Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Investigating the determinants of timing of first childbirth among women of the reproductive age In Mozambique

Maxwell Paganga

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Pietro COLETTI

Prof. dr. Niel HENS

2018
2019

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Investigating the determinants of timing of first childbirth among women of the reproductive age In Mozambique

Maxwell Paganga

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Pietro COLETTI

Prof. dr. Niel HENS

Investigating The Determinants Of Timing of First Childbirth Among Women Of The Reproductive Age In Mozambique

Master Thesis Biostatistics/ICP (3770-1819)
2018-2019

2nd year Master of Statistics
Hasselt University

Student:
Maxwell Paganga (1747149)

Submission Date: 14^{th} June, 2019

Supervisors:
Dr. Pietro Coletti
Prof. dr. Niel Hens

Acknowledgements

Firstly I would like to thank my supervisors, Dr. Pietro Coletti and Prof. dr. Niel Hens for their constant guidance throughout this essay. My great appreciation goes to Dr. Pietro Coletti for his effort, patience, enthusiastic encouragement, astounding guidance tips and useful critiques throughout the course of this research work.

I would also like to acknowledge with much appreciation the crucial role played by my colleagues (all the 2017-2019 Biostatistics students) who tirelessly offered their support throughout this program and special thanks goes to my group members Nancy Mumbua Musili (my wife), Murih Pusparum and Boniventura January for helping me throughout this program. I would like to thank my family and friends for their moral support throughout this study, it was not easy but they keep on encouraging me to carry on. Credit also goes to all lecturers at UHasselt Statistics Department for their intellectual support. A big thank you goes to VLIR-OUS for awarding me the scholarship, creating a dream for me and making my stay at the institution enjoyable and comfortable. A special mention goes to my daughter Grace Kimberly Paganga, for she is the pillar of my strength.

All protocol observed and above all the LORD GOD ALMIGHTY for guiding me throughout my endeavor to accomplish the results of this study by His grace which is sufficient always.

THANK YOU!!!

Abstract

This study is aimed at investigating the determinants of timing of first childbirth among women of the reproductive age (15-49) in Mozambique. The response variable is age at first childbirth which is defined as the age, in years, at which a woman gives birth to her first child. The cross-sectional dataset used in this study was obtained from the Demographic and Health Surveys (DHS) Program 2011 and contains a total of 22 357 women. Survival analysis was used to model the determinants of age at first childbirth. Kaplan-Meier survival curves were used during exploratory data analysis to come up with some descriptive statistics while the proportional hazards (PH) model was used to model the timing of first childbirth at 5% level of significance. Prior to fitting the multivariate Cox PH model, the least absolute shrinkage and selection operator (LASSO) method was used for variable selection. The factors that were found to have an effect on age at first childbirth were educational attainment of a women, region of residence, religion and ethnicity, the knowledge of contraceptive use and the age at which the woman first engaged in sexual intercourse. The logistic regression model was used to model pregnancy outcome and stepwise method was used for variable selection. The results showed that variables like smoking, working and knowledge of contraceptive use among women of the reproductive age had a positive effect on the response while age at first sexual intercourse had a negative effect on pregnancy outcome. Other factors that had an effect on pregnancy outcome were region of residence, woman's wealth index and educational attainment. Woman's marital status and age at which she had given birth to her first child were not related to pregnancy termination. The researcher thus recommends the government of Mozambique to implement policies and programs that seek to increase educational opportunities, especially for the girl child, so as to reduce the cases of early age at first childbirth. Pregnant women are also encouraged not to smoke or do strenuous jobs as these can increase the risk of pregnancy termination as a result of spontaneous abortions or miscarriages

Keywords: *Age at first childbirth, Determinants, Logistic regression, Survival analysis, Mozambique*

Contents

List of Figures

1 Introduction

Mozambique is a country which has a total area of 801,590 square kilometers and lies in the south-eastern part of Africa. Based on the latest United Nations estimates, the country has a population of around 31 million with median age of 17.3 years. Since the signing of the Rome Peace Accords in 1992, there has been some major achievements in some areas of health development and the country has made significant progress in recovering from its war torn past. However, even though there is progress in terms of positive socio-economic developments and the availability of health resources, there still remain some issues in the health sector (WHO, 2017). Many children and women continue to die from preventable causes related to maternity. According to the World Health Organization (WHO, 2017), the infant mortality rate is estimated to be 124/1000 live births and this figure strongly concur with the latest results from the United Nations International Children's Emergency Fund (UNICEF). The child mortality rates, which are usually considered to be a proxy indicator of socio-economic development, are estimated to be 178/1000 births. The maternal mortality rate in Mozambique is considered as one of the highest in the region and is estimated to be 408/100,000 live births. Most of these maternal deaths are a result of pregnancy complications or childbirth because most women do not deliver at health facilities.

According to WHO (2017), on average only 48% of deliveries take place at health care facilities and about 1.23 per 500,000 of mothers have access to essential obstetric and newborn care. The proportion of births taking place in health institutions was also found to vary significantly by province and so does most of the health indicators. This maybe as a result of differences in economic factors, traditional factors, accessibility to health care facilities and over-reliance on traditional birth attendants, especially woman in the rural areas (Roro et al., 2014). There is also a high incidence of childhood marriages in Mozambique which contribute to the young age of first-time mothers (UNICEF, 2015). According UNICEF (2015), these child marriages and teenage pregnancies are a major cause of high maternal and infant mortality rates, underage birth weights and postpartum bleeding among other problems. Early motherhood may also imply giving up on basic education and thus contributing to the the country's underdevelopment. These problems has been the major motive of conducting the present research of modelling age at first child birth in Mozambique. It is hoped that some findings of this study will contribute to inform the broader policy discussion in Mozambique on the issues of age at first childbirth, adolescent pregnancy, child marriage and girls' education.

1.1 Background

One of the most significant event in a woman's life is giving birth to her first child. The birth of the first child can be viewed as the first visible outcome of the fertility process and may play a big role in the future life of a woman. It is an event of great social and individual significance whose importance is recognised in all human societies, and marks a womans transition into motherhood. First childbirth signifies the transition of a couple

into a new social status, i.e. parenthood together with its expectations, responsibilities and deliverables (Rindfuss and St. John, 1983). This transition may result in a woman dropping some of her career-building roles and education for motherhood and parenthood. Some studies have shown that the time at which parenthood roles begin may affect the childbearing behavior of women as they advance towards reproductive ages. Many studies have reported that age of initiation of childbearing among women has strong effects on the demographic behavior of women and the entire population (Rajaretnam, 1990; Macro et al., 2014; Fagbamigbe and Idemudia, 2016). According to Rajaretnam (1990) and the Macro et al. (2014), an early beginning of childbirth may be associated with negative effects on a woman's socio-economic well-being in later years. This is because once a woman begins to give birth, she may drop some roles to do motherly roles that are usually demanding with regard to her time and resources. The negative effects of this might be in the areas of career development, educational or occupational attainment, marital stability, asset possessions as well as her health (Rindfuss and St. John, 1983). It also affects type of care and opportunity available to such women and their off-springs, social change, fertility trends and the state of the economy (Luc et al., 1993; Kumar et al., 2006). On the other hand, delayed age at first childbirth may reduce the rate of maternal and child morbidity and mortality (Sarkar, 2010).

Some authors have noted that the age at first childbirth has a direct relationship with fertility, as it may have an influence on the number of children a woman can bear throughout her reproductive period in the absence of any active fertility control (Bumpass et al., 1978; Fagbamigbe and Idemudia, 2016). According to Bumpass et al. (1978), age at first birth plays an important role on social changes, state of economy and period fertility trends. Timing of first birth may also affect how kids are raised, their intellectual development and the kind of opportunities available to them (Bumpass et al., 1978). For example, In Sub-Saharan African Countries, where contraceptive use is relatively low, it was observed that women who give birth to first child at a younger age tend to have many children in later years (Sarkar, 2010). There are many factors which affect age at first childbirth which include: individual factors, family factors, societal factors, economic factors as well as national and international factors among others (Fagbamigbe and Idemudia, 2016). Because of these consequences of timing of first birth, this study thus seeks to model age at first childbirth as well as investigate the determinants of timing of first childbirth among women of the reproductive age (15-49 years) in Mozambique.

Several studies have suggested that the well-being of a family is conditioned by age at first childbirth and how rapidly it proceeds. A large number of studies have concluded that early age at first childbirth can result in faster subsequent child bearing, influence the complete family sizes and has an increased chance of unwanted births (Finnäs and Hoem, 1980; Ford, 1984; RAO and Balakrishnan, 1988). In a study conducted by Bumpass et al. (1978) on age and marital status at first birth in the USA, it was concluded that early age at first childbirth might be associated with high fertility. On the other hand, delaying time to first childbirth was found to reduce completed family size, i.e. delaying age at first birth is associated with reduced fertility rates in some countries (Kohler et al., 2001). However, other studies have found fertility to be low despite early age at first childbirth and a lack

of efficient contraceptive methods (Turner, 1992). The explanation for this situation may be attributed to proper child spacing and abortion which has become the primary means of limiting fertility in many Eastern European countries (Dennehy et al., 1995). A similar result was also found in a study conducted by Konogolo (1985) in Kenya, who reported that there was an inverse relationship between age at first childbirth and fertility and the majority of women who start child bearing at early ages were likely to have lower levels of education and were either rural residents or urban poor. This means that fertility was low despite early age at first childbirth. Trussell and Reinis (1989) also supported this by adding that these women would usually comply more to traditional patterns of birth spacing which results in long birth intervals and hence reduced fertility.

Other studies done in several countries have shown that women's education has a significant effect on age at first childbirth. According to Weinberger et al. (1989), education was found to be the most important socio-economic variable associated with greater occupational differentiation which can influence the reproductive behaviour in many ways. In a study conducted by Maxwell (1987) in the United States of America, it was observed that there was a positive relationship between education and age at first childbirth. In this study, women with higher education were found to have longer time to first childbirth compared to those with no education. This was attributed to the fact that more educated women are more likely to postpone marriage and even child bearing within marriage as they will be focusing on their careers and taking advantage of new opportunities. Studies by Gaisie (1984) and Konogolo (1985) also confirmed these results by finding that post-primary schooling, especially of 9 or more years, has a strong effect in delaying first childbirth often by 3 to 4 years. However, on the contrary, education and employment were found to have a dual way relationship with age at first childbirth. In their study "Early female marriage in the developing world", Jensen and Thornton (2003) concluded that combining teenage motherhood with education is very difficult. They argued that the determination to acquire higher education might indeed lead to postponement of first childbirth, as women who are employed before first birth often delay the birth so as to work longer. On the other hand, women who began childbearing too early without any career development may find themselves unemployable in the future.

Studies done in India (Bloom and Reddy, 1986), Tanzania (Mturi, 1997) and Nigeria (Fagbamigbe and Idemudia, 2016) have shown that religious beliefs and cultural practices is another important variable which can influence marriage and child bearing, and therefore is a vital background variable for age at first childbirth. In India, Hindus are well known to marry and bear children at younger ages than non-Hindus, while in Tanzania the Islam are well known to encourage early marriage and thus resulting in early age at first childbirth. Other cultural practices such as forcing teenage girls to marry can also adversely affect timing of first childbirth. The inability to pay huge marriage rites, as practiced in certain parts of sub Saharan Africa, may also prolong the age at first marriage which can be a proxy for the age at first childbirth (Adebowale et al., 2012). Religion affiliation can also affect age at first childbirth by influencing the level of contraceptive use. Many religions in sub Saharan Africa forbid the use of modern contraceptives while others have a more

liberal stand. This implies that religions which are liberal to contraceptive use have a lower likelihood of early first childbirth and long intervals between subsequent birth compared to those which forbid the use of contraceptives. This result was confirmed by a study carried out in Sierra Leone by Gage (1986), who noted that Catholics had a lower age at first childbirth than Muslims because the Catholics have a negative perspective towards the use of contraceptives, although the doctrine emphasizes abstinence outside marriage.

According to Ohadike (1979), ethnicity can play a major role in the timing of first childbirth. He argued that even though age at first childbirth is primarily determined by biological processes, it can also be altered by socio-cultural factors so as to maintain the biological continuity of the members of the society. He further explained that different ethnics groups have different norms, beliefs and values as well as the practices that are likely to influence the reproductive performance of a given society and thus affect timing of first childbirth. In a study conducted by Bongaarts et al. (1984), on the proximate determinants of fertility in sub-Saharan Africa in Tanzania, it was observed that the timing of marriage and childbirth varies among the various ethnic groups. For instance, the inhabitants of the Lake regions and Southern zones were found to marry early and start child bearing at an early age. Fagbamigbe and Idemudia (2016) studied the timing of first childbirth among women in Nigeria and modeled factors affecting it using survival analysis techniques. They also observed that ethnicity was one of the significant factors affecting age at first childbirth among others.

Place of residence was also found to be a determinant of age of first childbirth. Previous research on rural-urban differentials in marriage and fertility timing have shown a significant difference in the incidence of first childbirth in rural areas compared to urban areas (Zabin et al., 1986; Luc et al., 1993). According to Zabin et al. (1986), women who reside in urban areas tend to experience a more dispersed marriage, cohabitation distribution and longer time to first childbirth than those in the rural areas. This finding was also supported by Luc et al. (1993), in their study "Selected determinants of fertility in Vietnam: age at marriage, marriage to first birth interval and age at first birth". In this study, 4 172 women in the reproductive age (15-49) were interviewed in the 1988 Vietnamese Demographic and Health Survey (VDHS). The obtained data was then used to examine age at marriage, marriage to first birth intervals and age at first childbirth. The results showed significant differences between rural and urban area residence, whereby women from the North had significantly higher age at first childbirth than women from the South. They also reported that most women based in rural areas with little or no education tend to marry at significantly younger ages than urban women and those with secondary education. The issue of rural-urban differentials in the timing of first childbirth was also observed in a study conducted in Nigeria by Fagbamigbe and Idemudia (2016). Women who reside in urban areas were found to be less likely to have early first childbirth compared to those based in the rural areas.

In their study, RAO and Balakrishnan (1988) found that age at first marriage had a significant impact on fertility and is regarded as the most significant determinant of age at first

childbirth. Their study confirmed that an increase in the ages at first marriage and first childbirth can result in a decrease in the incidence of high fertility after the first marriage. They also concluded that age at first marriage is essentially similar to age of entry into sexual relations and thus a major determinant of age at first childbirth. However, on the contrary, Sarkar (2010) argued that age at first marriage may have a limited influence on the age at first childbirth because most unmarried adolescents, especially in sub Saharan Africa, tend to engage in premarital sex with limited contraceptive use. This premarital sexual behaviour among the youth was hypothesised to be a result of the breakdown of traditional social controls by elders over the sexual behaviour of adolescents and was confirmed by a survey conducted in Kenya in which over 60% of the respondents were in support of the hypothesis (i.e. they believed that the rules and norms restricting premarital and extramarital sex are no longer applicable these days) (Ochalla-Ayayo et al., 1990). Therefore, due to the fact that most sexual activity in some societies is not confined to marriage and women may bear children before marriage, it is therefore wise to use the age at first sexual intercourse and date of first birth as more appropriate indicators of age to first childbirth than the age at first marriage.

Peer pressure was also identified by several studies as one of the factors which can lead to early timing to first childbirth. In a study to model the determinants of age at first birth done by Sarkar (2010) in Bangladesh, it was observed that many young girls were encouraged to marry due to pressure from peer groups, parents and even the society as a whole. Peer groups were found to have a positive relation with early timing of first childbirth as many young girls are pressurized to engage in activities such as smoking, alcohol drinking and use of drugs which can lead to degraded situations that could probably increase the chances of young girls into early motherhood. Sarkar (2010) used multiple logistic regression analysis to assess the proximate determinants of age at first union as the dependent variable. Apart from peer pressure, most women in Bangladesh engage in sexual activities at an early stage before the age of 15 and factors like current age, place of residence, education status, contraceptive use, religion and the incidence of primary sterility had a significant influence on timing of first childbirth. Other studies have found that many other factors are also associated with age at first childbirth. Some of these major proximate determinants of age at first childbirth are age at age at first sexual relations and age at menarche (Udry, 1979; Zelnik, 1981), wealth index (Agaba et al., 2010), cigarette smoking, alcohol consumption and use of contraceptives Sarkar (2010) and some biological factors such as the ability of females to get children and miscarriage before first childbirth (Rindfuss and St. John, 1983; Sarkar, 2010).

1.2 Objectives

The fundamental aim of this study is to explore several factors that are potentially associated with age at first child birth among women of the reproductive age (15-49) in Mozambique. The objective of the research is to answer the following questions:

i) What are the factors that determine when women of the reproductive age give birth

to their first child?

ii) To what extend do each of these factors contribute on age at first childbirth?

iii) Does age at first child birth affect pregnancy termination?

1.3 Data Description and Summary of the Applied Methodologies

The data for this study is a cross-sectional dataset obtained from the Demographic and Health Surveys (DHS) Program 2011 (DHS, 2011). The DHS Program is a 5 year project which assist institutions in collecting and analyzing data to be used in various researches. The DHS Program is aimed at providing refined information through appropriate partnerships in data collection, analysis and evaluation at national and international levels. It is also aimed at improving data collection and analysis tools and methodologies as well as enhancing the dissemination and utilization of data. DHS surveys collect primary data using several types of questionnaires to produce both raw and recode data formats. Raw data files are generally not distributed because they include the data as they were collected, without any structural changes, while recode data files are generated from the raw data and can be distributed to many institutions for analyses purposes. This implies that all variables in the raw data file are represented in the recode data file in a standardized format, with the same structure across countries participating in each DHS phase. This standardization is meant to facilitate comparisons across surveys (DHS, 2011).

From the DHS recode file of 2011, some variables of interest were extracted in order to model timing of first childbirth. The extracted data set had a total of 22 357 women aged 15-49 years, and some of the extracted information was on the women's background characteristics, sexual and reproductive history and knowledge of contraceptive use. The response variable in this study is age at first childbirth while education level, religion, place of residence, ethnicity and region are some of the explanatory variables (covariates). Also included as covariates are responses on whether the woman ever smoked, whether she had a terminated pregnancy or not and whether she has ever used something to prevent pregnancy. Women's current educational attainment was used as a proxy for education as of the time of first childbirth, the justification being that educational status rarely changes for persons who had primary or no education because primary education is mostly attained at or below the age of 12.

Survival analysis was used to model the determinants of age at first childbirth by fitting the proportional hazard model. Kaplan-Meier survival curves were used in the exploratory data analysis phase by performing univariate analyses of time to first childbirth and the independent variables. The Cox PH model was used in multivariate analysis and to compute the hazard ratios of the exposed groups relative to the reference group. The least absolute shrinkage and selection operator (LASSO) method was used for variable selection into the multivariate PH model, together with survival curves. The logistic regression model was used to model the probability of a woman ever having had a pregnancy termination. Finally, all analyses were performed in R version 3.4.1 and the statistical analysis system

software (SAS software version 9.4.). Table 1 shows some variables of interest which were extracted from the DHS recode file of 2011 together with their codes and description.

