 2021; Aleye et al., 2024). Most of the research has focused on first birth because of its advantages; women do not forget details of their first pregnancy, and the delay in the menstrual cycle that occurs after subsequent fertilization is not observed (Bagheri and Saadati, 2021). Furthermore, if the first birth interval is short and occurs at a young age, subsequent pregnancies may happen faster and the fertility rate will be increased (Aleye et al., 2024).

1.2 Analysis methods used for birth interval data

Until very recently, a vast array of researchers have demonstrated diverse techniques for analyzing birth interval data to understand the patterns, drivers, and implications of birth spacing. Hobcraft et al. (1983) conducted a comprehensive analysis of the impact of numerous birth intervals, including previous and following intervals. This study used counts of births in periods before and after the index kid, allowing for many short intervals. The analysis was based on retrospective birth histories in 26 World Fertility Surveys using multiple logistic regression. Additionally, the study was among the first to account for several confounding variables, including socioeconomic status, birth order, mother age, and past deaths. According to this study, children whose mothers had given birth to a previous child within the previous two to six years had the lowest relative risk of passing away at the age of under five. The relative risk rose as the number of births in the $0-2$ and $2-6$ year prior periods increased. When another child was born within 30 months, the same thing happened. The highest risk of mortality under the age of 5 years occurred in

children who had 3 prior siblings 2 to 6 years earlier, a prior sibling within the preceding 2 years, and a subsequent sibling in the 30 months following the birth of the index child.

However, these long-held results have been called into doubt in the last five years by several research on the impacts of birth spacing, comparing siblings who are discordant on birth interval length (Klebanoff, 2017). The reasoning behind this sibling comparison approach is that the effects of interval length itself can be isolated, net of risk factors shared among siblings that may be correlated with the length of birth intervals, by controlling for otherwise unobserved shared factors within the family (Kalbfleisch and Schaubel, 2023). Ball et al. (2014), using data from Australia, is the first known study to apply a sibling fixed effects analysis to the longitudinally linked database. The Women under study were singleton pregnant women with two birth intervals (the first between the birth and the conception of the second child, and the second between the birth and the conception of the third). Ball et al. (2014) discovered that after applying sibling fixed effects, the association between short birth interval (defined as 0–5 months) and the risk of preterm birth, LBW, and SGA was almost completely eliminated. A study utilizing Swedish data similarly discovered that, when utilizing a fixed effects analysis, short birth intervals again, defined as 0-5 months-were no longer linked to the risk of LBW or SGA, but they did raise the likelihood of premature birth (Class et al., 2017).

Extensive research has sought to understand the various factors that influence birth interval using various statistical methods (Hailu et al., 2016; Aleye et al., 2024; Aychiluhm et al., 2020; Ahammed et al., 2019). Logistic regression has been used in several studies to examine birth interval determinants (Hailu et al., 2016; Aleye et al., 2024; Aychiluhm

et al., 2020). (Aleye et al., 2024) employed logistic regression in a community-based case-control study in assessing determinants of short intervals among 440 married in Chinaksen district Ethiopia. The results of the study demonstrated that the number of live births, parental education, women's work, age at first marriage, and use of contraceptives all had a substantial impact on birth intervals. Another case-control study conducted from February to April 2014 by (Hailu et al., 2016) showed that several factors, such as a woman's calendar period, marriage age, contraceptive method, educational level, employment, place of residence, and household income influenced women's first, second and third birth intervals in Ethiopia.

A community-based cross-sectional study designed by Gelagay et al. (2023) uses also logistic regression to determine determinants of the birth interval among 1262 women of childbearing age in Dabat district, Ethiopia. This study shows that women who were rural residents, students, and daily workers in husbands' occupations, breastfeeding for 7 to 12, 13 to 23, and 24 and above months, nonuse of modern contraceptives, and highest wealth quartile were found to be significant predictors of short birth interval length. Aychiluhm et al. (2020) used information from the Demographic and Health Survey program to evaluate the short birth interval and its causes in Ethiopia's four developing regions. The study included 2683 women in the 15–49 age range who had at least two live children. The results of the multilevel multivariable logistic regression model indicated that the following factors were associated with a shorter birth interval: breastfeeding duration, having six or more children, living in a rural area, attending secondary education or above, not being exposed to the media, and preferring to wait two years or longer before giving birth.

Exavery et al. (2012) applied the Chi-Square test of association and multilevel logistic model to longitudinal data collected in the Rufiji Health and Demographic Surveillance System (HDSS) from January 1999 to December 2010 to investigate levels of birth spacing in Tanzania and factors affecting non-adherence to the WHO recommended minimum inter-birth interval length among women aged 15-49 years in Rufiji district. Non-adherence was considered as an inter-birth interval length below 33 months. Thus, a binary outcome variable was created and logistic regression was then used to construct a multivariate model and find correlates of non-adherence.

Nonetheless, some research has looked at identifying the birth interval determinants using multiple linear regression or Cox proportion regression (Ahammed et al., 2019; Bagheri and Saadati, 2021; Khan et al., 2016). A cross-sectional study carried out in Al-Oyaynah, Saudi Arabia (Hajian-Tilaki et al., 2009) is one of the studies that takes multiple linear regression into account in assessing the effect of social-demographics. The continuous and normally distributed birth interval was the study's underlying presumption. Maternal age, duration of breast feeding, sex of index child, history of still births, history of infant mortality of the index child were found to be important predictors of the birth interval in this study.

The Cox proportional hazard model was used by Ahammed et al. (2019) in assessing key factors affecting birth interval length in Bangladesh. In this study, 504 eligible women (women who had at least two children) were selected. The results indicated that work status, contraceptive use, and survival status of the index child were consistently significant while age at first marriage and place of residence were significant in some birth intervals.

According to a study by Fallahzadeh et al. (2013), the birth interval is predicted by the age of marriage, and a woman's education. The study aimed to assess the duration and determinants of inter-birth intervals among 400 women of reproductive age in Yazd, Iran.

A cross-sectional fertility survey in Tehran conducted by Bagheri and Saadati (2021) is the one of studies that accounts for correlations between birth intervals in the Cox model. This study employed 610 married women, aged 15-49 years, and analyzed three births using Survival Recurrent Event(SRE). According to Bagheri and Saadati (2021) study, the birth interval can be predicted by a mother's age, employment level, and exposure to mass media. Another study by Khan et al. (2016) also accounts for correlation in the Cox model in determining trends of the determinant of birth interval in Bangladesh using Bangladesh Demographic and Health Surveys (BDHS 2004, 2007, and 2011). However, data analysis was based on the parametric gamma frailty model. This study shows that maternal age, educational level of the mother, family composition and wealth index, unobserved community, and mother effects have a considerable impact on the birth interval in Bangladesh.

1.3 Problem statement

Multiple regression and logistic regression have been used to analyse birth interval data, however, establishing causal relationships between birth intervals and their determinations remains challenging due to the potential for unobserved confounding factors. Birth interval data often exhibit complex relationships with its predictors, including non-monotonic effects and time-varying associations. Moreover, birth interval data may be censored, meaning

that the observation period ends before all women have experienced the event of interest (subsequent birth). This can introduce bias and complicate the analysis, especially if the censoring mechanism is non-random. Statistical methods such as logistic regression and multiple regression struggle to capture these dynamics adequately, highlighting the need for flexible modeling techniques to capture complex relationships over time.

1.4 Study Objectives

The main objective is to apply parametric and semiparametric survival mixed effects regressions(shared frailty models) to model the length of time to successive second births within the reproductive mothers in Malawi.

The specific objectives are to:

1. To estimate the impact of natural clustering neighbourboods of women in the analysis of determinants of mothers' birth intervals in Malawi using mixed survival models

2. To compare the efficiency of normal and gamma probability distributions when used to describe the random effects in the mixed survival regression models

3. To compare the performance of Cox, Weibull, and Exponential mixed-effects survival models in the analysis of women's birth interval data

4. To analyse the fit of the mixed-effects survival regression method to the birth interval data using some selected post-estimation statistics.

1.5 Significance of this Study

This study will contribute to the discussion on how the socio-demographic and socio-cultural factors of a woman affect the timing of subsequent second births. The shared frailty model technique will be applied, which considers unobserved heterogeneity in testing the stability of the estimated coefficients by comparing the models with or without unobserved heterogeneity. This study will provide new knowledge on birth interval and reproductive health to help achieve set targets in the United Nations Sustainable Development Goals (SDGs) 2030, that is, to end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 deaths per 1,000 live births and under-5 mortality to at least as low as 25 deaths per 1,000 live births. Also, it will provide information for planning, monitoring, and evaluation of programs on reproductive health in Malawi.

1.6 Thesis structure

The thesis is structured as follows: Chapter 2 gives literature review of semiparametric and parametric survival mixed effects, particularly the shared frailty model and its estimation methods. Chapter 3 presents the methodology of the study. Chapter 4, results of the shared frailty models. Chapter 5 provides discussion and conclusion.

Chapter 2

Review of semi-parametric and parametric survival mixed-effects models

This chapter discusses some of the standard failure time models for clustered populations. It describes and explains some basic and general concepts in survival analysis to lay the foundation of semiparametric and parametric survival mixed effects, particularly the frailty model. It also reviews the estimation methods used in modeling the frailty model. The properties and theoretical bases of survival mixed effects distributions are considered here only briefly.

2.1 Time-to-event concept and related probability functions

A key characteristic of survival data is that the response variable is a non-negative discrete or continuous random variable and represents the time from a well-defined origin to a well-defined event (Hosmer et al., 2008; Collett, 2023; Cox, 2018). Another characteristic feature of these data sets is censoring observations (Moore, 2016). Censored data arises when the starting or ending events are not precisely observed. Possible censoring schemes

are right censoring, where the final endpoint is only known to exceed a particular value; left censoring, where events are known to have occurred before a certain time; or interval censoring, where the only information is that the event occurs within some interval (Klein and Moeschberger, 2003; Duchateau and Janssen, 2008).

The distribution of survival time is characterized by three main functions, namely the probability density function, survival function, and hazard function (Hosmer et al., 2008). Let f be the probability density function (p.d.f) of the event time T. The distribution function of T is then given by

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{2.1}$$

and represents the probability that the survival time is less than some value t (Collett, 2023). The survival function, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t, and so from Equation 2.1,

$$S(t) = P(T > t) = 1 - F(t). \tag{2.2}$$

The hazard function $h(t)$ is obtained from the conditional probability that an event occurs in the interval $t, t+\Delta t$ given that the event did not occur yet before time t. This conditional probability is then expressed as a probability per unit time by dividing by the time interval, Δ, to give a rate. The hazard function is then the limiting value of this ratio, as Δt tends to

zero, so that (Hosmer et al., 2008; Moore, 2016; Alvares et al., 2021)

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.3}$$

The function $h(t)$ is also referred to as the hazard rate, the instantaneous death rate, the intensity rate, or the force of mortality. The hazard function is related to the PDF and survival functions by

$$h(t) = \frac{f(t)}{S(t)} \tag{2.4}$$

That is, the hazard at time t is the probability that an event occurs in the neighborhood of time t divided by the probability that survival time is greater than or equal to t. The cumulative hazard function is defined as the area under the hazard function up to time t, that is,

$$H(t) = \int_0^t h(u)du \tag{2.5}$$

The survival function can be rewritten in terms of the cumulative hazard function. First, using

$$h(t) = -\frac{d}{dt} log(S(t)) \tag{2.6}$$

and after integrating and exponentiating, survival function becomes (Collett, 2023)

$$S(t) = exp\left(-\int_0^t h(u)du\right) = exp(-H(t)) \tag{2.7}$$

The form of likelihood expression is determined by the type of data that is available. If

there is no censoring, the likelihood function takes the general form

$$L(t_i) = f(t_1) \times f(t_2) \times ... \times f(t_n) = \prod_{i=1}^{n} f(t_i) \tag{2.8}$$

If some observations are censored, the expression 2.8 is adjusted (Alvares et al., 2021; Bagheri and Saadati, 2021). But for a right-censored observation, the survival function is inserted, indicating that the observation is known only to exceed a particular value. The likelihood in general then takes the form (Hosmer et al., 2008)

$$L(t_i) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \tag{2.9}$$

where δ_i is the censoring indicator. This expression means that when t_i is an observed event, the censoring indicator is $\delta_i = 1$, and then likelihood is a probability density function. When t_i is a censored observation, the censoring indicator is $\delta_i = 0$ then likelihood is a survival function.

2.2 Cox PH mixed-effects survival model

The Cox proportional hazards (PH) model (Cox, 1972) is a popular approach for analyzing survival data, especially when assessing multiple predictors' effect on the hazard rate. The model is given as (Cox, 1992; Collett, 2023)

$$h(t|x) = h_0(t) exp(x^{'} \beta) \tag{2.10}$$

where $h_0(t)$ is an unspecified baseline hazard rate, β is a parameter vector, and x is a covariate vector.

Mixed-effects models, also known as random-effects or hierarchical models, are used when data is hierarchical or clustered structure, where observations are not independent but rather nested within groups (Brown, 2021; Konstantopoulos and Hedges, 2019). Combining these two approaches results in the Cox PH mixed-effects survival model, which allows for the incorporation of both fixed and random effects into the survival analysis (Duchateau and Janssen, 2008; Wienke, 2003; Hosmer et al., 2008). In practice, fitting a Cox PH mixed-effects model involves specifying the fixed effects (covariates) that influence the hazard rate, as well as specifying the random effects structure, which captures the clustering or grouping in the data (Bagheri and Saadati, 2021; Khan et al., 2016; Klein and Moeschberger, 2003). The model is given as (Therneau et al., 2015)

$$h(t|x) = h_0(t)exp(x_{ij}^{'}\beta + b_i) \tag{2.11}$$

where the b_i are drawn from a statistical distribution. Like the standard Cox PH model, the mixed-effects version assumes that censoring is independent of the survival times after adjusting for covariates Klein and Moeschberger (2003). The hazard function can be rewritten as (Collett, 2023)

$$h(t|x) = h_0(t)w_i exp(x_{ij}^{'}\beta) \tag{2.12}$$

where $w_i = exp(b_i)$. This shows that hazard also depends on an unobservant random

variable, w_i, which acts multiplicatively on the hazard rate (Klein et al., 2014).

Cox PH mixed-effects survival model is particularly useful when there are recurrent events or clustering within the data that need to be accounted for Klein and Moeschberger (2003); Duchateau and Janssen (2008). The survival recurrent event model is used when the event of interest occurs more than one time for each subject Klein et al. (2014). Various survival recurrent events have been developed and suggested in the literature (Therneau et al., 2000; Kalbfleisch and Prentice, 2011). Therneau et al. (2000) presents a more mathematical and computationally detailed discussion of recurrent event data, with examples. Su et al. (2022) discusses recurrent events and compares them to generalized linear models, such as Poisson and logistic regression. Kalbfleisch and Prentice (2011) treat it at a fairly theoretical level using the counting process approach.

Another extension of the standard modeling situation occurs when subgroups of responses are correlated due to study design (Hosmer et al., 2008; Hougaard, 1995). This lack of independence of response can occur in any setting in which survival time is influenced by unmeasured factors that are the same within groups of subjects and are thought to have significant group-to-group variability (Wienke, 2003; Klein and Moeschberger, 2003). When these factors are present in the usual normal errors linear model setting, they are called random effects (Brown, 2021; Konstantopoulos and Hedges, 2019). Survival analysis models incorporating such factors are called shared frailty models (Hosmer et al., 2008; Collett, 2023; Duchateau and Janssen, 2008).

2.2.1 Common distributions used for random effects

Various distributions for random effects have been suggested in the literature and any distribution with a positive random variable can be used to model frailty (Hosmer et al., 2008; Hougaard, 1986b; Vaupel et al., 1979; McGilchrist and Aisbett, 1991; Clayton, 1978). The frailty distributions most often applied are the gamma distribution (Clayton, 1978; Vaupel et al., 1979; Oakes, 1989; Hougaard and Hougaard, 2000; Wienke, 2003), the positive stable distribution (Hougaard, 1986a), the power Variance function (PVF) distribution (Hougaard, 1986b), the inverse Gaussian distribution (Hougaard, 1984), the compound Poisson distribution (Aalen, 1988) and the log-normal distribution (McGilchrist and Aisbett, 1991).

2.2.2 Estimation of regression parameters in Cox PH mixed model

The frailty model estimations that have been suggested in the literature are mainly maximum likelihood estimates, although some of them are based on various approximations (Duchateau and Janssen, 2008; Wienke, 2010). (Duchateau and Janssen, 2008) presented a review of both a semi-parametric and a parametric approach to estimating the risk coefficients and the frailty parameters for the gamma frailty, the positive stable, and the lognormal. To derive the general form of the likelihood function, it is assumed that the common factor causes dependence between individuals in a given group, and conditional on that, all individuals within the group are independent (Duchateau and Janssen, 2008). Thus, for one group of n individuals, the conditional joint survival distribution of failure times $T_1, T_2, ..T_n$ is given

by

$$\begin{aligned} P(T_1 > t_1, ..., T_n > t_n|w) &= P(T_1 > t_1|w)...P(T_n > t_n|w) \\ &= exp\left\{-wH_0(t)exp(x^{'}\beta)\right\} \end{aligned} \tag{2.13}$$

From 2.13, the likelihood function for one group can be derived as follows:

$$\begin{aligned} L_j(\beta, \theta, w) &= P(T_j = t_j, T_1 > t_1, T_2 > t_2....) \\ &= \prod_{j=1}^{n} \left(h_0(t)we^{x^{'}\beta}\right)^{\delta_j} exp\left(-H_0(t)we^{x^{'}\beta}\right) \end{aligned} \tag{2.14}$$

If there is an assumption that $h_0(t)$ has a parametric form, the estimation can be handled in the usual way by differentiating the log-likelihood function (Klein et al., 2014; Collett, 2023). In Cox PHMM, however, the more classical approach is to assume that the hazard function is the product of an unspecified baseline hazard $h_0(t)$ and a parametric function of the covariates (Klein et al., 2014). In terms of survival frailty, such models are called semiparametric frailty models (Hosmer et al., 2008; Kelly and Lim, 2000). For such a model direct maximization of the marginal likelihood is no longer possible (Klein et al., 2014; Duchateau and Janssen, 2008). Semiparametric hazard models without frailty terms are fitted by maximization of the partial likelihood (Cox, 1972). For semiparametric frailty models, however, there is a need to account for the contribution of the unobserved frailty terms (Duchateau and Janssen, 2008; Therneau et al., 2000). The solution is to use the partial likelihood function or full likelihood function in combination with the Expectation-Maximisation (EM) algorithm (Collett, 2023; Duchateau and Janssen, 2008; Hosmer et al.,

2008; Therneau et al., 2015; Klein et al., 2014; Wienke, 2010).

The EM algorithm is used for estimation since the frailty can be as covariates (Klein and Moeschberger, 2003). EM algorithm alternates between finding expected values for the frailties based on current estimates of β and θ using these expected values to find updated estimates for β and θ. The EM algorithm iterates between an expectation and a maximization step. In the expectation step, the expected values of the unobserved frailties conditional on the observed information and the current parameter estimates are obtained. In the maximisation step, these expected values are considered to be the true information, and new estimates of the parameters of interest are obtained by maximisation of the full likelihood, given the expected values. The collection of frailty terms is the unobserved information Duchateau and Janssen (2008). The full likelihood is defined as the product of the conditional and the density of frailties.

$$L_{ij}(\theta, \beta, w_i) = \prod_{i=1}^{g} \prod_{j=1}^{n_i} \left(h_0(t) w_i e^{x_{ij}^{'} \beta} \right)^{\delta_{ij}} exp \left(-H_0(t) w_i e^{x_{ij}^{'} \beta} \right) f(w; \theta) \tag{2.15}$$

2.2.2.1 Expectation and maximisation for the gamma frailty model

The EM algorithm estimates the regression parameters for each of a fixed set of values of the variance parameter, θ, of the gamma frailty distribution (Klein and Moeschberger, 2003; Lee and Wang, 2003; Collett, 2023). The solution is the one yielding the largest value of a profile log-likelihood, whose equation is shown below in 2.16. The equation for

the profile log-likelihood (Klein and Moeschberger, 2003), equation is

$$
\begin{aligned}
l\left(\theta, \hat{\beta}\right) = & \sum_{i=1}^{g} d_i log(\theta) - log\left[\Gamma\left(\frac{1}{\theta}\right)\right] + log\left[\Gamma\left(\frac{1}{\theta} + d_i\right)\right] \\
& + \sum_{j=1}^{n_i} \delta_{ij}\left\{x_{ij}^{'}\hat{\beta} + log\left[h_{f0}(t_i j)\right]\right\} \\
& - \left(\frac{1}{\theta} + d_i\right) log\left[1 + \theta \sum_{j=1}^{n_i} log\left[1 + \theta \hat{H}_f\left(t_{ij}, x_{ij}, \hat{\beta}\right)\right]\right]
\end{aligned} \tag{2.16}
$$

where in i cluster there are d_i failure event-times, each denoted by δ_{ij}, so that $d_i = \sum_{j=1}^{n_i} \delta_{ij}$.

