Sexual debut: An analysis of the Birth to Twenty data

Author:
Rashmika Singh

Supervisor:
Professor Glenda Matthews
Co-Supervisor:
Mr Jordache Ramjith

Submitted in fulfilment of the academic requirements for the degree of Master of Science in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus.

25 August 2015

As the candidate's supervisor I have approved this dissertation for submission.

Signed: _____ Name: _____ Date: _____

Abstract

According to the literature, it is widely accepted that the early timing of first sex among adolescents is related to long-term health effects and current and future risky sexual behaviour (Sandfort et al., 2008). Despite the importance of youth sexual behaviour for sexual and reproductive health, and the severity of the Human Immunodeficiency Virus (HIV) and the Acquired Immune Deficiency Syndrome (AIDS), there exists relatively little empirical research on sexual debut in Southern Africa (Muula, 2008). The aim of this dissertation is to utilize survival analysis techniques to determine significant predictors of early sexual debut in a South African context.

A collaboration with the Human Sciences Research Council (HSRC) was fostered and access to the Birth to Twenty (Bt20) data was arranged. The data set consists of 3273 respondents who were followed from birth. Sexual exposure measures were recorded in six collection waves, namely 11-12, 13, 14, 15, 16 and 17-18 years.

Multivariate analyses were initially run by employing a standard survival analysis technique, namely Cox proportional hazards regression survival analysis for sexual debut. Analyses were run separately for males and females. A log-rank test showed that there was a significant difference between the survivor curves for voluntary sexual debut and involuntary sexual debut. This result prompted consideration to explore a competing risks regression model with voluntary sexual debut as the event of interest and involuntary sexual debut as the competing risk event.

SPSS was used to run exploratory analyses and Cox Regression (IBM Corp, 2012). Regression diagnostic plots were run in SAS (SAS Institute Inc, 2004). Competing risks regression was performed according to the method of Fine & Gray (1999) by evoking the STCRREG command in STATA and the validity of the proportional subhazards assumption was tested by including time interaction variables in the model (StataCorp, 2013). Where violations of the proportional subhazards assumption were found, the vary-

ing effect of the hazard functions on the time to sexual debut was interpreted accordingly.

Keywords

baseline functions, censoring, competing risks, Cox proportional hazards model, cumulative incidence, distribution function, maximum likelihood, protective factor, regression diagnostics, regression modeling, risk factor, risk set, semiparametric, sexual coercion, sexual debut, subdistribution hazard, survival analysis, survivor function.

Contents

List of Figures

List of Tables

Declaration - Plagiarism

I, _____ declare that

1. The research reported in this dissertation, except where otherwise indicated, is my original research.
2. This dissertation has not been submitted for any degree or examination at any other university.
3. This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    a. Their words have been rewritten but the general information attributed to them has been referenced.
    b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: _____

Acknowledgements

I would like to express my most sincere gratitude towards my supervisors, Mr Jordache Ramjith (co-supervisor) and Professor Glenda Matthews. Firstly, to Mr Ramjith, your faith in me has always been so reassuring. You were ever so ready to extend a helping hand with your abundant insight, however at the same time you allowed me sufficient room to express my own ideas. You have shared your knowledge with me and I have learnt so much from you. Every meeting with you has left me inspired and motivated. To Professor Matthews, I am thankful for your direction and support. You possess a wealth of knowledge and experience in the field of Statistics and I am truly grateful to have had you in my team for this dissertation.

To my sponsor, South African Centre for Epidemiological Modelling and Analysis, if it were not for your financial and academic support I would not have been able to successfully complete this dissertation. I also thank you for the opportunity to present my work at the annual research day at SACEMA during both years of study during this dissertation. The critique offered by distinguished members including Dr Jo Barnes who is an epidemiologist in the Division of Community Health at the Faculty of Medicine and Health Sciences at Stellenbosch University has provided me with invaluable insight and has allowed me to deliver a dissertation of value.

To my parents, Romilla and Roy, and to my friends Pravesh and Sugen, I thank you for always taking an interest in the progress of my work. You all have supported and motivated me to do more than just what is sufficient. You have inspired me to keep going and to produce a piece of work that I am proud of.