Table 1: Variable description

Variable (Code)	**Description**
Education (hv106)	Highest level of education the household member attended (1-Don't know, 2-Pre-school, 3-Primary, 4-Secondary, 5-Higher).
HH_MS (hv115)	Marital status of the household member (1-Never married, 2-Divorced, 3-Widowed, 4-Married).
Wealth_Index (hv270)	The wealth index is a composite measure of a household's cumulative living standard (1-Poorest, 2-Poorer, 3-Middle, 4-Richer, 5-Richest).
Smoking (ha35)	Smoking status (0-No, 1-Yes).
Religion (v130)	Country-specific religion (1-Catholics, 2-Muslims, 3-Protestants, 4-Non-religious, 5-Other).
Ethnicity (v131)	Country-specific ethnicity (1-Emakhuwa, 2-Portuguese, 3-Xichangana, 4-Cisena, 5-Elomwe, 6-Echuwabo, 7-Others).
Pregnancy (v213)	Whether the respondent is currently pregnant (0-No, 1-Yes).
PregTerm (v228)	Whether the respondent ever had a pregnancy that terminated in a miscarriage, abortion, or still birth (0-No, 1-Yes).
Working (v714)	Whether the respondent is currently working (0-No, 1-Yes).
Contraceptive (v301)	Knowledge of any method of contraceptive use (1-Modern, 2-Traditional, 3-Folkloric, 4-None).
KnowOfCycle (v217)	Knowledge of the ovulatory cycle indicates when during her monthly cycle the respondent thinks a woman has the greatest chance of becoming pregnant (1-after period ended, 2-at any time, 3-before period begins, 4-don't know, 5-during her period, 6-middle of the cycle, 7-other).
Region (v101)	De facto region of residence. Region in which the respondent was interviewed (1-Cabo Delgado, 2-Gaza, 3-Inhambane, 4-Manica, 5-Maputo Cidade, 6-Maputo Provincia, 7-Nampula, 8-Niassa, 9-Sofala, 10-Tete, 11-Zambezia).
Age@FirstSex (v525)	Age at first sexual intercourse. Respondents who had never had sex are coded 0.
Age@Firstbirth (v212)	Age of the respondent at first birth.

2 Methodology

2.1 Survival Analysis

Survival analysis is a terminology used to describe the analysis of data in the form of times from a well-defined time origin until a pre-specified event or end point of interest occurs (Collett, 2015). Survival data is also referred to as time to event data, time to failure/failure time data or survival time data and the response variable is the time until a particular event

occur, which is often called survival time, event time, or failure time (Rizopoulos, 2012). Survival analysis has been a very active research field for several decades and is frequently used in clinical trials, epidemiological studies, economics, marketing, industries, sociology, psychological experiments and many other disciplines. In these settings, the event of interest may be the development of some disease, the appearance of a tumor, **age at first childbirth**, duration of a first marriage, the length of subscription to a newspaper and many others (Rizopoulos, 2012; Collett, 2015).

Survival analysis data has special features that are not amenable to standard statistical procedures that are used in data analysis. Generally, it deals with censored data and the outcome of interest is whether or not an event occurs and when that event occurs (Collett, 2015). Censored data is one in which the end point of interest is not fully observed on all subjects under study. The end point of interest is not fully observed because some subjects may not have experienced the event of interest at the end of the study (administrative censoring), or the event may be death from a cause that is known to be unrelated to treatment or lost to follow up (i.e. individual whereabouts cannot be traced) (Collett, 2015). Standard statistical procedures such as linear and logistic regression models are therefore not suited to model such outcomes as they are not equipped to handle censoring. They tend to underestimate the true time to event thus producing invalid results (Collett, 2015). The analysis of censored data depends on the nature of the censoring mechanism which differ with regard to their relative positioning on the time axis to the true event times. There exist three main types of censoring namely right censoring, left censoring and interval censoring which are described in Table 2 below (Rizopoulos, 2012; Collett, 2015).

Table 2: Types of censoring (Collett, 2015)

Censoring	**Description**
Right	This is a type of censoring such that for a subset of the subjects under study, the event of interest is only known to occur after a certain time point. This means that the exact event time, say T, is not observed but only the lower bound for the time is observed, say C, and therefore $T > C$.
Left	Left censoring is encountered when the event of interest for some subjects under study is only known to occur before a certain time point. This means only the upper bound for the time is observed and the actual survival time for a subject is therefore less than that observed ($T < C$).
Interval	This type of censoring is encountered when the event of interest is only known to occur between two certain time points, i.e. only the time interval ($C_L < T < C_U$) is known, where C_L is the lower limit, C_U is the upper limit and T is the true survival time.

In survival analysis, they may exist situations in which some subjects have a lifetime that is smaller than some value and may not be observed at all. This is called truncation and differs from left censoring because for left censored data the subject exists but for a truncated data

the researcher may be completely unaware of the subject existence. Therefore, censoring implies sampling from the whole population, but getting only a partial information about T from some individuals, while truncation implies sampling from a conditional distribution or sub-population (Rizopoulos, 2012). Censoring mechanisms can also be classified as informative or noninformative, depending on whether or not the probability of a subject being censored depends on the failure process. Informative censoring is when a subject withdraws from the study for reasons directly related to the expected failure time while noninformative censoring occurs when a subject withdraws from the study for reasons not related to expected failure time, but it can depend on other covariates or prognostic factors. Noninformative censoring is similar to the missing at random (MAR) assumption, because the event times are independent of censoring mechanism (Rizopoulos, 2012). An important assumption that is usually made when analysing censored survival data is that the true survival time, T, is independent of any mechanism that causes that subject's survival time to be censored at time c, where $c < T$. This means that any subject censored at time c is representative for all other subjects exposed to the event at that time and having the same values of the covariate(s), given the censoring process operates randomly. This is known as independent censoring otherwise dependent censoring (Collett, 2015). In this study, it will be assumed that censoring is independent and non-informative.

In summarising survival data, the three most important functions used are the survival function, probability density function and the hazard function. Given a non-negative random variable, T, associated with the actual survival time, t, of a subject, the survival function, $S(t)$, gives the probability that a subject will survive beyond time t without the event. Thus, $S(t)$ can be expressed as:

$$S(t) = P(T \geq t) = 1 - F(t) = 1 - \int_0^t f(x)dx,$$

where $F(t)$ denotes the cumulative distribution function of T with corresponding probability density function $f(t)$. The probability density function can be expressed as

$$f(t) = \frac{d}{dt}[F(t)] = -\frac{d}{dt}[S(t)].$$

The hazard function, $h(t)$, represents the conditional or instantaneous rate at which events occur during a short interval of time, t to $t + \delta t$, given the subject had no previous events per interval width, δt. The hazard function can be expressed as:

$$h(t) = \lim_{\delta t \to 0^+} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} = \frac{f(t)}{S(t)}. \tag{1}$$

Equation 1 shows that these three functions of survival data analysis are all related because given one of the survival functions, the other two can easily be derived. The cumulative hazard function, $H(t)$, is obtained by integrating $h(u)$ over $(0; t)$;

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

In relation to the present study, the population at "risk" are all women of the reproductive age (15-49) since they are all likely to give birth one time or the other. The survival function thus gives the probability that a woman "survives" longer than some specified time t without giving birth to first child, while the hazard function gives the instantaneous potential per unit time (in years) to give birth for the first time after time t, given that the woman had not given birth up to time t. The "survival time" is therefore the the age of the women at first childbirth, while the survival time for those with no birth as of the time of the survey was their current age. A censor variable was created and those women who had given birth to first child were given a censoring index coded "1" and "0" otherwise.

2.1.1 Estimating the survivor and hazard functions

As previously mentioned, many standard statistical analysis techniques do not apply to survival data due to the issue of censoring and the fact that the distributions of survival data tend to be positively skewed and far from normal. Because of these problems, unique parametric and nonparametric methods are therefore used to model survival time (Therneau and Grambsch, 2013). There are many nonparametric methods that are used for modeling or estimating $S(t)$ and $H(t)$ in survival data. Some of these methods include the empirical survival function, the actuarial estimate (life table), the Nelson-Aalen survival function and Kaplan-Meier or product-limit method. Among these, the most frequently used is the Kaplan-Meier or product-limit method, derived by Kaplan and Meier (1958) using maximum likelihood arguments. It assumes that every subject follows the same survival function (no covariates or other individual differences). By making assumptions of survival time distribution, parametric methods such as exponential, Weibull, Gamma, or log-normal can also be used to model survival time. More detail regarding these can be found in many survival analysis textbooks, such as Klein and Moeschberger (2006),Therneau and Grambsch (2013), Tableman and Kim (2003), Rizopoulos (2012), Collett (2015) and many others.

The Kaplan Meier (KM) estimator, developed for scenarios where survival time is measured on a continuous scale, whereby only intervals containing an event contribute to the estimate, is commonly used to estimate the survivor function of ungrouped censored survival data. A series of steps of declining magnitude is obtained from the plot of KM estimate of the survivor function. The KM estimator approaches the true survivor function for the population when the sample size is large enough with respect to the population under study (Collett, 2015). Given $S(t)$, the probability that an individual will not have recurrence of an event after time t, the observed times until death of n sample members is denoted as $t_1 < t_2 < t_3 < \cdots < t_n$. The non parametric KM estimator of the survivor function is then estimated by:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right),$$

where $t_1, t_2, \ldots, t_n$ are the observed survival times for n subjects. The risk set size at t_j (number of subject observed at time t_j) is denoted by n_j and d_j is the number of events at t_j. The KM is applied when there is censored data and a series of time intervals are formed such that each interval has one event time and is taken to occur at the start of

the interval. Here, no interval should begin at a censored time and if it happens that an event and a censored observation occurs at the same time, the censored time is taken to occur immediately after an event when computing n_j (Collett, 2015). According to Collett (2015), the advantage of the KM method over other methods, e.g the life table method for analyzing survival and failure time data, is that the resulting estimates do not depend on the grouping of the data into a certain number of time intervals.

The simplest estimator of the standard error of the estimated survival probability is obtained by using an approximation based on the binomial distribution. However, in this case the standard error tend to increase with time i.e. smaller precision due to a smaller risk-set, and thus the Greenwood formula can be used to overcome this problem. Initially the variance of the KM estimate is estimated using a couple of Taylor series approximations (Collett, 2015). This is given by:

$$var(\hat{S}(t)) = [\hat{S}(t)]^2 \prod_{j=1}^{k} \left(\frac{d_j}{n_j(n_j - d_j)} \right), \tag{2}$$

where n_j and d_j are defined as above. The standard error of the KM estimate of the survivor function is the the square root of the variance in Equation 2 and is given by:

$$se(\hat{S}(t)) = \hat{S}(t) \prod_{j=1}^{k} \left(\frac{d_j}{n_j(n_j - d_j)} \right)^{1/2}, \tag{3}$$

for $t_{(k)} \leq t < t_{(k+1)}$. The result in Equation 3 above is known as the Greenwood formula and the $100(1-\alpha)\%$ confidence interval (CI) for $S(t)$ for the KM estimate of the survival function is given by:

$$\hat{S}(t) \pm Z_{1-\frac{\alpha}{2}} se(\hat{S}(t)),$$

where $Z_{1-\frac{\alpha}{2}}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. These pointwise confidence intervals are valid for the single fixed time at which the inference is made. However, one drawback with this way of calculating the confidence intervals is that they may extend below zero or above one. The easy solution to this problem is to replace any limits greater than one or lower than zero with one and zero, respectively. There are also more sophisticated ways to solve this by transforming S(t) to the $(-\infty, \infty)$ interval using the log-log transformation, calculate the confidence intervals for the transformed value and then back-transform everything (Collett, 2015).

2.2 Comparing survival functions

It is often of great interest to compare the chances of survival for different groups, for example subjects education level on time to first childbirth. From the survivor function produced by the KM estimator, they may seem to be differences between the groups but however, the observed differences may have occurred by chance and so a formal test must be carried out to show if the differences were indeed statistically significant (Klein and

Moeschberger, 2006). The null and alternative hypotheses for this test are as follows:

$$H_0 : S_1(t) = S_2(t) = \cdots = S_k(t) \text{ for all}$$
$$H_1 : S_i(t) \neq S_j(t) \text{ for at least one pair } i, j$$

for k different groups. The null hypothesis states that the true survival functions in all groups are the same, while the alternative states that the functions are different in at least one pair. The main idea behind this test is to compare the observed number of events in each group to the expected events under the null hypothesis (Mantel and Haenszel, 1959). This is done by computing a vector, say z, where the element of the j^{th} group, for $j = 1, \ldots, k$, is given by:

$$z_j = \sum_{i:t_{(i)}<\tau} W(t_{(i)}) \left(d_{ij} - \frac{n_{ij} d_i}{n_i} \right),$$

where n_{ij} is the risk set in the j^{th} group, $W(t_{(i)})$ is a weight function which defines the importance of the observation(s) at time $t_{(i)}$ and τ is the time at which all of the risk set of all groups is at least five. The test statistic will then be computed as $z^\intercal \Sigma^{-1} z$, where the covariance matrix Σ for z is estimated from the data (Klein and Moeschberger, 2006). Under the null hypothesis, this test statistic approximately follows a χ^2 distribution with $k-1$ degrees of freedom. Different types of this test can be obtained by varying the weight function, and the most common type is the log-rank test, where $W(t_{(i)}) = 1$ for all time points (Mantel and Haenszel, 1959). The test statistics for the log-rank test is given by:

$$\frac{[\sum_{k=1}^{s}(O_{ik} - E_{ik})]^2}{\sum_{k=1}^{s} var(O_{ik} - E_{ik})},$$

where O_{ik} and E_{ik} are the observed number of events in each group and the expected events under the null hypothesis, respectively. The weight function in not in the test statistics because of equal weighting of the failure times.

Another type is the Wilcoxon-Gehan test (also called Breslow test), which is more sensitive to early differences and thus gives more weight to early survival times than late survival times, i.e. $W(t_{(i)}) = R_i$ (Klein and Moeschberger, 2006; Collett, 2015). In case the null hypothesis is rejected, pairwise testing should be conducted in order to figure out the group(s) that are deviating, and by doing so it is then important to adjust for the multiple comparisons to avoid increasing the chance of finding a falsely significant difference above the desired confidence level (Kutner et al., 2005). It should be emphasized that the decision about the choice of the test should be made before seeing the data to avoid potential bias if the point is chosen a posteriori, and in this current study the log-rank test was chosen.

2.3 Proportional Hazards Regression Model

An alternative approach to modeling survival data is the Proportional Hazards (PH) model (Cox, 1972), which assumes that the effect of the covariates is to increase or decrease the

hazard function by a proportionate amount at all durations (i.e. proportion of hazards are constant from time to time). This model is the most widely used method for modeling survival data. Several non-parametric methods like the Kaplan-Meier estimator and the log-rank test can be used to examine differences in survival for specified groups but not if survival is effected by some continuous covariate, or in examining the chances of survival as a function of two or more predictors. These univariable analysis methods can also suffer from several drawbacks such as confounding or bias during comparison of some survival curves, due to unequal distribution of another factors. They can also be problematic if there are too many strata because there will be too small sample size per stratum, and it is therefore important to adjust the analysis to regression methods suitable for survival analysis models (Collett, 2015). The PH model is one way to deal with this limitation. It is a semi-parametric model for fitting survival data in which the effect of a unit increase in covariate is multiplicative with respect to hazard rate. The model gives an expression for the hazard at time t for a subject with a given specification of a set of independent variables, $\boldsymbol{X}$, to predict individuals' hazard. The basic PH model is given by:

$$h(t|X) = h_0(t)\exp(\boldsymbol{X}^\intercal\boldsymbol{\beta}),$$

where $h_0(t)$ is the baseline hazard corresponding to the hazard function for a subject for whom all the covariates included in the model equal to zero, $\boldsymbol{\beta}$ is a vector of covariate coefficients and $\boldsymbol{X}$ is the covariate vector. The covariates may be time-dependent (in this case the model is known as the Cox's Model), but are here assumed to be fixed at the start of study. The baseline hazard can be group specific, but the coefficients are assumed to be constant throughout the study regardless of group, hence the notation semi-parametric. The sign of the coefficient indicates how a covariate affects the hazard rate.

If two subjects with covariate values $\boldsymbol{X}$ and $\boldsymbol{X^*}$ respectively are compared, then the ratio of their hazard rates (known as the hazard ratio) at any time point is given by:

$$\frac{h(t|X)}{h(t|X^*)} = \frac{h_0(t)\exp(\sum_{k=1}^{p}\boldsymbol{\beta_k}\boldsymbol{X_k})}{h_0(t)\exp(\sum_{k=1}^{p}\boldsymbol{\beta_k}\boldsymbol{X_k^*})} = \exp\left(\sum_{k=1}^{p}\boldsymbol{\beta_k}(\boldsymbol{X_k}-\boldsymbol{X_k^*})\right). \tag{4}$$

The hazard ratio in Equation 4 does not depend on time t, i.e. it is proportional (constant) throughout the study. This assumption is known as the proportional hazard assumption and it greatly facilitates the interpretation of covariate effects. Because the hazard ratio for two subjects with fixed covariate vectors $\boldsymbol{X}$ and $\boldsymbol{X^*}$ is constant over time, the model is known as the proportional hazards model (Collett, 2015). This does not however imply that the absolute difference between the two individuals discussed above is constant, the exponentiated covariates act multiplicatively on a baseline hazard which may vary freely over time. Estimation of parameter estimates, $\boldsymbol{\beta}$, is based on maximising the partial likelihood function introduced by Cox (1972), as opposed to the likelihood function. The partial likelihood is given by:

$$\mathcal{L}(\mathcal{B}) = \prod_{k=1}^{K}\left\{\frac{\exp(\boldsymbol{X_k^\intercal}\boldsymbol{\beta})}{\sum_{l\in R(t_{(k)})}\exp(\boldsymbol{X_l^\intercal}\boldsymbol{\beta})}\right\}^{\delta_k}, \tag{5}$$

where $\boldsymbol{X_k}$ is the vector of covariates (risk/prognostic factors) for a woman who has given birth to first child at the k^{th} ordered time $t_{(k)}$, $R(t_{(k)})$ is the risk set at time t_k and δ_k is an event indicator, which is 0 if the k^{th} survival time is right censored, and 1 otherwise. The maximum likelihood estimates of the $\boldsymbol{\beta}$ parameters can be obtained by maximising the log-likelihood function using numerical methods (Collett, 2015). However, it should be noted that the likelihood function in Equation 5 can be modified slightly if there exist tied events and methods such as Breslows method, Efrons method and exact method can be applied to handle tied events. In this current study, many ties where expected as the event of interest is age at first childbirth, i.e. many women are expected to give birth to first children at the same age. Among the three mentioned methods of handling ties, the exact method is the most precise but however, it is computationally expensive (numerically complex) hence the Efron method was used. According to Therneau and Grambsch (2013), the Efron method is a good approximation to the exact method provided that the number of ties is not too large compared to the Breslow approximation. They also reported that there is little difference between these approaches when the number of ties is small. In general, when using the Efron approximation the likelihood becomes:

$$\mathcal{L}(\mathcal{B}) = \prod_{k=1}^{K} \frac{\exp(\boldsymbol{X_k}^{\intercal}\boldsymbol{\beta})}{\prod_{r=0}^{\delta_k - 1}\{\sum_{l \in R(t_{(k)})} \exp(\boldsymbol{X_l}^{\intercal}\boldsymbol{\beta} - r\bar{w})\}},$$

where $\bar{w} = \delta_k^{-1} \sum_{l \in D_k}$ is the average weight over the set D_k of subjects who fail at time k. More detail regarding handling of ties can be found in many survival analysis textbooks, such as Klein and Moeschberger (2006), Therneau and Grambsch (2013), Kalbfleisch and Prentice (2011), Tableman and Kim (2003), Rizopoulos (2012), Collett (2015) and many others.