The specific steps are as follows Duchateau and Janssen (2008); Kelly and Lim (2000); Collett (2023); Hosmer et al. (2008):

Step 1: The proportional hazards model (without frailty) is fitted containing the covariates of interest. Following the fit, the estimate of the baseline cumulative hazard function is obtained for each subject, $\hat{H}_0(t_i)$. This is used to obtain the estimate of the cumulative hazard for each subject,

$$
\hat{H}(t_i, x_i, \hat{\beta}) = \hat{H}_0(t_i) exp\left(x_i^{'}\beta\right) \tag{2.17}
$$

Step 2: A set of possible values for the variance parameter θ is created. Typically, this set is constructed by beginning with a small value, for example, 0.25, and increasing by 0.25 until reaching some maximum value, for example, 4 or 5. For each of the values of $theta$ in this set, steps 3, 4, and 5 are followed.

Step 3: The estimation step (E) consists of computing, for each subject, an estimate of the value of their frailty as

$$
\hat{w_i} = \frac{1 + \theta \times \delta_i}{1 + \theta \times \hat{H}(t, x_i, \hat{\beta})}. \tag{2.18}
$$

Step 4: The maximization step (M) consists of fitting the proportional hazards model with the same covariates, but including $\hat{w}_i$ in the hazard function, as follows

$$\begin{aligned} h_f(t, \hat{w}_i, x_{ij}, \beta) &= h_0(t)\hat{w}_i exp(x_{ij}^{'}\beta) \\ &= h_0(t) exp\left[x_{ij}^T\beta + log(\hat{w}_i)\right] \end{aligned} \tag{2.19}$$

Following the fit of the model in 2.19, the baseline hazard that includes the estimated frailty term is estimated by

$$\hat{h}_{f0}(t_i) = \frac{\delta_j}{\sum_{l \in R(t_{il})} \hat{w}_i exp(x_{il}^{'}\hat{\beta})} \tag{2.20}$$

the cumulative baseline hazard function

$$\hat{H}_{f0}(t_i) = \sum_{t_j \leq t_i} \hat{h}_{f0}(t) \tag{2.21}$$

and the cumulative hazard containing the frailty, namely

$$\hat{H}_{f0}(t, x_i, \hat{\beta}) = \hat{H}_{f0}(t_i) exp(x_i^{'}\hat{\beta}) \tag{2.22}$$

The E- and M-steps are repeated until convergence is achieved.

Step 5: Evaluate the log-likelihood in 2.25 using the specified value of θ and the results from the fit in the M-step at convergence. Steps 1 to 5 are performed for all chosen values of θ. The maximum likelihood estimate of θ is the value that maximizes equation 2.25. Significance tests for model coefficients may be performed either using Wald tests or a profile likelihood ratio test (Hosmer et al., 2008; Klein and Moeschberger, 2003). The Wald test uses estimated standard errors from a covariance matrix obtained from the negative of

the inverse of the matrix of second derivatives of the profile log-likelihood (Duchateau and Janssen, 2008; Klein and Moeschberger, 2003). Let **H** denote the Hessian matrix of the second partial derivatives of the marginal loglikelihood function $l(\theta, \beta)$. The negative of the expected value of the Hessian matrix is known as the Fisher information. Let $\mathcal{I}$ denote the Fisher information matrix given as Collett (2023); Duchateau and Janssen (2008)

$$\mathcal{I}(\theta, \beta) = -E(\mathbf{H}(\theta, \beta)) :$$

The observed information matrix, I, is then the negative of the Hessian matrix.

$$I(\theta, \beta) = -\left[\mathbf{H}(\theta, \beta)\right]$$

Therefore, the asymptotic variance-covariance matrix of the estimates is obtained by taking the inverse of the Fisher information matrix, and the estimated variance-covariance matrix is obtained by taking the inverse of the observed information matrix by evaluating it at the actual values of the maximum likelihood estimators (Collett, 2023).

2.2.2.2 The penalized partial likelihood for the normal random effects density

Arguably another popular way of fitting semiparametric shared frailty models is via the penalised partial likelihood method (Therneau et al., 2000; Collett, 2023; Klein and Moeschberger, 2003; Duchateau and Janssen, 2008). The penalised partial likelihood approach leads to the same estimates as the EM algorithm in the case of the semiparametric gamma frailty model (Duchateau and Janssen, 2008). The penalised partial likelihood approach is commonly

used in estimating log-normal frailty model (Therneau et al., 2015). The conditional partial likelihood for log-normal frailty by cluster i is given by (Duchateau and Janssen, 2008):

$$L_i(\beta_i, b_i|t_i, X_{ij}) = \prod_{i=1}^{d_i} \left[\frac{exp(X_{il}^T\beta + b_i)}{\sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i)} \right] \tag{2.23}$$

where in i cluster there are d_i failure event-times, each denoted by δ_{ij}, so that $d_i = \sum_{j=1}^{n_i} \delta_{ij}$. Using ideas developed in the original Cox model (Cox, 1992), let $t_{i1} < t_{i2} < ... < t_{ir_i}$ be the r_i ordered observed event-times in cluster i, with corresponding covariates $X_{i1}, X_{i2}, ..., X_{ir_i}$. Then $R(t_{il})$, $l = 1, 2, ..., r_i$ is the set of individuals who are at risk of failure at time t_{ij}, called the risk set.

The full joint partial likelihood of the data and random effects is the product of the conditional likelihood Equation 2.23 over all clusters and the product likelihood function of density functions of random effects, $f(b)$ where $b_i = (b_1, b_2, ..., b_G)$. This is given by:

$$L(\beta, b|t, X) = \prod_{i=1}^{G}\prod_{j=1}^{n_i} \left[\frac{exp(X_{il}^T\beta + b_i)}{\sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i)} \right] \times (2\pi\sigma^2)^{\frac{-G}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{G} b_i \right) \tag{2.24}$$

The full joint partial log-likelihood function is:

$$\begin{aligned} l(\beta, b|t, X) = &\sum_{i=1}^{G}\sum_{j=1}^{n_i} \left[(X_{il}^T\beta + b_i) - ln \sum_{s\in R(t_{il})} exp(X_{is}^T\beta + b_i) \right] \\ &+ ln\left[(2\pi\sigma^2)^{\frac{-G}{2}} \right] + \left[-\frac{1}{2\sigma^2}\sum_{i=1}^{G} b_i \right] \end{aligned} \tag{2.25}$$

The score functions for β and b follow from the log-likelihood Equation 2.25 and are,

respectively, given by:

$$U_\beta = \frac{\partial l(\beta, b)}{\partial \beta} = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \left[X_{ij} - \frac{\sum_{s \epsilon R(t_{il})} X_{ij} exp(X_{il}^T \beta + b_i)}{\sum_{s \epsilon R(t_{il})} exp(X_{is}^T \beta + b_i)} \right] \tag{2.26}$$

and,

$$U_b = \frac{\partial l(\beta, b)}{\partial b} = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \left[1 - \frac{\sum_{s \epsilon R(t_{il})} X_{ij} exp(Z_{il}^T \beta + b_i)}{\sum_{s \epsilon R(t_{il})} exp(X_{is}^T \beta + b_i)} \right] - \frac{1}{2\sigma^2} \sum_{i=1}^{G} b_i. \tag{2.27}$$

The maximization process proceeds iteratively by starting with a provisional estimate of σ^2 and finding the estimates of the βs and the b_is that maximize $l(\beta, b|t, X)$. Next, a marginal log-likelihood, log $L_m(\beta, \sigma_b^2)$, is obtained by integrating the log-likelihood in Expression 2.25 over the random effects b_i. Ripatti et al. (2002) show that, at estimates $\hat{\beta}$ of β, this is well-approximated by

$$\log L_m(\sigma_b^2) = \log L_{pen}(\hat{\beta}, \hat{b}, \sigma_b^2) - \frac{1}{2} \log(\sigma_b^{2n}) - \frac{1}{2} \log |I_b|, \tag{2.28}$$

where I_b is the $n \times n$ observed information matrix for the random effects, formed from the negative second partial derivatives of log $L_{pen}(\beta, b, \sigma_b^2)$ with respect to the $b_i, i = 1, 2, ..., n$, evaluated at $\hat{\beta}$, $\hat{b}$, and $|I_b|$ is the determinant of that matrix. Using the estimates of the βs, the marginal log-likelihood in Equation 2.28 is then maximized with respect to σ_b^2 to give a revised estimate of σ_b^2. This process is repeated until the difference between two successive estimates of σ_b^2 is sufficiently small.

Estimates of the random effects, $\hat{b}_i$, and hence the frailty terms $\hat{w}_i = exp(\hat{b}_i)$, are also

obtained from this iterative process. However, estimates is based on the method of restricted maximum likelihood (REML) estimation (Collett, 2023; Duchateau and Janssen, 2008). REML estimates are obtained from a likelihood function that is independent of β, and the REML estimate of the variance of the random frailty term is

$$\tilde{\sigma}_b^2 = n^{-2}\left\{\sum_{i=1}^{n}\tilde{b}_i^2 + \text{trace}(\tilde{V}_b)\right\} \tag{2.29}$$

where $\tilde{V}_b$ is the estimated variance-covariance matrix of the REML estimates of the b_i, $\tilde{b}_i$. The trace of this matrix is just the sum of the estimated variances of the $\tilde{b}_i$. This is the preferred estimate of the variance of a normally distributed random frailty effect. The standard error of $\tilde{\sigma}_b^2$ can be found from

$$se(\tilde{\sigma}_b^2) = \sqrt{(2\tilde{\sigma}_b^2)}\left[n + \frac{1}{\tilde{\sigma}_b^4}\text{trace}(\tilde{V}_b\tilde{V}_b) - 2\frac{1}{\tilde{\sigma}_b^2}\text{trace}(\tilde{V}_b)\right]^{-1/2} \tag{2.30}$$

2.2.3 Residual analysis in Cox PHMM

In Cox proportional hazards mixed-effects models, residual analysis involves examining the discrepancies between the observed data and the model's predictions (Moore, 2016; Collett, 2023). This analysis is crucial for assessing the adequacy of the model fit, identifying influential observations, and detecting potential violations of model assumptions (Collett, 2023). The use of residuals for model checking has been well-developed in linear regression theory (James et al., 2023). The residuals are plotted versus some quantity, such as a covariate value, and the observed pattern is used to diagnose possible problems with the

fitted model. Some residuals have the additional property of not only indicating problems but also suggesting remedies (Moore, 2016; Kuitunen et al., 2021). That is the pattern of the plotted residuals may suggest an alternative model that fits the data better Collett (2023). Many of these residuals have been generalized to survival mixed effects analysis. In addition, the fact that survival data evolves over time, and requires special assumptions such as proportional hazards, makes it necessary to develop additional diagnostic residual methods (Kuitunen et al., 2021).

2.2.3.1 Martingale Residual

Martingale residuals are defined as the difference between the observed event indicator (0 or 1) and the corresponding predicted hazard at each event time (Collett, 2023). If there are no time-dependent covariates and if the survival times are right-censored, this is given by (Therneau et al., 2000; He et al., 2020)

$$m_{ij} = \delta_{ij} - \hat{H}_0(t_{ij}) exp(x_{ij}^{'} \hat{\beta}) \tag{2.31}$$

Martingale residuals are useful for detecting outliers and influential observations. The asymmetry of the residuals might appear to be a disadvantage, but Therneau et al. (2000) show that a plot of these residuals versus the covariate x_{ij} can reveal not only discrepancies in the model but also the actual functional form of this covariate.

2.2.3.2 Deviance Residuals

Deviance residuals are calculated based on the logarithm of the likelihood ratio, comparing the likelihood of the observed data under the fitted model to the likelihood of the saturated model (Collett, 2023; Jiang and Guterman, 2024). It is defined in terms of the martingale residual as follows Collett (2023); He et al. (2020)

$$D_{ij} = \text{sign}(m_{ij}) \times \{-2 \times [m_{ij} + \delta_{ij} log(\delta_{ij} - m_{ij})]\}^{1/2} \tag{2.32}$$

Deviance residuals are symmetrically distributed with an expected value of 0 (if the fitted model is correct). The sum of squares of these residuals is the value of the likelihood ratio test, which is analogous to the deviance from generalized linear model theory. While the properties of deviance residuals might seem preferable to those of martingale residuals, only the latter have the property that shows the functional form of a covariate (Collett, 2023).

2.2.3.3 Schoenfeld Residuals

Schoenfeld residuals are based on the covariate values at the event times and the differences between observed and expected covariate values (Cheung et al., 2021). This residual differs from those considered previously in one important respect. This is that there is not a single value of the residual for each individual, but a set of values, one for each explanatory variable included in the fitted Cox regression model(Collett, 2023; He et al., 2020). The

ij-th Schoenfeld residual for X_{ij}, the p-th explanatory variable in the model is given by

$$r_{S_{ijp}} = \delta_{ij}\{x_{ij} - \frac{\sum x_{ijp} exp(x_{ij}^T \hat{\beta})}{\sum_{s \in R(il)} exp(x_{is}^T \hat{\beta})}$$

where $R(t_{ij})$ is the set of all individuals at risk at time t_{ij}. These residuals are used to assess the proportional hazards assumption, as they should not exhibit any systematic patterns over time if the assumption holds (Moore, 2016; Collett, 2023).

2.2.3.4 Score Residuals

Score residuals are calculated as the difference between the partial derivative of the log-likelihood function with respect to the fixed effects parameters and their expected values under the fitted model (Kaombe and Manda, 2023). It is given by:

$$v_{ij} = \int_o^\infty \left[X_{ijp}(t) - \bar{X}_p(\hat{\beta}, t) \right] dm(t_{ij})$$

where $m(t_{ij})$ is residual of ij-th unit at time t_{ij}, also called martingale residual, which measures the excess number of events; and p denotes the number of covariates; while $(\bar{X})_p = \frac{\sum x_{ijp} exp(x_{ij}^T \hat{\beta})}{\sum_{s \in R(il)} exp(x_{is}^T \hat{\beta})}$ is the weighted average of covariate risk sets. Score residuals can be used to identify potential outliers and influential observations (Collett, 2023)

2.3 Parametric mixed-effects survival models

The previous sections focused on the use of semiparametric models for the analysis of censored survival time data. The rationale for using these techniques, in particular the

semiparametric proportional hazards regression model, was to avoid having to specify the hazard function completely. The utility of the proportional hazards model stems from the fact that a reduced set of assumptions is needed to provide hazard ratios that are easily interpreted and clinically meaningful. However, there may be settings in which the distribution of survival time, through previous research, has a known parametric form that justifies the use of a fully parametric model to address the goals of the analysis better (Duchateau and Janssen, 2008; Wienke, 2010). There are several texts that discuss parametric survival time models (e.g.(Bagheri and Saadati, 2021; Jiang and Guterman, 2024; Emmert-Streib and Dehmer, 2019)). This section shall briefly discuss parametric distributions for baseline hazards that are relevant to the subject of mixed effects in survival analysis.

2.3.1 Exponential mixed-effects survival models

Considering an jth individual in ith group for whom the values of the p explanatory variables in the model of equation 2.12 are all equal to zero. The hazard function for such an individual is $w_i h_0(t)$. If the survival time of this individual has an exponential distribution with parameter λ, then their hazard function is such that

$$h_0(t) = \lambda \tag{2.33}$$

Considering covariates, hazard rate $h_{ij}(t|X_{ij})$ given by Jiang and Guterman (2024); Klein et al. (2014); Moore (2016)

$$h_{ij}(t|x_{ij}) = w_i \lambda exp(x_{ij}^{'}\beta) \tag{2.34}$$

The cumulative or integrated hazard rate is

$$H_0(t) = \lambda t \tag{2.35}$$

The corresponding survivor function is

$$\begin{aligned} S_{ij}(t) &= exp\left(-w_i \int_0^t \lambda du\right) \\ &= e^{-w_i \lambda t} \end{aligned} \tag{2.36}$$

and so the implied probability density function of the survival times is

$$f(t) = w_i \lambda exp(x_{ij}^{'}\beta) e^{-w_i \lambda exp(x_{ij}^{'}\beta)t} \tag{2.37}$$

2.3.2 Weibull mixed-effects survival model

In practice, the assumption of a constant hazard function, or equivalently of exponentially distributed survival times, is rarely tenable. A more general form of hazard function is such that (Duchateau and Janssen, 2008; Moore, 2016)

$$h_0(t) = \lambda \alpha t^{\alpha - 1} \tag{2.38}$$

with $\lambda > 0$, $\alpha > 0$. A scale parameter (λ) is a parameter that provides information on the way the hazard (or density) is stretched out and the shape parameter (α) is a parameter that allows the density to take a variety of shapes, depending on the value of the shape

parameter. For $\alpha = 1$, the hazard is constant over time (the exponential hazard). When making this parametric assumption 2.38 for the baseline hazard, it follows that the event times are Weibull distributed. Hence

$$h_{ij} = \lambda\alpha t^{\alpha-1} w_i exp(x_{ij}^{'}\beta) \tag{2.39}$$

From the form of this hazard function 2.38, the survival time of the jth individual in i group has a Weibull distribution with scale parameter $\lambda exp(x_i^{'}\beta)$ and shape parameter α. This again is a manifestation of the proportional hazards property of the Weibull distribution. This result shows that the effect of the explanatory variates in the model is to alter the scale parameter of the distribution, while the shape parameter remains constant. It follows that

$$\begin{aligned} S_{ij}(t) &= exp\left(-\int_0^t \lambda\alpha s^{\alpha-1} w_i exp(x_{ij}^{'}\beta)ds\right) \\ &= exp(-\lambda t^{\alpha} w_i exp(x_{ij}^{'}\beta)) \end{aligned} \tag{2.40}$$

and

$$f_{ij}(t) = \lambda\alpha t^{\alpha-1} w_i exp(x_{ij}^{'}\beta) exp(-\lambda t^{\alpha} w_i exp(x_{ij}^{'}\beta)) \tag{2.41}$$

2.3.3 Estimation of regression parameters in selected Parametric mixed survival models

In the literature of statistics, several methods are used to estimate unknown parameters in a frailty model. These methods include maximum likelihood methods that are only applicable in parametric models, the EM algorithm, penalized partial likelihood, Markov

Chain Monte Carlo(MCMC), and best linear unbiased predictor (BLUP) method (Munda et al., 2012; Hosmer et al., 2008; Ripatti et al., 2002). Maximum likelihood estimation is not straightforward for models that involve intractable integration. The maximum likelihood estimation in the gamma frailty model is straightforward as the frailties can easily integrated into the likelihood function and obtain the parameter estimates using classical maximum likelihood techniques.