I started this dissertation in 2013. At the time I was a full time Masters student. In 2014 I was fortunate enough to secure a place in the prestigious Standard Bank Quantitative Modeling Graduate Programme. Whilst this was precisely the career starter I had desired, it also placed severe strain on completion of my Masters as I discovered how difficult it was to give off my best at work and dedicate my efforts to my Masters simultaneously. I am indebted to my line managers Wendy Nicolae, Akesh Munnilal and Elvera Van Wyk who have approved my study leave so that I may give my undivided attention to the completion of this dissertation.

I would also like to take this opportunity to thank my previous sponsors; Investec and the Institute of Applied Statistics. Your support has provided the building blocks necessary for my success to this stage. Words alone cannot quantify the gratitude that I feel towards you. Finally, I would like to thank Studietrust; the bursary administrators of the Investec bursary. Specifically to Dr Hofmeyr and Zama Mojalefa, I thank you for your support and kindness.

Chapter 1

Introduction

For quite some time, early initiation of sexual debut has been a fundamental area of interest to psychologists, sex researchers and well-being health specialists. According to the literature, it is widely accepted that the early timing of first sex among adolescents is related to long-term health effects and current and future risky sexual behaviour (Sandfort et al., 2008). The risky behaviour may include multiple sex partners, sexual relations with casual partners and a disregard for formal contraceptive measures (Harrison et al., 2005). Driven by concerns of an increase in unwanted teenage pregnancies and sexually transmitted diseases including HIV, much research is concentrated around understanding the factors that are associated with initial adolescent sexual experiences in an attempt to devise programmes and strategies to influence adolescents to delay first sex (Berry & Hall, 2009).

Several factors have reportedly been listed as precursors to early sexual debut among adolescents. Biological factors include age, gender, pubertal timing and testosterone levels (Lammers et al., 2000). Gender is an important factor to consider as many studies that have examined issues involving sexual debut have shown that it is likely that the relationship between early sexual debut and the associated predictors of early sexual debut differ for males and females (Zaba et al., 2004). Nnko et al. (2004) suggest that females tend to under-report sexual debut while males tend to do the opposite. In particular, in South Africa, data recorded in nationally representative surveys indicate that the median age of reported sexual debut is approximately 16 years for male respondents

and 17 years for female respondents (Richter et al., 2005; Pettifor et al., 2005). However, another study conducted in a rural area in Kwa-Zulu Natal indicated that the median age of sexual debut was 18.5 for females and 19.5 for males (McGrath et al., 2009). Age at first sex has been found to vary from study to study and is dependant upon a host of factors including area of residency (Zaba et al., 2004; Lammers et al., 2000). In a paper exploring growth and pubertal timing, Rogol et al. (2000) agree with the general result that on average, girls enter and complete each stage of puberty earlier than do boys.

Social factors that are likely to affect age at sexual debut include religiosity, socioeconomic status, academic performance, parental supervision and parents' level of education. Lammers et al. (2000) found that a greater religious affiliation, higher socioeconomic status, better academic performance and greater parental supervision were associated with delaying sexual debut. The strength of the associations differed across gender and age. Social factors were more strongly associated with delaying sexual intercourse among younger age groups compared to the older age groups. A higher socioeconomic status was found to be associated with delaying sexual debut. Lammers et al. (2000) explain that it is possible that households with a higher socioeconomic status may have more resources to contribute to supervision. Hogan & Kitagawa (1985) found that it was more difficult for families with a low socioeconomic status to provide supervision to adolescents. According to Lammers et al. (2000), there is a strong association between sexual debut and performance at school, however the mechanism by which sexual debut is affected by school performance is unclear. In a study aimed at identifying risk and protective factors (including school connectedness) on adolescent health (including sexuality), Resnick et al. (1997) showed that school connectedness may be the mediating variable which links better school performance to delaying sexual debut. Perhaps performing well at school gives adolescents a higher self-esteem or equips them with better long-range planning skills, thus enabling them to make safer sexual decisions (Resnick et al., 1997).