In order to check the validity of the regression parameters in the case of ordinary maximum likelihood, there are three main methods to test the global hypothesis $H_0 : \beta = \beta_0$. These are the Wald's test, the Likelihood Ratio test and the Score Test and all three are also applicable to test hypotheses about β derived from the Cox partial likelihood (Klein and Moeschberger, 2006). The Wald's test is based on the fact that for large samples, the maximum partial likelihood estimate, $\hat{\boldsymbol{\beta}}$, is p-variate normally distributed with mean β and covariance matrix estimated by the inverse of the information matrix, $\boldsymbol{I}$. Since the sum of squared standard normal variables is chi-square distributed, the test statistic, χ^2_W, is approximately chi-square distributed with p degrees of freedom if H_0 is true, and is given by:

$$\chi^2_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^{\intercal} \boldsymbol{I} \hat{\boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}).$$

The Score Test is based on the scores, $\boldsymbol{U(\beta)}$, and for large samples $\boldsymbol{U(\beta)}$ has a p-variate normal distribution with mean 0 and covariance matrix $\boldsymbol{I(\beta)}$ under H_0. Hence, the test statistic

$$\chi^2_S = \boldsymbol{U(\beta_0)}^{\intercal} \boldsymbol{I^{-1}(\beta_0) U(\beta_0)}$$

is approximately chi-square distributed with p degrees of freedom if H_0 is true. The test

statistic for the Likelihood ratio test is

$$\chi^2_L = 2\left[log(L_p(\hat{\boldsymbol{\beta}})) - log(L_p(\boldsymbol{\beta_0}))\right],$$

which is also chi-square distributed with p degrees of freedom for large samples. The three tests do not have to produce the exactly same result but they are asymptotically equivalent (Collett, 2015). According to Therneau and Grambsch (2013) and Klein and Moeschberger (2006), the likelihood statistic has the best statistical properties and should be used when in doubt, but one needs to be careful when dealing with missing data when applying this test as the involved models should be fitted to the same data.

As aforementioned, the PH regression model makes two important assumptions. The first one is the assumption of constant hazard ratio over time and the second one is the existence of a log-linear relationship between the hazard and its covariates. It is therefore of paramount importance to check the validity of these assumptions and model adequacy. Just like in linear regression, diagnostic procedures for checking model accuracy are based on residuals along with the dfbeta, but the definition of residuals for the PH regression model is not as intuitive as that of linear regression as there are a number of different residuals which serves different purposes. But in general, residuals measure the difference between the observed data and the expected data under the assumption of the model (Collett, 2015). Three types of residuals which will be used in this current study are the deviance residuals, the martingale residuals and the Schöenfeld residuals and are explained briefly in the next subsections. Martingale residuals were used to examine the functional form for continuous covariates and the hazard, the Schöenfeld residuals for examining the PH assumption and the deviance residuals for examining the overall fit of the model (Therneau and Grambsch, 2013; Collett, 2015).

2.3.1 Martingale Residuals

These residuals are derived from the counting process formulation of Fleming and Harrington (1991). The martingale residual process is defined as:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)e^{\hat{\beta}^\intercal X_i(u)}du, \quad i = 1, \dots n,$$

where $N_i(t)$ and $Y_i(u)$ are the event counting process and the "at risk" process, respectively. Under the PH model and for fixed covariates, the residuals takes the simple form given by:

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i)e^{\hat{\beta}^\intercal X_i} \equiv \delta_i - \hat{E}_i, \tag{6}$$

where δ_i is the event indicator and t_i is the observed time for the i^{th} subject. Looking at Equation 6, martingale residuals can be defined as the difference between the observed and the conditionally expected number of events for the i^{th} subject in the interval $(0 - t_i)$, given the fitted model. In general, martingale residuals ranges from $-\infty$ to $+1$, have a mean 0, sum to zero, are positively skewed (not symmetrically distributed around 0 even if fitted model is correct) and are asymptotically uncorrelated (Therneau and Grambsch,

2013; Collett, 2015). Plots of martingale residuals (obtained from the null PH model) against ordered survival time (age at first childbirth) for each continuous covariate can be used to determine the functional form that is required for the covariate. A straight line is an indication that a linear functional form of the covariate is needed, otherwise some transformations may be required. The LOWESS (locally weighted scatter-plot smoother) smoothing was used to smooth the curve fitted to the scatter plot, as graphs obtained here are usually quite "noisy" (Cleveland, 1979; Collett, 2015).

2.3.2 Deviance Residuals and Delta-betas

According to (Therneau and Grambsch, 2013) deviance residuals are a modification or transformation of martingale residuals so as to produce symmetric residuals about 0 when the fitted model is appropriate. For fixed covariates, the deviance residuals are defined as:

$$\hat{r}_{D_i} = sign(\hat{M}_i)\sqrt{-\hat{M}_i - \delta_i log(\delta_i - \hat{M}_i)},$$

where $\hat{M}_i$) are the martingale residuals for the i^{th} subject and $sign(\hat{M}_i)$ is a sign function which ensures that these two types of residuals have the same sign. Deviance residuals are symmetrically distributed around 0 and plots of these residuals against the risk score ($\boldsymbol{\beta}^\intercal \boldsymbol{X}$) can be used to reveal individual outliers (Therneau and Grambsch, 2013; Collett, 2015). These residuals follow a standard normal distribution and thus residuals which lie outside the range -2 and +2 are considered as outlying observations (outliers). The delta-betas were used to further assess the model accuracy by identifying influential observations, i.e. those that can have a huge impact on the model-based inferences. Given the j^{th} parameter estimate $\hat{\beta}_j$ for $j = 1, \ldots, p$, the delta-beta can be defined as $\Delta_i\hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_{j(i)}$, which is the change in $\hat{\beta}_j$ after omitting the i^{th} observation, for $i = 1, \ldots, n$. An observation is considered *influential* if its removal from the data will result in a huge increase or decrease of the relative hazard, i.e. if $\Delta_i\hat{\beta}_j$ is large in absolute value. (Collett, 2015).

2.3.3 Schöenfeld Residuals

Schöenfeld residuals are defined for each unique event time and, unlike martingale and deviance residuals, they do not depend on the survival time and the cumulative hazard function (Collett, 2015). Schöenfeld residuals are obtained for every covariate that is included in the PH model, implying that a set of these residuals can be obtained for each subject. The j^{th} Schöenfeld residual is given by:

$$r_{s_{ji}} = \delta_i \left\{ x_{ji} - \left(\frac{\sum_{l \in R(t_{(i)})} x_{ji} \exp(\boldsymbol{X}_{\boldsymbol{l}}^\intercal \boldsymbol{\beta})}{\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{X}_{\boldsymbol{l}}^\intercal \boldsymbol{\beta})} \right) \right\},$$

where δ_i is the event indicator as described earlier, $R(t_{(i)})$ is the risk set at time t_i and x_{ji} is the value of the j^{th} covariate for the i^{th} subject. According to Collett (2015), these residuals have mean 0, sum up to 0 and are uncorrelated. They are used to detect departures from the PH assumption but are computationally expensive. A more effective way to test the

PH assumption is to use scaled Schöenfeld residuals which are easy and straight forward to compute (Therneau and Grambsch, 2013). These residuals are given by:

$$r^*_{s_i} = d[var(\hat{\beta})]r_{s_i},$$

where d is the number of events and $var(\hat{\beta})$ is the variance-covariance matrix of the parameter estimates of the fitted model. According to Hosmer Jr et al. (2008), violation of the PH assumption has huge impact on the hazard ratio as it can no longer have a real meaning and interpretation. So to enable more precise interpretation, violation should be taken into account by appropriate modification of the model such as researching intensively on variable selection; stratifying by the "culprit" variable(s) or extending the model by including time-dependent variable(s) (Collett, 2015). In this study, the violation was visualized by constructing plots of scaled Schöenfeld residuals against time for each covariate. These residuals are independent of time under the PH assumption so the variable(s) causing violation of the PH assumption can easily be identified.

2.3.4 Least Absolute Shrinkage and Selection Operator (LASSO)

Prior to fitting the multivariate PH model, the LASSO was used for variable selection. As explained by James et al. (2013), the LASSO regression analysis is a shrinkage and variable selection method for linear regression models as well as Cox PH models. The goal here was to obtain the subset of predictors that have huge impact on age at first childbirth as well as getting rid of those variables with no or little impact on the response. LASSO does this by imposing a constraint on the model parameters that causes regression coefficients for some variables (those with little impact) to shrink toward zero, and thereby excluded from the model. Variables with non-zero regression coefficients are most strongly associated with the response variable, hence the LASSO regression analysis helps determine which predictors are most important (Tibshirani et al., 2015). The objective of the LASSO in Cox PH regression modelling is to estimate the $\beta-$parameters by maximizing the partial likelihood function $\mathcal{L}(\mathcal{B})$ in Equation 5, subject to the constraint that $\sum_{j=1}^{p} |\beta_j| \leq s$, for some value of s known as the "budget" (James et al., 2013). The parameter estimates maximize;

$$\mathcal{L}_\lambda(\mathcal{B}) = \mathcal{L}(\mathcal{B}) - \lambda \sum_{j=1}^{p} |\beta_j|, \tag{7}$$

where $\lambda \sum_{j=1}^{p} |\beta_j|$ is known as the LASSO penalty, $\sum_{j=1}^{p} |\beta_j|$ is the L_1-norm and λ is a tuning parameter which controls the amount of shrinkage and is determined by cross-validation (James et al., 2013; Collett, 2015). There exist several types of cross-validation but in this study K-fold cross-validation was used. In general, cross-validation involves dividing the data set into K random data subsets or folds of roughly equal sizes, without replacement, and then use the K-1 folds as the training set while the remaining fold is used as the validation set. This is repeated K times until each fold has been used in both training and validation sets (James et al., 2013; Tibshirani et al., 2015). For each iteration and a chosen value of λ, the partial likelihood deviance is computed and stored and the average or mean and standard error values chosen after all iterations are done. The process is then

repeated using a different value of λ until all λ values in the range have been used. Finally, the value of λ which yields the minimum and one standard error of the partial likelihood deviance, denoted lambda.min and lambda.1se, respectively, is selected as the best value to use (Tibshirani et al., 2015).

2.4 Logistic Regression Model

The logistic regression model sometimes called the logistic model or the logit model, is a special case of a generalized linear model which analyzes the relationship between multiple independent variables and a categorical response variable (Agresti, 2018). Logistic regression estimates the probability of occurrence of an event by fitting data to a logistic curve and the response variable may be quantitative, categorical, or a mixture of the two. It is common practice to assume that the outcome variable, denoted as Y, is a categorical and a dichotomous variable having either a success or failure as the outcome (Agresti, 2018). There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical (Agresti, 2018). In a study of determinants of pregnancy outcome, for example, the response variable may be normal births vs adverse birth outcome (stillbirths, miscarriages or other abnormalities) at the time of the survey. In this kind of situation, the standard multiple regression analysis becomes inappropriate as the response and predictors cannot be related through a linear relationship. When the response variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed. The general logistic regression model for the i^{th} subject with binary response variable y_i, which is Bernoulli distributed, can be written as:

$$\pi_i = \frac{\exp(\alpha + \boldsymbol{x_i^\intercal}\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x_i^\intercal}\boldsymbol{\beta})}, \tag{8}$$

where α is an intercept term, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of odds ratio parameters (Agresti, 2018). The variable y_i takes on the value 1 with probability $\pi_i = P(y_i = 1|x_i) = 1 - P(y_i = 0|x_i)$, where $x_i = (x_1, \ldots, x_p)$ is the subjects covariate vector, and value 0 with probability $1 - \pi_i$. Equivalently, from Equation 7, the logit (log odds) has a linear relationship given by:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + \boldsymbol{x_i^\intercal}\boldsymbol{\beta} = \alpha + \sum_{i=1}^{p} \beta_i x_i. \tag{9}$$

The odds in the logistic regression model is transformed using the natural logarithm in order to give values of $\alpha + \sum_{i=1}^{p} \beta_i x_i$ that fall between 0 and 1 since it calculates the probability of an event occurring over the probability of an event not occurring (Agresti, 2018). The regression coefficient, say β_1 is interpreted as the estimated increase or decrease in the log odds of the response per unit increase in the value of the predictor variable. In other words, exponentiating both sides of Equation 9 shows that the odds are an exponential function of x_i and $\exp \beta_1$ can be interpreted as the odds ratio (OR) associated with a one

unit increase in the independent variable x_1, keeping other variables constant. The OR can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome. An $OR = 1$ indicates exposure does not affect odds of outcome, $OR > 1$ indicates exposure associated with higher odds of outcome and $OR < 1$ indicates exposure associated with lower odds of outcome (Agresti, 2018). This binary logistic regression was fit to the data to study the determinants of the pregnancy outcome (Pregnancy Terminated).

2.4.1 The Hosmer and Lemeshow Goodness-of-Fit Test

In logistic regression, there are various possible methods of assessing goodness-of-fit of a model, which include the deviance statistics, the Pearson statistics and the Hosmer and Lemeshow (HL) statistics (Agresti, 2018; Hosmer and Lemeshow, 2000). The goodness-of-fit test of all these three statistics are based on the likelihood-ratio test between the fitted model and the full or most saturated model. In this current study, the HL statistic was used because the deviance and Pearson goodness-of-fit tests require sufficient replication within sub-populations to be valid. Also, the presence of one or more continuous variables in the model will cause the data to be too sparse to use these statistics (are affected by the number of trials per row in the data). The HL statistics which was proposed by Hosmer and Lemeshow (2000) through simulations and available only for binary response models, does not suffer from these drawbacks and hence was used in this study. The HL goodness of fit test is based on splitting the sample observations according to their predicted probabilities, or risks from the logistic regression model and forming the number of groups, g. These groups are formed in such a way that the first group consists of the observations with the smallest predicted probabilities and the last group consists of the observations with the largest predicted probabilities. To be specific, the predicted values are arranged in increasing order, and then separated into several groups of approximately equal size. According to Hosmer and Lemeshow (2000), the standard recommended number of groups is 10 if the number of of variables is less than 10, and there must be at least three groups in order for the statistic to be computed. For each group, both the observed and the expected number of events and non-events are calculated. The expected number of events is obtained by summing the predicted probabilities for all the individuals in the group, and the expected number of non-events is the difference between the group size and the expected number of events. To assess how well the model fits the data, the HL statistic compares the observed and expected frequencies of events and non-events, i.e. the statistic is obtained by calculating the Pearson chi-square statistic from the $g \times 2$ table of observed and expected frequencies, where g is the number of groups (Hosmer and Lemeshow, 2000). The statistic is given by:

$$G^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - n_i\bar{\pi}_i)^2}{n_i\bar{\pi}_i(1-\bar{\pi}_i)}, \tag{10}$$

where n_i is the total frequency of subjects in the i^{th} group, O_i is the total frequency of event outcomes in the i^{th} group, and $\bar{\pi}_i$ is the average estimated probability of an event

outcome for the i^{th} group. Given the model is correctly specified, the G^2_{HL} test statistic approximately follows a chi-squared distribution on g-2 degrees of freedom. A large value of G^2_{HL} (and a small p-value) is an indication that the predicted probabilities deviate from the observed probabilities meaning that the model is a poor fit. However, this statistic may suffer from drawbacks such as failing to take overfitting into account which may result in low power and different values of g may result in changes in p-values.

The stepwise procedure was used to select variables to be included in the model, whereby the initial step is to start with a model with no predictors and then either enter or remove a predictor based on the partial F-tests (t-tests for the slope parameters) at each step along the way. The stopping rule is when no more predictors can be justifiably entered or removed from the model, resulting in the final model. The significance levels to enter and remove variables from the model are 0.15 and 0.05, respectively (Kutner et al., 2005).

It should be noted that prior to fitting both the PH model and the multivariate logistic regression model, cross tabulations were done using χ^2 testing and the issue of collinearity was checked by performing association tests among the covariates. This was done because if the regression models are fit in the presence of multicollinearity, the results can be hugely affected. One-Way ANOVA, a test of independence, was used to analyze the relationships between numerical and categorical variables, while a χ^2 test was used to determine if there was a statistically significant relationship between two categorical variables (Kutner et al., 2005). The strength of association among these variables was then measured using Cramer's V association statistics (based on Pearson's χ^2 statistic), which gives a value between 0 and 1. A value of Cramer's V statistics that is above 0.7 is an indication that the variables are strongly associated (highly correlated) which may imply the presence of collinearity (Goodman and Kruskal, 1979). Cramer's V association statistic is given by:

$$V = \sqrt{\frac{\chi^2}{n\mathcal{V}}},$$

where n is the number of observations, $\mathcal{V} = min(r-1, c-1)$ is the degrees of freedom, and r and c are the number of rows and columns in the contingency table, respectively.

Variance inflation factor (VIF), a formal test for detecting the presence of multicollinearity, was computed for all explanatory variables that were included in the final model(s). These factors measure how much the variances of the estimated regression coefficients are inflated as a result of multicollinearity in the model. A VIF value that is greater than 10 indicates the presence of multicollinearity (Kutner et al., 2005). Since the calculation of VIF is not based on the response variable for a model, linear regression procedure was used to calculate these factors. The formula for VIF is as follows;

$$(VIF)_j = (1 - R^2_j)^{-1} \quad j = 1, 2, \ldots, p-1,$$

where R^2_j is the coefficient of multiple determination when covariate X_j is regressed on all the other covariates in the model.

3 Results

3.1 Data cleaning

A cross-sectional representative data obtained from DHS 2011 for Mozambique was used for this study. Several data manipulations and cleaning were done prior to analysis in order to remove some unrealistic values such as zero age at first sex, BMI values greater than 60kg/m^2 and some unrealistic values of weight and height. Some new variables were created from the existing ones, for example variables *"survtime"* and *"survevent"*, which represent the survival time and event of interest, respectively. The variable *"survevent"* refers to the occurrence of the event of interest, i.e. giving birth to first child in this case. The *"survtime"* variable refers to the survival time which is defined as the the age of the women at first childbirth, or the current age of women who have not yet given birth as of the time of the survey. A censor variable was therefore created whereby women who had given birth to first child were given a censoring index coded "1", "0" otherwise. Variables *"Contraceptives"* and *"Knowledge of cycle"* were dichotomized as some strata had very small sample sizes which were not sufficient for analysis and comparison purposes. For the exploratory analysis, the variables *"age at first sex"* and *"age at first birth"* were categorised in order to study the distribution of respondents by age groups.

3.2 Exploratory Data Analysis

After data cleaning, a total of 18 989 subjects was used for the final analyses. Among these, 3 933(20.7%) subjects did not have a child and were right censored, while 15 056(79.3%) subjects had experienced the event of interest (had given birth to first child). Table 3 shows the distribution of respondents by characteristics, birth history and summaries of the age at first childbirths. The median overall survival time (i.e. age at first childbirth) was 19 years, obtained using the Kaplan-Meier estimate of the survival function. Just above half (50.62%) of the total population had their first child within the age interval 16 to 21 years, 38.49% within age interval 21-30 years, 6.97% at 30 years or later while 3.92% before attaining the age 16 years.