Following ideas from section 2.1, the conditional likelihood function for the jth subject in ith cluster, with $H_0(t)$ the cumulative baseline hazard, is

$$L_{ij}(\psi, \beta|w_i) = \prod_{i=1}^{G} \left(h_0(t) w_i e^{x'\beta} \right)^{\delta_{ij}} exp\left(-H_0(t) w_i e^{x'\beta} \right) \tag{2.42}$$

where ψ is a vector of parameters of the baseline hazard. For instance, in case of a Weibull baseline hazard, $\psi = (\lambda, \alpha)$. Parameter estimation in the gamma model is straightforward (Duchateau and Janssen, 2008). The frailty term can be integrated out and an explicit representation of the unconditional survival function exists 2.42, which can be used to derive the likelihood function.

The full likelihood function for the ith group with gamma frailty is given by (Klein and Moeschberger, 2003)

$$L_i(\psi, \theta, \beta) = \prod_{i=1}^{n} \int_0^{\infty} \left(h_0(t) w_i e^{x'\beta} \right)^{\delta_{ij}} exp\left(-H_0(t) w_i e^{x'\beta} \right) \frac{w^{1/\theta - 1} e^{-w/\theta}}{\Gamma(1/\theta)\theta^{1/\theta}} dw. \tag{2.43}$$

Rearranging the terms in 2.43, the following expression is obtained

$$L_i(\psi,\theta,\beta) = \prod_{i=1}^{n} h_0(t)^{\delta_{ij}} e^{(X'\beta\delta_{ij})} \int_0^\infty \frac{w^{1/\theta+d_i-1} e^{-w/\theta} exp\left(-\sum_{i=1}^{n_i} H_0(t) w e^{X'\beta}\right)}{\Gamma(1/\theta)\theta^{1/\theta}} dw \tag{2.44}$$

$$= \prod_{i=1}^{n} h_0(t)^{\delta_{ij}} e^{(X'\beta\delta_{ij})} \frac{\Gamma(\frac{1}{\theta}+d_i)\theta^{1/\theta+d_i}}{\Gamma(1/\theta)\theta^{1/\theta}} \int_0^\infty \frac{w^{\frac{1}{\theta}+d_i-1} exp\left(-w(\frac{1}{\theta}+\sum_{i=1}^{n_i} H_0(t) e^{X'\beta})\right)}{\Gamma(1/\theta+d_i)\theta^{1/\theta+d_i}} dw$$

where $d_i = \sum_{i=1}^{n} \delta_{ij}$.

The term under the integral is the moment generating function (mgf) of a gamma distribution with a pdf $\Gamma(1/\theta + d_i, 1/\theta)$. Using this fact, the expression for the marginal likelihood function is obtained by integrating out the frailty term w.

$$L_i(\psi,\theta,\beta) = \prod_{i=1}^{n} h_0(t)^{\delta_{ij}} e^{(X'_{ij}\beta\delta_{ij})} \frac{\Gamma(1/\theta+d_i)\theta^{(1/\theta+d_i)}}{\Gamma(1/\theta)\theta^{1/\theta}\theta^{(1/\theta+d_i)}\left(1/\theta+\sum_{i=1}^{n} H_0(t) e^{X'_{ij}\beta}\right)^{(1/\theta+d_i)}}$$
$$\int_0^\infty \frac{w^{1/\theta+d_i-1} exp\left(-w\left(1/\theta+\sum_{i=1}^{n} H_0(t) e^{X_{ij}\beta}\right)\right)\left[1/\theta+\sum_{i=1}^{n} H_0(t) e^{X'_{ij}\beta}\right]^{(1/\theta+d_i)}}{\Gamma(1/\theta+d_i)} dw \tag{2.45}$$

It is easy to see that the term under the integral is the pdf of $\Gamma(1/\theta, 1/\theta + \sum_{i=1}^{n} H_0(t) e^{X'_{ij}\beta})$ which integrates to 1. Therefore, the obtained marginal likelihood function is (Klein and

Moeschberger, 2003; Duchateau and Janssen, 2008)

$$L_i(\psi, \theta, \beta) = \frac{\Gamma(1/\theta + d_i) \prod_{i=1}^{n} h_0(t)^{\delta_{ij}} e^{X_{ij}^{'} \beta \delta_{ij}}}{\left(1/\theta + \sum_{i=1}^{n} H_0(t) e^{X_{ij}^{'} \beta}\right)^{(1/\theta + d_i)} \Gamma(1/\theta) \theta^{1/\theta}} \tag{2.46}$$

Taking the logarithm of this expression and summing over the clusters, the marginal loglikelihood function is obtained as (Duchateau and Janssen, 2008), $l(\psi, \theta, \beta)$.

$$\begin{aligned} l(\psi, \theta, \beta) = & \sum_{i=1}^{n} [d_i log(\theta) - log(\Gamma(1/\theta)) + log(\Gamma(1/\theta + d_i)) - (1/\theta + d_i) log(1 + \\ & \theta \sum_{i=1}^{n} H_0(t) e^{X_{ij}^{'} \beta}) + \sum_{i=1}^{n} \delta_{ij} (X_{ij}^{'} \beta + log(h_0(t)))] \end{aligned} \tag{2.47}$$

The maximum likelihood estimates for ψ, θ, and β can be obtained by setting each of the first-order derivatives to 0 and solving for the parameters of interest.

2.3.3.1 Estimating survival using Exponential Distribution with Gamma Frailty Model

Suppose that the survival times of n individuals, $t_1, t_2, ..., tn$, are assumed to have an exponential distribution with mean λ^{-1}. For the exponential distribution,$h_0(t) = \lambda$ and $H_0(t) = \lambda t$ and on substituting into expression 2.47, the log-likelihood function observations is given by (Collett, 2023)

$$\begin{aligned} l(\lambda, \theta, \beta) = & \sum_{i=1}^{n} [d_i log(\theta) - log(\Gamma(1/\theta)) + log(\Gamma(1/\theta + d_i)) - (1/\theta + d_i) log(1 + \\ & \theta \sum_{i=1}^{n} \lambda t e^{X_{ij}^{'} \beta}) + \sum_{i=1}^{n} \delta_{ij} (X_{ij}^{'} \beta + log(\lambda))] \end{aligned} \tag{2.48}$$

The maximum likelihood estimates can be obtained by setting each of the first-order derivatives to 0 and solving for the parameter of interest Collett (2023); Klein and Moeschberger (2003); Duchateau and Janssen (2008).

The first partial derivative of the scale parameter λ is given by;

$$\begin{aligned}\frac{\partial l(\lambda,\theta,\beta)}{\partial \lambda} &= \frac{-(d_i+1/\theta)\theta\sum_{j=1}^{n_i} t_{ij}exp(x_{ij}{}'\beta)}{1+\theta\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)} + d_i\lambda^{-1} \\ &= \frac{-(d_i\theta+1)\sum_{j=1}^{n_i} t_{ij}exp(x_{ij}{}'\beta)}{1+\theta\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)} + d_i\lambda^{-1}\end{aligned} \tag{2.49}$$

The first partial derivative of the variance parameter estimate θ is given by

$$\begin{aligned}\frac{\partial l(\lambda,\theta,\beta)}{\partial \theta} &= \frac{-(d_i+\theta^{-1})\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)}{1+\theta\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)} + \theta^{-2}log\left(1+\theta\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)\right) \\ &- I(d_i>0)\sum_{l=0}^{d_i-1}\left(\theta-l\theta^2\right)^{-1} + d_i\theta^{-1}\end{aligned} \tag{2.50}$$

The partial first derivative of the regression parameter estimate β is given by;

$$\frac{\partial l(\lambda,\theta,\beta)}{\partial \beta} = \frac{-(d_i\theta+1)\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)x_{ij}}{1+\theta\sum_{j=1}^{n_i}\lambda t_{ij}exp(x_{ij}^{'}\beta)} + \sum_{j=1}^{n_i}\delta_{ij}x_{ij} \tag{2.51}$$

The asymptotic variance-covariance matrix can also be derived from the loglikelihood expression (Duchateau and Janssen, 2008; Collett, 2023). The asymptotic variance-covariance matrix of the vector of parameter estimates is the inverse of the Fisher information matrix; the estimated variance-covariance matrix is obtained as the inverse of the observed information

matrix.

The second derivatives with respect to the parameters are given by

$$\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\lambda^2} = \frac{(d_i + 1/\theta)\left(\sum_{j=1}^{n_i} t_{ij} exp(x_{ij}{}'\beta)\right)^2}{\left(1/\theta + \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}'\beta)\right)^2} + d_i\lambda^{-2} \tag{2.52}$$

$$\begin{aligned}\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\theta^2} =& I(d_i > 0)\sum_{l=0}^{d_i-1}\frac{1+2l\theta}{(l\theta^2+\theta)} - \frac{2log\theta - 3}{\theta^3} \\ &- \frac{2}{\theta} log\left(\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta) + 1/\theta\right) \\ &- \frac{3/\theta + d_i + (4+2\theta d_i)\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta)}{\left(\theta + \theta^2\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta)\right)^2}\end{aligned} \tag{2.53}$$

$$\begin{aligned}\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\beta^2} =& \frac{(d_i + \theta^{-1})}{\left(\theta^{-1} + \sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta)\right)^2} \\ &\left[\left(\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta)\right)^2 - \sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta) x_{ij}^2\left(1/\theta + \sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}'\beta)\right)\right]\end{aligned} \tag{2.54}$$

and so the asymptotic variance of $\hat{\lambda}$ is

$$var(\hat{\lambda}) = \left\{-E\left(\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\lambda^2}\right)\right\}^{-1}$$

Consequently, the standard error of $\hat{\lambda}$ is given by

$$se(\hat{\lambda}) = \sqrt{var(\hat{\lambda})}$$

This result could be used to obtain a confidence interval for the mean survival time. In particular, the limits of a $100(1-\alpha)\%$ confidence interval for λ are $\hat{\lambda} \pm z_{\alpha/2} se(\hat{\lambda})$, where $z_{\alpha/1}$ is the upper $\alpha/2$-point of the standard normal distribution.

2.3.4 Post-estimation analysis for parametric mixed survival model

Diagnostic procedures for the assessment of model adequacy are as important in parametric modeling as they are when the Cox mixed effects regression model is used in the analysis of survival data. This is followed by a summary of graphical procedures for assessing the suitability of models fitted to data that are assumed to have a Weibull, log-logistic, or lognormal distribution (Collett, 2023; Klein and Moeschberger, 2003).

The graphical checks serve as a means of rejecting clearly inappropriate models, not to prove that a particular parametric model is correct (He et al., 2020). In fact, in many applications, several parametric models may provide reasonable fits to the data and provide quite similar estimates of key quantities. The key tool is to find a function of the cumulative hazard rate which is linear in some function of time (Collett, 2023). The basic plot is made by estimating the cumulative hazard rate by the Nelson-Aalen estimator (Klein and Moeschberger, 2003). To illustrate this technique, consider a check of the appropriateness of the Weibull distribution. The cumulative hazard rate for loglogistic is $H(x) = log(1 +$

ax^b). This implies that, for the log logistic model,

$$logexp[H(x)] - 1 = loga + blogx,$$

so, a plot of $log\{exp[\hat{H}(x)] - 1\}$ versus $log(x)$ should be approximately linear Klein and Moeschberger (2003); Collett (2023). If the curves do not appear linear for each figure, this is evidence that the parametric model does not provide an adequate fit to the data. The slope of the line gives a crude estimate of b and the y intercept gives a crude estimate of $log(a)$. Here, $\hat{H}$ is the Nelson-Aalen estimator.

For the parametric regression problem, analogs of the residual plots can be made with a redefinition of the various residuals to incorporate the parametric form of the baseline hazard rates. The first such residual is the Cox-Snell residual that provides a check of the overall fit of the model (Collett, 2023; Klein and Moeschberger, 2003). The Cox-Snell residual, r_{ij}, is defined by $r_{ij} = \hat{H}(T_{ij}|X)$,where $\hat{H}$ is the fitted model. If the model fits the data then the $r_{ij}^{'}$s should have a standard (λ = 1) exponential distribution, so that a hazard plot of r_{ij} versus the Nelson-Aalen estimator of the cumulative hazard of the r_{ij} should be a straight line with slope 1. The Cox-Snell residuals are given by (Klein and Moeschberger, 2003)

Exponential $r_{Cij} = T_{ij}exp(\hat{\beta}^T X_{ij})$

Weibull $r_{Cij} = exp(\hat{\beta}^T X_{ij})T_{ij}^{\hat{\gamma}}$

The martingale and deviance residuals were defined for Cox regression models. For a parametric model, the martingale residual is defined by (Therneau et al., 2000)

$$M_{ij} = \delta_{ij} - r_{ij}$$

and the deviance residual by (Klein and Moeschberger, 2003)

$$D_{ij} = sign[M_i j]\{-2[M_{ij} + \delta_{ij} log(\delta_{ij} - M_{ij})]\} l/2$$

As for the Cox model, the martingale residual is an estimate of the excess number of deaths seen in the data, but not predicted by the model. In the parametric case, note that the derivation of M_{ij} as a martingale does not hold but, because the residuals are similar in form to those for the Cox model, the name carries through (Klein and Moeschberger, 2003; Therneau et al., 2000). The deviance residuals are an attempt to make the martingale residuals more symmetric about 0. If the model is correct, then, the deviance residuals should look like random noise (Klein and Moeschberger, 2003). Plots of either the martingale or deviance residuals against time, observation number, or acceleration factor provide a check of the model's adequacy (Collett, 2023; Therneau et al., 2000,?).

2.4 Examples of studies that applied mixed-effects Cox PHMM or mixed-effects parametric survival model to some clustered data

This section provides a collection of examples from the literature dealing with interesting frailty models. The main emphasis is to give an overview of the variety of applications of frailty models in the solution of different statistical problems.

Ghavami et al. (2021) modeled time-dependent effects on survival in breast cancer using a modified gamma frailty model with piecewise constant baseline hazards. After 20 years of follow-up, it was demonstrated that the effects of tumor size, lymph node status, and histologic grade are still significant predictors; however, the impacts of these risk variables gradually fade with time from prognosis. Frailty models are one possibility to allow for such kind of attenuation of prognostic effects.

Several shared frailty models with Weibull baseline hazard were used to the recurrence of breast cancer by Feleke et al. (2022). The data set includes 322 women who received treatment for breast cancer at the Felege Hiwot Comprehensive Specialized Hospital (FHCSH). Age and the disease's stage at the time of the first treatment were the two explanatory factors used. The length of time until recurrence following treatment is the outcome of interest. The covariate state, which is a factor with three levels, was incorporated into the modeling of repeated durations. Other causes of death were viewed as appropriate censorship and were therefore considered to be non-informative. The analysis takes into

account using three models; the Cox proportional hazard model, the accelerated failure time model, and shared frailty models. Adding a frailty term to the model results in a significant rise in the log-likelihood. The discrete four-point frailty model fits the data better than the gamma and log-normal frailty models, respectively, even if formal likelihood ratio tests are inappropriate.

Using time dependent random effects and nonparametric random effect distribution, Manda (2020) applied a shared frailty model with observable covariates to child survival data.. For both approaches, full Bayesian inference using the Gibbs sampler algorithm was used for computing posterior parameter for the mixing distribution and regression coefficients.

The impact of the frailty distribution on the relationship between survival times under the discrete k-point, gamma, inverse Gaussian, and Poisson distributions was examined by Bardo and Unkel (2023). Bardo and Unkel (2023) demonstrates how the frailty distribution's specification affects the association's pattern and strength.

A spatially correlated frailty model was applied to Kenya infant mortality data by Daniel et al. (2021). Spatial correlations between clusters are modeled using a correlation matrix for the frailties (correlated frailty model). A couple of studies looking into different models of spatial fragility are Azevedo et al. (2023) and Frévent et al. (2024).

Researchers in dentistry frequently deal with the issue of clustered event times. There are, however, few instances of patient-focused dental research using frailty models. Fagbamigbe et al. (2021) examined the amount of time until caries to develop. The Cox proportional hazard frailty model was used by Zhang et al. (2019) to investigate sealant retention. A

multilevel log-normal frailty model was used by Molina et al. (2021) to analyze Chinese schoolchildren's lifetime treatment restorations of carious lesions.

Chapter 3

Research Methods

This chapter presents methodological approaches used to model clustered women's birth intervals. It describes the research design used to model clustered women's birth intervals. It further looks at model appropriateness and statistical analysis used.

3.1 Data

3.1.1 Research Design

The study is quantitative and uses secondary data from a well-rated nationwide survey. The information-gathering approach is via structured interviews administered by well-trained research assistants or field workers.

3.1.2 Data Sources and Description

This study obtained and analyzed data from the Demographic and Health Surveys (DHS) across all districts in Malawi. Malawi is a country in south-eastern Africa that borders Tanzania to the north, Zambia to the west, and Mozambique to its east, south, and west. The country is divided into 28 administrative districts, with a total population of just over 20 million people (NSO, 2017). The 2015-16 Malawi Demographic and Health

Survey (2015-16 MDHS) was conducted between October 2015 and February 2016 by the National Statistical Office (NSO) of Malawi in joint collaboration with the Ministry of Health (MoH) and the Community Health Services Unit (CHSU). Malawi conducted its first DHS in 1992 and again in 2000, 2004, and 2010. The 2015-16 MDHS is the fifth in the series. The DHS Program was funded by the Malawi government, the U.S. Agency for International Development (USAID), the National AIDs Commission (NAC), the United Nations Children's Fund (UNICEF), the United Nations Population Fund (UNFPA), and UN Women. The survey is based on a nationally representative sample that provides estimates at the national and regional levels and for urban and rural areas with key indicator estimates at the district level. The survey included 26,361 households, 24,562 female respondents, and 7,478 male respondents. The data are available at www.DHSprogram.com.

The survey used a two-stage stratified sampling design, with small enumeration areas(SEAs) as primary sampling units and households as secondary units (NSO, 2017). Each district was stratified into urban and rural areas; this yielded 56 sampling strata. Samples of SEAs were selected independently in each stratum in two stages. In the first stage, 850 SEAs, including 173 SEAs in urban areas and 677 in rural areas, were selected with probability proportional to the SEA size and with independent selection in each sampling stratum. The SEA size is the number of residential households in the SEA. A household listing operation was implemented in all the selected SEAs during August-October 2016. The resulting lists of households served as a sampling frame for the selection of households in the second stage. In the second selection stage, a fixed number of 30 households per urban cluster and 33 per rural cluster were selected with an equal probability of systematic

selection from the newly created household listing. All women aged 15-49 who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey were eligible to be interviewed.

3.1.3 Study outcome

The main outcome of this study was time in months to successive second births among productive women in Malawi. The event time is defined as the interval between the date of first birth and a second live birth or censoring. Women who had a second birth child were deemed to have had the event(a second birth) and were assigned the number 1. Women who gave first birth but not second birth or those who were not even impregnated during the survey period are considered as censored for second birth interval analysis. Those who did not have a second child before the period of study were censored and assigned the number 0.

3.1.4 Explanatory Variables

The data are right censored and all the variables considered were categorical except continuous age. The explanatory variables were educational status of woman(no education=0, primary=1, secondary=2, tertiary=3),region(northern=0, central=1, southern=0), sex of first child(male = 0, female = 1), age at first birth, survival status of index child(live = 1, dead = 0), wealth index (poor=0, middle=1, rich=2), and contraceptive method(no method=0, traditional Method =1, Modern Method=2) were the independent variables (Aychiluhm et al., 2020; Ahammed et al., 2019; Bagheri and Saadati, 2021). The factor variables with more than

two categories were coded as dummy variables. The set of values of the explanatory variables in the survival mixed effects model is represented by the vector x_{ij} so that x = (Age, Alive, Sex, Central, Southern, Rural, Primary, Secondary, Higher, Middle, Rich, Traditional, Modern). These included independent variables in the model were selected based on recommendations from the literature. In this study, 850 sample enumeration areas were considered as clusters for the analysis.

3.2 Statistical methods

Descriptive graphs of the survivor function would be used for comparing the events experienced at a time by two or more groups and the survival quantities of covariates to describe the survival experience of women at specific times. The Kaplan–Meier estimator of the survival curve give the estimate of survivor function among different groups of covariates to make comparisons. In general, the pattern of one survivorship function lying above another means the group defined by the upper curve has better survival than the group defined by the lower curve (Feleke et al., 2022; Jiang and Guterman, 2024).