According to a study of American youth conducted by Mueller et al. (2008), ex-

posure to formal sex education is one of the most influential tools that can be utilized in affecting positive and safe adolescent sexual behaviors. Formal sex education was defined as any education that assists in making safe, healthy and informed decisions about sex. This could be via parents, schools, communities or peers. The overall results of the study suggested that exposure to formal sex education was associated with abstinence from sexual intercourse, delayed initiation of sexual intercourse and an increased usage of contraception at first sex. In contrast, several population-based studies have shown that sex education had almost no effect on reducing the likelihood of adolescents engaging in sexual intercourse however it did appear to have some impact in the contraceptive decisions of youth (Marsiglio & Mott, 1986; Dawson, 1986). Mueller et al. (2008) reason that the positive associations between receiving formal sex education and postponing sexual initiation is attributable to the fact that the study on American youth allows to control for the sequence of events, that is, it is known whether adolescents received sex education before or after first sex. Many other similar studies do not have this kind of information. Furthermore, sex education is now being offered to more adolescents at earlier ages, this could also perhaps be a reason for the positive findings between receiving sex education and more responsible sexual behaviors. Evidence from intervention efficacy research shows that certain sex education curricula can effectively decrease risky sexual behavior in adolescents (Manlove et al., 2004). Although the appropriate content is debated, most sex and health experts do support some form of sex education for adolescents (Mueller et al., 2008). While some researchers advocate abstinence-only sex education, others strongly support a more holistic approach in the form of a comprehensive sex education. Both approaches have arguably been able to influence sexual decisions of adolescents. Factors such as societal beliefs regarding sex and social cultures affect which type of approach is used. Future research conducted on the association between formal sex education and youth's engagement in sexual activities should consider the prevalence of evidence-based sex education programs, the extent to which these are implemented and their overall efficacy (Mueller et al., 2008).

Recent evidence suggests that adolescents are becoming an important group

in shaping the HIV epidemic (Joint United Nations Programme on HIV/AIDS, 2013). Upon the emergence of HIV, the world has seen an unprecedented HIV prevalence. According to the Joint United Nations Programme on HIV/AIDS (2013), 35.3 million [32.2 million $-$ 38.8 million] people worldwide were living with HIV in 2012. Sub-Saharan Africa continues to shoulder most of this burden, accounting for a staggering 70% of all new HIV infections in 2012. In particular, in 2009, an estimated 2 million adolescents (aged $10 - 19$) were living with HIV (United Nations Publication, 2011). Young people need to be specifically targeted for HIV prevention and intervention as early sexual debut is associated with greater sexual risk behaviors in comparison to older individuals (Berry & Hall, 2009).

The remainder of this chapter is focused on introducing the Birth to Twenty study. Details leading to the study inception are reported and is followed by a description of the data which includes all important detail regarding the data and data collection.

Chapter 2 presents the variables to be considered in the study and also gives insight as to how they were constructed. Frequency tables and crosstabulations are used to describe the general nature of the data.

The Birth to Twenty sexual debut survival analysis methodology is discussed in Chapter 3. Firstly, key concepts and definitions which are central to survival analysis are defined. Regression modeling and proportional hazards regression models are introduced. Two models are discussed. Firstly, a popular and standard method of analysing survival data is explored, namely the Cox proportional hazards regression model which was first proposed by Cox (1972). A theoretical background on the model is discussed in detail and also includes regression diagnostic procedures. Secondly, the idea of survival analysis in the presence of competing risks is then addressed and the proportional hazards model for the subdistribution of a competing risk put forward by Fine & Gray (1999) is considered. Estimation for this model and regression diagnostics are not explicitly discussed as they are analogous to that of the Cox model.

Chapter 4 first presents the results and discussion of the Cox proportional hazards model. Risk factors for time to sexual debut were investigated and regression diagnostics were performed to determine the adequacy of the model fit and the validity of the proportional hazards assumption. Next, the competing risks regression results and discussion are presented according to the method of Fine & Gray (1999). The inclusion of time interaction variables into the model served as both a test of the subhazards proportionality assumption and a remedy in the case of a violation in the assumption. Thereafter, a conclusion of the results is given.

A conclusion is presented in Chapter 5. This chapter also addresses limitations of the study, discusses key results and implications and provides suggestions for possible future research.

1.1 The Birth to Twenty study

Richter et al. (2007) is the main reference for this section.

The latter years of the 1980's were a time of significant sociopolitical turmoil in South Africa. South African law was characterized by the Apartheid regime which curtailed the rights of Black inhabitants and maintained White supremacy but this state was crumbling. Black Africans began to dismiss laws of segregation that dictated where they lived and worked. Very rapid urbanization arose in areas that were previously known to be classified as White areas. It was expected that this rapid unplanned urbanization would result in significant effects on the health and development of children. Movement to urban areas meant improved access to education, better work opportunities and higher quality health care which could reduce preventable childhood morbidity and mortality. However, the government's inability to cater for the needs of this excess growth in the population in urban areas could in fact worsen the state of existing infectious diseases, such as HIV and tuberculosis. It could also have

led to a rise in non-infectious conditions which are related to the lifestyle, urban stressors and socio-cultural changes such as substance abuse and obesity.