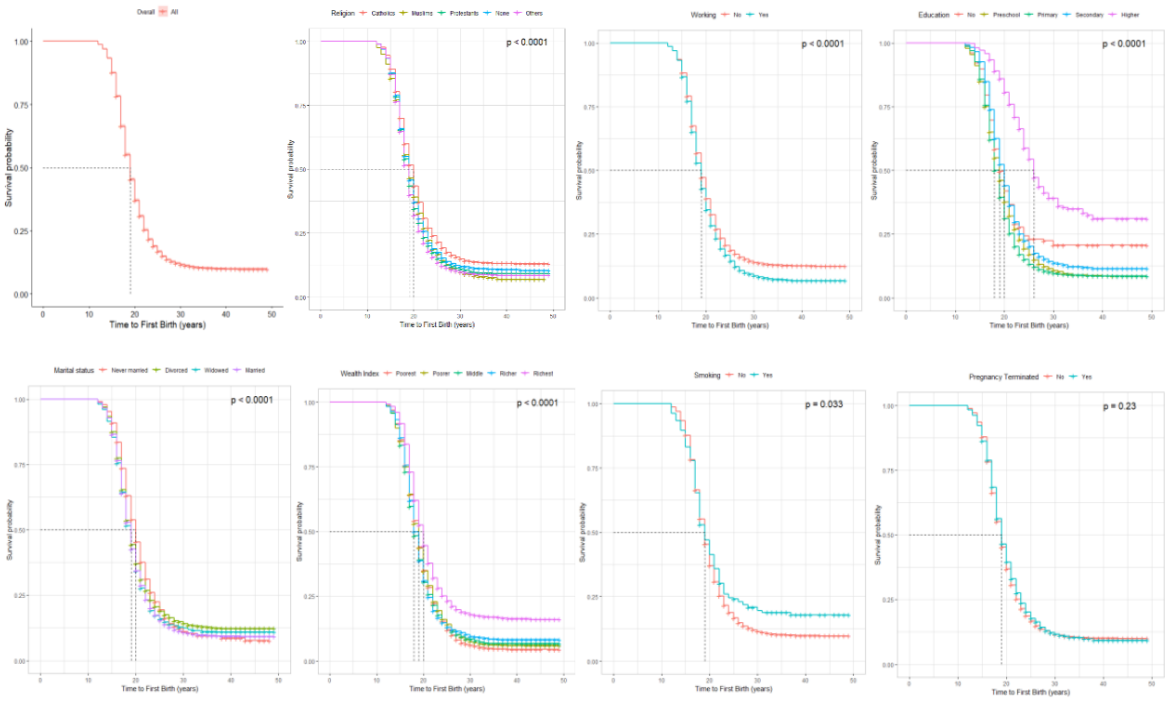
Among women with low education (*Pre-school*), 87.13% had given birth to at least a child with median age at first birth of 19 years, while 54.55% of women with higher education had given birth to at least a child, with median age at first birth of 26 years. Increasing education status, from primary to higher education, seems to prolong age at first childbirth. About 63% of women who were never married experienced the event of interest compared to about 85% of women who were married, with median survival time of 20 and 19 years, respectively. Looking at the wealth index categories, 87.84% of the poorest group had given birth to at least a child with median age at first birth of 19 years, compared to 71.20% of the the richest group with median age at first birth of 20 years. Catholics seem to have a higher median survival time of 20 years compared to other religions and non-religious groups. The *Ethnicity* categories show similar trends in terms of the percentage to experience the event of interest except the Portuguese ethnic group, which had a higher median survival time (median=22 years) compared to other groups (median=19 years).

Table 3: Summary Statistics

Characteristic	Total (N)	Ever given Birth (% Events)	Timing of First Childbirth among Ever given Birth				Median (years)	95% C.I. (Greenwood s.e)
			< 16	16-21	21-30	30+		
Education								
Don't know	175	135(77.14)	7(5.19)	66(48.89)	57(42.22)	5(3.70)	19	(19-20)
Preschool	4281	3730(87.13)	198(5.31)	1714(45.95)	1447(38.79)	371(9.95)	19	(19-19)
Primary	9314	7655(82.19)	298(3.89)	4181(54.62)	2774(36.24)	402(5.25)	18	(18-19)
Secondary	4812	3314(68.87)	86(2.60)	1616(48.76)	1407(42.46)	205(6.19)	20	(20-20)
Higher	407	222(54.55)	1(0.45)	44(19.82)	110(49.55)	67(30.18)	26	(25-27)
HH_MS								
Never married	4709	2956(62.77)	103(3.48)	1455(49.22)	1196(40.46)	202(6.83)	20	(20-20)
Divorced	2011	1690(84.04)	61(3.61)	870(51.48)	638(37.75)	121(7.16)	19	(19-19)
Widowed	821	721(87.82)	34(4.72)	363(50.35)	277(38.42)	47(6.52)	19	(18-19)
Married	11448	9689(84.63)	392(4.05)	4933(50.91)	3684(38.02)	680(7.02)	19	(19-19)
Wealth Index								
Poorest)	2031	1784(87.84)	62(3.48)	845(47.37)	702(39.35)	175(9.81)	19	(19-19)
Poorer)	2473	2111(85.36)	105(4.97)	1021(48.37)	796(37.71)	189(8.95)	19	(19-19)
Middle)	2918	2463(84.41)	130(5.28)	1331(54.04)	839(34.06)	163(6.62)	18	(18-19)
Richer)	4126	3400(82.40)	153(4.50)	1846(54.29)	1236(36.35)	165(4.85)	18	(18-19)
Richest)	7441	5298(71.20)	140(2.64)	2578(48.66)	2222(41.94)	358(6.76)	20	(20-20)
Smoking								
Yes	208	167(80.29)	14(8.38)	84(50.30)	58(34.73)	11(6.59)	19	(18-20)
No	18781	14889(79.28)	576(3.87)	7537(50.62)	5737(38.53)	1039(6.98)	19	(19-19)
Religion								
Catholics	4583	3393(74.03)	105(3.09)	1658(48.87)	1336(39.38)	294(8.66)	20	20-20
Muslims	2229	1806(81.02)	117(6.48)	837(46.35)	690(38.21)	162(8.97)	19	19-19
Protestants	4429	3642(82.23)	132(3.62)	1942(53.32)	1371(37.64)	197(5.41)	19	19-19
Non-religious	4023	3253(80.86)	120(3.69)	1675(51.49)	1258(38.67)	200(6.15)	19	19-19
Others	3725	2962(79.52)	116(3.92)	1509(50.95)	1140(38.49)	197(6.65)	19	19-19
Age at First Sex (Categorized)								
< 14	1982	1780(89.81)	467(26.24)	1011(56.80)	258(14.49)	44(2.47)	15	(15-15)
14-19	15275	11930(78.10)	120(1.01)	6573(55.10)	4486(37.60)	751(6.30)	19	(19-19)
19-25	1658	1311(79.07)	3(0.23)	37(2.82)	1050(80.09)	221(16.86)	22	(22-22)
25+	74	35(47.30)	0(0.00)	0(0.00)	1(2.86)	34(97.14)	32	(30-NA)
Ethnicity								
Emakhuwa	2269	1861(82.02)	88(4.73)	942(50.62)	675(36.27)	156(8.38)	19	(19-19)
Portuguese	2662	1555(58.41)	12(0.77)	682(43.86)	686(44.12)	175(11.25)	22	(21-22)
Xichangana	4496	3621(80.54)	83(2.29)	1823(50.35)	1570(43.36)	145(4.00)	19	(19-19)
Cisena	2021	1698(84.02)	90(5.30)	854(50.29)	623(36.69)	131(7.71)	19	(18-19)
Elomwe	654	560(85.63)	29(5.18)	261(46.61)	214(38.21)	56(10.00)	19	(19-19)
Echuwabo	615	512(83.25)	36(7.03)	256(50.00)	177(34.57)	43(8.40)	19	(18-19)
Others	6272	5249(83.69)	252(4.80)	2803(53.40)	1850(35.24)	344(6.55)	18	(18-19)
Pregnancy Terminated								
Yes	1932	1555(80.49)	75(4.82)	752(48.36)	601(38.65)	127(8.17)	19	(19-19)
No	17057	13501(79.15)	515(3.81)	6869(50.88)	5194(38.47)	923(6.84)	19	(19-19)
Knowledge of Cycle (Dichotomized)								
Yes	17246	13737(79.65)	551(4.01)	6940(50.52)	5269(38.36)	977(7.11)	19	(19-19)
No	1743	1319(75.67)	39(2.96)	681(51.63)	526(39.88)	73(5.53)	19	(19-19)
Knowledge of contraceptives								
Modern method	18461	14662(79.42)	570(3.89)	7448(50.80)	5641(38.47)	1003(6.84)	19	(19-19)
No method	513	388(75.63)	20(5.15)	168(43.30)	153(39.43)	47(12.11)	20	(19-20)
Folkloric method	9	5(55.56)	0(0.00)	4(80.00)	1(20.00)	0(0.00)	19	(14-NA)
Traditional method	6	1(16.67)	0(0.00)	1(100.00)	0(0.00)	0(0.00)	NA	NA
Working								
Yes	7665	6585(85.91)	238(3.61)	3304(50.17)	2512(38.15)	531(8.06)	19	(19-19)
No	11324	8471(74.81)	352(4.16)	4317(50.96)	3283(38.76)	519(6.13)	19	(19-19)
Region								
Cabo Delgado	1156	932(80.62)	32(3.43)	469(50.32)	354(37.98)	77(8.26)	19	(19-19)
Gaza	1940	1524(78.56)	33(2.17)	715(46.92)	712(46.72)	64(4.20)	19	(19-19)
Inhambane	1756	1368(77.90)	51(3.73)	712(52.05)	489(35.75)	116(8.48)	19	(19-19)
Manica	1380	1185(85.87)	39(3.29)	730(61.60)	366(30.89)	50(4.22)	18	(18-18)
Maputo Cidade	2981	2142(71.86)	57(2.66)	1006(46.97)	914(42.67)	165(7.70)	20	(20-20)
Maputo Provincia	2238	1694(75.69)	36(2.13)	889(52.48)	691(40.79)	78(4.60)	19	(19-19)
Nampula	1009	825(81.76)	28(3.39)	438(53.09)	285(34.55)	74(8.97)	19	(18-19)
Niassa	999	844(84.48)	93(11.02)	410(48.58)	285(33.77)	56(6.64)	18	(18-19)
Sofala	2694	2138(79.36)	104(4.86)	1073(50.19)	777(36.34)	184(8.61)	19	(19-19)
Tete	1287	1116(86.71)	36(3.23)	585(52.42)	446(39.96)	49(4.39)	19	(18-19)
Zambezia	1549	1288(83.15)	81(6.29)	594(46.12)	476(36.96)	137(10.64)	19	(19-19)
Total	18989	15056 (79.3)	590(3.92%)	7621(50.62%)	5795(38.49%)	1050(6.97%)	19	(19-19)

A similar trend is also observed for the different regions although some differ slightly in terms of the median survival time which varies from 18 to 20 years among all regions, with

Maputo Cidade having the largest value while Manica and Niassa having the smallest values. The median age was the same (19 years) irrespective of the categories of the variables *Smoking*, *PregTerm*, *KnowOfCycle* and *Working*. As expected, the median survival time for the categorized *age at first sex* increases with increasing age group. It can be observed that most subjects had their first children between the ages of 14 and 19 years and only a few after the age of 25 years. All these explanations can also be seen in Figure 1 which shows the KM survival plots for all the variables together with the overall survival function. The p-values shown in the plots are as a result of the log-rank test to compare the survival distributions across groups/categories for each variable. A statistically significant p-value ($p < 0.05$) is an indication that the time to first childbirth might be different for at least one pair, otherwise the same for all categories.



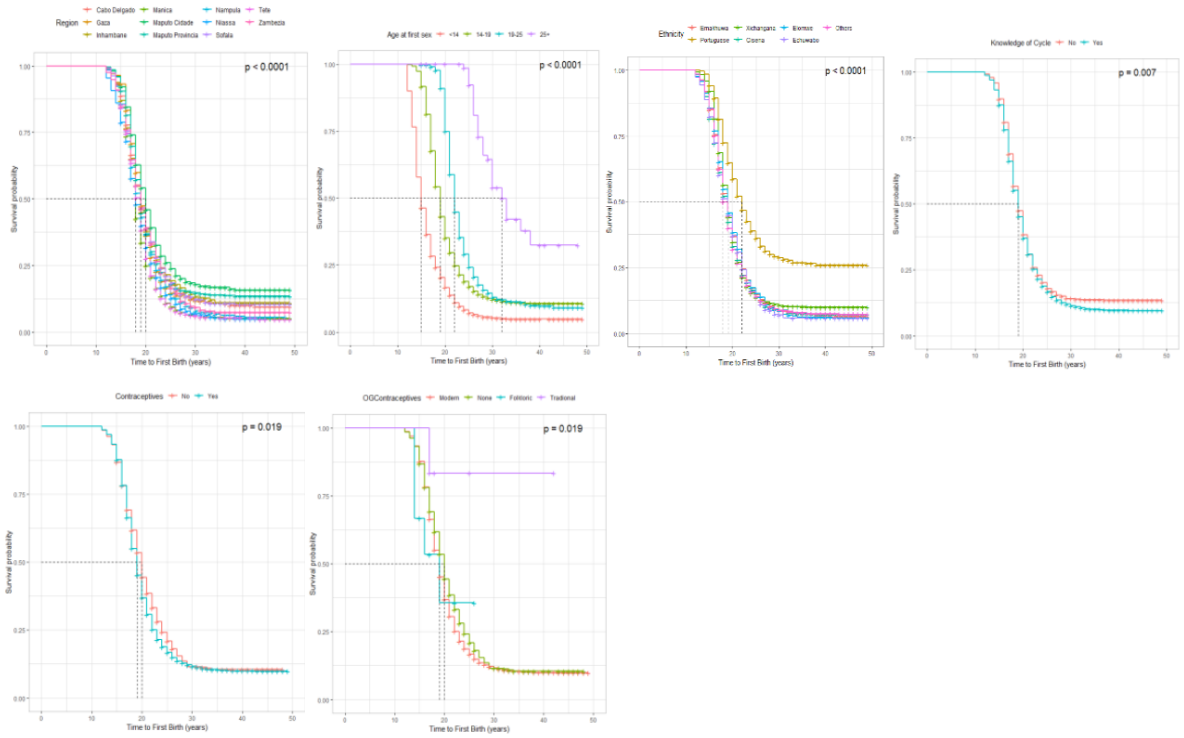


Figure 1: KM Survival functions for patterns of Timing to First Childbirth

It should however, be emphasised that some of the estimated median values may not be precise (unstable) due to the high number of censored observations (Collett, 2015). Also, in examining the chances of survival as a function of two or more predictors, these univariate analysis methods may suffer from several drawbacks such as confounding or bias when comparing some survival curves due to unequal distribution of some factors. They can also be problematic if there are too many strata because there will be too small sample size per stratum as observed in the case of knowledge on the use of contraceptive-original variable. It was therefore important to adjust the analysis to regression methods suitable for survival analysis models (Collett, 2015). The Cox Proportional Hazard model is one way to deal with this limitation and will be discussed next.

3.3 The Proportional Hazards Model (PH Model)

Prior to fitting the multivariate Cox PH model, variable selection was done in order to select those variables which could have a huge impact on the response. This was done by first choosing variables on the basis of association with the response according to exploratory data analysis. Variables like *"PregTerm"*, *"BMI"* and *"Smoking"* were not associated with the response and thus were not included in the LASSO, which was the second procedure used for further selection. The remaining variables were then included in the LASSO model.

3.3.1 LASSO Variable Selection

Figure 2 shows the standardized LASSO coefficients as a function of the L_1-norm, while Figure 3 shows the LASSO plot showing the best lambda selected by 10-fold cross-validation. As indicated in Figure 2, each curve show the value of a particular coefficient for the corresponding L_1-norm value. The plot indicates the order in which the variables are shrunk to zero as the value of L_1-norm or model size becomes small. The first vertical line in Figure 3 indicates the value of λ at which the best minimum cross validation error or minimum partial likelihood deviance (lambda.min) is obtained, while the second vertical line is the maximum value of cross validation error within one standard error of the minimum (lambda.1se). The y-axis is the partial likelihood deviance and the upper x-axis show numbers of estimated non-zero coefficients in model (model size) which correspond to each value of lambda chosen in the lower x-axis (log(Lambda)). By selecting the tuning parameter λ as lambda.1se, coefficients (5, 15, 21 and 27) seem to be more influential while coefficients (1, 3, 4, 6, 7, 8, 10, 12, 13, 16, 17, 18, 25, 29, 31 and 32) seem to be less influential and hence were shrunk to zero (see Appendix Table A.1). The non-zero coefficients were then used as covariates for the multivariate PH model. However, it should be noted that the variable marital status (*HH_MS*) was the only one dropped from the model as all its category coefficients were shrunk to zero while other variables were returned into the model as at least one category had a non-zero coefficient.

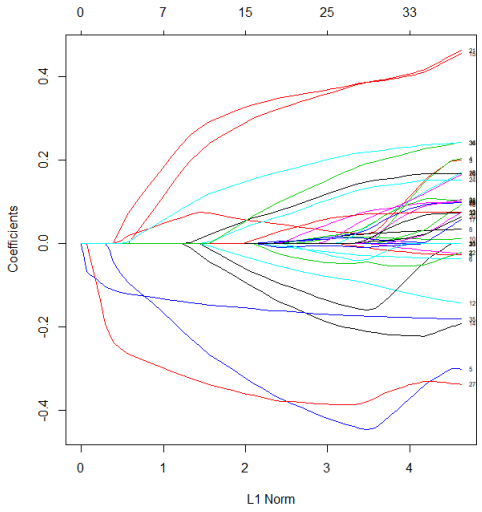


Figure 2: Lasso variable selection

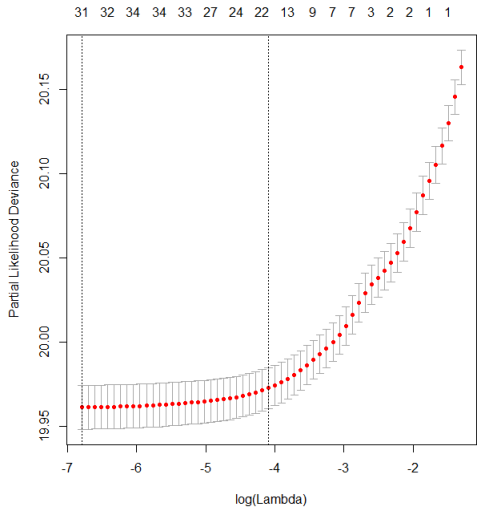


Figure 3: 10-fold cross validation

After performing variable selection, some cross tabulations (see Tables A.3, A.4, A.5 and A.6 in the Appendix) were done using χ^2 testing and the issue of collinearity was checked by performing association tests among the covariates included in the model. A One-Way ANOVA for testing of independence between numerical and categorical variables showed

that there were no association between the continuous variable *age at first sex* and all other categorical variables ($p < 0.05$), except for the variable *Wealth index* ($p = 0.088$). The χ^2 test of association among categorical variables showed significant associations among all other covariates except for the variables *Knowledge of cycle* and *Working* ($p = 0.4258$), so there was need to test for strength of association among these variables before fitting the Cox PH regression analysis. Cramer's V association statistics was used to test for strength of association and no value exceeded 0.7, an indication that none of the variables were highly correlated (see Table A.7 in the Appendix). To further confirm this, variance inflation factor (VIF) was computed for all explanatory variables that were included in the model and there was no indications of serious multicollinearity problems as all the standardised GVIF were far less than 10, for both models (see Table A.8 in the Appendix).

3.3.2 Model checking in Cox PH model

After performing variable selection, the remaining variables were then used to fit the PH model. The Efrons method (Efron, 1977) was used to account for occurrence of ties. As aforementioned, the PH regression model makes two important assumptions. The first assumption is that of a constant hazard ratio over time and the second one is the existence of a log-linear relationship between the hazard and its covariates. It was therefore of paramount importance to check the validity of these assumptions and model adequacy. Diagnostic procedures for checking model accuracy were based on three types of residuals, namely martingale residuals, Schöenfeld residuals and deviance residuals. Martingale residuals were used to examine the functional form for continuous covariates and the hazard, the Schöenfeld residuals were used to examine the PH assumption and the deviance residuals were used to examine the overall fit of the model. To check for influential observations, *delta-betas* were used.