The differences in the survival distribution between the groups for each covariate are tested using the log-rank test. The null hypothesis to be tested is that there is no difference between the probabilities of an event(having a second child) occurring at any time point for each group. This test is performed using the function `survdiff` from the `survival` package (Therneau et al., 2015) in R.4.3.3 software. The survival difference between the groups is statistically significant when the p-value is significant(p-value < 0.05).

3.2.1 Cox PH Mixed-Effects Survival Model with Normally Distributed Random Effects

The following model was fitted:

$$\begin{aligned} h_{ij}(t) =& h_0(t) w_i exp(x_{ij}^{'}\beta) \\ =& h_0(t) w_i exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive) \end{aligned} \tag{3.1}$$

where $h_0(.)$ is the unknown baseline hazard, β is the vector of fixed-effects coefficients and w_i is the vector of random effects coefficients (Therneau et al., 2015). Notice that there is no constant term in the linear component of this Cox proportional hazards mixed-effects model. If a constant term b_0, say, were included, the baseline hazard function could simply be rescaled by dividing $h_0(t)$ by $exp(\beta_0)$, and the constant term would cancel out. It is generally more convenient to work with an alternative representation of the frailty effect, obtained by setting $w_i = exp(b_i)$ (Therneau et al., 2015). The model in Equation 3.2 can then be written

$$\begin{aligned} h_{ij}(t) =& h_0(t) exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive + b_i) \end{aligned} \tag{3.2}$$

where $b_i \sim Normal\,(0, \sigma^2)$. In this model, $b_i = log w_i$ is a random effect in the linear component of the proportional hazards model.

3.2.1.1 Estimation method for Cox PH Mixed Model with Normally Distributed Random Effects.

The estimation procedure is a generalization of the results developed by (McGilchrist and Aisbett, 1991). To find the best linear unbiased predictor, it is necessary to maximize the penalized partial log-likelihood which has two components. The first component is the partial log-likelihood of the failure times taking the random effects fixed and the second component is related to the distribution associated with the random effects. The penalized partial log-likelihood is

$$l_{pen}(\beta, b) = \sum_{i=1}^{G} \delta_{ij} \left(x_{ij}^{'}\beta + b_i - log \sum_{l \in R(t_{il})} exp(x_{il}^{'}\beta + b_l) \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{G} b_i^2 \qquad (3.3)$$

where δ is the event indicator and $R(t_{il})$ is the set of women without second birth.

The maximization process proceeds iteratively by starting with a provisional estimate of σ^2 and finding the estimates of the βs and the b_is that maximize $l_p(\beta, b|t, X)$. Next, a marginal log-likelihood, log $L_m(\beta, \sigma_b^2)$, is obtained by integrating the log-likelihood in Expression 3.3 over the random effects b_i. Ripatti et al. (2002) show that, at estimates $\hat{\beta}$ of β, this is well-approximated by

$$\log L_m(\sigma_b^2) = \log L_{pen}(\hat{\beta}, \hat{b}, \sigma_b^2) - \frac{1}{2}\log(\sigma_b^{2n}) - \frac{1}{2}\log|I_b|, \qquad (3.4)$$

where I_b is the $n \times n$ observed information matrix for the random effects, formed from the negative second partial derivatives of log $L_{pen}(\beta, b, \sigma_b^2)$ with respect to the $b_i, i = 1, 2, ..., n$,

evaluated at $\hat{\beta}$, $\hat{b}$, and $|I_b|$ is the determinant of that matrix. Using the estimates of the βs, the marginal log-likelihood in Equation 3.4 is then maximized with respect to σ_b^2 to give a revised estimate of σ_b^2. This process is repeated until the difference between two successive estimates of σ_b^2 is sufficiently small.

Estimates of the random effects, $\hat{b}_i$, are also obtained from this iterative process based on the method of restricted maximum likelihood (REML) estimation (Collett, 2023). REML estimates are obtained from a likelihood function that is independent of β, and the REML estimate of the variance of the random effects term is

$$\tilde{\sigma}_b^2 = n^{-2}\left\{\sum_{i=1}^{n}\tilde{b}_i^2 + \text{trace}(\tilde{V}_b)\right\} \tag{3.5}$$

where $\tilde{V}_b$ is the estimated variance-covariance matrix of the REML estimates of the b_i, $\tilde{b}_i$. The trace of this matrix is just the sum of the estimated variances of the $\tilde{b}_i$. The standard error of $\tilde{\sigma}_b^2$ can be found from

$$se(\tilde{\sigma}_b^2) = \sqrt{(2\tilde{\sigma}_b^2)}\left[n + \frac{1}{\tilde{\sigma}_b^4}\text{trace}(\tilde{V}_b\tilde{V}_b) - 2\frac{1}{\tilde{\sigma}_b^2}\text{trace}(\tilde{V}_b)\right]^{-1/2} \tag{3.6}$$

These standard errors is used to obtain approximate confidence intervals for the unknown β-parameters (Collett, 2023). In particular, a 95% confidence interval for a parameter β is the interval with limits $\hat{\beta} \pm 1.96se(\hat{\beta})$, where $\hat{\beta}$ is the estimate of β, and 1.96 is the upper 0.025-point of the standard normal distribution. If a 95% confidence interval for β does not include zero, this is evidence that the value of β is non-zero. The corresponding estimate of the hazard ratio is e^{β}. A 95% confidence interval for the hazard ratio, e^{β}, can

be found simply by exponentiating the confidence limits for β. The value of greater than 1 of the hazard ratio for an explanatory variable x_{ij} implies increased birth intensity (shorter birth interval between first and second child) relative to the baseline for whom $x = 0$. Conversely, a less than 1 hazard ratio depicts a longer birth interval. To examine the main hypothesis about the heterogeneity of risks between subjects, a likelihood ratio test is used to test the null hypothesis that the frailty term has zero variance, that is, σ^2 = 0. These estimates and tests are performed using the function `coxph` from the `survival` package (Therneau et al., 2015) in R.4.3.3 software.

3.2.2 Cox PH Mixed-Effects Survival Model with Gamma Distributed Random Effects

The following model was fitted:

$$\begin{aligned} h_{ij}(t) =& h_0(t) w_i exp(x_{ij}^{'}\beta) \\ =& \alpha\lambda w_i t^{\alpha-1} exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive) \end{aligned} \tag{3.7}$$

where $w_i \sim \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$, $h_0()$ is unknown hazard function. When the random variable associated with random effects w_i is assumed to have a gamma distribution with unit mean and variance θ, the corresponding random effects, $b_i = \log w_i$, have an exp-gamma distribution.

3.2.2.1 Estimation method for Cox PH Mixed Model with Gamma Distributed Random Effects.

The log-likelihood function for the Cox proportional hazard mixed model with gamma-distributed random effects is

$$\begin{aligned} l(\beta, b) = &\sum_{i=1}^{G} \delta_{ij} \left(x_{ij}^{'}\beta + b_i - log \sum_{l \in R(t_{il})} exp(x_{il}^{'}\beta + b_l) \right) \\ &+ n/\theta \text{log}\theta - n\text{log}\Gamma(1/\theta) - \sum_{i=1}^{n} \{(1/\theta)e^{b_i} - b_i\} \end{aligned} \tag{3.8}$$

in which $\sum_{i=1}^{n}\{(1/\theta)e^{b_i} - b_i\}$ is the penalty term. This log-likelihood function is used to estimate the components of β and b, and so terms involving θ alone are omitted to give the penalized partial log-likelihood function

$$l_{pen}(\beta, b) = \sum_{i=1}^{G} \delta_{ij} \left(x_{ij}^{'}\beta + b_i - log \sum_{l \in R(t_{il})} exp(x_{il}^{'}\beta + b_l) \right) - \sum_{i=1}^{n} \{\theta e^{b_i} - b_i\} \tag{3.9}$$

To obtain estimates of β, b, and θ, estimates of β and b are first taken to be the values that maximize Equation 3.9 for a given value of θ. The marginal log-likelihood for θ, at estimates $\hat{\beta}$ and $\hat{b}$ of β and b, is then used to obtain a revised estimate of θ. This has been shown by Therneau et al. (2000) to be given by

$$\begin{aligned} \text{log}L_m(\theta) = &\text{log}L_{pen}(\hat{\beta}, \hat{b}, \theta) + n/\theta(1 - \text{log}\theta) \\ &- \sum_{i=1}^{n} \left\{ (1/\theta + \delta_i)\text{log}(1/\theta + \delta_i) - \text{log}\left[\frac{\Gamma(1/\theta + \delta_i)}{\Gamma(1/\theta)} \right] \right\} \end{aligned} \tag{3.10}$$

where δ_i is the event indicator for the i-th women. Maximising $logL_m(\theta)$in Equation 3.10 with respect to θ, leads to $\hat{\theta}$. This estimate is then used with Equation 3.9 to obtain updated

estimates of β and b, and so on until the process converges.

The standard errors of the estimates are approximated from the inverse of the Hessian matrix for the second partial derivatives of the marginal loglikelihood function $\log L_m(\theta)$. These standard errors are used to obtain approximate 95% confidence intervals for the unknown β-parameters. To examine the main hypothesis about the heterogeneity of risks between subjects, a likelihood ratio test is used to test the null hypothesis that the frailty term has zero variance, that is, $\sigma^2 = 0$. These estimates and tests are performed using the function `coxph` from the `survival` package (Therneau et al., 2015) in R.4.3.3 software.

3.2.3 Exponential Mixed-Effects Survival Model with Normally Distributed Random Effects

The following model is fitted:

$$\begin{aligned} h_{ij}(t) =& h_0(t) exp(x_{ij}^{'}\beta + b_i) \\ =& \lambda exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive + b_i) \end{aligned} \tag{3.11}$$

where $b_i \sim normal\,(0, \sigma^2)$, $h_{ij}(t)$ is the hazard function for the jth woman from the ith cluster with vector of covariate x_{ij}, β is the covariate effect, and b_i is the random effects for cluster i.

3.2.3.1 Estimation of Exponential Mixed-Effects Survival Model with Normally Distributed Random Effects

No explicit form of unconditional likelihood exists because of the intractable Laplace transform. The likelihood of the data is of the form

$$L_(\beta, \lambda, \sigma^2) = \lambda exp(x_{ij}^{'}\beta + b_i) exp(-\lambda exp(x_{ij}^{'} + b_i)t) d\Phi(b_i) \tag{3.12}$$

with Φ as the cumulative distribution function of the normal distribution with mean zero and variance σ^2. No explicit form exists for the marginal likelihood for the lognormal frailty model. Instead of integrating out the random effects analytically, the saddlepoint approximation is used.

The maximum likelihood estimates for λ, σ^2, and β are obtained by setting each of the first-order derivatives of marginal likelihood to 0 and solving for the parameters of interest. The standard errors are given by the square root of the diagonal elements of the inverse of the information matrix of the marginal likelihood evaluated at ($\hat{\lambda}$,$\hat{\sigma}^2$,$\hat{\beta}$). To examine the hypothesis about the heterogeneity of risks between subjects, a likelihood ratio test is used to test the null hypothesis that the frailty term has zero variance-that is, σ^2 = 0. These tests are performed using the function `parfm` from the `parfm 2.77` package in R.4.3.3 software (Munda et al., 2012).

3.2.4 Exponential Mixed-Effects Survival Model with Gamma Distributed Random Effects

The following exponential mixed-effects survival model with gamma-distributed random effects is fitted:

$$\begin{aligned} h_{ij}(t) =& h_0(t) w_i exp(x_{ij}^{'} \beta) \\ =& \lambda w_i exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive) \end{aligned} \tag{3.13}$$

where $w_i \sim \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$, $h_{ij}(t)$ is the hazard function for the jth woman from the ith cluster with vector of covariate x_{ij}, β is the covariate effect, and w_i the random effects term for cluster i.

3.2.4.1 Estimation of the Exponential Mixed-Effects Survival Model with Gamma Distributed Random Effects

The full likelihood function for the ith group with gamma random effects is given by (Klein and Moeschberger, 2003)

$$L_i(\psi, \theta, \beta) = \prod_{i=1}^{n} \int_0^{\infty} \left(\lambda w_i e^{x^{'}\beta}\right)^{\delta_{ij}} exp\left(-\lambda t w_i e^{x^{'}\beta}\right) \frac{w^{1/\theta - 1} e^{-w/\theta}}{\Gamma(1/\theta)\theta^{1/\theta}} dw. \tag{3.14}$$

Marginal log-likelihood is obtained by integrating the log-likelihood in Expression 3.14 over the random effects w_i. Maximum likelihood estimators for λ, β are found by maximizing

this marginal loglikelihood.

$$
\begin{aligned}
l(\lambda,\theta,\beta) = &\sum_{i=1}^{G}[d_i log(\theta) - log(\Gamma(1/\theta)) + log(\Gamma(1/\theta + d_i)) - (1/\theta + d_i)log(1+ \\
&\theta \sum_{j=1}^{n_i} \lambda t e^{X_{ij}^{'}\beta}) + \sum_{j=1}^{n_i} \delta_{ij}(X_{ij}^{'}\beta + log(\lambda))]
\end{aligned}
\tag{3.15}
$$

The maximum likelihood estimates of parameters is obtained by setting each of the first-order derivatives to 0 and solving for the parameter of interest Collett (2023); Moore (2016).

The first partial derivative of the scale parameter λ is given by;

$$
\begin{aligned}
\frac{\partial l(\lambda,\theta,\beta)}{\partial \lambda} &= \frac{-(d_i + 1/\theta)\theta \sum_{j=1}^{n_i} t_{ij} exp(x_{ij}{}'\beta)}{1 + \theta \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)} + d_i \lambda^{-1} \\
&= \frac{-(d_i\theta + 1) \sum_{j=1}^{n_i} t_{ij} exp(x_{ij}{}'\beta)}{1 + \theta \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)} + d_i \lambda^{-1}
\end{aligned}
\tag{3.16}
$$

The first partial derivative of the variance parameter estimate θ is given by

$$
\begin{aligned}
\frac{\partial l(\lambda,\theta,\beta)}{\partial \theta} = &\frac{-(d_i + \theta^{-1}) \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)}{1 + \theta \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)} + \theta^{-2} log\left(1 + \theta \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)\right) \\
&- I(d_i > 0) \sum_{l=0}^{d_i - 1} \left(\theta - l\theta^2\right)^{-1} + d_i \theta^{-1}
\end{aligned}
\tag{3.17}
$$

The partial first derivative of the regression parameter estimate β is given by;

$$
\frac{\partial l(\lambda,\theta,\beta)}{\partial \beta} = \frac{-(d_i\theta + 1) \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta) x_{ij}}{1 + \theta \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)} + \sum_{j=1}^{n_i} \delta_{ij} x_{ij} \tag{3.18}
$$

The asymptotic variance-covariance matrix is also derived from the loglikelihood expression (Duchateau and Janssen, 2008; Collett, 2023). The asymptotic variance-covariance matrix of the vector of parameter estimates is the inverse of the Fisher information matrix; the estimated variance-covariance matrix is obtained as the inverse of the observed information matrix.

The second derivatives with respect to the parameters are given by

$$\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\lambda^2} = \frac{(d_i + 1/\theta)\left(\sum_{j=1}^{n_i} t_{ij} exp(x_{ij}{}'\beta)\right)^2}{\left(1/\theta + \sum_{j=1}^{n_i} \lambda t_{ij} exp(x_{ij}^{'}\beta)\right)^2} + d_i\lambda^{-2} \tag{3.19}$$

$$\begin{aligned}\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\theta^2} =& I(d_i > 0)\sum_{l=0}^{d_i-1}\frac{1+2l\theta}{(l\theta^2+\theta)} - \frac{2log\theta - 3}{\theta^3} \\ &- \frac{2}{\theta} log\left(\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta) + 1/\theta\right) \\ &- \frac{3/\theta + d_i + (4+2\theta d_i)\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta)}{\left(\theta + \theta^2\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta)\right)^2}\end{aligned} \tag{3.20}$$

$$\begin{aligned}\frac{\partial^2 l(\lambda,\theta,\beta)}{\partial\beta^2} =& \frac{(d_i + \theta^{-1})}{\left(\theta^{-1} + \sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta)\right)^2} \\ &\left[\left(\sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta)\right)^2 - \sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta) x_{ij}^2\left(1/\theta + \sum_{j=1}^{n_i}\lambda t_{ij} exp(x_{ij}^{'}\beta)\right)\right]\end{aligned} \tag{3.21}$$

and so the asymptotic variance of $\hat{\lambda}$ is

$$var(\hat{\lambda}) = \left\{ -E\left(\frac{\partial^2 l(\lambda, \theta, \beta)}{\partial \lambda^2} \right) \right\}^{-1}$$

Consequently, the standard error of $\hat{\lambda}$ is given by

$$se(\hat{\lambda}) = \sqrt{var(\hat{\lambda})}$$

These standard errors are used to obtain 95% confidence intervals for the unknown β-parameters (Collett, 2023). A 95% confidence interval for the hazard ratio, e^{β}, is found simply by exponentiating the confidence limits for β. To examine the hypothesis about the heterogeneity of risks between subjects, a likelihood ratio test is used to test the null hypothesis that the frailty term has zero variance-that is, θ = 0. These estimates and tests are performed using the function `frailtyPenal` from the `frailtypack` package in R.4.3.3 software (Munda et al., 2012).

3.2.5 Weibull Mixed-Effects Survival Model with Normally Distributed Random Effects

The following Weibull mixed-effects survival model with normally distributed random effects is fitted:

$$\begin{aligned} h_{ij}(t) =& h_0(t)exp(x_{ij}^{'}\beta + b_i) \\ =& \alpha\lambda t^{\alpha-1}exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive + b_i) \end{aligned} \tag{3.22}$$

where $w_i \sim normal\,(0, \sigma^2)$, $h_{ij}(t)$ is the hazard function for the jth woman from the ith cluster with vector of covariate x_{ij}, β the covarite effect, and w_i the frailty term for cluster i.

3.2.5.1 Estimation of Weibull Mixed-Effects Survival Model with Normally Distributed Random Effects

No explicit form of unconditional likelihood exists because of the intractable Laplace transform. The likelihood of the data is of the form

$$L_(\beta, \theta, \sigma^2) = \alpha\lambda exp(x_{ij}^{'}\beta + b_i)t^{\alpha-1}exp(-\lambda exp(x_{ij}^{'} + b_i)t^{\alpha})d\Phi(b_i) \tag{3.23}$$

with Φ as the cumulative distribution function of the normal distribution with mean zero and variance σ^2. No explicit form exists for the marginal likelihood of the lognormal frailty model. Instead of integrating out the random effects analytically, the saddlepoint

approximation is used.

The maximum likelihood estimates for λ, α, θ, and β are obtained by setting each of the first-order derivatives of marginal likelihood to 0 and solving for the parameters of interest.

The standard errors are given by the square root of the diagonal elements of the inverse of the information matrix of the marginal likelihood evaluated at ($\hat{\lambda}$, $\hat{\alpha}$,$\hat{\theta}$,$\hat{\beta}$). To examine the hypothesis about the heterogeneity of risks between subjects, a likelihood ratio test is used to test the null hypothesis that the frailty term has zero variance-that is, σ^2 = 0. These estimates and tests are performed using the function `parfm` from the `parfm 2.77` package in R.4.3.3 software (Munda et al., 2012).

3.2.6 Weibull Mixed-Effects Survival Model with Gamma Distributed Random Effects

The following Weibull mixed-effects survival model with gamma-distributed random effects is fitted:

$$\begin{aligned} h_{ij}(t) =& h_0(t) w_i exp(x_{ij}^{'} \beta) \\ =& \alpha \lambda w_i t^{\alpha-1} exp(\beta_1 Age + \beta_2 Live + \beta_3 Sex + \beta_4 Region + \beta_5 Rural + \\ & \beta_6 Educational + \beta_7 WealthIndex + \beta_8 Contraceptive) \end{aligned} \tag{3.24}$$

where $w_i \sim \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$, $h_{ij}(t)$ is the hazard function for the jth woman from the ith cluster with vector of covariate x_{ij}, β is the covarite effect, and w_i the frailty term for cluster i.