As a direct result of these concerns, in 1988, Noel Cameron from the University of the Witwatersrand and Derek Yach from the South African Medical Research Council (MRC) approached Andries Brink, who was the MRC President at the time, requesting funds to start a birth cohort study in the Soweto-Johannesburg (Gauteng) area. The study aimed to follow a group of urban children across the first decade of their life and was thus named *Birth to Ten* (Bt10). Once the study duration had ended, the study committee decided to then extend the follow up period by ten years, and the study was then renamed *Birth to Twenty* (Bt20). However, today, more than twenty years since the start of the study, the study is still active and is presently in its twenty-fifth year of follow up. The study was colloquially termed *Mandela's Children* because the subjects were born within seven weeks following the release of Nelson Mandela from prison and were the first South African cohort born into a democratic South Africa.

The first round of data collection was in 1989/1990 where the pregnant women were surveyed about demographic information and their pregnancy conditions. Additionally, as of October 2005, the second generation of children had started to be born. The first young mother was only 14 years old when her baby was delivered. Birth to Twenty is a multidisciplinary longitudinal study which tracks the growth, health and education progress of the respondents. For the necessary *time to sexual debut* survival analysis, this dissertation will focus only on data pertaining to sexual behaviour of the original cohort.

Running and maintaining such a large-scale study requires a significant amount of support. From the inception of Birth to Twenty, it has been supported by the South African Medical Research Council. Additional funders include the Institute for Behavioural Sciences at the University of South Africa. As of 1998, a major source of funding has been the Wellcome Trust, with further support from the Human Sciences Research Council (HSRC) of South Africa, the Medical Research Council , the University of the Witwatersrand, the Mellon Foundation,

the South-African Netherlands Programme on Alternative Development and the Anglo American Chairman's fund. For this dissertation, a collaboration has been fostered with the HSRC to obtain access to the Birth to Twenty data.

Birth to Twenty is the largest longitudinal study in South Africa. It is unique and has been the source of reference for a number of significant policy decisions in the country. In fact, from 1997 to 1999, the Minister of Health used results from the study pertaining to children's recognition of cigarette brands to help pass tobacco legislation that prevents the public advertisement of cigarettes and the sale of cigarettes to minors.

1.2 Data description

Birth to Twenty data were collected at several sites including clinics, stipulated study sites, households and schools of respondents in Soweto-Johannesburg in South Africa. The original cohort area was approximately $400km^2$. However, upon the emergence of a democratic South Africa, respondents were no longer restricted by laws that governed where they lived and the urban landscape changed considerably. Thereafter, the study tracked respondents throughout the Gauteng province covering an area of $17000km^2$ (Richter et al., 2007). The complete data set records data from as early as when the mother of the respondent is pregnant. These initial interviews took place at public antenatal clinics in the study area.

Over 2000 pregnant women participated in the first interview at the antenatal clinics. However, as a result of a hospital strike, the cohort enrolment dates had to be changed and only 1594 of the women interviewed gave birth during the revised cohort enrolment dates. Another selection criterion for admission into the study was that both the baby and the mother were to stay in the area for at least six months after the baby was born. The reason for this criterion was that the pilot studies had shown that several women came from rural areas to deliver their babies in urban areas, which means that soon after delivering

their babies they would leave the urban areas.

Even though all births are documented in the municipal area through a local ordinance, to make certain that records were not missing, mortuaries were checked and infants who came to hospitals for their 6 week postnatal check-up were backtracked by Bt20 staff. Based on these records, 5449 births were registered throughout the 7 week enrolment phase. Of these respondents, 3273 met the entrance criterion of residency in the area. Only 2216 of this population qualify to be entered into the survival analysis for sexual debut based on whether the respondent had a recording of the time to event (sexual debut) and a status (a response to whether or not they had engaged in first sex) as recordings on both variables are necessary for survival analysis.