3.3.3 Formal test for PH assumption (Schöenfeld's test)

Table 4 (left) displays the results for the tests of proportional hazard assumption on the original model for each covariate. The p-values of most covariates are statistically significant implying that they do not satisfy the proportional hazard assumption. The p-value associated with a Global Test of non-proportionality suggested strong evidence of non-proportionality for the entire model ($p < 0.001$). The Global Test test the global null hypothesis that the PH assumption is valid. The parameter ρ is the Pearson's correlation between the scaled Schöenfeld residuals and some function of time. If the variable is time-invariant then the slope of the plotted line should be zero. The χ^2 tests are used for testing if the variables are time-invariant and a low p-value is therefore an indication that the PH assumption has been violated (i.e. the Schöenfeld residuals are not constant over time). Plots of scaled Schöenfeld residuals against time were constructed for each covariate in order to determine which covariates were violating the PH assumption. The variables that were deemed most likely to contribute to non-proportionality were *wealth index*, *current*

working status and *age at first sex*, with non constant residuals over time.

Table 4: Schoenfeld's test for PH assumption

Variable	Original model			Modified final model		
	rho (ρ)	Chisq	P-value	rho (ρ)	Chisq	P-value
Pre-school	0.01726	4.52	0.0334	0.0151	3.47	0.0625
Primary	0.0146	3.22	0.0727	0.0169	4.35	0.0371
Secondary	0.0225	7.68	0.0056	0.0134	2.75	0.0970
Higher	0.0428	27.50	< 0.001	0.0081	1.01	0.3138
Poorer	-0.0223	7.55	0.0060	-	-	-
Middle	-0.0506	38.70	< 0.001	-	-	-
Richer	-0.0631	60.80	< 0.001	-	-	-
Richest	-0.0775	92.90	< 0.001	-	-	-
Gaza	0.0032	0.16	0.693	0.0094	1.39	0.2383
Inhambane	0.0016	0.04	0.841	0.0124	2.41	0.1206
Manica	-0.0380	22.20	< 0.001	0.0060	0.57	0.4506
Maputo Cidade	-0.0041	0.25	0.614	0.0069	0.74	0.3905
Maputo Provincia	-0.0043	0.28	0.597	0.0049	0.38	0.5401
Nampula	-0.0297	13.40	< 0.001	0.0028	0.12	0.7276
Niassa	0.0164	4.11	0.0427	0.0071	0.77	0.3810
Sofala	-0.0359	19.70	< 0.001	0.0017	0.04	0.8322
Tete	-0.0320	15.70	< 0.001	0.0085	1.13	0.2870
Zambezia	-0.0046	0.33	0.569	0.0132	2.68	0.1013
Portuguese	-0.0212	6.90	0.0086	-0.0023	0.08	0.7722
Xichangana	-0.0133	2.71	0.0999	-0.0084	1.08	0.2988
Cisena	-0.0114	1.97	0.160	-0.0001	$< .01$	0.9927
Elomwe	-0.0019	0.05	0.818	-0.0073	0.80	0.3722
Echuwabo	-0.0019	0.05	0.817	-0.0067	0.68	0.4082
Others	-0.0127	2.46	0.117	-0.0084	1.09	0.2973
Muslims	0.0264	10.70	0.0011	0.0090	1.27	0.2597
Non-religious	-0.0115	2.00e	0.157	-0.0114	1.97	0.1607
Others	-0.0162	3.94	0.0470	-0.0108	1.78	0.1821
Protestants	-0.0134	2.69	0.101	-0.0091	1.26	0.2625
KnowOfCycle1	0.0021	0.07	0.792	0.00656	0.64	0.4230
Contraceptive1	-0.0192	5.57	0.0183	-0.0111	1.87	0.1715
AgeAt1stSex	0.4809	3840	< 0.001	0.0090	1.20	0.2742
Working1	0.0322	15.70	< 0.001	0.0154	3.60	0.0577
AgeAt1stSex*log(time)	-	-	-	-0.0087	1.12	0.2895
GLOBAL	NA	4330	**< 0.001**	NA	29.40	**0.4439**

Figure 4 and Figure 5 shows plots of the scaled Schöenfeld residuals against time for checking or examining the PH assumption by diagnosing a time-varrying effect (i.e. checking if residuals are constant over time). Figure 4, obtained from the initial model, clearly shows that the residuals are not constant over time indicating that age at first sex may be time-dependent. This was confirmed by the fact that the interaction term was statistically significant with p-value less than 0.0001 (Table 5). Figure 5, obtained after introducing

a *AgeAtFirstSex*time* interaction to the initial model clearly shows that the residuals are now constant over time thus a time-dependent covariate of age at first sex was introduced in the model. The model was further extended by stratifying by variables *wealth index* and *working status* in order to satisfy the PH assumption. Figures A.1, A.2, A.3, A.4, A.5 and A.6 show plots of the Schöenfeld residuals against time, of other covariates other than *age at first sex*, for examining the PH assumption by checking if residuals are constant over time (see Appendix). Table 4 (right) displays the results for the tests of proportional hazard assumption for the modified final PH model. The global test now shows that the assumption is valid for the entire model with a small violation by the variable *Education* (Primary category). A major drawback here is that the stratified variables (*Wealth index* and *working status*) can no longer be used for inference.

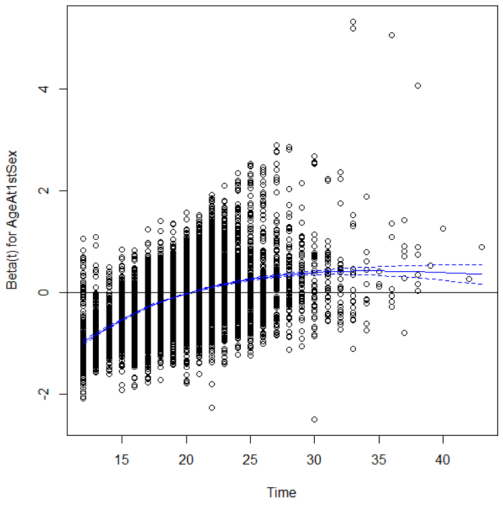


Figure 4: Scaled Schöenfeld residuals against time with smooth spline (Initial model)

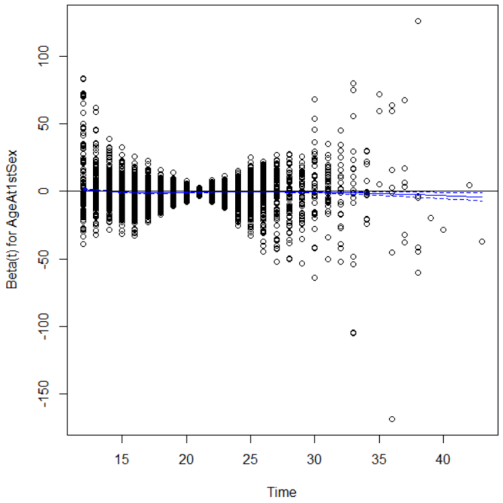


Figure 5: Scaled Schöenfeld residuals against time with smooth spline (Extended model)

3.3.4 Exploring the Functional Form

Martingale residuals, obtained from fitting a PH model that contains no covariates, were used to check the functional form of continuous covariate *"AgeAtFirstSex"*. These residuals were plotted against the values of *"AgeAtFirstSex"* as shown in Figure 6. The LOWESS smoothing was used to smooth the curve (black line) fitted to the scatter plot. The smoothed curve displays a linear functional form of covariate i.e. almost a straight line plot, suggesting that a linear term of *age at first sex* is appropriate to use in the model. The downward slope indicates a negative coefficient of this covariate, implying that engaging in first sexual intercourse in later ages prolongs the timing of first childbirth.

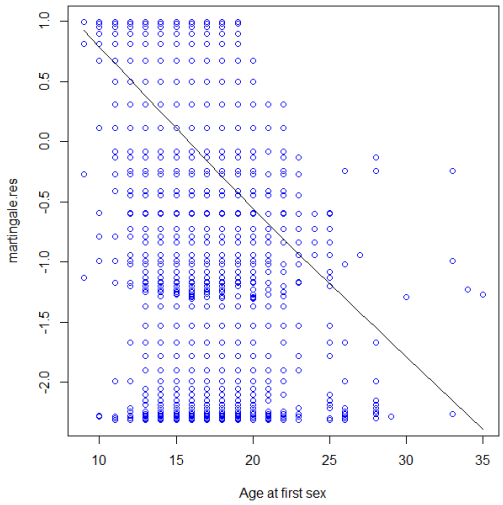


Figure 6: Functional form of Age at first sex

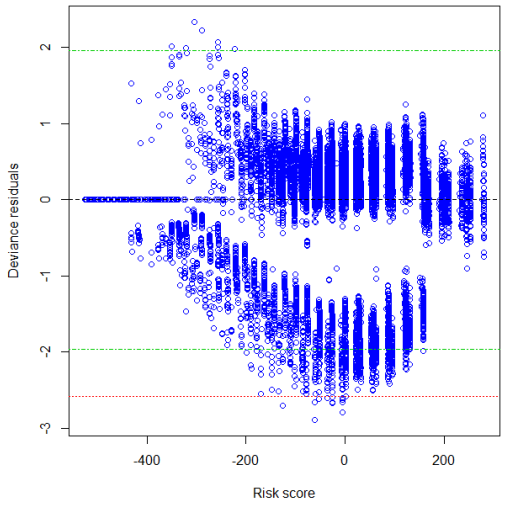


Figure 7: Deviance residuals against values of the risk score

3.3.5 Assessing the Overall Fit of the Model

Deviance residuals (plotted against the risk score) were used to assess the overall fit of the model (Collett, 2015). As shown in Figure 7, the residuals are spread around zero indicating that the model fits the data well. However, there are several observations that lie outside the region -2 and 2 (green lines) which can be seen as potential outliers. About 22 observations lie outside the region -2.5 and 2.5 (red lines) and these can be seen as potential influential observations. To identify these potential influential observations, *delta-betas* were used. This was done by comparing the maximum absolute *delta-betas* of each covariate by their corresponding standard errors from the fitted model. Here, an observation is considered influential if the maximum absolute *delta-betas* of each covariate is greater than their corresponding standard errors (Therneau and Grambsch, 2013). Table A.2 (see Appendix) shows the obtained results and none of the maximum absolute *delta-betas* of each covariate exceeded their corresponding standard errors showing that there were no influential observations. As a result, a model with all observations was considered as the final model.

3.3.6 Proportional Hazard Model Interpretation

The results of the modified final proportional hazard model are displayed in Table 5. The variables *Education status* (Secondary and Higher), *Regions* (Inhambane and Sofala), *Religions* (Protestants, Others and Non-religious), the Portuguese ethnic group, *Contraceptive1*

and *AgeAt1stSex* were statistically significant, while the remaining covariates were not statistically significant at 5% level. Testing for non-proportionality of the hazards is equivalent to testing if the time-dependent coefficient is significantly different from zero, and in this case the interaction of *AgeAt1stSex* and time was statistically significant. From the PH model, the Global Likelihood Ratio (also Wald and Score) test statistics are used for testing global null hypothesis: $\boldsymbol{\beta} = \mathbf{0}$ (i.e. all of the covariates have no "influence" on survival time). All tests yield p-values less than 0.0001, suggesting the survival rates may differ when varying at least one of the variables in the model.

The hazard ratio of first birth among women with secondary education relative to that of women with no education was 0.7210(95% CI: 0.6050; 0.8592), and that of women with higher education relative to that of women with no education was 0.7177(95% CI: 0.574; 0.8965). These hazard ratios and the negative coefficients imply that women with secondary and higher education seem to prolong age at first childbirth compared to those with no education. The p-values of women with pre-school and primary education were not statistically significant, meaning that there was no difference in the timing of first childbirth from those with no education. In the presence of the other variables in the model, Inhambane and Sofala regions appear to have an effect on age at first childbirth. The hazard ratio of a woman who resides in Inhambane was 0.8295(95% CI: 0.7386; 0.9316) relative to that of a woman who resides in Cabo Delgado, and that of a woman who resides in Sofala relative to that of a woman who resides in Cabo Delgado was 0.8596(95% CI: 0.7663; 0.9643). These hazard ratios together with the negative coefficients imply that women who reside in Inhambane and Sofala seem to give birth to first children at a older age (delay age at first childbirth) compared to those who reside in Cabo Delgado. The p-values of women from other regions were not statistically significant, meaning that their age at first childbirth were not different from those who reside in Cabo Delgano.

With respect to the variable *Religion*, Protestants, other religions and the non-religious category appear to have an effect on age at first childbirth, in the presence of other variables. Relative to the reference category, the Catholics religion, the hazard of early first birth was slightly higher among Protestants, Other religions and the Non-religious category with hazard ratios of 1.0760(95% CI: 1.0188; 1.1363), 1.0840(95% CI: 1.0274; 1.1446) and 1.0850(95% CI: 1.0280; 1.1444), respectively. These hazard ratios together with the positive coefficients imply that women who belong to these three religions seem to give birth to first children at an early age compared to those who belong to the Catholic religion. There was no difference between the Muslims and the Catholics in terms of timing of first birth as indicated by the hazard ratio of 0.9747, which close to 1, and a p-value of 0.4499. Among all ethnic groups, only the Portuguese seem to have an effect on age at first childbirth, in the presence of other variables, as indicated by the significant p-value of less than 0.0001. The negative coefficient of -0.2331 implies that women who belong to the Portuguse ethnic group tend to delay early first birth. The hazard ratio of a woman who belong to this ethnic group was 0.7920(95% CI: 0.7101; 0.8835) relative to that of a woman who belong to the Emakhuwa ethnic group. The other ethnic groups had large p-values ($p > 0.05$) indicating that the timing of first childbirth by women belonging to these ethnic groups was similar

to that of their counterparts from the Emakhuwa ethnic group.

Table 5: Proportional hazard model estimates

Variable	coef	exp(coef)	se(coef)	z value	Pr(>|z|)	95% C.I(exp(coef))
Education						
No-education	(Reference)	1.0000				
Pre-school	0.1282	1.1370	0.0894	1.434	0.1516	(0.9540; 1.3546)
Primary	-0.0437	0.9573	0.0880	-0.496	0.6198	(0.8055; 1.1376)
Secondary	-0.3272	0.7210	0.0895	-3.656	0.0003	(0.6050; 0.8592)
Higher	-0.3317	0.7177	0.1135	-2.922	0.0035	(0.574; 0.8965)
Region						
Cabo Delgado	(Reference)	1.0000				
Gaza	-0.0369	0.9639	0.0647	-0.568	0.5698	(0.8491; 1.0942)
Inhambane	-0.1869	0.8295	0.0592	-3.157	0.0016	(0.7386; 0.9316)
Manica	0.0252	1.0260	0.0606	0.416	0.6773	(0.9107; 1.1549)
Maputo Cidade	-0.0433	0.9576	0.0627	-0.692	0.4892	(0.8469; 1.0827)
Maputo Provincia	-0.1024	0.9026	0.0636	-1.611	0.1072	(0.7968; 1.0224)
Nampula	-0.0597	0.9420	0.0515	-1.159	0.2465	(0.8516; 1.0421)
Niassa	0.0503	1.0520	0.0515	0.977	0.3284	(0.9507; 1.1632)
Sofala	-0.1513	0.8596	0.0586	-2.580	0.0099	(0.7663; 0.9643)
Tete	0.0455	1.0470	0.0599	0.760	0.4472	(0.9307; 1.1768)
Zambezia	-0.0820	0.9212	0.0648	-1.267	0.2052	(0.8115; 1.0459)
Religion						
Catholics	(Reference)	1.0000				
Muslims	-0.0256	0.9747	0.0339	-0.756	0.4499	(0.9121; 1.0416)
Non-religious	0.0812	1.0850	0.0274	2.968	0.0030	(1.0280; 1.1444)
Others	0.0811	1.0840	0.0276	2.941	0.0033	(1.0274; 1.1446)
Protestants	0.0732	1.0760	0.0278	2.630	0.0085	(1.0188; 1.1363)
Ethnicity						
Emakhuwa	(Reference)	1.0000				
Portuguese	-0.2331	0.7920	0.0558	-4.182	< 0.0001	(0.7101; 0.8835)
Xichangana	0.0348	1.0350	0.0553	0.628	0.5300	(0.9289; 1.1540)
Cisena	0.0524	1.0540	0.0561	0.934	0.3504	(0.9441; 1.1762)
Elomwe	0.0420	1.0430	0.0730	0.575	0.5654	(0.9038; 1.2034)
Echuwabo	0.0160	1.0160	0.0693	0.230	0.8178	(0.8871; 1.1638)
Others	0.0198	1.0200	0.0457	0.433	0.6654	(0.9326; 1.1155)
Knowledge of cycle						
KnowOfCycle0	(Reference)	1.0000				
KnowOfCycle1	0.0589	1.0610	0.0308	1.912	0.0558	(0.9985; 1.1267)
Contraceptives						
Contraceptive0	(Reference)	1.0000				
Contraceptive1	0.2477	1.2810	0.0533	4.644	< 0.0001	(1.1539; 1.4222)
AgeAt1stSex	-0.7225	0.4855	0.0689	-10.488	< 0.0001	(0.4242; 0.5557)
AgeAt1stSex*log(time)	0.2409	1.2720	0.0228	10.589	< 0.0001	(1.2169; 1.3304)

In the presence of the other variables in the model, the variables *Contraceptive1* (those who had knowledge of contraceptive use) and the age at first sexual intercourse also had

a significant effect on the timing of first childbirth both with p-values less than 0.0001. Women who had knowledge of contraceptive use had a higher hazard of early first birth compared to those who did not. The hazard ratio of a woman who had knowledge of contraceptive use, relative to that of a woman who had no knowledge was 1.2810(95% CI: 1.1539; 1.4222). The hazard of early first birth was 0.4855 per unit increase in woman's age at first sexual intercourse. Also, since the coefficient was negative ($\beta = -0.7225$), then the hazard ratio was decreasing over log(time), implying that a unit increase in woman's age at first sexual intercourse will prolong the timing of first childbirth.

3.4 Logistic Regression

A binary logistic regression was fit to the data and the probability of having ever had a pregnancy termination (*PregTerminated1*) was modelled. Prior to fitting the model, stepwise method was used for variable selection and the variable *age at first childbirth* was also included in the final model as it was the main variable of interest. The Hosmer and Lemeshow (HL) test was used to test for the overall goodness of fit of the model and the predicted values were partitioned into 10 groups of approximately equal size as shown in Table A.9 (see Appendix). The results of the HL goodness of fit test are shown in Table 6.

Table 6: HL G.O.F Test

Chi-Square	d.f	p-value
13.5466	8	0.0944

The p-value is not statistically significant (p=0.0944), suggesting that the model fits the data well and there is no need to include any interactions or non-linear terms in the model. The results of the logistic regression are shown in Table 7 and variables like *AgeAt1stSex*, *Smoking*, *Working status*, *Contraceptives*, *Region*, *Ethnicity*, *Wealth index* and *Education* had significant effect on pregnancy outcome. The variables marital status (*HH_MS*) and age at first childbirth (*AgeAt1stBirth*) had no effect on pregnancy outcome. An odds ratio above 1 is an indication that the variable is more likely to result in the woman having experienced a pregnancy termination, holding other variables constant.

Keeping other variables fixed, the conditional odds ratio of having had a pregnancy termination (vs no pregnancy termination) for those who had knowledge of contraceptives was almost 3 times more (OR 2.999, 95% CI 1.714-5.249, p< .0001) as compared to those who had no knowledge. The variable *Working status* was significantly associated with having had a pregnancy termination, with an odds ratio of 1.906 (95% CI 1.718-2.113, p< .0001) in subjects who were working compared to those who were not. This means that women who were working had almost double the risk of having had a pregnancy termination than those who were not working. The fitted model also suggest that, holding other variables at a fixed value, the odds of pregnancy termination for women who smoked over the odds of those of non-smokers was 1.619 (95% CI 1.071-2.449, p=0.0233). In terms of percentage change, the odds of having had a pregnancy termination for women who smoke was 61.9% higher than that of non-smokers.