Maximum likelihood estimators for $\alpha, \lambda, \alpha, and \lambda$ were found by maximizing the following loglikelihood.

$$l(\lambda, \alpha\theta, \beta) = \sum_{i=1}^{G}[d_i log(\theta) - log(\Gamma(1/\theta)) + log(\Gamma(1/\theta + d_i)) - (1/\theta + d_i)log(1+ \theta \sum_{i=1}^{n_i} \lambda t^{\alpha} e^{X_{ij}^{'}\beta}) + \sum_{j=1}^{n_i} \delta_{ij}(X_{ij}^{'}\beta + log(\lambda\alpha t^{\alpha-1}))] \tag{3.25}$$

3.2.6.1 Estimation of the Weibull Mixed-Effects Survival Model with Gamma Distributed Random Effects

The full likelihood function for the ith group with gamma frailty is given by (Klein and Moeschberger, 2003)

$$L_i(\psi, \theta, \beta) = \prod_{i=1}^{n} \int_0^{\infty} \left(\lambda\alpha t^{\alpha-1} w_i e^{x^{'}\beta}\right)^{\delta_{ij}} exp\left(-\left(\lambda t w_i e^{x^{'}\beta}\right)^{\alpha}\right) \frac{w^{1/\theta-1} e^{-w/\theta}}{\Gamma(1/\theta)\theta^{1/\theta}} dw. \tag{3.26}$$

Marginal log-likelihood is obtained by integrating the log-likelihood in Expression 3.26 over the random effects w_i. The marginal loglikelihood function is obtained as (Duchateau and Janssen, 2008),

$$l(\psi, \theta, \beta) = \sum_{i=1}^{n}[d_i log(\theta) - log(\Gamma(1/\theta)) + log(\Gamma(1/\theta + d_i)) - (1/\theta + d_i)log(1+ \theta \sum_{i=1}^{n} \lambda t^{\alpha} e^{X_{ij}^{'}\beta}) + \sum_{i=1}^{n} \delta_{ij}(X_{ij}^{'}\beta + log(\lambda\alpha t^{\alpha-1}))] \tag{3.27}$$

The maximum likelihood estimates for λ, α, θ, and β is obtained by setting each of the first-order derivatives to 0 and solving for the parameters of interest. Standard errors

are computed as the square roots of the diagonal elements of the observed information matrix of the marginal likelihood evaluated at ($\hat{\lambda}$, $\hat{\alpha}$,$\hat{\theta}$,$\hat{\beta}$). To examine the hypothesis about the heterogeneity of risks between subjects, a likelihood ratio test is used to test the null hypothesis that the frailty term has zero variance-that is, $\theta = 0$. These tests are performed using the function `frailtyPenal` from the `frailtypack` package in R.4.3.3 software (Munda et al., 2012).

3.3 Model Selection procedure

In many cases, random effect term acts as a nuisance term in the model. The methods based on the change in the $-2\log\hat{L}$ statistic, is used to compare parametric models that incorporate frailty (Duchateau and Janssen, 2008). Then, model selection procedures is applied to decide which components to include in a parametric model in the context of frailty. The fitted models is subsequently used to draw inferences about their effects, allowing for random effects.

However, the situation is more complicated for the Cox PHMM model. This is because a penalised likelihood does not behave in the same way as the usual likelihood or partial likelihood functions, since it contains unobservable random effects. Instead, the maximised marginal log-likelihood statistic is used, which is $\log L_m(\hat{\sigma}_b^2)$ or $\log L_m(\hat{\theta})$ for lognormal and gamma frailties, respectively. Differences in the values of $-2\log L_m(\hat{\sigma}_b^2)$ or $-2\log L_m(\hat{\theta})$, is compared with percentage points of a χ_1^2 distribution in the usual manner. In the case of lognormal frailty effects, the marginal likelihood based on the maximum likelihood

estimates of the β's is used, rather than that based on the REML estimates.

3.3.1 Testing for the presence of random effect

For a more precise determination of the significance of the random effect, the result that the asymptotic distribution of the change in $-2\log\hat{L}$ is used, when a frailty term is added to a fully parametric model, is an equally weighted mixture of a χ^2_0 and χ^2_0 distributions, written $0.5(\chi^2_0 + \chi^2_1)$ (Moore, 2016; Duchateau and Janssen, 2008; Wienke, 2010). A χ^2_0 distribution has a point mass at zero, and so if the random variable W has a $0.5(\chi^2_0 + \chi^2_0)$ distribution, $P(W = 0) = 0.5$ and $P(W > w) = 0.5P(\chi^2_1 > w)$ for $w > 0$. Consequently, to apply this result, the P-value for testing the hypothesis that the frailty variance is zero is just half of that obtained from using a χ^2_1 distribution.

In the case of the Cox PHMM model, fitted using the method of penalised partial log-likelihood the significance of a frailty term is assessed using the $-2\log\hat{L}_m$ statistic formed from the maximised marginal log-likelihood. For normal random effects, the maximised marginal log-likelihood for the model with random effect, obtained from maximum likelihood estimation and not REML, can be compared directly with the maximised partial log-likelihood for the corresponding Cox regression model without random effect. A similar process is used when gamma frailty effects are assumed, but here the total number of events, $\sum_{i=1}^{n} \delta_i$, has to be added to the maximised marginal log-likelihood. As with parametric modelling, this test procedure is only approximate, and bootstrapping is required for a more reliable assessment of the magnitude of the random effect (Jiang and Guterman, 2024; Collett, 2023).

3.4 Post-estimation statistics

The martingale and deviance residuals were used to assess the goodness of the fit. Martingale residuals are defined as the difference between the observed event indicator (0 or 1) and the corresponding predicted hazard at each event time (Collett, 2023). These residuals are given by (Therneau et al., 2000; He et al., 2020)

$$m_{ij} = \delta_{ij} - \hat{H}_0(t_{ij})exp(x_{ij}^T\hat{\beta}) \tag{3.28}$$

Martingale residuals are useful for detecting outliers and influential observations. A plot of these residuals versus the covariate x_{ij} reveals discrepancies in the model and the actual functional form of this covariate.

Deviance residuals are calculated based on the logarithm of the likelihood ratio, comparing the likelihood of the observed data under the fitted model to the likelihood of the saturated model (Collett, 2023). It is defined in terms of the martingale residual as follows Collett (2023)

$$d_{ij} = \text{sign}(m_{ij}) \times \{-2 \times [m_{ij} + \delta_{ij}log(\delta_{ij} - m_{ij})]\}^{1/2} \tag{3.29}$$

Deviance residuals are symmetrically distributed with an expected value of 0 if the fitted model is correct).

If the structure of the fitted model is particularly sensitive to one or more observations in the data set, such observations is detected using diagnostics that are designed to highlight observations that influence the complete set of parameter estimates in the risk score. The

influence of observation on the overall fit of the model is examined by which the value of minus twice the logarithm of the maximized likelihood, $-2\log\hat{L}$, under a fitted model, changes when observation, in turn, is left out. Write $\log L(\hat{\beta})$ for the value of the maximized log-likelihood when the model is fitted to all n observations, and $\log L(\hat{\beta}(i))$ for the value of the maximized log-likelihood of the n observations when the parameter estimates are computed after omitting the ith observation from the fit. The diagnostic

$$2\{\log L(\hat{\beta}) - \log L(\hat{\beta}(i))\} \tag{3.30}$$

is useful in the study of influence (Collett, 2023; He et al., 2020).

Chapter 4

Research Results

This chapter illustrates the application of the shared frailty model on clustered women's birth intervals data. It presents the corresponding results and compares the main methods: Cox PH mixed-effects survival model with normal and gamma distributed random effects, Exponential mixed-effects survival model with normal and gamma distributed random effects, and Weibull mixed-effects survival model with normally and gamma distributed random effects.

4.1 Data Summary

Table 1 shows demographic information for the birth interval data and variables considered for the individual level. A total of 3829(20.2%) mothers did not have a second child while 15159(79.8%) mothers had a second child making 18988 participants in the study. The mean ages of mothers who did not have a second child and those with a second child were 19.3 and 18.3 years, respectively. The majority of mothers (74.4%) without a second child and (81.9%) with a second child live in rural places. About 19.2% of mothers without a second child and 19.6% of mothers with a second child were from the northern region, and 34.1% of mothers without a second child and 33.5% of mothers with a second child were from central. The majority of mothers who did not have a second child (56.3%) and

those who had a second child (64.3%) had primary education, and the minority 4.9% of mothers who did not have a second child and those who had a second child 1.6% had tertiary education. About 46.6% of mothers without second child and 42.3% of mothers with second child were rich and 46.6% of mothers and 42.3% of the uncensored were poor. Almost all (95.5%) of the mothers without a second child and 86.1% of the mothers with a second child had a first live child. Around 51.9% of mothers who did not have a second child and 40.5% of the mothers who had a second birth did not utilize contraceptives during the preceding birth. Regarding the sex of the previous child, 50.2% of mothers without a second child and 50.4% of mothers with a second child had a female child as their first child.

4.2 Results for Univariate Analysis

The Kaplan-Meier and log-rank tests were used to analyze univariate associations of the birth intervals for proximate determinants, and socioeconomic and demographic characteristics. Figure 1 shows Kaplan-Meier for the place of residence, region, education level of mother, and status of first child. The estimated survivor functions for place of residence show that the birth interval is shorter for rural women than for urban women. This means that at any time t, the estimated probability of not having a second child beyond t is greater for urban women than for rural women. Because the estimated survival function for urban women does not go to zero, the largest observation was a woman without a second child. The fig. 1 also shows a separation of the survival functions for the three regions. The estimated survival functions for the region are not completely separated. Generally, the pattern of

one survival function lying above another means the group defined by the upper curve has a longer birth interval than the group represented by the lower curve. The statistical question of whether the observed difference in survival curves is statistically significant is tested by log-rank test. Based on a classical log-rank test given in table 1, the P-values were <0.001, <0.001,<0.001, and <0.001 for the place of residence, region, education level of the mother, and status of the first child respectively

The Kaplan-Meier survivor function estimates for the sex of the first child, wealth index, and current contraceptive method, are plotted in section 4.2. This figure shows that the estimated survivor function for those women who are rich and also middle is always greater than that of poor women. The survival curves for the sex of the first child are not separated indicating that there is no difference in birth interval between women who had female first and those who had male first child. The figure also shows that survivor function for the women who do not use contraceptive methods always lies above that for women who use contraceptive methods. Based on a classical log-rank test, the P-values were 0.800, <0.001, and <0.001 for the sex of the first child, wealth index, and current contraceptive method respectively.

From Table 1, the following variables appeared to be significantly related to second successive births and were considered in multivariable analysis: age of mother at first birth, place of residence, region, education level of mother, the status of the first child, wealth index, and contraceptive method.

Table 1: Background characteristics of study respondents (N = 18988) presented in means and percentages for overall data

Variables	Mothers without Second Child	Mothers with Second Child	Overall	Log-rank P-value
	(N=3829)	(N=15159)	(N=18988)	
Age of Mother at 1st birth				
Mean (SD)	19.3 (3.3)	18.3 (3.0)	18.5 (3.1)	
[Min, Max]	[12.0, 40.1]	[8.1, 36.9]	[8.10, 40.1]	
Place of Residence				<0.001
Urban*	979 (25.6%)	2747 (18.1%)	3726 (19.6%)	
Rural	2850 (74.4%)	12412 (81.9%)	15262 (80.4%)	
Region				<0.001
Northern*	736 (19.2%)	2977 (19.6%)	3713 (19.6%)	
Central	1307 (34.1%)	5072 (33.5%)	6379 (33.6%)	
Southern	1786 (46.6%)	7110 (46.9%)	8896 (46.9%)	
Education				<0.001
No Education*	191 (5.0%)	2465 (16.3%)	2656 (14.0%)	
Primary	2162 (56.5%)	9815 (64.7%)	11977 (63.1%)	
Secondary	1287 (33.6%)	2630 (17.3%)	3917 (20.6%)	
Tertiary	189 (4.9%)	249 (1.6%)	438 (2.3%)	
Sex of Child				0.800
Male*	1923 (50.2%)	7637 (50.4%)	9560 (50.3%)	
Female	1906 (49.8%)	7522 (49.6%)	9428 (49.7%)	
Death of First Child				<0.001
Dead*	173 (4.5%)	2108 (13.9%)	2281 (12.0%)	
Alive	3656 (95.5%)	13051 (86.1%)	16707 (88.0%)	
Wealth Index				<0.001
Poor*	1405 (36.7%)	5812 (38.3%)	7217 (38.0%)	
Middle	640 (16.7%)	2940 (19.4%)	3580 (18.9%)	
rich	1784 (46.6%)	6407 (42.3%)	8191 (43.1%)	
Contraceptive Method				<0.001
No Method*	1986 (51.9%)	6139 (40.5%)	8125 (42.8%)	
Traditional Method	25 (0.7%)	176 (1.2%)	201 (1.1%)	
Modern method	1818 (47.5%)	8844 (58.3%)	10662 (56.1%)	

* Reference variable

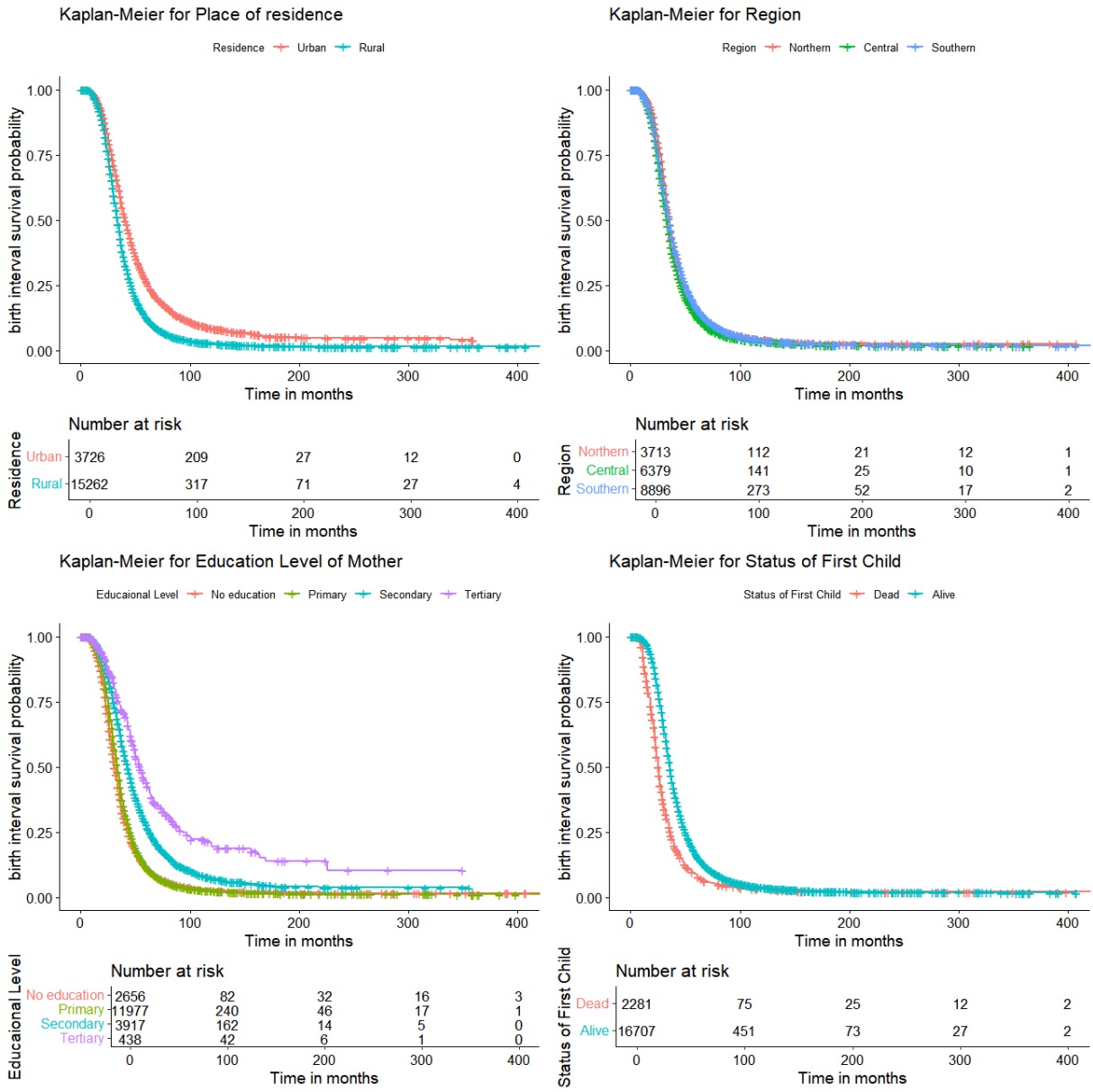

Figure 1: Plots of the Kaplan-Meier curve estimates comparing waiting time to second childbirth across levels of mother characteristics.

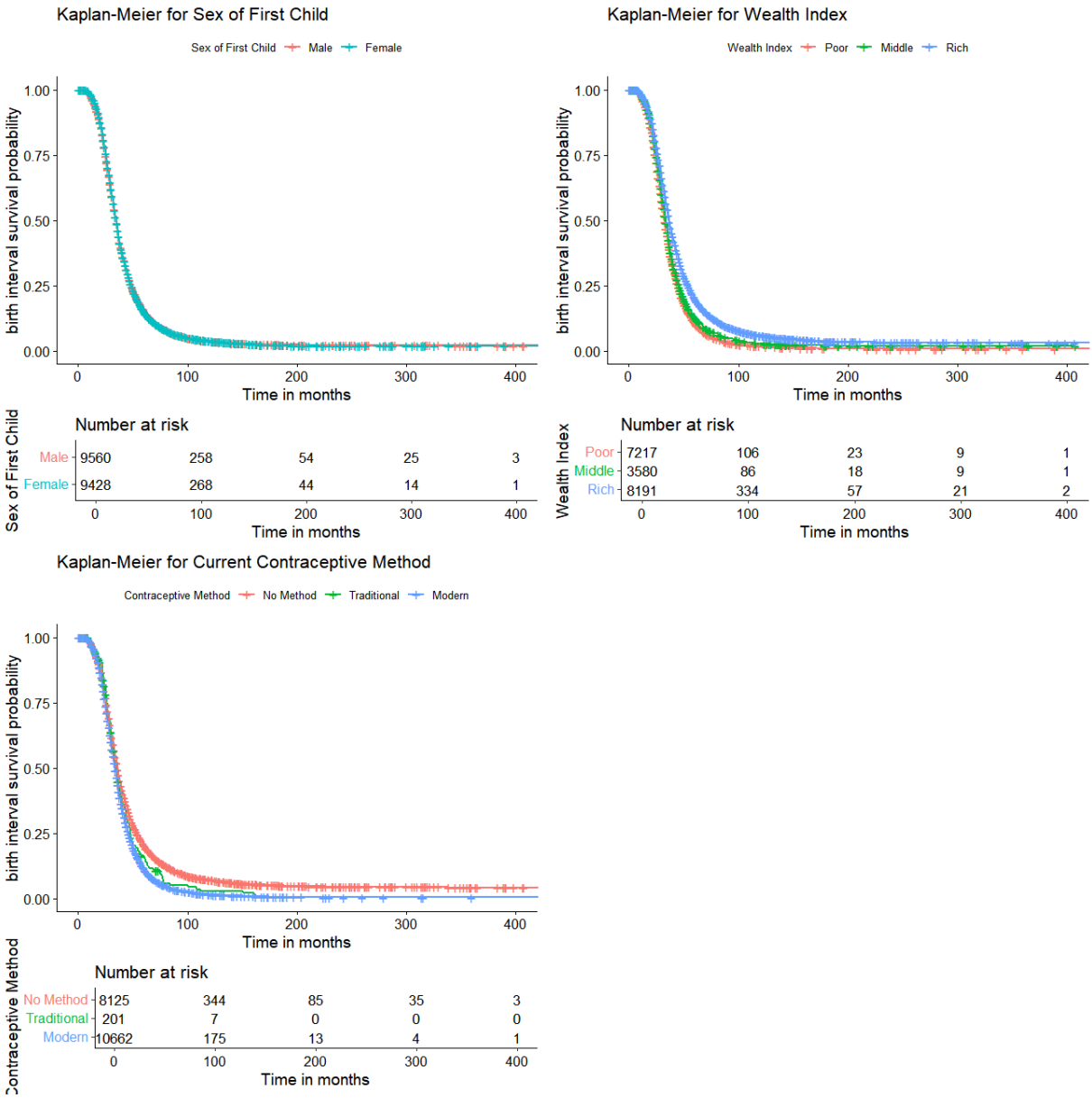

Figure 2: Plots of the Kaplan-Meier curve estimates comparing waiting time to second childbirth across levels of mother characteristics.