Sexual behaviour measures were recorded in six data collection waves: 11-12 years, 13 years, 14 years, 15 years, 16 years and 17-18 years. Participants were followed once during the 11-12 year data collection wave and then biannually for the subsequent years. Initially, during the 11-12, 13 and 14 year collection waves, respondents were questioned by experienced Bt20 interviewers. For the 15 and 16 year data collection waves, respondents completed questionnaires via secret ballot and more recently for the 17-18 year wave, questionnaires were completed through a computer-assisted self-interview (CASI) system. A standard set of questions were asked and were repeated over the data collection waves. These questions related to first reported experience of foreplay, oral sex, anal sex and sexual intercourse. The questionnaire also included whether the sexual behaviours had been voluntary or involuntary and recorded the partner's age. As expected, in a longitudinal study, reported age of first sexual behaviours were inconsistent at each data collection wave. One possible reason is recall bias. To deal with this, the first report of sexual behaviours was taken to be the age at first sex. Additionally, the reporting of ages below 12 years at first sexual experiences were assumed to be involuntary.

Chapter 2

Exploratory data analysis

2.1 Introduction

This study focuses on factors which affect the timing of sexual debut among adolescents. Many aspects are taken into account and in particular, a main focus is to explore longitudinally the age at first experience of sexual intercourse in a prospective South African birth cohort.

Exposure measures were considered from the 13 year data collection wave to the 17 - 18 year data collection wave. Demographic data (age, gender and anthropometric indicators), social measures (religiosity, maternal education and father presence) and a household measure (asset index) were routinely recorded. The asset index was obtained by summing eight household assets and three categories were formed, namely low, middle and high.

Race was categorized by four groups according to Apartheid racial classification. The groups are; Black, Coloured, Asian and White. The majority of the respondents in the sample are of Black ethnicity. Distribution of the sample according to race was roughly representative of the South African population, except for an initial under-representation of White respondents. According to Richter et al. (2007), the reason for this under-representation was two-fold. Firstly, at the time, most White families used private health care systems and only children registered through public health care systems were included in this study. Secondly, White respondents tend to show higher attrition than other respon-

dents. This was because White families tend to be more wealthy than other families and thus see little or no value from participating in such studies. In an attempt to deal with the under-representation of White participants, a supplementary sample of 120 White children from a bone health study were recruited into the Bt20 study at age 10 years. These children were born during the cohort enrolment dates but not in the area. This has allowed the sample to then be roughly representative of the South African population.

Maternal education was classified into three groups. These are; no formal or primary education, secondary education and post-school training.

Father presence was classified as “father present” if the respondent was either living in the same household as the father or seeing the father on a regular basis if they did not live with the father. Alternatively, father presence was classified as “minimal or no contact” if the respondent was not living with their father and saw their father rarely or not at all.

Religiosity was assessed through responses to questions involving the reported importance of religion in the respondent's life, how often they attend religious services and the frequency with which their family prays together. According to this, three groups were established, namely not at all religious, somewhat religious and very religious.

The Tanner staging of breast (females) and genital (males) development was used to assess the sexual maturation of the respondents and was then classified into three groups. These are; prepubertal, early pubertal and late pubertal development (Richter et al., 2007).

Height was classified into three groups by using z-scores of height-for-age from the World Health Organization growth standards for males and females (de Onis et al., 2007). The groups are; stunted ($z < -2$), average height ($-2 < z < 2$) and tall for age ($z > 2$). Note that it is simply referred to as ”height” throughout this dissertation but the definition applied is height-for-age as defined above.

The status variable indicates whether the participant had engaged in sexual intercourse or not. "Time" is the amount of time in years until first experience of sexual intercourse. For censored observations, "time" is taken to be the decimal age of the participant since the decimal age is the exact age of the participant and has been recorded in the study.

2.2 Sample characteristics

The first sex survival analysis data set consists of 2216 observations of which 51.85% are female. Frequency tables, crosstabulations and bar charts were used to describe the nature of the data.