Table 7: Logistic Regression ML Parameter estimates

Parameter	d.f	Estimate	s.e	Wald χ^2	p-value	OR	95% Wald C.I.
Intercept	1	-2.1327	0.2714	61.7576	< .0001	-	-
Education							
No-education	(Reference)						
Pre-school	1	-0.2261	0.0804	7.9025	0.0049	0.776	(0.486-1.239)
Primary	1	-0.0477	0.0642	0.5536	0.4569	0.928	(0.590-1.460)
Secondary	1	0.0431	0.0695	0.3835	0.5357	1.016	(0.641-1.610)
Higher	1	0.2037	0.1265	2.5905	0.1075	1.193	(0.702-2.028)
Wealth Index							
Poorest	(Reference)						
Poorer	1	-0.1020	0.0716	2.0327	0.1539	0.943	(0.749-1.187)
Middle	1	-0.1717	0.0665	6.6668	0.0098	0.880	(0.698-1.109)
Richer	1	0.0277	0.0561	0.2445	0.6210	1.074	(0.859-1.342)
Richest	1	0.2895	0.063	21.1323	< .0001	1.102	(1.102-1.766)
Region							
Zambezia	(Reference)						
Cabo delgado	1	0.5358	0.1103	23.6093	< .0001	3.095	(2.085-4.592)
Gaza	1	0.1951	0.1016	3.6881	0.0548	2.201	(1.501-3.227)
Inhambane	1	-0.1357	0.0902	2.2648	0.1323	1.581	(1.096-2.282)
Manica	1	0.1023	0.0981	1.0882	0.2969	2.006	(1.389-2.897)
Maputo Cidade	1	0.1717	0.0819	4.3955	0.0360	2.150	(1.503-3.077)
Maputo Provincia	1	0.5656	0.0803	49.6766	< .0001	3.188	(2.229-4.561)
Nampula	1	0.2334	0.1395	2.7996	0.0943	2.287	(1.481-3.532)
Niassa	1	-0.3187	0.1400	5.1829	0.0228	1.317	(0.854-2.030)
Sofala	1	0.0141	0.0907	0.0242	0.8764	1.837	(1.302-2.590)
Tete	1	-0.7698	0.1437	28.7063	< .0001	0.839	(0.548-1.285)
Ethnicity							
Emakhuwa	(Reference)						
Portuguese	1	-0.0569	0.0859	0.4385	0.5079	0.916	(0.676-1.241)
Xichangana 3	1	-0.1207	0.086	1.9708	0.1604	0.859	(0.633-1.165)
Cisena 4	1	-0.2477	0.1119	4.8961	0.0269	0.757	(0.532-1.075)
Elomwe 5	1	-0.5016	0.2121	5.5944	0.0180	0.587	(0.336-1.024)
Echuwabo 6	1	0.8427	0.1244	45.8815	< .0001	2.251	(1.523-3.328)
Others 7	1	0.0529	0.0707	0.561	0.4538	1.022	(0.782-1.335)
Contraceptives							
Contraceptive0	(Reference)						
Contraceptive1	1	0.5492	0.1427	14.8018	0.0001	2.999	(1.714-5.249)
Working status							
Working0	(Reference)						
Working1	1	0.3224	0.0264	149.1642	< .0001	1.906	(1.718-2.113)
Marital status							
Never married	(Reference)						
Divorced	1	0.0341	0.0626	0.2961	0.5864	1.260	(1.055-1.504)
Widowed	1	0.1252	0.0878	2.0334	0.1539	1.380	(1.078-1.767)
Married	1	0.0377	0.0435	0.7526	0.3857	1.265	(1.114-1.435)
Smoking							
smoking0	(Reference)						
Smoking1	1	0.2410	0.1055	5.2191	0.0223	1.619	(1.071-2.449)
AgeAt1stSex	1	-0.0320	0.0129	6.1456	0.0132	0.969	(0.944-0.993)
AgeAt1stBirth	1	0.0034	0.0043	0.6322	0.4265	1.003	(0.995-1.012)

The variable *age at first sex* had a negative effect on pregnancy outcome, with about 3% decrease (OR 0.969, 95% CI 0.944-0.993, p=0.0132) in the odds of having had a pregnancy termination for a one year increase in age. The conditional odds ratio of having a pregnancy termination (vs no pregnancy termination) for women who had pre-school education was 0.776 (95% CI 0.486-1.239, p=0.0049) as compared to those who had no education. Woman who belonged to the middle and richest wealth index categories were significantly associated with pregnancy outcome. The odds ratio of having had a pregnancy termination was 12% lower (OR 0.880, 95% CI 0.698-1.109, p=0.0098) for middle class women and 10% higher (OR 1.102, 95% CI 1.102-1.766, p< 0.0001) for the richest women, all compared to the poorest women. With respect to the variable *Region*, the odds ratio of having had a pregnancy termination was about 3 times higher for Cabo Delgado (OR 3.095, 95% CI 2.085-4.592, p< 0.0001) and Maputo Provincia (OR 3.188, 95% CI 1.481-3.532, p< 0.0001), 2 times higher for Maputo Cidade (OR 2.150, 95% CI 1.503-3.077, p=0.0360), 32% higher for Niassa (OR 1.317, 95% CI 0.854-2.030, p=0.0228) and 16% lower for Tete (OR 0.839, 95% CI 0.548-1.285, p< 0.0001), all compared to the Zambezia region.

In the case of the variable *Ethnicity*, the conditional odds ratio of having had a pregnancy termination was 0.757 for Cisena (95% CI 0.532-1.075, p=0.0269), 0.587 for Elomwe (95% CI 0.336-1.024, p=0.0180) and 2.251 for Echuwabo (95% CI 1.523-3.328, p< 0.0001), all compared to the Emakhuwa ethnic group. The estimated coefficient of age at first childbirth is very close to 0, and the odds ratio of pregnancy termination vs no pregnancy termination was almost 1, implying that this variable is unrelated to pregnancy outcome (p=0.4265). Moreover, Table A.10 (see Appendix) displays the Type 3 analysis of effects based on the Wald test and it can be seen that the variables marital status (*HH_MS*) and age at first childbirth (*AgeAt1stBirth*) are not statistically significant (p=0.1700 and p=0.4265, respectively). This is an indication that there no evidence that the pregnancy outcome is related to either marital status or age at first childbirth.

4 Discussion

In this study, survival analysis was used to investigate the determinants of timing of first childbirth among women of the reproductive age (15-49 years) in Mozambique, while multiple logistic regression was used to model pregnancy termination. In survival analysis, the Kaplan-Meier survival curves were used as the initial univariate step in the analysis of age at first childbirth to generate unbiased descriptive statistics and to estimate the probability of time to event (age at first childbirth). The Cox PH model was used in multivariate analysis to model the timing to first childbirth and to determine which factors had huge impact on the response variable. Prior to fitting the PH model, variable selection was done on the basis of the exploratory data analysis and the least absolute shrinkage and selection operator (LASSO) method was used for further variable selection. The selected variables were then used in the multivariate PH model and inference was done. In binary logistic regression, stepwise model selection was used to select variables in the model and the Hosmer and Lemeshow statistic was used to test for the overall model fit.

In studying the relationship between age at first childbirth and education level, it was observed that the risk of first childbirth for women with secondary and higher education were 18% and 19% lower, respectively, compared to women with no education. This implies that women with secondary and higher education tend to delay first childbirth compared to women with no education, which is an indication that the risk of bearing first child early is reduced as the level of education increases. This finding was in agreement with some results reported in literature, i.e. the outcomes of the United States of America, Vietnam and Nigeria studies which found women with secondary education to have a significantly higher age at first birth than those with little or no education (Maxwell, 1987; Weinberger et al., 1989; Luc et al., 1993; Fagbamigbe and Idemudia, 2016). Studies by Gaisie (1984) and Konogolo (1985) also confirmed this result by finding that post-primary schooling (secondary and higher education) had a strong effect in delaying first childbirth. These results therefore, provide observational evidence that the educational attainment of a woman is an important determinant of timing to first childbirth in Mozambique, even in the presence of other variables. The lower risk of early first childbirth among educated women may be attributed to the fact that educated women tend to postpone child bearing to further their education careers, finding matching partners or waiting to first get good jobs. Another logical explanation, although not found in this study, may be that most educated women usually have knowledge of family planning methods and use of contraceptives which can prolong time to first childbirth, as compared to uneducated women or those with low education. This finding therefore, suggests the implementation of policies and programs in Mozambique that support the education of young women so as to prolong the timing of first childbirth and reduce the high incidence of childhood marriages which contribute to the young age of first-time mothers which, according to UNICEF (2015), is a major cause of high maternal mortality and infant mortality rates, underage birth weights and postpartum bleeding among other problems.

Region of residence had a significant effect on age at first childbirth in Mozambique. This

means that the delay of first childbirth was associated with where a woman resides. In this regard, the risk of early first childbirth was 17% lower for the women in Inhambane region and 14% lower for women in Sofala region, all compared to women who resides in Cabo Delgado region. The risk of early first childbirth was not different for women in the other regions (Gaza, Manica, Maputo Cidade, Maputo Provincia, Nampula, Niassa, Tete, and Zambezia), compared to the women who resides in Cabo Delgado region. This finding was in line with some of the existing literature such as the study conducted by Luc et al. (1993) to find determinants of fertility in Vietnam. In this study, the result showed differences between women from the North and those from the South, whereby women from the North had significantly higher age at first childbirth than women from the South. Fagbamigbe and Idemudia (2016) also found similar results although they focused on place of residence based on the issue of rural-urban differentials. In this regard, women who reside in rural areas were found to have a high likelihood of early first childbirth compared to those based in the urban areas. These differences may be attributed to the fact that several regions differ in terms of resources, people who live there (ethnicity and religions), e.g. Inhambane region is one of the richest/wealthy regions in Mozambique so this may have contributed to the reduction in the risk of early first childbirth in this region. The same reason may also apply to Sofala region whose capital city of Beira, has the busiest port in Mozambique which is crucial for the trade in oil and other valuables.

As expected, age at first sexual intercourse was one of the most significant determinant of age at first childbirth. There was a positive association between these two whereby women who engaged in early first sexual intercourse had first childbirth earlier compared to those who delayed first sexual intercourse. This result was also reported by Zelnik (1981) who identified age at first sexual intercourse as one of the major proximate determinants of age at first childbirth. In Mozambique, age at first sexual intercourse is essentially similar to age at first marriage because there is an issue of early marriages in the country, whereby 48% of girls are married before the age of 18 and 14% are married before the age of 15, despite 18 years has been established as the legal age of consent to marriage (UNICEF, 2017). Early marriages imply early age at first sexual intercourse thus increasing the chances of early first childbirth. This explains the reason why the overall median age at first childbirth was as low as 19 years. On the other hand, an increase in the ages at first sexual intercourse can result in a lower likelihood of early first childbirth. This concurs with the research done by Bongaarts et al. (1984), who found that age at first marriage and age at first sexual intercourse had a significant impact on fertility and are regarded as the most significant determinants of age at first childbirth.

There were also differences in age at first childbirth along religious affiliations. Women who were affiliated to other religions, the Protestants and the non-religious were more likely to have earlier first childbirth than the Catholics. There was no difference between the Muslims and the Catholics in terms of timing of first birth. Catholic women prolong the timing of first childbirth maybe because the doctrine has more emphases on abstinence outside marriage as compared to other religious groups. However, contradictory findings have been documented by several researchers mentioned in the literature who noted that Catholics

had a lower age at first childbirth than Muslims and other religious groups because the Catholics have a negative perspective towards the use of contraceptives compared to other religious groups which have a more liberal stand towards contraceptive use (Bloom and Reddy, 1986; Gage, 1986; Mturi, 1997). The fact that these researches have been carried out in different countries may have resulted in these variations.

A lower rate of early first childbirth was found among women who belong to the Portuguese ethnic group than those who belong to other ethnic groups. This result maybe from the fact that different ethnics groups have different norms, beliefs and values as well as different practices that are likely to influence the reproductive performance of a given society and thus affect timing of first childbirth (Ohadike, 1979). However in this regard, it was found, through cross tabulations, that 91% of the Portuguese belonged to the richest wealth index category and over 97% belongs to at least the richer category. This implies that this result may have been highly confounded by the variable *Wealth index*. Moreover, it was also observed that over 70% of the Portuguese had at least secondary education, and this proportion was far much larger than the other ethical groups. Furthermore, about 64% of the Portuguese were residents of Maputo (Cidade and Provincia), which happens to be the capital city of Mozambique. A cross tabulation showed that about 96% of people who who live in Maputo Cidade and about 80% of people who live in Maputo Provincia belongs to the richest wealth index class. Combining all these findings, it can be concluded that the confounding effect of *Wealth index* and *Education* may be the reason why the Portuguese ethnic group had an effect on timing of first childbirth.

The fitted Cox PH model suggested a higher rate of early first childbirth among women who had knowledge of contraceptive use than those who did not. This was a strange result as women who know how to use contraceptives are expected to prolong age at first childbirth by delaying pregnancy, unless if they really want to get pregnant. In a study to model the determinants of age at first birth done by Sarkar (2010) in Bangladesh, it was observed that women who used contraceptives tend to prolong age at first childbirth compared to those who did not, which is a contradiction to the results of this current study. The reason for this contradiction can be attributed to the fact that in this current study, women were asked if they had knowledge of any method of contraceptive use and not if they were actually using them. However, knowing any method of contraceptive is not enough and does not imply using them, so the variable *"Knowledge of contraceptives"* which was used in this study may not be a good proxy for contraceptive use. Knowledge of the ovulation cycle was found to have no effect on the timing of first childbirth among women in Mozambique.

The results of the binary logistic regression showed that variables *AgeAt1stSex*, *Smoking*, *Working status*, *Contraceptives*, *Region*, *Ethnicity*, *Wealth index* and *Education* were related to having had a pregnancy termination. Smoking was found to have a negative effect with an increase of 61% in the odds ratio of having had a pregnancy termination in smokers compared to non-smokers. This result concurs with that of a pilot study done by Žilaitienė et al. (2007) who found that smoking was among the factors related to life style that affect fertility by increasing the risk of miscarriage (spontaneous abortion). Other studies

have also shown that smoking has a huge impact on pregnancy outcome and can result in women who smoke having spontaneous abortions, stillbirth, vaginal bleeding, disruption of the placental and giving birth to premature babies and to children with birth defects and low birth weight (Sandahl, 1989).

The risk of having had a pregnancy termination among women who were working was almost twice the risk of those who were not working. Similar findings were reported by Axmon et al. (2006), in a study to determine factors affecting time to pregnancy, who reported that working was allowed for pregnant women, but some work related factors such as working long hours, doing strenuous jobs that require lifting too heavy things, prolonged standing and walking continuously are not allowed as they may cause stress which disturbs the menstrual performance and increases the risk of miscarriage. Also working long hours can cause tiredness among pregnant women, which may lead to contraction that can damage the unborn baby resulting in spontaneous abortion during early pregnancy or premature births (Bonde et al., 2013).

It was also found that women who had knowledge of contraceptives were almost 3 times more likely to have a pregnancy termination as compared to those who had no knowledge of contraceptives. It is however unclear how the knowledge of contraceptive can result in an increased risk of pregnancy termination and this may require further research. Different regions were found to have different effects on pregnancy outcome. Women who reside in regions like Cabo Delgado, Maputo Provincia, Maputo Cidade and Niassa had a higher risk of pregnancy termination, while women who reside in the Tete region had a lower risk of pregnancy termination, all compared to those who reside in the Zambezi region. This variation can be attributed to differences in life style and geographical condition in various regions, so future studies should be conducted to further assess these factors. A similar result was observed for the variable *Ethnicity* were women who belong to Cisena and Elomwe ethnics group had a lower risk of pregnancy termination, while those who belong to the Ehuwabo ethnic group had a higher risk of pregnancy termination, all compared to those from the Emakhuwa ethnic group.

Age at first sex was also related to a pregnancy outcome, with women who engage in first sexual intercourse at an earlier age more likely to having had a pregnancy termination. According to UNICEF (2015), the high incidence of early age at first sexual intercourse that result from childhood marriages is a major cause of high maternal mortality and infant mortality rates, underage birth weights, postpartum bleeding and abortions among other problems. Also, women who engage in early sexual activities are more likely to have unwanted pregnancies when not yet ready for motherhood, so they may end up terminating the pregnancy through induced abortion. Looking at the variable *Education*, it was found that only women with pre-school education had a significant impact on pregnancy outcome. However, in general, there seems to be an increase in the risk of pregnancy termination with increasing educational attainment (although not statistically significant). This result may be attributed to the fact that educational interventions can help increase knowledge of family planning and and child spacing, which can result in the abortion of

unintended pregnancies. Wealth index had a similar trend with the chances of pregnancy termination increasing with increasing wealth status, i.e. from poorer to richest. Age at first childbirth was found to have no effect on pregnancy termination.

In data analysis and results interpretation, the following limitations were kept in mind. The fact that the data used for this study was a recode data format (secondary data) may have affected the accuracy of the results due to bias that may have been induced when generating it from the raw data format. Also, some values in the recode data set had been imputed, prior to this research, and this may have further affected the accuracy of the results. In addition, this data was not specifically meant for this type of analysis and as a result there was a limited choice of variables for the researcher to include in the analysis as some important variables were missing. This led to the researcher using some variables as proxy variables, e.g. using variable *knowledge of contraceptive* as a proxy for contraceptive use, which may not be a perfect substitute and thus affecting the study outcome. Also, inference of some variables of interest could not be performed as they were used as stratifying variables in order not to violate the PH model assumption. Further studies can be done using different methods like accelerated failure time models (AFT models), which do not require any stratification and allow more precise inferences. Furthermore, instead of considering different regions in the analysis, it would be wise to use rural-urban differentials in the timing of first childbirth in Mozambique, i.e. splitting all regions into two categories, rural or urban, and study how timing of first childbirth and pregnancy outcome differs in these two categories. This categorisation will help come up with meaningful conclusions which can easily be compared to other similar researches.

5 Conclusion and Recommendations

In determining the timing to first childbirth among the reproductive women in Mozambique, early first childbirth was found to be mostly influenced by education and age at first sexual intercourse. The other variables which affected early first childbirth were *Region* (Inhambane and Sofala), *Religion* (Protestants, Others and Non-religious), the Portuguese ethnic group and the knowledge of contraceptive use. This study also found that most women start child bearing at an early age (median years of 19) and so there is a need for programs that seek to increase womens age at first childbirth. Since education was found to be one of the most influencing factor of timing of first childbirth, the researcher recommends the government of Mozambique to implement policies and programs that seek to increase educational opportunities, especially for girls so as to reduce the cases of early age at first childbirth. Awareness campaigns should be done, particularly to young women with little or no education, to inform them on reproductive health, the advantages of postponing timing of first childbirth and to provide them with basic life skills to enable them to avoid early first childbirth. This will also help reduce maternity and infant mortality rates and other problems which were found to be related to early age at first childbirth by other studies mentioned in the literature. Factors such as *AgeAt1stSex*, *Smoking*, *Working status*, *Contraceptives*, *Region*, *Ethnicity*, *Wealth index* and *Education* were related to the outcome pregnancy termination. In light of these findings, pregnant women are therefore encouraged not to smoke or do strenuous jobs as these can increase the risk of pregnancy termination as a result of spontaneous abortions or miscarriages. Although not included in this study, several literature have shown that early marriages are major cause of early age at first sex, early-age pregnancies and early age at first childbirth in sub-Saharan Africa, so policies must be put in place to delay early marriages.

References

Adebowale, A. S., Yusuf, B. O., and Fagbamigbe, A. F. (2012). Survival probability and predictors for woman experience childhood death in nigeria:analysis of north–south differentials. *BMC Public Health*, 12(1):430.