4.3 Results for multivariable analysis

To avoid making specific assumptions about the form of the underlying baseline hazard function, a Cox PHMM regression model was fitted that includes the mother's age at first birth, place of residence, region, education level of the mother, the status of the first child, wealth index, and contraceptive method. To test the hypothesis that all random effects are zero in the presence of the explanatory variables, the value of the $-2\log\hat{L}$ statistic for the fitted Cox model, which is 266734 compared with values of the maximized marginal log-likelihood statistic, $-2\log\hat{L}_m$, for Cox models with random effect.

When a normal random effect term is added, the value of the maximized marginal log-likelihood statistic, $-2\log L_m(\hat{\sigma}_b^2)$, is 266358.8, and this leads to a change in the value of the $-2\log\hat{L}_m$ statistic of 375.2, which is significant (P = <0.001). On fitting models with gamma random effects, a likelihood ratio test for the frailty is the difference between the twice log partial-likelihood with the random effect term integrated out which is 266723, and the twice loglikelihood of a no-frailty model. The results is 11 and it has one degree of freedom with a p-value = 0.001.

To compare the parameter estimates for the various terms and their standard errors, Table 2 shows the estimates for the fitted Cox models, including the estimated variances of a normal random effect and gamma random effects. The Cox PHMM with normal random effects chooses the estimate of the random effect using a REML criterion. REML method has a slightly larger estimated random effect variance of 0.014 as compared to the gamma random effect variance of 0.005. There is a slight difference in the hazard ratio and standard

error obtained when either a normal or gamma random effect is included.

An exponential proportional hazards model was fitted that contains the mother's age at first birth, place of residence, region, education level of the mother, the status of the first child, wealth index, and contraceptive method. The value of the $-2\log\hat{L}$ statistic for this model is 146070.1. A normally distributed random effect is introduced, by adding it to the linear component of the model and maximizing the likelihood function. The Exponential model with normal random effect has a$-2\log\hat{L}$ value of 146117.4, which is larger than exponential model without random effects by 47.2. This means that there is no substantial heterogeneity between the birth intervals of the women. The variance of the normal random effect corresponding to the exponential is 0.009. When the gamma random effect is introduced, the corresponding variance is 0.008. The exponential model with gamma random effects has a $-2\log\hat{L}$ value of 146070, and so the addition of the frailty term leads to a reduction in the value of the 146070 statistic of 0.1. This change is not significant (P = 0.751).

On fitting a Weibull model containing age of mother at first birth, place of residence, region, education level of mother, status of first child, wealth index, and contraceptive method, the value of the $-2\log\hat{L}$ statistic is 139670.2. To take account of cluster effects, a normal shared random effect is added, and the $-2\log\hat{L}$ statistic decreases to 138942.3, and the change in $-2\log\hat{L}$ on adding the random effect is 727.9. Comparing this reduction with percentage points of a χ^2_1 distribution, the random effect is highly significant, with a P-value of less than 0.001. This means that there is substantial heterogeneity between the birth intervals of the women in this study. A less conservative estimate of the significance

Table 2: Maximum likelihood estimates from Cox survival mixed effects models with normal and gamma distributed random effects for the effects of women characteristics on birth interval, 2015-16 MDHS.

	Cox PHMM with Normal RE				**Cox PHMM with Gamma RE**			
Covariates	**HR**	**SE**	**95% CI**	**P-val**	**HR**	**SE**	**95% CI**	**P-val**
Age of Mother	0.998	0.003	(0.991, 1.004)	0.380	0.998	0.003	(0.993, 1.003)	0.334
Place of Residence								
Urban*								
Rural	1.261	0.026	(1.203, 1.321)	<0.001	1.264	0.025	(1.204, 1.327)	<0.001
Region								
Northern*								
Central	1.083	0.026	(1.032, 1.136)	0.002	1.084	0.025	(1.033, 1.138)	0.002
Southern	0.926	0.025	(0.886, 0.968)	0.003	0.927	0.023	(0.886, 0.971)	0.003
First Child Status								
Dead*								
Live	0.613	0.024	(0.578, 0.651)	<0.001	0.611	0.024	(0.583, 0.641)	<0.001
Level of Education								
No education*								
Primary	0.945	0.023	(0.896, 0.996)	0.009	0.943	0.023	(0.901, 0.987)	0.011
Secondary	0.637	0.031	(0.597, 0.680)	<0.001	0.636	0.031	(0.599, 0.676)	<0.001
Tertiary	0.459	0.070	(0.400, 0.527)	<0.001	0.458	0.070	(0.399, 0.524)	<0.001
Wealth Index								
Poor*								
Middle	0.968	0.023	(0.925, 1.013)	0.150	0.965	0.023	(0.923, 1.009)	0.143
Rich	0.948	0.022	(0.909, 0.989)	0.013	0.945	0.021	(0.906, 0.985)	0.016
Contraceptive Method								
No Method*								
Traditional	1.272	0.077	(1.093, 1.481)	0.002	1.271	0.077	(1.093, 1.477)	0.002
Modern	1.350	0.017	(1.306, 1.396)	<0.001	1.346	0.017	(1.302, 1.392)	<0.001
Variance	0.014			<0.001	0.005			0.001
$-2\hat{L}$	266358.8				266724			

(*) = Reference category, HR = Hazard ratio, SE = Standard error, CI = Confidence interval, P-val = P-value

Table 3: Maximum likelihood estimates from Exponential survival mixed effects models with normal and gamma distributed random effects for the effects of women characteristics on birth interval, 2015-16 MDHS.

	Exponential MM with Normal RE				**Exponential MM with Gamma RE**			
Covariates	**HR**	**SE**	**95% CI**	**P-val**	**HR**	**SE**	**95% CI**	**P-val**
Age of Mother	0.995	0.003	(0.990, 0.999)	0.037	0.995	0.003	(0.990, 0.999)	0.042
Place of Residence								
Urban*								
Rural	1.174	0.025	(1.118, 1.234)	<0.001	1.175	0.024	(1.119, 1.235)	<0.001
Region								
Northern*								
Central	1.042	0.025	(0.992, 1.096)	0.103	1.042	0.024	(0.992, 1.095)	0.08
Southern	0.955	0.024	(0.911, 1.001)	0.053	0.955	0.022	(0.911, 1.001)	0.033
First Child Status								
Dead*								
Live	0.749	0.024	(0.715, 0.785)	<0.001	0.749	0.024	(0.715, 0.785)	<0.001
Level of Education								
No education*								
Primary	0.955	0.023	(0.913, 0.999)	0.048	0.957	0.023	(0.914, 1.001)	0.053
Secondary	0.720	0.031	(0.678, 0.765)	<0.001	0.719	0.030	(0.677, 0.764)	<0.001
Tertiary	0.539	0.070	(0.470, 0.619)	<0.001	0.539	0.069	(0.470, 0.619)	<0.001
Wealth Index								
Poor*								
Middle	0.975	0.023	(0.932, 1.020)	0.277	0.976	0.023	(0.933, 1.021)	0.204
Rich	0.967	0.021	(0.927, 1.008)	0.112	0.965	0.021	(0.926, 1.007)	0.072
Contraceptive Method								
No Method*								
Traditional	1.261	0.077	(1.084, 1.467)	0.003	1.261	0.077	(1.084, 1.467)	0.002
Modern	1.280	0.017	(1.238, 1.323)	<0.001	1.280	0.017	(1.239, 1.323)	<0.001
Variance	0.009	0.007			0.008	0.004		
λ	0.027	0.002			0.026	0.002		
$-2\hat{L}$	146117.4				146070			

(*) = Reference category, HR = Hazard ratio, SE = Standard error, CI = Confidence interval, P-val = P-value

of the effect is found by referring the change in $-2\log\hat{L}$ to a $0.5(\chi_0^2 + \chi_1^2)$ distribution, equivalent to halving the P-value from a comparison with a χ_1^2 distribution, and it is then even more significant. The variance of the normal random effect is 0.009.

When a Weibull model is fitted together with a gamma random effect, the gamma random effect has variance $\hat{\theta} = 0.008$. The Weibull model with gamma random has a $-2\log\hat{L}$ value of 138900.8, and so addition of the frailty term leads to a reduction in the value of the $-2\log\hat{L}$ statistic of 769.2. This change is highly significant (P < 0.001), and greater than that when a normal random effects is introduced.

As with the Weibull mixed effects model, the standard errors are slightly larger than in Cox PHMM and exponential model, but the hazard ratio estimates in table 2, table 3, and table 4 are broadly similar. As for the Weibull model, the gamma frailty model leads to the smallest value of the $-2\log\hat{L}$ statistic and is a best fit to the data than five models. Therefore, the post-estimation and results interpretation are based on the Weibull regression model with gamma distributed random effects.

In the presence of gamma random effect, a hazard ratio (HR) for the mother's age at first birth is 0.995. However, the 95% confidence interval for this estimate ranges from 0.990 to 1.001, reflecting the much greater uncertainty about the age at first birth effect when account is taken of gamma random effect. The hazard ratio for the woman whose first child is 0.709 which is significant. This shows that a woman whose first child is alive has a 29.1% lower hazard of having a second child compared to a woman whose first child is dead. This shows that women with a deceased first child went on to have a second

Table 4: Maximum likelihood estimates from Weibull survival mixed effects models with normal and gamma distributed random effects for the effects of women characteristics on birth interval, 2015-16 MDHS.

	Weibull MM with Normal RE				**Weibull MM with Gamma RE**			
Covariates	**HR**	**SE**	**95% CI**	**P-value**	**HR**	**SE**	**95% CI**	**P-val**
Age of Mother	0.995	0.003	(0.990, 1.001)	0.093	0.995	0.003	(0.990, 1.001)	0.067
Place of Residence								
Urban*								
Rural	1.357	0.039	(1.258, 1.463)	<0.001	1.356	0.052	(1.258, 1.461)	<0.001
Region								
Northern*								
Central	1.116	0.041	(1.030, 1.210)	0.007	1.118	0.046	(1.032, 1.212)	0.007
Southern	0.933	0.038	(0.865, 1.007)	0.075	0.934	0.036	(0.866, 1.009)	0.082
First Child Status								
Dead*								
Live	0.715	0.026	(0.680, 0.752)	<0.001	0.709	0.018	(0.674, 0.746)	<0.001
Level of Education								
No education*								
Primary	1.048	0.025	(0.998, 1.101)	0.060	1.043	0.026	(0.993, 1.096)	0.096
Secondary	0.731	0.033	(0.685, 0.780)	<0.001	0.725	0.024	(0.680, 0.773)	<0.001
Tertiary	0.500	0.074	(0.432, 0.578)	<0.001	0.497	0.037	(0.430, 0.574)	<0.001
Wealth Index								
Poor*								
Middle	0.938	0.024	(0.895, 0.984)	0.008	0.938	0.023	(0.895, 0.984)	0.009
Rich	0.895	0.023	(0.856, 0.937)	<0.001	0.897	0.021	(0.857, 0.938)	<0.001
Contraceptive Method								
No Method*								
Traditional	1.489	0.080	(1.274, 1.741)	<0.001	1.487	0.118	(1.270, 1.739)	<0.001
Modern	1.588	0.018	(1.532, 1.645)	0.001	1.582	0.029	(1.527, 1.640)	<0.001
Variance	0.116	0.009		<0.001	0.110	0.008		<0.001
λ	0.001	0.0001			0.001	0.0001		
α	1.798	0.011			1.805	0.011		
$-2\hat{L}$	138942.3				138900.8			

(*) = Reference category, HR = Hazard ratio, SE = Standard error, CI = Confidence interval, P-val = P-value

birth faster. This study shows that mothers with secondary and tertiary levels of education had a lower hazard rate of 0.725 and 0.497 respectively compared to those with no formal education. There is no significant difference in birth interval for women with primary education and those with no formal education. The place of residence of mothers who reside in the urban areas was set as the reference category. Comparing the rural to urban, it can be seen that, rural mothers have a hazard ratio of 1.117. Hence rural mothers have increased birth intensity (shorter birth interval) as compared to urban mothers. Women from the central region have shorter birth intervals than women from the central with a hazard ratio of 1.118. There is no significant difference in birth interval between women from the northern and southern regions. This study also shows that there is a significantly lower hazard rate for mothers who are rich and middle compared to those who are poor. Finally, the study shows that the hazard rate of having a second child is higher in women who are currently using modern and traditional contraceptive methods when compared to mothers who are not using contraceptive methods. These estimated hazard ratios are conditional on shared frailty, and so refer to a woman with a specific random effect value. For Weibull with gamma random effect, the hypothesis that variance $\theta = 0$ can be rejected, meaning the effects of unobserved heterogeneity are not negligible. This parameter was estimated taking into consideration the variation between enumeration areas in Malawi. The shared parameter of the women indicates that those grouped into the same cluster or area have similar characteristics or share the same birth interval but differ between or from cluster to cluster. In other words, the probability of having a second child may be similar within a cluster but different between clusters due to some characteristics that were not or could not be measured.

4.4 Assessment of model fit

The fitness of the Weibull mixed-effects model with gamma-distributed random effects and outliers is examined from the martingale and deviance residuals plots. Figure 3 illustrates the martingale residuals against covariates. This plot displays the functional form required for the covariate helped by superimposing a smoothed curve that is fitted to the scatterplot. The residual scatter ranges between -24 to 1. The residuals that are close to unity correspond to women who have shorter birth intervals than estimated by the model. Meanwhile, negative residuals indicate that women have long birth intervals yet are expected to have a second child earlier. The smoothed curve indicates that there is no need for anything other than a linear term in covariates.

Furthermore, the plot of the deviance residual against linear predictor(risk score variable) is presented fig. 4. The risk score variable for every sample is constructed based on the given covariates. This is done by multiplying the covariate value with the corresponding regression coefficient estimated from the full likelihood approach and then taking the sum over the products. The plot that the deviance residual against the linear predictor has a distinct pattern in that the deviance residual values tend to be more dispersed downward with the increased values of the risk score. This shows that there are influential observations in the data.

Upon identifying the influential observations to the model, their impact on regression parameter estimates is analyzed by re-fitting the model to data without the detected influential observations and observing the changes in parameter estimates and $-2\log\hat{L}_m$. This is done

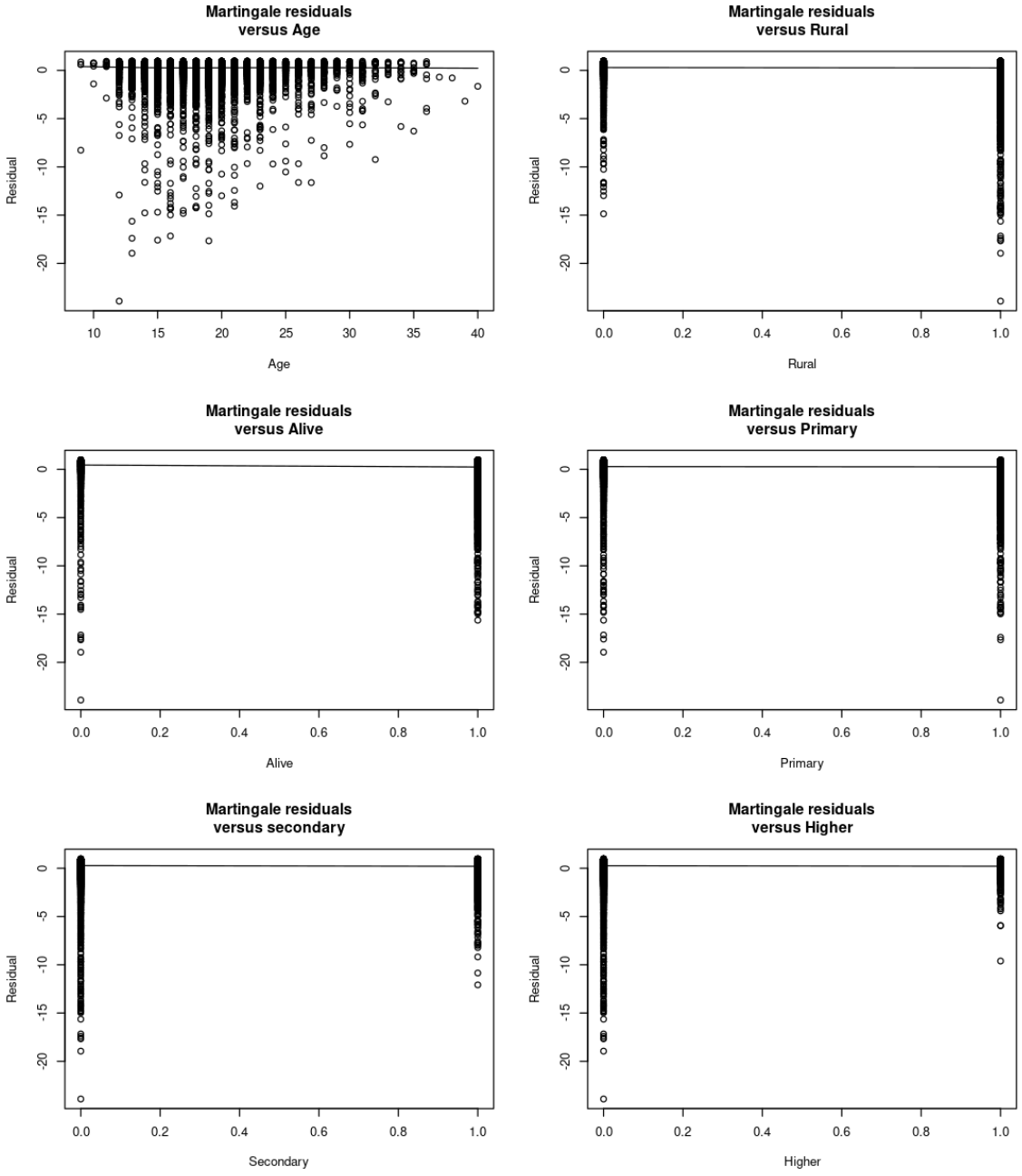

Figure 3: Plot of the Martingale residuals against explanatory variables

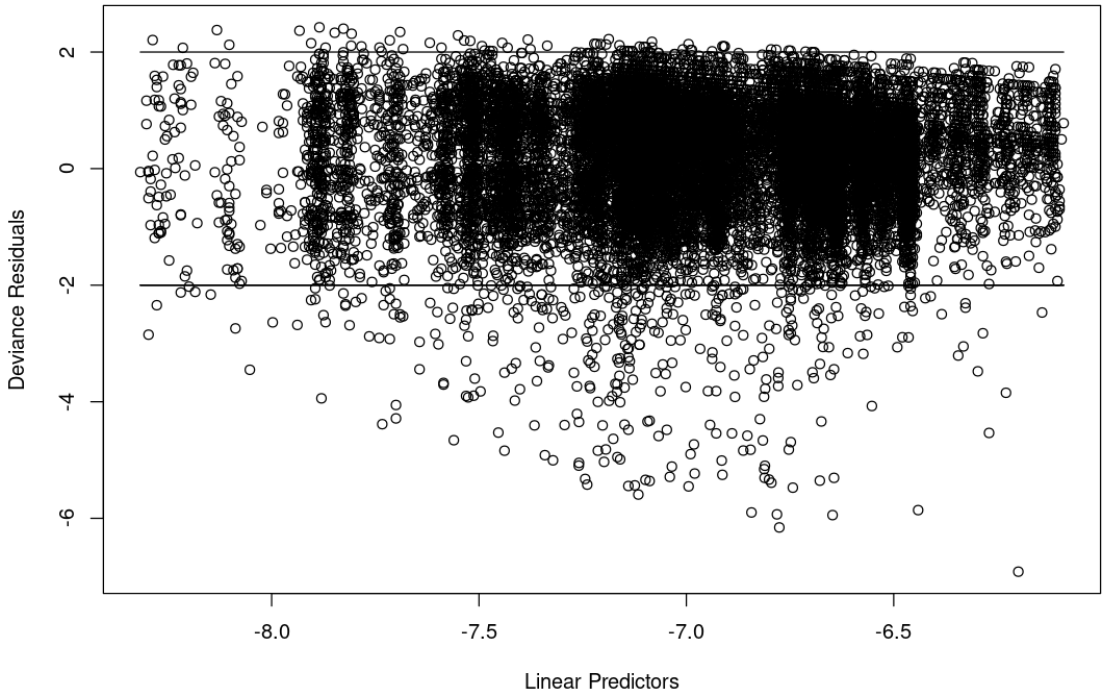

Figure 4: Plot of the deviance residuals against linear predictor

at a cutoff of ± 3 and ± 2 in which 160 and 463 observations are dropped respectively. This is done to make residuals roughly symmetrically distributed around the zero line.