Table 2.1 Sample characteristics by gender

Characteristic		Sample		Female		Male	
		Count	%	Count	%	Count	%
Racial classification group	Black	1800	81.23	934	81.29	866	81.16
	Coloured	288	13.00	151	13.14	137	12.84
	White	70	3.16	35	3.05	35	3.28
	Asian	58	2.62	29	2.52	29	2.72
Maternal education	No formal/primary education	258	11.64	132	11.49	126	11.81
	Secondary education	1574	71.03	820	71.37	754	70.67
	Post-school training	197	8.89	99	8.62	98	9.18
Socioeconomic status	Low	769	34.70	386	33.59	383	35.90
	Middle	397	17.92	229	19.93	168	15.75
	High	362	16.34	194	16.88	168	15.75
Father presence	Minimal or no contact	599	27.03	336	29.24	263	24.65
	Father present	1089	49.14	547	47.61	542	50.80
Religiosity	Not at all	110	4.96	29	2.52	81	7.59
	Somewhat	349	15.75	178	15.49	171	16.03
	Very	1232	55.60	678	59.01	554	51.92
Height	Stunted	700	31.59	313	27.24	387	36.27
	Normal	1508	68.05	832	72.41	676	63.36
	Tall	8	0.36	4	0.35	4	0.37
Pubertal status	Prepubertal	237	10.69	60	5.22	177	16.59
	Early pubertal	1264	57.04	636	55.35	628	58.86
	Late pubertal	342	15.43	261	22.72	81	7.59
Foreplay	Engaged	1549	69.90	768	66.84	781	73.20
	Did not engage	667	30.10	381	33.16	286	26.80
Oral Sex	Engaged	528	23.83	235	20.45	293	27.46
	Did not engage	1688	76.17	914	79.56	774	72.54

Table 2.1 presents the characteristics of the sample for females, males and for

the sample as a whole. 81.23% of the sample are of Black ethnicity. Only 8.89% of the children had mothers who had some kind of education after matriculation. In other words, the majority of the respondents had mothers whose highest level of education was no formal/primary school or secondary school. Table 2.1 also shows that most of the respondents had a low socioeconomic status. The characteristics are roughly evenly distributed between females and males for race, maternal education, socioeconomic status and father presence. Overall, a greater portion of females tend to be more religiously inclined than males. A higher proportion of males experienced stunted growth than females. Substantially more females showed faster pubertal development. This is expected as, on average, girls enter and complete each stage of puberty earlier than do boys (Rogol et al., 2000). It is also evident that a greater portion of males had engaged in foreplay and oral sex than females.

From Table 2.1, it is calculated that socioeconomic status had 31% missing data, father presence had 24% missing data and religiosity had 24% missing data. Logistic regression was employed to impute the missing values based on the relationship between the exposure variables. It was found that maternal education, race and father presence were significant predictors of socioeconomic status. A multinomial regression model was then used to impute the missing values for socioeconomic status and all analyses including socioeconomic status hereafter are done so inclusive of the imputed values. Race was found to be the only significant predictor of father presence. Thus, father presence was not used in the model due to the significant amount of missing values. Gender was found to be the only significant predictor of religiosity. Furthermore, the proportional odds assumption for the logistic regression model was violated so the missing values were not imputed for religiosity.

Table 2.2 shows that for the Black, Coloured and Asian respondents, the majority have mothers who have a secondary education. The White group is the only group who have a majority of mothers having a post-school education and is also the only group that does not have any mothers with highest level of education being no formal/primary school education. These results were not uncommon

Table 2.2 Race and maternal education crosstabulation

		Maternal education					
		No formal/primary		Secondary		Post-school training	
		Count	%	Count	%	Count	%
Race	Black	236	14.59	1312	81.09	133	8.22
	Coloured	21	8.43	209	83.94	19	7.63
	White	0	0.00	19	38.00	31	62.00
	Asian	1	2.04	34	69.39	14	28.57

during the Apartheid era, where White citizens in South Africa generally had greater access to a better education as compared to other race groups due to overt racist policies (Nnadozie, 2013).

Table 2.3 Race and socioeconomic status crosstabulation

		Socioeconomic status					
		Low		Middle		High	
		Count	%	Count	%	Count	%
Race	Black	957	54.07	468	26.44	345	19.49
	Coloured	78	30.12	86	33.20	95	36.68
	White	1	1.52	3	4.55	62	93.94
	Asian	10	18.87	8	15.09	35	66.04

The majority of the Coloured, White and Asian respondents hold a high socioeconomic status while it is the opposite for the Black group, as shown in Table 2.3. The uneven distribution of socioeconomic status within the race groups is most substantial for White respondents followed by Asian respondents with 93.94% and 66.04% falling into the high category respectively. Only a minute proportion (1.52%) of White respondents had a low socioeconomic status according to the asset index method. As with the crosstabulation between race and maternal education, the inequalities between the race groups in terms of socioeconomic status can be attributed to overt racist policies of the Apartheid regime where Black citizens tend to be worse off compared to other race groups (Nnadozie, 2013).