Agaba, P., Atuhaire, L. K., and Rutaremwa, G. (2010). Determinants of age at first marriage among women in western uganda. In *European population conference*.

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

Axmon, A., Rylander, L., Albin, M., and Hagmar, L. (2006). Factors affecting time to pregnancy. *Human reproduction*, 21(5):1279–1284.

Bloom, D. E. and Reddy, P. H. (1986). Age patterns of women at marriage, cohabitation, and first birth in india. *Demography*, 23(4):509–523.

Bonde, J. P. E., Jørgensen, K. T., Bonzini, M., and Palmer, K. T. (2013). Risk of miscarriage and occupational activity: a systematic review and meta-analysis regarding shift work, working hours, lifting, standing and physical workload. *Scandinavian journal of work, environment & health*, 39(4):325.

Bongaarts, J., Frank, O., and Lesthaeghe, R. (1984). The proximate determinants of fertility in sub-saharan africa. *Population and Development Review*, pages 511–537.

Bumpass, L. L., Rindfuss, R. R., and Jamosik, R. B. (1978). Age and marital status at first birth and the pace of subsequent fertility. *Demography*, 15(1):75–86.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.

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

Dennehy, E. B., Edwards, C. A., and Keller, R. L. (1995). Aids education intervention utilizing a person with aids: Examination and clarification. *AIDS education and prevention*.

DHS (2011). Demographic and health surveys (dhs) program. https://www.dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm. [Online; accessed 8-May-2019].

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565.

Fagbamigbe, A. F. and Idemudia, E. S. (2016). Survival analysis and prognostic factors of timing of first childbirth among women in nigeria. *BMC pregnancy and childbirth*, 16(1):102.

Finnäs, F. and Hoem, J. M. (1980). Starting age and subsequent birth intervals in cohabitational unions in current danish cohorts, 1975. *Demography*, 17(3):275–295.

Fleming, T. and Harrington, D. (1991). Counting processes and survival analysis john wiley & sons. *Inc. New York*.

Ford, K. (1984). Timing and spacing of births.

Gage, A. (1986). Child spacing and fertility in greater freetown, sierra leone. *Unpublished M. Phil Thesis*.

Gaisie, S. K. (1984). The proximate determinants of fertility in ghana.

Goodman, L. A. and Kruskal, W. H. (1979). Measures of association for cross classifications. In *Measures of association for cross classifications*, pages 2–34. Springer.

Hosmer, D. W. and Lemeshow, S. (2000). Applied logistic regression. john wiley & sons. *New York*.

Hosmer Jr, D. W., Lemeshow, S., and May, S. (2008). *Applied survival analysis: regression modeling of time-to-event data*, volume 618. Wiley-Interscience.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jensen, R. and Thornton, R. (2003). Early female marriage in the developing world. *Gender & Development*, 11(2):9–19.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

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

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kohler, H.-P., Skytthe, A., Christensen, K., et al. (2001). The age at first birth and completed fertility reconsidered: Findings from a sample of identical twins. Technical report, MPIDR Working Paper, WP-2001-006.

Konogolo, L. (1985). Variation in entry into motherhood and length of effective reproductive life among women in kenya. In *Studies in African and Asian Demography. CDC Annual Seminar*.

Kumar, G., Danabalan, M., et al. (2006). Determinants of delayed first birth. *Indian J Community Med*, 31(4):272–3.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. (2005). *Applied linear statistical models*, volume 5. McGraw-Hill Irwin Boston.

Luc, N., Thang, N. M., Swenson, I., and San, P. B. (1993). Selected determinants of fertility in vietnam: age at marriage, marriage to first birth interval and age at first birth. *Journal of biosocial science*, 25:303–310.

Macro, I., Commission, N. P., et al. (2014). Nigeria demographic and health survey 2013.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748.

Maxwell, N. L. (1987). Influences on the timing of first childbearing. *Contemporary Economic Policy*, 5(2):113–122.

Mturi, A. J. (1997). Patterns of fertility and contraceptive use in tanzania. *DISSERTATION ABSTRACTS INTERNATIONAL*, 58(3-C):844.

Ochalla-Ayayo, A. et al. (1990). Sexual practices and the risk of the spread of stds and aids in kenya. task force on sex practices and the risk of stds and aids in kenya. *World Health Organization, and population studies and research institute, University of Nairobi, Kenya*.

Ohadike, P. O. (1979). Socio-economic cultural and behavioral factors in natural fertility variations.

Rajaretnam, T. (1990). How delaying marriage and spacing births contributes to population control: an explanation with illustrations. *Journal of Family Welfare*, 36(4):3–13.

RAO, K. V. and Balakrishnan, T. (1988). Age at first birth in canada: a hazards model analysis. *Genus*, pages 53–72.

Rindfuss, R. R. and St. John, C. (1983). Social determinants of age at first birth. *Journal of Marriage and the Family*, pages 553–565.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.

Roro, M. A., Hassen, E. M., Lemma, A. M., Gebreyesus, S. H., and Afework, M. F. (2014). Why do women not deliver in health facilities: a qualitative study of the community perspectives in south central ethiopia? *BMC research notes*, 7(1):556.

Sandahl, B. (1989). Smoking habits and spontaneous abortion. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 31(1):23–31.

Sarkar, P. (2010). Determinants of age at first birth in Bangladesh. *J Mod Math Stat*, 4:1–6.

Tableman, M. and Kim, J. S. (2003). *Survival analysis using S: analysis of time-to-event data*. Chapman and Hall/CRC.

Therneau, T. M. and Grambsch, P. M. (2013). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

Tibshirani, R., Wainwright, M., and Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Trussell, J. and Reinis, K. I. (1989). Age at first marriage and age at first birth. *Population bulletin of the United Nations*, (26):127–185.

Turner, R. M. (1992). Russian fertility is low, despite early age at first birth and lack of effective contraceptive methods. *Perspectives on Sexual and Reproductive Health*, 24(5):236.

Udry, J. R. (1979). Age at menarche, at first intercourse, and at first pregnancy. *Journal of Biosocial Science*, 11(4):433–441.

UNICEF (2015). Child marriage and adolescent pregnancy in mozambique: Causes and impact. https://www.girlsnotbrides.org/resource-centre/child-marriage-and-adolescent-pregnancy-in-mozambique-causes-and-impact/. [Online; accessed 20-May-2019].

UNICEF (2017). The economic impacts of child marriage. https://www.girlsnotbrides.org/child-marriage/mozambique/. [Online; accessed 25-May-2019].

Weinberger, M. B., Lloyd, C., and Blanc, A. K. (1989). Women's education and fertility: A decade of change in four latin american countries. *International Family Planning Perspectives*, pages 4–28.

WHO (2017). Mozambique's health system. https://www.who.int/countries/moz/areas/health_system/en/index1.html. [Online; accessed 20-May-2019].

Zabin, L. S., Smith, E. A., Hirsch, M. B., and Hardy, J. B. (1986). Ages of physical maturation and first intercourse in black teenage males and females. *Demography*, 23(4):595–605.

Zelnik, M. (1981). Determinants of fertility behaviour among us female aged 15-19, 1971 and 1976. final report. *Contract NO1-HD-82848. Baltimore: Johns Hopkins University.*

Žilaitienė, B., Diržauskas, M., Preikša, R., and Matulevičius, V. (2007). Cigarette smoking and waiting time to pregnancy: results of a pilot study. *Medicina*, 43(12):959.

APPENDIX

Table A.1: Parameter estimates of LASSO

	Number	Variable	Coefficient
	1	(Intercept)	0.0000
Education	2	**Pre-school**	**-0.2078**
	3	Primary	0.0000
	4	Secondary	0.0000
	5	**Higher**	**-0.2622**
M.S	6	Divorced	0.0000
	7	Widowed	0.0000
	8	Married	0.0000
W.I	9	**Poorer**	**-0.0331**
	10	Middle	0.0000
	11	**Richer**	**0.0357**
	12	Richest	0.0000
Region	13	Gaza	0.0000
	14	**Inhambane**	**-0.0895**
	15	**Manica**	**0.3337**
	16	Maputo Cidade	0.0000
	17	Maputo Provincia	0.0000
	18	Nampula	0.0000
	19	**Niassa**	**-0.0545**
	20	**Sofala**	**0.0936**
	21	**Tete**	**0.2335**
	22	**Zambezia**	**-0.0839**
Religion	23	**Muslims**	**-0.1211**
	24	**None**	**0.0335**
	25	Others	0.0000
	26	**Protestants**	**0.0527**
Ethnicity	27	**Portuguese**	**-0.3776**
	28	**Xichangana**	**0.0439**
	29	Cisena	0.0000
	30	**Elomwe**	**-0.0229**
	31	Echuwabo	0.0000
	32	Others	0.0000
	33	**KnowOfCycle1**	**0.0793**
	34	**Contraceptive1**	**0.0251**
	35	**AgeAt1stSex**	**-0.2028**
	36	**Working1**	**0.1615**

*Parameter estimates for the LASSO: The variable Marital status (HH_MS) was dropped from the model as all its categories were shrunk to zero.

Table A.2: Test for influential observations

Covariate	max|*delta-betas*|	s.e
Pre-school	0.0156	0.0894
Primary	0.0155	0.0880
Secondary	0.0158	0.0895
Higher	0.0155	0.1135
Gaza	0.0068	0.0647
Inhambane	0.0078	0.0592
Manica	0.0071	0.0606
Maputo Cidade	0.0076	0.0627
Maputo Provincia	0.0078	0.0636
Nampula	0.0044	0.0515
Niassa	0.0053	0.0515
Sofala	0.0069	0.0586
Tete	0.0069	0.0599
Zambezia	0.0078	0.0648
Muslims	0.0024	0.0339
Non-religious	0.0016	0.0274
Others	0.0014	0.0276
Protestants	0.0021	0.0278
Portuguese	0.0052	0.0558
Xichangana	0.0051	0.0553
Cisena	0.0053	0.0561
Elomwe	0.0077	0.0730
Echuwabo	0.0111	0.0693
Others	0.0043	0.0457
KnowOfCycle1	0.0028	0.0308
Contraceptive1	0.0072	0.0533
AgeAt1stSex	0.0101	0.0689
AgeAt1stSex:log(time)	0.0035	0.0228

*Influence of observations on a parameter estimate: A comparison of max|*delta-betas*| and their corresponding standard errors of the variables from the fitted model.

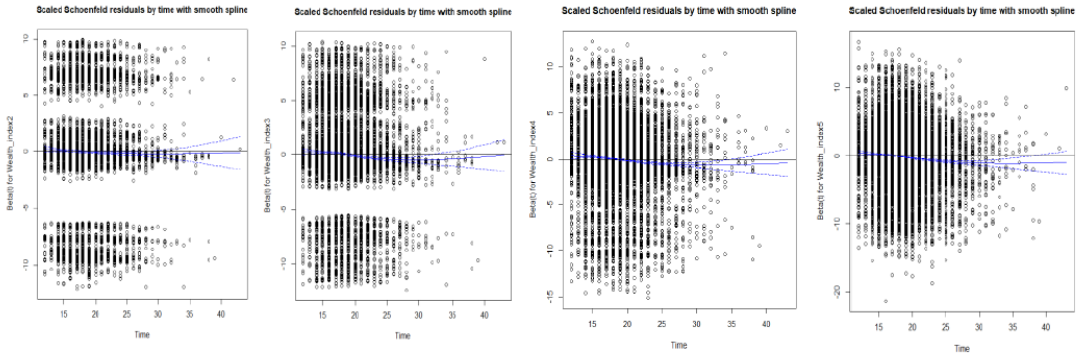


Figure A.1: Schöenfeld residuals plots for Wealth index categories

*From left to right: Wealth_index2 (Poorer), Wealth_index3 (Middle), Wealth_index4 (Richer) and

Wealth_index5 (Richest)

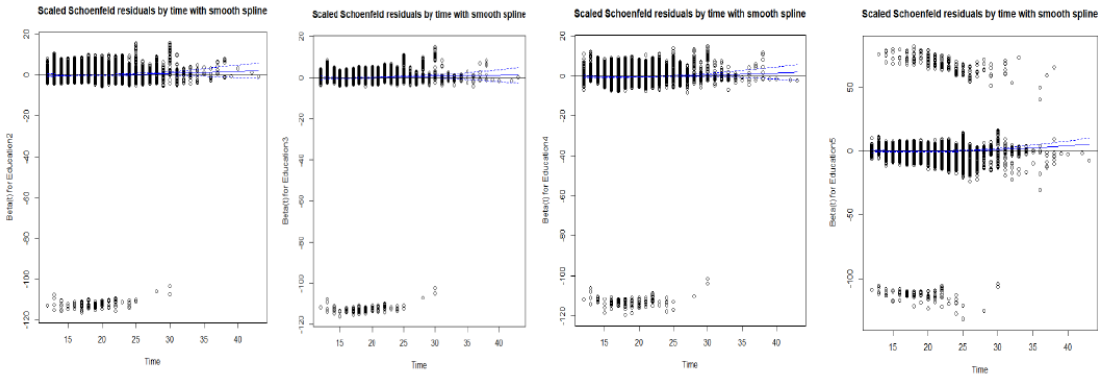


Figure A.2: Schöenfeld residuals plots for Education categories

*From left to right: Education2 (Pre-school), Education3 (Primary), Education4 (Secondary) and Education5 (Higher)

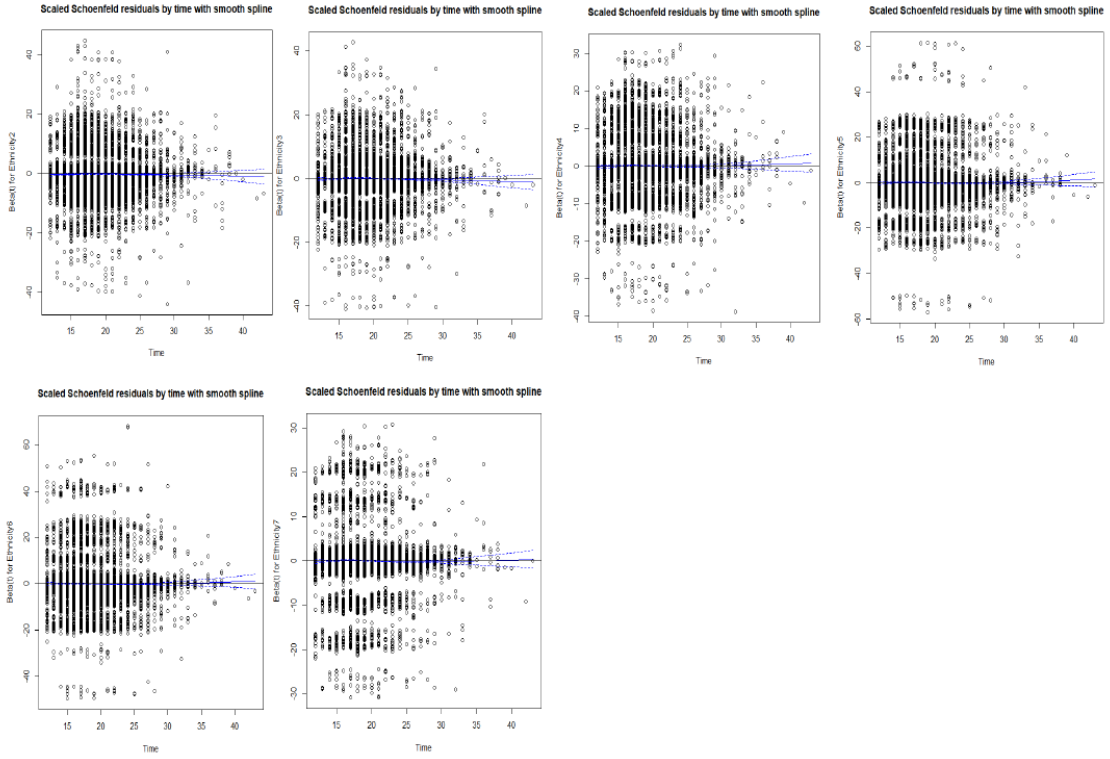


Figure A.3: Schöenfeld residuals plots for Ethnicity categories

*From left to right (1^{st} row): Ethnicity2 (Portuguese), Ethnicity3 (Xichangana) and Ethnicity4

(Cisena). (2^{nd} row): Ethnicity5 (Elomwe), Ethnicity6 (Echuwabo) and Ethnicity7 (Others)

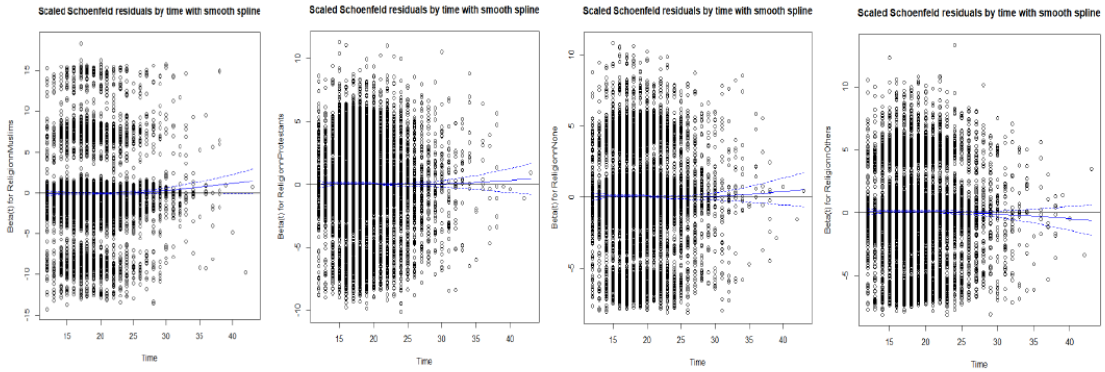


Figure A.4: Schöenfeld residuals plots for Religion categories

*From left to right: Muslims, Protestants, Non-Religious and Others

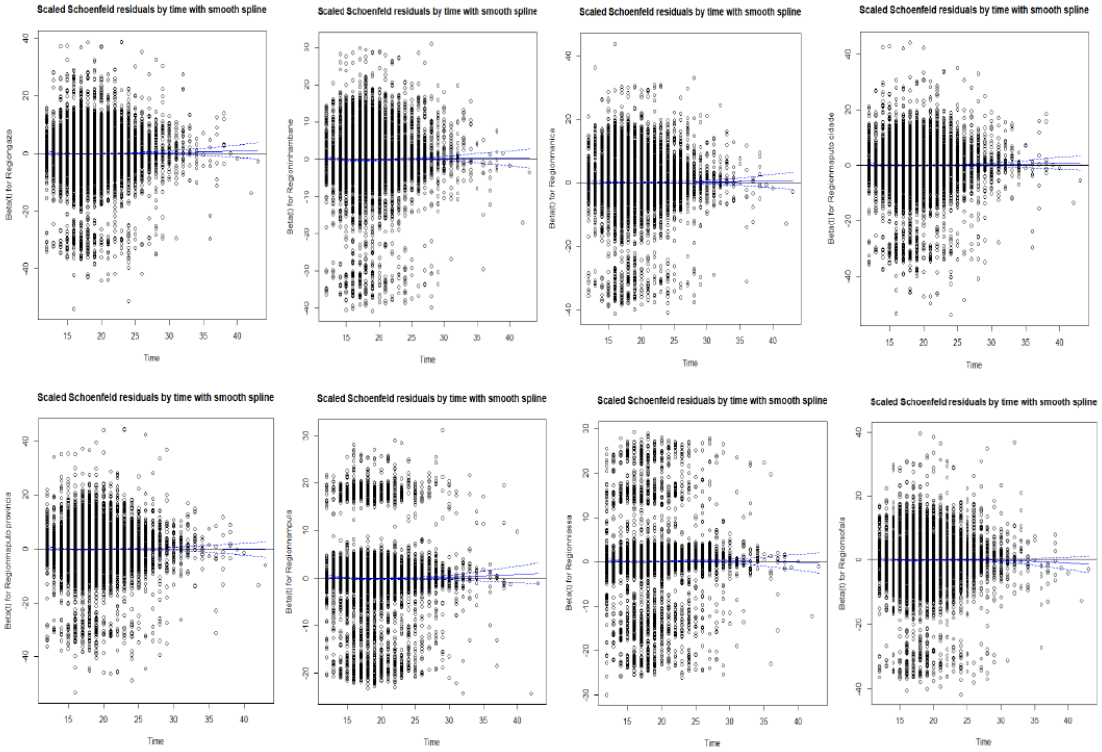


Figure A.5: Schöenfeld residuals plots for Region categories

*From left to right (1^{st} row): Gaza, Inhambane, Manica and Maputo Cidade. (2^{nd}) row: Maputo