The Table 5 shows results for the Weibull model with gamma random effects without identified influential observations. The findings show that the removal of 160 influential observations from analysis resulted in a reduction in $-2\log\hat{L}_m$ by 5558.5. There is also a reduction in the variance of the gamma random effect. The estimated hazard ratio of the mother's age at first birth is significant and the hazard ratio for wealth index is not significant which can be seen in Table 5 when fitting the model without observation 160. Removing 463 influential observations from analysis resulted in further reduction $-2\log\hat{L}_m$ in $-2\log\hat{L}_m$ by 10127.5. The variance of gamma random effects also got lower when 463 influential observations were removed. It is shown in Table 5 that the impact of 463 influential observations on the estimate of predictors of the birth interval was so huge compared to that of 163 influential observations.

Table 5: Maximum likelihood estimates from Weibull survival mixed effects models with gamma-distributed random effects without influential observations for the effects of women characteristics on birth interval, 2015-16 MDHS.

	Without 160 Observations				Without 463 Observations			
Covariates	**HR**	**SE**	**95% CI**	**P-value**	**HR**	**SE**	**95% CI**	**P-value**
Age of Mother	0.999	0.003	(0.994, 1.005)	0.0791	1.010	0.003	(1.005, 1.015)	<0.001
Place of Residence								
Urban*								
Rural	1.422	0.029	(1.335, 1.516)	<0.001	1.377	0.042	(1.295, 1.463)	<0.001
Region								
Northern*								
Central	1.082	0.036	(1.013, 1.156)	0.019	1.098	0.036	(1.030, 1.670)	0.004
Southern	0.905	0.029	(0.850, 0.963)	0.002	0.904	0.028	(0.852, 0960)	0.001
First Child Status								
Dead*								
Live	0.570	0.014	(0.543, 0.599)	<0.001	0.526	0.013	(0.501, 0.552)	<0.001
Level of Education								
No education*								
Primary	0.933	0.027	(0.890, 0.979)	0.005	0.948	0.023	(0.904, 0.995)	0.030
Secondary	0.580	0.019	(0.545, 0.618)	<0.001	0.573	0.018	(0.538, 0.610)	<0.001
Tertiary	0.329	0.024	(0.285, 0.380)	<0.001	0.327	0.024	(0.283, 0.379)	<0.001
Wealth Index								
Poor*								
Middle	0.981	0.023	(0.937, 1.029)	0.441	0.981	0.023	(0.935, 1.026)	0.389
Rich	0.953	0.022	(0.912, 0.996)	0.034	0.975	0.022	(0.933, 1.019)	0.258
Contraceptive Method								
No Method*								
Traditional	1.226	0.098	(1.048, 1.434)	0.011	1.270	0.102	(1.085, 1.486)	<0.003
Modern	1.418	0.024	(1.370, 1.467)	0.001	1.369	0.023	(1.323, 1.416)	<0.001
Variance	0.056	0.006		<0.001	0.046	0.005		<0.001
λ	0.0004	0.0001			0.00002	0.0001		
α	2.13	0.015			2.352	0.014		
$-2\hat{L}$	133342.3				128773.3			

(*) = Reference category, HR = Hazard ratio, SE = Standard error, CI = Confidence interval, P-val = P-value

Chapter 5

Discussion and Conclusion

5.1 Discussion of results

Short or very long birth intervals have consequences for maternal and child health as well as population fertility rate and growth. So, it is imperative to regularly study variations and determinants of the length of birth intervals, especially in a country like Malawi, which has been experiencing below-replacement fertility levels over the past two decades (Kopp et al., 2018). Using recent data from the 2015-2016 Malawi Demographic Health Survey, this study studied differentials in the length of birth intervals using a survival mixed effects model. The major advantage of using survival models in analyzing birth interval data, especially first birth, is using all of the data, even women who are censored. But, in other statistical methods, these data must be excluded. In this study, the percentage of censoring was 20.2%.

Under the Cox PHMM, the factors that increased birth intensity (shorter birth interval) were: rural residence, central region, and contraceptive region. southern region, live first child, education level, rich women have the opposite effect of decreasing birth intensity (longer birth interval) as a result of the less than hazard ratio under the Cox model. The age of the mother and middle wealth index were not significant predictors of the birth interval

as their confidence interval contains one. The Cox PHMM with normal random effect has $-2\log\hat{L}$ of 266358.8 providing a better fit than the Cox PHMM with gamma random effects which has $-2\log\hat{L}$ of 266724. In both the Cox PHMM with gamma and normal random effects inclusion of random effects were all significant.

The results of fitting parametric exponential mixed effects models show stronger effects on the education level of the mother, age of the mother at first birth, and live first child, (increase of birth intervals with increase in education and mother's age at first birth) and rural residence, and contraceptive method have shorter birth intervals. For the region of residence, only exponential with gamma random effect shows a significant difference in birth intervals between the southern and northern regions of residence. For both exponential mixed effects models, the hypothesis that between-cluster variance is zero cannot be rejected. This means that the effects of unobserved heterogeneity are negligible in exponential models with normal and gamma random effects. The -2 log-likelihood suggests that the exponential mixed effect model with gamma random effects provides a better fit than exponential models with normal random effects.

Also, fitting parametric Weibull models with gamma and normal random effects shows that mother‘s education and rich women, southern region, and live first birth residence decrease birth intensity (longer birth interval) while rural residence, central region, and contraceptive use gives the opposite effect of increasing birth intensity (shorter birth interval). The age of the mother at first birth and middle wealth index are insignificant in determining birth interval. In both the Weibull models with gamma and normal random effects inclusion of random effects were all significant. The Weibull model with gamma random effects is a

better fit compared to the Weibull model with normal random effects. Of all six models, the Weibull model with gamma random effects is the best-fitted model because this model gives the smallest $-2\log\hat{L}$ value.

In the Weibull model with gamma random effect, the survival status of the previous child emerged as a significant protective factor for the birth intervals. This study finds that the hazard of having a second child is 29.1% lower for women whose first child is alive compared to those whose first child is dead. This can be explained by the fact that mothers who experience the death of their first child might feel a stronger urge to have another child sooner, driven by grief and the desire to fill the void left by the loss. Mothers with a healthy first child might choose to space their children further apart to provide adequate care and attention. This finding is in line with many other studies like Erfani et al. (2018); Bagheri and Saadati (2021).

The hazard rates of having short birth intervals were higher among mothers who had no formal education as compared to their women counterparts who attended secondary and tertiary education. This finding is consistent with evidence from a study conducted in (Bagheri and Saadati, 2021; Seyedtabib et al., 2020). This can be partly explained by the fact that educated women are well-informed about optimal healthcare choices and have greater autonomy to make decisions and use quality healthcare services.

Place of residence had a significant effect on the second birth interval. Mothers from rural have a 35.6% higher hazard of having a second child compared to urban mothers. This is consistent with the study by Erfani et al. (2018) who showed that women who lived in

completely- developed regions in Tehran have their second child later than ones who lived in developing regions. This can be explained by the fact that urban mothers have better access to healthcare services, including family planning and reproductive health services, allowing women to better plan their pregnancies than rural mothers.

The wealth index of the mother was also a strong predictor of birth interval. The significantly lower hazard rate for wealthier mothers compared to poorer mothers can be explained by the fact that wealthier mothers have better access to healthcare, including reproductive health services and family planning. This access allows them to plan and space their pregnancies more effectively. this is consistent with evidence from a study done by Shifti et al. (2023). In contrast, a study conducted in Yohannes et al. (2011) showed that the length of birth interval increased with increasing wealth index.

A higher hazard rate for women using contraceptives compared to those not using them suggests these women might be having second children sooner than those not using any contraception. This can be partly explained by the fact women might use contraceptives to temporarily delay their next pregnancy to achieve the desired spacing between children. Once they decide to have a second child, they discontinue use, leading to a concentrated period of higher fertility. This is inconsistent with evidence from studies done by Erfani et al. (2018); Bagheri and Saadati (2021) who showed that mothers who did not utilize modern contraceptive methods before they got pregnant with their last child were more likely to experience short birth intervals than those who used one.

The development of the Weibull model with gamma random effect is continued by checking

the adequacy of the model using martingale and deviance residuals. In this study, the martingale residuals plot against the explanatory variable indicates a linear term in covariates. He et al. (2020); Kalbfleisch and Schaubel (2023); Dauda (2022) are examples of studies that used the martingale residual plot to compare the performance of semi-parametric and parametric models. The plot of deviance residuals against linear predictor indicates the inadequacy of the model as the scatter plot is not symmetric around the zero line. One possible reason for the poor fit is a high level of influential observation He et al. (2020). This is supplemented by analyzing the impact of influential observation on regression parameter estimates by re-fitting the model to data without the influential observation and observing the changes in parameter estimates.

5.2 Study limitations

In this study only the gamma and the normal densities have been used as random effects distribution. It might be of interest to consider other random effect distributions like positive stable, compound poisson and inverse gaussian in future studies. This study also used martingale and deviance residual to assess model adequacy although other residual like Cox-snell residuals could been used. Another type of frailty modeling, which was not addressed in this thesis is multi-level frailty modeling. A multi-level frailty model is when more than one frailty term is used in the model.

5.3 Conclusions

The primary objective of this project was to apply parametric and semiparametric survival mixed effects regressions (shared frailty models) to model the length of time to successive second births within the reproductive health of mothers in Malawi. Specifically, Cox PH mixed-effects survival model with normal and gamma distributed random effects, Exponential mixed-effects survival model with normal and gamma distributed random effects, and Weibull mixed-effects survival model with normal and gamma distributed random effects were fitted to observed birth interval data with an objective to compare the parameter estimates and assess the estimated heterogeneity parameters. Weibull mixed-effects survival model with gamma distributed random effects were found to be a better fit for describing the birth interval data than other models.

The data used in this study were the 2015-16 Malawi demographic and health survey (MDHS) data, where multi-stage cluster sampling was used as the sampling design. The clusters were defined as 850 enumeration area and it is assumed that birth intervals for the women from the same cluster are correlated because they share the same environment. This study shows that some important covariates, which have been used in the models and varied among the birth intervals. The study have found that the covariates place of residence, mothers education, wealth index, current contraceptive use and survival status of previous birth are significantly associated with birth interval. In this study, the observations within specific area are found to be correlated.

5.4 Recommendations

Most studies on birth spacing have been done using various methodologies to arrive at their conclusions. Timing in birth spacing is very crucial as incorrect information in terms of time interval between births can lead to wrong conclusions In view of this, the following recommendations are made:

a) Birth spacing has to do with timing which is a continuous variable. Dichotomization of the time between successive births may make interpretation of the model in terms of odd ratios easy but will end up with loss of information. This is because a continuous variable such as time is converted into a binary outcome.

b) For correct conclusions one has to analyze the original time until birth using survival models. These models allow for the incorporation of censoring which do not discard censored times in the analysis as would be the case if traditional linear models are used.

c) If birth interval is clustered within mothers, responses to the overall effect of the covariate on the birth intensity cannot be assumed to be independent. Hence using the cox regression will lead to erroneous conclusions. To address such a question one should consider using a survival mixed-effects model which allows for the correction of the effects of unobserved heterogeneity. The frailty model allows for the introduction of an observed frailty term that relaxes the independent assumption.

d) After a model has been fitted to an observed set of survival data, the adequacy of

the model needs to be assessed. Indeed, the use of diagnostic procedures for model checking is an essential part of the modelling process.

e) Citizens of Malawi must be educated on the birth interval. They must be made aware of its wide-ranging socio-economic, socio cultural and demographic effects on health issues such as birth spacing.

f) Health institutions in Malawi should not relent in their efforts to intensify education in maternity and child health in order to reduce drastically maternal and infant mortality.

Bibliography

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in medicine*, 7(11):1121–1137.

Ahammed, B., Kabir, M. R., Abedin, M. M., Ali, M., and Islam, M. A. (2019). Determinants of different birth intervals of ever married women: Evidence from bangladesh. *Clinical Epidemiology and Global Health*, 7(3):450–456.

Aleye, B., Usso, A. A., Mengistie, B., Dessie, Y., Adem, H. A., Alemu, A., Yuya, M., and Mohammed, A. (2024). Determinants of short birth interval among married multiparous women in chinaksen district, eastern ethiopia: a case-control study. *Frontiers in Global Women's Health*, 4:1278777.

Alvares, D., Lázaro, E., Gómez-Rubio, V., and Armero, C. (2021). Bayesian survival analysis with bugs. *Statistics in Medicine*, 40(12):2975–3020.

Aychiluhm, S. B., Tadesse, A. W., Mare, K. U., Abdu, M., and Ketema, A. (2020). A multilevel analysis of short birth interval and its determinants among reproductive age women in developing regions of ethiopia. *Plos one*, 15(8):e0237602.

Azevedo, D. R., Prates, M. O., and Bandyopadhyay, D. (2023). Alleviating spatial confounding in frailty models. *Biostatistics*, 24(4):945–961.

Bagheri, A. and Saadati, M. (2021). Determinants of birth intervals using prentice-williams-peterson-gap time model: Tehran case study. *International Journal of Fertility & Sterility*, 15(3):234.

Ball, S. J., Pereira, G., Jacoby, P., De Klerk, N., and Stanley, F. J. (2014). Re-evaluation of link between interpregnancy interval and adverse birth outcomes: retrospective cohort study matching two intervals per mother. *Bmj*, 349.

Bardo, M. and Unkel, S. (2023). The shape of the relative frailty variance induced by discrete random effect distributions in univariate and multivariate survival models. *arXiv preprint arXiv:2303.04915*.

Blake, J. (2022). *Family size and achievement*, volume 3. Univ of California Press.

Bongaarts, J. and Hodgson, D. (2022). *Fertility transition in the developing world*. Springer Nature.

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in r. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920960351.

Cheung, C. C., Vittinghoff, E., Marcus, G. M., and Gerstenfeld, E. P. (2021). Beware of the hazards: limitations of the proportional hazards assumption. *EP Europace*, 23(12):2048–2048.

Class, Q. A., Rickert, M. E., Oberg, A. S., Sujan, A. C., Almqvist, C., Larsson, H., Lichtenstein, P., and D'Onofrio, B. M. (2017). Within-family analysis of interpregnancy interval and adverse birth outcomes. *Obstetrics and gynecology*, 130(6):1304.

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(1):141–151.

Collett, D. (2023). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Conde-Agudelo, A., Rosas-Bermúdez, A., and Kafury-Goeta, A. C. (2007). Effects of birth spacing on maternal health: a systematic review. *American journal of obstetrics and gynecology*, 196(4):297–308.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, D. R. (1992). Regression models and life-tables. breakthroughs in statistics. *Stat. Soc*, 372:527–541.

Cox, D. R. (2018). *Analysis of survival data*. Chapman and Hall/CRC.

Daniel, K., Onyango, N. O., and Sarguta, R. J. (2021). A spatial survival model for risk factors of under-five child mortality in kenya. *International Journal of Environmental Research and Public Health*, 19(1):399.

Dauda, K. (2022). Development and modeling of decision tree for survival data with multiple events using deviance and cox-snell residuals within node homogeneity technique. *International Journal of Advanced Health Science and Technology*, 2(3):183–190.

Duchateau, L. and Janssen, P. (2008). *The frailty model*. Springer.

Emmert-Streib, F. and Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038.

Erfani, A., Nojomi, M., and Hosseini, H. (2018). Prolonged birth intervals in hamedan, iran: variations and determinants. *Journal of biosocial science*, 50(4):457–471.

Exavery, A., Mrema, S., Shamte, A., Bietsch, K., Mosha, D., Mbaruku, G., and Masanja, H. (2012). Levels and correlates of non-adherence to who recommended inter-birth intervals in rufiji, tanzania. *BMC pregnancy and childbirth*, 12:1–8.

Fagbamigbe, A. F., Karlsson, K., Derks, J., and Petzold, M. (2021). Performance evaluation of survival regression models in analysing swedish dental implant complication data with frailty. *PLoS One*, 16(1):e0245111.

Fallahzadeh, H., Farajpour, Z., and Emam, Z. (2013). Duration and determinants of birth interval in yazd, iran: a population study. *Iranian journal of reproductive medicine*, 11(5):379.

Feleke, B., Tesfaw, L. M., and Mitku, A. A. (2022). Survival analysis of women breast cancer patients in northwest amhara, ethiopia. *Frontiers in Oncology*, 12:1041245.

Frévent, C., Ahmed, M.-S., Dabo-Niang, S., and Genin, M. (2024). A shared-frailty spatial scan statistic model for time-to-event data. *Biometrical Journal*, 66(5):e202300200.

Gelagay, A. A., Worku, A. G., Bashah, D. T., Tebeje, N. B., Gebrie, M. H., Yeshita, H. Y., Cherkose, E. A., Ayana, B. A., Lakew, A. M., Asmamaw, D. B., et al. (2023). Factors affecting birth interval among mothers in dabat district, amhara regional state, northwest ethiopia: A community-based cross-sectional study, 2022. *Heliyon*, 9(6).

Ghavami, V., Mahmoudi, M., Foroushani, A. R., Baghishani, H., Yaseri, M., and Shandiz, F. H. (2021). A competing risks cure frailty model: An application to relapse-free

survival of breast cancer patients. *Pakistan Journal of Statistics and Operation Research*, pages 591–605.

Greenhalgh, S. (2023). Controlling births and bodies in village china. In *People's Republic of China, Volumes I and II*, pages Vol1–121. Routledge.

Gu, D., Andreev, K., and Dupre, M. E. (2021). Major trends in population growth around the world. *China CDC weekly*, 3(28):604.

Hailu, D., Gulte, T., et al. (2016). Determinants of short interbirth interval among reproductive age mothers in arba minch district, ethiopia. *International journal of reproductive medicine*, 2016.

Hajian-Tilaki, K., Asnafi, N., and Aliakbarnia-Omrani, F. (2009). The patterns and determinants of birth interval in multiparous women in babol, northern iran. *Southeast Asian journal of tropical medicine and public health*, 40(4):852.

He, B., Liu, Y., Wu, Y., Yin, G., and Zhao, X. (2020). Functional martingale residual process for high-dimensional cox regression with model averaging. *Journal of Machine Learning Research*, 21(207):1–37.