Respondents with no formal/primary or secondary level maternal education are more likely to hold a low socioeconomic status (Table 2.4). Those respondents whose mothers have obtained post-school training tend to hold a high socioeconomic status. $5.81\%, 24.33\%$ and 47.72% respectively of mothers with no

Table 2.4 Socioeconomic status and maternal education crosstabulation

		Socioeconomic status					
		Low		Middle		High	
		Count	%	Count	%	Count	%
Maternal education	No formal/primary	197	76.36	46	17.83	15	5.81
	Secondary	749	47.59	442	28.08	383	24.33
	Post-school training	54	27.41	49	24.87	94	47.72

formal/primary, secondary and post-school training possess high socioeconomic statuses.

2.3 Sexual debut

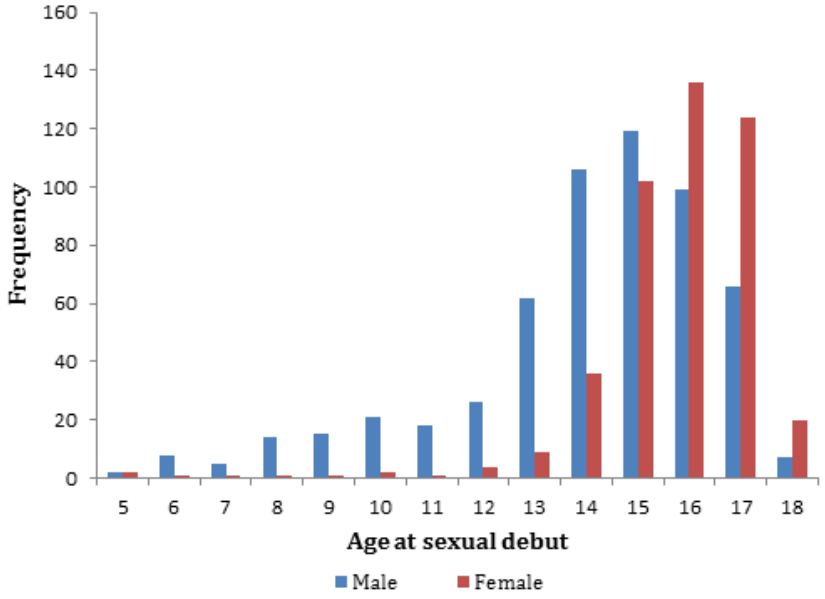

Figure 2.1: Bar chart displaying the distribution of age at sexual debut by gender

Figure 2.1 is a simple bar chart that displays the distribution of sexual debut for females and males at each age from age 5 years up to age 18 years. It is apparent that at early ages ranging from 5 years to 14 years for females, there are somewhat small increases in sexual debut. However, at age 15 there appears to be a dramatic increase in sexual debut, peaking at age 16, this slightly declines at age 17 and then reduces drastically at age 18 years. The distribution of sexual debut among male adolescents have approximately the same left-skewed shape as that of female adolescents but the changes from one age wave to the

next are less pronounced and relatively more gradual. For male adolescents, during young ages ranging from 5 years to 12 years, increases in sexual debut are quite subtle. At age 13 years and 14 years there are sharp increases in sexual debut. Sexual debut for males peaks at age 15 years.

Figure 2.1 shows that male adolescents reportedly engage in sexual debut earlier than their female counterparts. To formally test whether there are differences in ages at sexual debut across gender, an independent samples t-test was employed. The null hypothesis that there was no difference in the average age of sexual debut for males compared to females was tested. The alternative hypothesis was that females have a higher average age of sexual debut relative to males. The test produced a p-value less than 0.0001 indicating sufficient evidence to conclude that the average age of sexual debut is higher for females than it is for males.

Table 2.5 Cumulative incidence for adolescent females and males engaging in sexual debut up to age 18 years

		Sexual debut			
		Engaged		Did not engage	
Age	Gender	Count	%	Count	%
12 years or younger	Female	13	1.13	1136	98.87
	Male	109	10.22	958	89.78
	Sample	122	5.51	2094	94.49
$\leq$ 13 years	Female	22	1.93	1118	98.07
	Male	171	16.16	887	83.84
	Sample	193	8.78	2005	91.22
$\leq$ 14 years	Female	58	5.12	1075	94.88
	Male	277	26.41	772	73.59
	Sample	335	15.35	1847	84.65
$\leq$ 15 years	Female	160	14.25	963	85.75
	Male	396	38.19	641	61.81
	Sample	556	25.74	1604	74.26
$\leq$ 16 years	Female	296	26.64	815	73.36
	Male	495	48.34	529	51.66
	Sample	791	37.05	1344	62.95
$\leq$ 17 years	Female	420	38.46	672	61.54
	Male	561	55.65	447	44.35
	Sample	981	46.71	1119	53.29
$\leq$ 18 years	Female	440	42.93	585	57.07
	Male	568	59.48	387	40.52
	Sample	1008	50.91	972	49.09