Cidade, Nampula, Niassa and Sofala

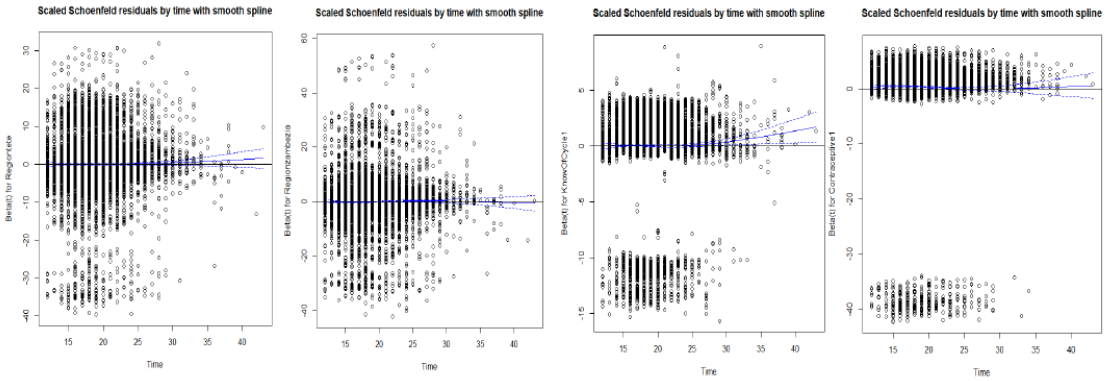


Figure A.6: Schöenfeld residuals plots

*From left to right: The first two are for remaining Region categories: Tete and Zambezia, the third plot is for KnowOfCycle1 and the last one is for Contraceptive1

Table A.3: Ethnicity and Wealth Index

		Wealth Index				
		Poorest	Poorer	Middle	Richer	Richest
Ethnicity	Emakhuwa	552(24.33)	645(28.43)	471(20.76)	334(14.72)	267(11.77)
	Portuguese	**3(0.11)**	**5(0.19)**	**36(1.35)**	**179(6.72)**	**2439(91.62)**
	Xichangana	35(0.78)	74(1.65)	295(6.56)	1210(26.91)	2882(64.10)
	Cisena	418(20.68)	482(23.85)	508(25.14)	439(21.72)	174(8.61)
	Elomwe	278(42.51)	175(26.76)	113(17.28)	66(10.09)	22(3.36)
	Echuwabo	151(24.55)	103(16.75)	92(14.96)	114(18.54)	155(25.20)
	Others	594(9.47)	989(15.77)	1403(22.37)	1784(28.44)	1502(23.95)

*The first number is the frequency and the one in brackets is the percentage frequency.

*A cross tabulation of *Ethnicity* and *Wealth index* . There seem to be some confounding effect of *Wealth index* on *Ethnicity*, especially the Portuguese as just above 91% of this group belong to the richest wealth index category and over 97% belongs to at least the richer category.

Table A.4: Ethnicity and Education

		Education				
		No	Pre-school	Primary	Secondary	Higher
Ethnicity	Emakhuwa	3(0.13)	800(35.26)	1183(52.14)	270(11.90)	13(0.57)
	Portuguese	**18(0.68)**	**43(1.62)**	**725(27.24)**	**1597(59.99)**	**279(10.48)**
	Xichangana	92(2.05)	536(11.92)	2410(53.60)	1407(31.29)	51(1.13)
	Cisena	1(0.05)	765(37.85)	1026(50.77)	223(11.03)	6(0.30)
	Elomwe	0(0.00)	191(29.20)	432(66.06)	31(4.74)	0(0.00)
	Echuwabo	0(0.00)	148(24.07)	341(55.45)	113(18.37)	13(2.11)
	Others	61(0.97)	1798(28.67)	3197(50.97)	1171(18.67)	45(0.72)

*The first number is the frequency and the one in brackets is the percentage frequency.

*A cross tabulation of *Ethnicity* and *Education*. There seem to be some confounding effect of *Education* on *Ethnicity*, especially the Portuguese as over 70% of this group had at least secondary education, which is a very large proportion compared to other ethical groups.

Table A.5: Ethnicity and Region

		Ethnicity						
		Emakhuwa	**Portuguese**	Xichangana	Cisena	Elomwe	Echuwabo	Others
Region	Cabo Delgado	792	19(1.64)	2	0	0	0	343
	Gaza	0	155(7.99)	1646	4	0	2	133
	Inhambane	0	159(9.05)	39	5	4	5	1544
	Manica	5	80(5.80)	49	143	11	28	1064
	Maputo Cidade	31	**1153(38.68)**	1417	7	2	37	334
	Maputo Provincia	12	**577(25.78)**	1297	11	0	19	322
	Nampula	949	47(4.66)	2	0	0	5	6
	Niassa	431	42(4.20)	3	4	11	15	493
	Sofala	9	215(7.98)	31	1470	27	44	898
	Tete	0	136(10.57)	7	114	0	41	989
	Zambezia	40	79(5.10)	3	263	599	419	146

*The first number is the frequency and the one in brackets is the percentage frequency.

*A cross tabulation of *Region* and *Ethnicity*. There is a possibility of a confounding effect of *Region* on *Ethnicity*, especially the Portuguese as about 64% of this group are residentsof Maputo (Cidade and Provincia), which happens to be the capital city of Mozambique. A cross tabulation Table A.6 below shows that about 96% of people who live in Maputo Cidade and about 80% of people who live in Maputo Provincia belong to the richest wealth index class.

Table A.6: Region and Wealth index

		Wealth index				
		Poorest	Poorer	Middle	Richer	Richest
Region	Cabo Delgado	247	411	196	163	139
	Gaza	24	59	257	891	709
	Inhambane	55(3.13)	131(7.46)	291(16.57)	743(42.31)	536(30.52)
	Manica	84	162	379	466	289
	Maputo Cidade	0	0	1	100	**2880(96.61)**
	Maputo Provincia	8	24	61	388	**1757(78.51)**
	Nampula	282	230	193	167	137
	Niassa	79	281	356	158	125
	sofala	339(12.58)	512(19.01)	670(24.87)	674(25.02)	499(18.52)
	Tete	299	256	263	242	227
	Zambezia	614	407	251	134	143

*The first number is the frequency and the one in brackets is the percentage frequency.

*A cross tabulation of *Region* and *Wealth Index*. There is a possibility of a confounding effect between the two variables just over 96% of people who live in Maputo Cidade and about 80% of people who live in Maputo Provincia belong to the richest wealth index class.

Table A.7: Cramer's V Association Statistics

	Education	Wealth index	KnowOfCycle	Marital status	Contraceptives	Religion	Ethnicity	Working	Region
Education	1	0.293	0.044	0.233	0.113	0.097	0.245	0.066	0.208
Wealth index		1	0.107	0.160	0.152	0.099	0.347	0.030	0.398
KnowOfCycle			1	0.032	0.005	0.041	0.199	0.066	0.294
Marital status				1	0.023	0.058	0.149	0.070	0.143
Contraceptives					1	0.037	0.123	0.051	0.185
Religion						1	0.319	0.066	0.362
Ethnicity							1	0.079	**0.603**
Working								1	0.348
Region									1

*Cramer's V Association Statistics to measure the strength of association between categorical variables. No value exceed 0.7 indicating that none of the variables were highly correlated.

Table A.8: Variance Inflation Factor (VIF)

	PH regression model			Logistic regression model		
	GVIF	Df	$GVIF^{(1/(2*Df))}$	GVIF	Df	$GVIF^{(1/(2*Df))}$
Education	1.739225	4	1.071629	1.903503	4	1.083788
Wealth index	3.342491	4	1.162810	3.025282	4	1.148407
Region	89.714791	10	1.252112	38.196970	10	1.199780
Religion	2.429655	4	1.117360	-	-	-
Ethnicity	39.494758	6	1.358454	27.085627	6	1.316421
KnowOfCycle	1.130168	1	1.063094	-	-	-
Contraceptive	1.052670	1	1.025997	1.019053	1	1.009481
Working	1.202658	1	1.096658	1.159537	1	1.076818
AgeAt1stSex	1.151845	1	1.073240	1.248245	1	1.117249
Marital status	-	-	-	1.263637	3	1.039770
AgeAt1stBirth	-	-	-	1.185524	1	1.088818

*Variance inflation factor (VIF) for all explanatory variables that were included in the two models. Clearly there are no indications of serious multicollinearity problems as all the standardised GVIF are far less than 10, for both models.

Table A.9: Partition for the Hosmer and Lemeshow Test

		PregTerminated1 (Yes)		PregTerminated0 (No)	
Group	Total	Observed	Expected	Observed	Expected
1	1899	54	50.13	1845	1848.87
2	1899	90	84	1809	1815
3	1899	130	114.46	1769	1784.54
4	1900	140	141.87	1760	1758.13
5	1899	134	162.91	1765	1736.09
6	1900	173	186.46	1727	1713.54
7	1899	196	215.27	1703	1683.73
8	1899	263	256	1636	1643
9	1899	320	306.14	1579	1592.86
10	1896	432	414.76	1464	1481.24

*Partition for the Hosmer and Lemeshow Test using the standard recommendation of 10 groups (g= 10) of approximately equal size. The table shows the observed and expected events in each decile of the predicted probabilities.

Table A.10: Type 3 Analysis of Effects

Effect	DF	Wald χ^2	p-value
Education	4	10.7362	0.0297
Wealth index	4	24.4776	$< .0001$
Region	10	123.2694	$< .0001$
Ethnicity	6	58.6076	$< .0001$
Contraceptive	1	14.8018	$< .0001$
Working status	1	149.1642	$< .0001$
Marital status	3	15.0817	0.1700
Smoking	1	5.2191	0.0223
AgeAt1stSex	1	6.1456	0.0132
AgeAt1stBirth	1	0.6322	0.4265

*Type 3 Analysis of Effects test for the overall effect of each variable on pregnancy outcome. It can be seen that variables *AgeAt1stSex*, *Smoking*, *Working status*, *Contraceptives*, *Region*, *Wealth index* and *Education* have a significant effect on pregnancy termination.

R and SAS Codes

```
                    ######## Important codes only #######
                    #########SURVIVAL ANALYSIS#########
###Setting working directory and reading data
setwd("C:\\Users\\maxyp\\Documents\\BIOSTATISTICS\\year 2\\semester 2\\Thesis\\My Thesis")
data<-read.table("datafile.csv",sep = ",", na.strings = "NA", quote = "\"", header = TRUE)

###Converting integers into factors (All categorical variables)
data$Education <- as.factor(data$Education)

###Changing age zeros to NA then delete them
data$AgeAt1stSex[data$AgeAt1stSex == 0] <- NA
x <- na.omit(data)
summary(x$AgeAt1stSex)
x$AgeAt1stSex[x$AgeAt1stSex > 49] <- NA
x1 <- na.omit(x)
summary(x1$AgeAt1stSex)

###Removing BMI values >60, (there exist some abnormal BMI values of 99)
x1$BMI[x1$BMI > 60] <- NA
y <- na.omit(x1)
summary(y$BMI)

### Removing abnormal weight values
y$Weight[y$Weight >= 110] <- NA
z <- na.omit(y)
summary(z$Weight)

### Removing abnormal height values
z$Height[z$Height >= 200] <- NA
z1 <- na.omit(z)
summary(z1$Height)

####categorising age at first sex
z1$age_grp <- z1$AgeAt1stSex
z1$age_grp <- ifelse((z1$AgeAt1stSex<14) , '1',z1$age_grp)
z1$age_grp <- ifelse((z1$AgeAt1stSex>=14 & z1$AgeAt1stSex<19) , '2',z1$age_grp)
z1$age_grp <- ifelse((z1$AgeAt1stSex>=19 & z1$AgeAt1stSex<25) , '3',z1$age_grp)
z1$age_grp <- ifelse((z1$AgeAt1stSex>=25) , '4',z1$age_grp)
z1$age_grp<-as.factor(z1$age_grp)
summary(z1$age_grp)

###Descriptive statistics (all variables)
Modelfit<-survfit(Surv(survtime, survevent) ~1, data=z1,type="kaplan-meier")
Modelfit1<-survfit(Surv(survtime, survevent) ~Education, data=z1,type="kaplan-meier")

###Comparing survival curves (Log-Rank test)
survdiff(Surv(survtime, survevent) ~Education,data=z1,rho =0)
```

```
###Plotting survival functions (All variables)
ggsurvplot(Modelfit1, data =z1, xlab="Time to First Birth (years)",
    pval = TRUE, pval.method = FALSE,  ggtheme =theme_light(), legend.title="Education",
    legend.labs =c("No", "Preschool", "Primary", "Secondary", "Higher"),
    pval.coord=c(40,1), surv.median.line="hv")

###LASSO variable selection.
x <- model.matrix( ~ Education +HH_M_S + Wealth_index + Region + Religionn + Ethnicity
                    + KnowOfCycle + Contraceptive + AgeAt1stSex + Working, z1)
y <- Surv(z1$survtime, z1$survevent)
fit <- glmnet(x, y, family="cox")
plot(fit, label=T)
set.seed(1)
cv.fit <- cv.glmnet(x, y, family="cox", alpha=1)
plot(cv.fit)
coef(cv.fit, s = "lambda.1se")

###Checking for association (all combinations)
assoc<-table(z1$Region,z1$Wealth_index)
chisq.test(assoc)
###Cross tabulations (all combinations)
source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")
crosstab(z1, row.vars = "Religion", col.vars = "Education", type = "f") #frequency
crosstab(z1, row.vars = "Region", col.vars = "Wealth_index", type = "r") #percentage

###Cramer's V association statistic (all combinations)
assocstats(table(Education, HH_M_S))#Cramer's V: 0.233

###checking independency between numerical and categorical variables (ANOVA)
aov1<-aov(z1$AgeAt1stSex~ z1$Education)
anova(aov1)

###Obtaining Martingale residuals from the null model
col.PH.1<-coxph(Surv(survtime, survevent) ~ 1, data=z1)
martingale.res<-resid(col.PH.1)
plot(z1$AgeAt1stSex, martingale.res, xlab="Age at first sex", col="blue")
lines(lowess(z1$AgeAt1stSex, martingale.res, iter=0, f=1))

###fit PH model (Initial model)
cox.PH.all<-coxph(Surv(survtime, survevent) ~  Education + Wealth_index + Region
                  + Religionn + Ethnicity+ KnowOfCycle
                  + Contraceptive+Working + AgeAt1stSex, data=z1,  ties = "efron")

###fit PH model (Final model)
col.PH.all<-coxph(Surv(survtime, survevent) ~Education + strata(Wealth_index) + Region
                  + Religionn + Ethnicity+ KnowOfCycle + Contraceptive +
                    strata(Working) + AgeAt1stSex*log(survtime), data=z1,  ties = "efron",
                  control = coxph.control(iter.max = 100))
summary(col.PH.all)
col.PHfit2.all3<-cox.zph(col.PH.all, transform = "identity")
```

```
###Examining the PH Assumption graphically (Plots of the Schoenfeld residuals against time)
for (i in 1:(nrow(col.PHfit2.all3$table)-1)){
  plot(col.PHfit2.all3[i], main="Scaled Schoenfeld residuals by time with smooth spline",
       col="blue")
  graphics::abline(a=0, b=0, col="black")
}

###Assessing GOF using Deviance Residuals
col.devres.all<-residuals(col.PH.all, type="deviance")
col.fitval.all<-predict(col.PH.all, type="lp")
plot(col.fitval.all, col.devres.all, col="blue", xlab ="Risk score", ylab="Deviance residuals")
abline(h = c(-1.96, 1.96), lty=4, col=3)
abline(h=c(-2.58, 2.58), lty=3, col=2)
abline(h=0, lty=2, col=1)

###Identifying potential outliers and influential observations
col.devres.all[col.devres.all< -2.5]#25 outliers
length(col.devres.all[col.devres.all> 2.5])#0

###Getting the dfbetas (delta-betas)
res<-resid(col.PH.all, type="dfbeta")
max(abs(res[,1]))
max(abs(res[,2]))

###Testing multicollinearity
cvif <- vif(col.PH.all)




                  #########LOGISTIC REGRESSION#########
### Defining full and null models and do step procedure (stepwise selection)
model.null = glm(PregTerminated ~ 1,
                 data=z1, family = binomial(link="logit"))
model.full = glm(PregTerminated ~ Education + HH_M_S + Wealth_index + Region
                 + Ethnicity+ Religionn + Smoking + KnowOfCycle + survtime
                 + Contraceptive + AgeAt1stSex + Working,
                 data=z1, family = binomial(link="logit"))

###Stepwise variable selection
step(model.null, scope = list(upper=model.full),
     direction="both", test="Chisq", data=z1)

finalmodel<-glm(PregTerminated ~ Region + Working + Ethnicity + survtime +
              Wealth_index + Contraceptive + HH_M_S + Smoking + AgeAt1stSex +
              Education, family = binomial(link = "logit"), data = z1)
summary(finalmodel)

###Testing for Multicollinearity
```

```
vif(finalmodel)

###exporting data from R
write.csv(z1, "C:\\Users\\maxyp\\Documents\\BIOSTATISTICS\\year 2\\semester 2\\
Thesis\\My Thesis\\thesis.csv")

proc import datafile="C:\Users\maxyp\OneDrive\Documents\BIOSTATISTICS\year 2\semester 2\
Thesis\My Thesis\thesis.csv" out=birth dbms=csv replace;
getnames=yes;
run;

proc print data=birth;
run;

/* Logistic regression   */
proc logistic data=birth;
class Education(ref="1")  Wealth_index(ref="1")  Region Ethnicity(ref="1") Contraceptive(ref="0")
Working(ref="0") HH_M_S(ref="1") Smoking(ref="0");
model PregTerminated(desc) = Education  Wealth_index  Region Ethnicity Contraceptive Working
HH_M_S Smoking AgeAt1stSex survtime;
output out=a pred=yhat;
proc ttest data=a; var yhat; run;


proc logistic data=birth;
class Education(ref="1")  Wealth_index(ref="1")  Region Ethnicity(ref="1") Contraceptive(ref="0")
Working(ref="0") HH_M_S(ref="1") Smoking(ref="0");
model PregTerminated(desc) = Education  Wealth_index  Region Ethnicity Contraceptive Working
HH_M_S Smoking AgeAt1stSex survtime
/selection=stepwise expb; run;

/* GOF ...hosmer-lemeshow */
proc logistic data=birth;
class Education(ref="1")  Wealth_index(ref="1")  Region Ethnicity(ref="1") Contraceptive(ref="0")
Working(ref="0") HH_M_S(ref="1") Smoking(ref="0");
model PregTerminated(desc) = Education  Wealth_index  Region Ethnicity Contraceptive Working
HH_M_S Smoking AgeAt1stSex survtime / lackfit;
run;
```