Hobcraft, J., McDonald, J. W., and Rutstein, S. (1983). Child-spacing effects on infant and early child mortality. *Population Index*, pages 585–618.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). Model development. *Applied survival analysis: Regression modeling of time-to-event data,*, 132168.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, 71(1):75–83.

Hougaard, P. (1986a). A class of multivariate failure time distributions. *Biometrika*, 73(3):671–678.

Hougaard, P. (1986b). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis*, 1:255–273.

Hougaard, P. and Hougaard, P. (2000). *Analysis of multivariate survival data*, volume 564. Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python*, pages 69–134. Springer.

Jiang, F. and Guterman, E. (2024). Survival analysis. *Statistical Methods in Epilepsy*, pages 124–142.

Kaba, A. J. (2020). Explaining africa's rapid population growth, 1950 to 2020: trends, factors, implications, and recommendations. *Sociology mind*, 10(4):226–268.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Kalbfleisch, J. D. and Schaubel, D. E. (2023). Fifty years of the cox model. *Annual Review of Statistics and Its Application*, 10(1):1–23.

Kaombe, T. M. and Manda, S. O. (2023). Detecting influential data in multivariate survival models. *Communications in Statistics-Theory and Methods*, 52(11):3910–3926.

Kelly, P. J. and Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in medicine*, 19(1):13–33.

Khan, J. R., Bari, W., and Latif, A. M. (2016). Trend of determinants of birth interval dynamics in bangladesh. *BMC public health*, 16:1–11.

Klebanoff, M. A. (2017). Interpregnancy interval and pregnancy outcomes: causal or not? *Obstetrics & Gynecology*, 129(3):405–407.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2014). *Handbook of survival analysis*. CRC Press Boca Raton, FL:.

Konstantopoulos, S. and Hedges, L. V. (2019). Statistically analyzing effect sizes: Fixed- and random-effects models. *The handbook of research synthesis and meta-analysis*, 3:245–280.

Kopp, D. M., Bula, A., Maman, S., Chinula, L., Tsidya, M., Mwale, M., and Tang, J. H. (2018). Influences on birth spacing intentions and desired interventions among women who have experienced a poor obstetric outcome in lilongwe malawi: a qualitative study. *BMC Pregnancy and Childbirth*, 18(1):1–12.

Kozuki, N., Lee, A. C., Silveira, M. F., Victora, C. G., Adair, L., Humphrey, J., Ntozini,

R., Black, R. E., Katz, J., and for Gestational-Age-Preterm Birth Working Group, C. H. E. R. G. S. (2013). The associations of birth intervals with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis. *BMC public health*, 13:1–9.

Kuitunen, I., Ponkilainen, V. T., Uimonen, M. M., Eskelinen, A., and Reito, A. (2021). Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. *BMC musculoskeletal disorders*, 22(1):489.

Lee, E. T. and Wang, J. (2003). *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons.

Manda, S. (2020). Flexible modeling of frailty effects in clustered survival data. *Computational and Methodological Statistics and Biostatistics: Contemporary Essays in Advancement*, pages 489–509.

McGilchrist, C. and Aisbett, C. (1991). Regression with frailty in survival analysis. *Biometrics*, pages 461–466.

Molina, K. C., Calsavara, V. F., Tomazella, V. D., and Milani, E. A. (2021). Survival models induced by zero-modified power series discrete frailty: Application with a melanoma data set. *Statistical Methods in Medical Research*, 30(8):1874–1889.

Moore, D. F. (2016). *Applied survival analysis using R*, volume 473. Springer.

Munda, M., Rotolo, F., and Legrand, C. (2012). parfm: Parametric frailty models in r. *Journal of Statistical Software*, 51:1–20.

Muza, C. and Mangombe, K. (2021). The fertility transition in africa. *Southern African Journal of Demography*, 21(1):69–103.

NSO, I. (2017). Malawi 2015–16 demographic and healthy survey: Key findings. *NSO and ICF*.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493.

Pimentel, J., Ansari, U., Omer, K., Gidado, Y., Baba, M. C., Andersson, N., and Cockcroft, A. (2020). Factors associated with short birth interval in low-and middle-income countries: a systematic review. *BMC pregnancy and childbirth*, 20:1–17.

Ranjbar, M., Rahimi, M. K., Heidari, E., Baharinia, S., Alimondegari, M., Lotfi, M. H., and Shafaghat, T. (2023). What factors influence couples' decisions for childbearing? evidence from a systematic scoping review.

Ripatti, S., Larsen, K., and Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated monte carlo em algorithm. *Lifetime Data Analysis*, 8:349–360.

Rutstein, S. and Winter, R. (2014). The effects of fertility behavior on child survival and child nutritional status: evidence from the demographic and health surveys, 2006 to 2012.

Seyedtabib, M., Moghimbeigi, A., Mahmoudi, M., Majdzadeh, R., and Mahjub, H. (2020). Pattern and determinant factors of birth intervals among iranian women: a semi-parametric multilevel survival model. *Journal of Biosocial Science*, 52(4):534–546.

Shapiro, D. and Hinde, A. (2020). Laggards in the global fertility transition. *Vienna Yearbook of Population Research*, 18:123–140.

Shifti, D. M., Chojenta, C., Hassen, T. A., and Harris, M. L. (2023). Short birth interval prevalence, determinants and effects on maternal and child health outcomes in asia-pacific region: a systematic review and meta-analysis protocol. *BMJ open*, 13(12):e076908.

Su, C.-L., Platt, R. W., and Plante, J.-F. (2022). Causal inference for recurrent event data using pseudo-observations. *Biostatistics*, 23(1):189–206.

Tesema, G. A., Worku, M. G., Alamneh, T. S., Teshale, A. B., Yeshaw, Y., Alem, A. Z., Ayalew, H. G., Liyew, A. M., and Tessema, Z. T. (2023). Estimating the impact of birth interval on under-five mortality in east african countries: a propensity score matching analysis. *Archives of Public Health*, 81(1):63.

Tesema, G. A., Worku, M. G., and Teshale, A. B. (2021). Duration of birth interval and its predictors among reproductive-age women in ethiopia: Gompertz gamma shared frailty modeling. *Plos one*, 16(2):e0247091.

Therneau, T. et al. (2015). Mixed effects cox models. *CRAN repository*.

Therneau, T. M., Grambsch, P. M., and Therneau (2000). *The cox model*. Springer.

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

Wienke, A. (2003). Frailty models. *Rostock, Germany Max Planck institute for demographic research*.

Wienke, A. (2010). *Frailty models in survival analysis*. Chapman and Hall/CRC.

Yohannes, S., Wondafrash, M., Abera, M., and Girma, E. (2011). Duration and determinants of birth interval among women of child bearing age in southern ethiopia. *BMC pregnancy and childbirth*, 11(1):1–6.

Zhang, W., Mulder, J., and Frencken, J. E. (2019). Is preventing micro-cavities in dentine from progressing with a sealant successful? *British Dental Journal*, 226(8):590–594.

Appendices

Appendix A

A.1 Ethical Approval

The Malawi National Statistical Office (NSO), along with the Measure DHS initiative, gathered secondary data for the study. The data owners said that they followed ethical guidelines when gathering data in the Malawi Demographic and Health Survey (MDHS) reports from 2015–16 (NSO, 2017). After receiving online consent from the Measure DHS program, these data were accessed via https://dhsprogram.com/data/dataset_admin/login_main.cfm. Every method was used in compliance with the applicable rules and regulations.

A.2 R code used for data analysis in the birth interval study

```
#Loading the required packages
library(tidyverse)
library(survival)
library(haven)
library(table1)
library(survminer)
library(frailtypack)
library(parfm)


#Setting working directory
setwd("/home/chiku/Documents/Data")
```

```
#Reading stata data
MWBR7AFL <- read_stata("MWBR7AFL.DTA")

#Selection required variable
data <- MWBR7AFL[c("caseid", "v004", "v008a", "
  ↪ v012", "v024", "v025", "v130", "v131", "v106"
  ↪ , "v190", "v212", "v312", "v313", "b1", "b2",
  ↪ "b4", "b5", "b6", "b7", "bidx", "bord", "b8",
  ↪  "b11", "b12", "b18", "b19", "sdist")]

#Selection second birth interval
data <- filter(data, bord ==1)

#Creating censoring variable(0 for women with no
  ↪ second child)
data$status <- ifelse(is.na(data$b12), 0, 1)

#Getting time to successive birth in months
t1 <- ifelse(is.na(data$b12), data$b19, data$b12)
t2 <- ifelse(is.na(data$b12), data$b6, data$b12)
time <- ifelse(is.na(t1),t2,t1)

#Setting 0.5 time in month for women with time 0
time <- ifelse(time==0,0.5,time)
time <- as.numeric(time) #Setting time to numeric

cluster <- data$v004 #Cluster variable

data$b5 <- as.factor(data$b5) #Child is alive(1=
  ↪ yes)
Age <- data$v212 #Age of the mother at first birth
data$b4 <- as.factor(data$b4) #Sex of the first
  ↪ child
data$v024 <- as.factor(data$v024) #Region
data$v025 <- as.factor(data$v025) #Place of
  ↪ residence
```

```
data$v106 <- as.factor(data$v106) #level of
  ↪ education

#Recoding Current contraceptive method
data$v313 <- ifelse(data$v313==2,1,data$v313)
data$v313 <- ifelse(data$v313==3,2,data$v313)
data$v313 <- as.factor(data$v313)

#Recoding Wealth index
data$v190 <- ifelse(data$v190==2,1,data$v190)
data$v190 <- ifelse(data$v190==3,2,data$v190)
data$v190 <- ifelse(data$v190==4,3,data$v190)
data$v190 <- ifelse(data$v190==5,3,data$v190)
data$v190 <- as.factor(data$v190)

#Plot of Kaplain-Meier and Log-rank for Place of
  ↪ Residence
par(mfrow = c(2,2))


t<- survfit(formula = Surv(time, status)~v025,
  ↪ data)
ggsurvplot(t, data = data,risk.table = TRUE,
  legend.title = "Residence",
  legend.labs = c("Urban", "Rural"),
  ylab="birth␣interval␣survival␣probability",
  xlab="Time␣in␣months",
  title = "Kaplan-Meier␣for␣Place␣of␣residence"
)
survdiff(Surv(time, status)~v024, data = data)

#Plot of Kaplain-Meier and Log-rank for
  ↪ Educational level
t<- survfit(formula = Surv(time, status)~v106,
  ↪ data)
ggsurvplot(t, data = data,risk.table = TRUE,
```

```
  legend.title = "Educaional Level",
  legend.labs = c("No education", "Primary", "
    ↪ Secondary", "Tertiary"),
  ylab="birth interval survival probability",
  xlab="Time in months",
  title = "Kaplan-Meier for Education Level of
    ↪ Mother"
)
survdiff(formula = Surv(time, status) ~ v106, data
  ↪  = data)

#Plot of Kaplain-Meier and Log-rank for Status of
  ↪ first Child
t<- survfit(formula = Surv(time, status)~b5, data)
ggsurvplot(t, data = data,risk.table = TRUE,
   legend.title = "Status of First Child",
   legend.labs = c("Dead", "Alive"),
   ylab="birth interval survival probability",
   xlab="Time in months",
   title = "Kaplan-Meier for Status of First Child
     ↪ "
)
survdiff(formula = Surv(time, status) ~ b5, data =
  ↪  data)

#Plot of Kaplain-Meier and Log-rank for Sex of
  ↪ first Child
t<- survfit(formula = Surv(time, status)~b4, data)
ggsurvplot(t, data = data,risk.table = TRUE,
  legend.title = "Sex of First Child",
  legend.labs = c("Male", "Female"),
  ylab="birth interval survival probability",
  xlab="Time in months",
  title = "Kaplan-Meier for Sex of First Child"
)
survdiff(formula = Surv(time, status) ~ b4, data =
```

```
  ↪ data)

#Plot of Kaplain-Meier and Log-rank for Wealth
  ↪ Index
t<- survfit(formula = Surv(time, status)~v190,
  ↪ data)
ggsurvplot(t, data = data,risk.table = TRUE,
   legend.title = "Wealth␣Index",
   legend.labs = c("Poor", "Middle", "Rich"),
   ylab="birth␣interval␣survival␣probability",
   xlab="Time␣in␣months",
   title = "Kaplan-Meier␣for␣Wealth␣Index"
)
survdiff(formula = Surv(time, status) ~ v190, data
  ↪ = data)


#Plot of Kaplain-Meier and Log-rank for
  ↪ Contraceptive Method
t<- survfit(formula = Surv(time, status)~v313,
  ↪ data)
ggsurvplot(t, data = data,risk.table = TRUE,
  legend.title = "Contraceptive␣Method",
  legend.labs = c("No␣Method", "Traditional", "
    ↪ Modern"),
  ylab="birth␣interval␣survival␣probability",
  xlab="Time␣in␣months",
  title = "Kaplan-Meier␣for␣Current␣Contraceptive
    ↪ ␣Method"
)
survdiff(formula = Surv(time, status) ~ v313, data
  ↪ = data)

#Creating descriptive table 1
table1(~as.factor(v025) + v212 + as.factor(v024) +
  ↪ as.factor(v106) + as.factor(v313) + as.
```

```
  ↪ factor(v190)+as.factor(b4) + as.factor(b5)|
  ↪ status, data=data)

#Creating dummy variables for factor variables
data <- model.matrix(~v024+v025+v106+v190+v313+b4+
  ↪ b5,data)[,-1]
data <- as.data.frame(data) #Setting data as data
  ↪ frame

#Redefining variable names
Sex <- data$b42
Alive <- data$b51
Central <- data$v0242
Southern <- data$v0243
Rural <- data$v0252
Primary <- data$v1061
Secondary <- data$v1062
Higher <- data$v1063
middle <- data$v1902
rich <- data$v1903
traditional <- data$v3131
Modern <- data$v3132

#Binding the data
data <- cbind(Age, Alive, Sex, Central, Southern,
  ↪ Rural, Primary, Secondary, Higher, middle,
  ↪ rich, traditional, Modern, cluster, status,
  ↪ time)
data <- as.data.frame(data) #Creating data frame


#Fitting Semiparamtric gamma frailty model
Semi_gamma <- coxph(formula = Surv(time, status)~
  ↪ Age + Alive + Central + Southern + Rural +
  ↪ Primary + Secondary + Higher + middle + rich
  ↪ + traditional + Modern+ frailty.gamma(cluster
```

```
  ↪ ), data = data)
summary(Semi_gamma)
Semi_gamma$loglik

#Fitting Semiparamtric normal frailty model
Semi_lognormal <- coxme(Surv(time, status)~Age +
  ↪ Alive + Central + Southern + Rural+ Primary +
  ↪  Secondary + Higher + middle + rich +
  ↪ traditional + Modern + frailty.gamma, data =
  ↪ data)
Semi_lognormal
Semi_lognormal$loglik

#Running parametric gamma frailty model with
  ↪ exponential baseline hazard
Exp_gamma <- parfm(Surv(time, status)~Age + Alive
  ↪ + Central + Southern + Rural + Primary +
  ↪ Secondary+ Higher + middle + rich+
  ↪ traditional + Modern, cluster = "cluster",
  ↪ data = data, dist = "exponential", frailty =
  ↪ "gamma")
Exp_gamma
ci.parfm(Exp_gamma, level = 0.05, digits = 3)

#Running parametric lognormal frailty model with
  ↪ exponential baseline hazard
Exp_lognormal <- parfm(Surv(time, status)~Age +
  ↪ Alive + Central + Southern + Rural + Primary
  ↪ + Secondary+ Higher + middle + rich+
  ↪ traditional + Modern,cluster = "cluster",
  ↪ data = data, dist = "exponential", frailty =
  ↪ "lognormal", method = "CG")
Exp_lognormal
ci.parfm(Exp_gamma, level = 0.05, digits = 3)
```

```
Wei_lognormal <-frailtyPenal(Surv(time, status)~
  ↪ Age + Alive + Central + Southern + Rural +
  ↪ Primary + Secondary + Higher + middle + rich+
  ↪  traditional + Modern+cluster(cluster), data
  ↪ = data, hazard = "Weibull", RandDist = "LogN"
  ↪ )
Wei_lognormal
summary(Wei_lognormal, d= 3)

Wei_gamma <-frailtyPenal(Surv(time, status)~Age +
  ↪ Alive + Central + Southern + Rural + Primary
  ↪ + Secondary + Higher + middle + rich+
  ↪ traditional + Modern+cluster(cluster), data =
  ↪  data, hazard = "Weibull", RandDist = "Gamma"
  ↪ )
Wei_gamma
summary(Wei_gamma, d=3)

# Predict martingale and deviance residuals and
  ↪ linear predictors
data$martingale <- residuals(Wei_gamma, type = "
  ↪ martingale")
data$deviance <- residuals(Wei_gamma, type = "
  ↪ deviance")
data$linear_p <- predict(Wei_gamma, type = "lp")

data$Residual <- data$martingale
attach(data)


#Plotting martingale residuals
par(mfrow=c(3,2))
plot(Residual $\sim$ Age)
lines(lowess(Residual~Age))
title("Martingale␣residuals␣versus␣Age")
```

```
plot(Residual $\sim$ Rural)
lines(lowess(Residual$\sim$Rural))
title("Martingale␣residuals␣versus␣Rural")

plot(Residual $\sim$ Alive)
lines(lowess(Residual$\sim$Alive))
title("Martingale␣residuals␣versus␣Alive")

plot(Residual $\sim$ Primary)
lines(lowess(Residual$\sim$Primary))
title("Martingale␣residuals␣versus␣Primary")

plot(Residual $\sim$ Secondary)
lines(lowess(Residual$\sim$Secondary))
title("Martingale␣residuals␣versus␣secondary")

plot(Residual $\sim$ Higher)
lines(lowess(Residual$\sim$Higher))
title("Martingale␣residuals␣versus␣Higher")

plot(Residual $\sim$ middle)
lines(lowess(Residual$\sim$middle))
title("Martingale␣residuals␣versus␣middle")

plot(Residual $\sim$ rich)
lines(lowess(Residual$\sim$rich))
title("Martingale␣residuals␣versus␣rich")

plot(Residual $\sim$ traditional)
lines(lowess(Residual$\sim$traditional))
title("Martingale␣residuals␣versus␣traditional")

plot(Residual $\sim$ modern)
lines(lowess(Residual$\sim$modern))
title("Martingale␣residuals␣versus␣modern")
```

```
data$two <- 2
data$twon <- -2
plot(x = data$linear_p, y = data$deviance,
xlab = "Linear Predictors", ylab = "Deviance
  ↪ Residuals")
lines(lowess(data$deviance ~ data$linear_p))
lines(x=data$linear_p, y = data$two)
lines(x=data$linear_p, y = data$twon)

#Drop observations based on residuals
dat <- data[data$deviance >= -3 & data$deviance <=
  ↪  3 ]

# Refit the Weibull model with gamma frailty on
  ↪ the filtered data
weibull_frailty3 <- frailtyPenal(Surv(time, status
  ↪ ) ~ Age + Alive + Central + Southern + Rural
  ↪ + Primary + Secondary + Higher + middle +
  ↪ rich + traditional + Modern + cluster(cluster
  ↪ ), data = dat, hazard = "Weibull", RandDist =
  ↪  "LogN")
# Summary of the final model
weibull_frailty3
summary(weibull_frailty3)

dat <- data[data$deviance >= -3 & data$deviance <=
  ↪  3 ]
# Refit the Weibull model with gamma frailty on
  ↪ the filtered data
weibull_frailty2 <- frailtyPenal(Surv(time, status
  ↪ ) ~ Age + Alive + Central + Southern + Rural
  ↪ + Primary + Secondary + Higher + middle +
  ↪ rich + traditional + Modern + cluster(cluster
  ↪ ), data = dat, hazard = "Weibull", RandDist =
  ↪  "LogN")
# Summary of the final model
```

```
weibull_frailty2
summary(weibull_frailty2)
```