Table 2.5 records the cumulative incidence at the associated age, of sexually

experienced and inexperienced adolescents, up to age 18 years. Each age category is subdivided by gender as we have hypothesized that responses differ for males and females (Lammers et al., 2000). By age 15 years, 14.25% of female adolescents and 38.19% of male adolescents had engaged in sexual intercourse. This comprised 25.74% of the sample. By age 18 years, 42.93% of females had engaged in first sex and 59.48% of males had engaged in first sex. Overall, by the 18 year data collection wave 50.91% of the sample had already engaged in sexual debut whereas 49.09% of the sample had maintained their virginity.

By age 12 years, there already were reports of sexual debut for both gender categories but for males these reports are significantly greater. There are 13 reports of sexual debut for females by age 12 and 109 for males. Reports of first sex prior to age 12 years are highly likely to have been coerced. The specific number of cases of sexual debut before age 12 can be seen separately for each data wave in Figure 2.1.

2.4 Sexual coercion

This section focuses on the type of sexual debut where type of sexual debut was either voluntary or involuntary. The terms involuntary sexual debut and coerced sexual debut are used interchangeably throughout this dissertation. It is defined as the act of persuading or forcing an individual to engage in first sex against his or her will (Agardh et al., 2011).

Table 2.6 Distribution of reported type of sexual debut across gender

		Voluntary		Involuntary		Total	
		Count	%	Count	%	Count	%
Gender	Female	339	79.58	87	20.42	**426**	43.43
	Male	406	73.15	149	26.85	555	56.57
Total		745	75.94	236	24.06		

Table 2.6 shows the reported voluntary and involuntary responses of sexual debut for females and males. Of the females, 79.58% reported that their sexual debut was voluntary and 20.42% had reported involuntary sexual debut whereas of the males, 73.15% had reported voluntary sexual debut and 26.85%

had reported involuntary sexual debut. A greater proportion of males appear to have been coerced into first sex. By looking at all those respondents who engaged in voluntary sexual debut we see that 45.50% are female and 54.50% are male while of those respondents who reported coerced sexual debut we see that 36.86% are female and 63.14% are male but this does not tell us much as a greater proportion of males than females engaged in sexual debut. Fisher's exact test was then employed to formally test whether there was an association between gender and the type of sexual debut. The test was conducted at the 0.05 level of significance and tested the null hypothesis that there was no association between gender and type of sexual debut against the alternative hypothesis that there was an association between gender and type of sexual debut. Fisher's exact test produced a p-value of 0.02 and thus the null hypothesis was rejected at the 0.05 level of significance indicating that there is an association between gender and type of sexual debut.

Next, the aim is to examine and compare reported voluntary and involuntary sexual debut across age. To make meaningful comparisons, the "age" variable was grouped into three broad categories, namely 14 years or younger, 15 to 16 years and 17 to 18 years.

Table 2.7 Distribution of reported type of sexual debut across age at sexual debut

		Voluntary		Involuntary		Total	
		Count	%	Count	%	Count	%
Age group at sexual debut	14 years or younger	220	67.90	104	32.10	324	33.03
	15 to 16 years	343	77.60	99	22.40	442	45.06
	17 to 18 years	182	84.65	33	15.35	215	21.92
Total		745	75.94	236	24.06		

Reporting of coerced sexual debut was less frequent for older respondents (Table 2.7). The younger the respondent, the more frequent was involuntary sexual debut. The reason for this is either that involuntary sexual debut occurs more often in younger adolescents or older respondents are less likely to report involuntary sexual debut. Of the adolescents who engaged in first sex, 32.10% of those 14 years or younger reported coerced first sex, 22.40% of the 15 to 16 year

old adolescents reported coerced first sex and 15.35% of the 17 to 18 year old adolescents reported coerced first sex. In total, of all the adolescents who had engaged in sexual debut and had reported the type of sexual debut, almost a quarter (24.06%) were reportedly coerced while 75.94% had reported voluntary sexual debut. Pearson's chi-square test for independence was used to test the null hypothesis that there was no association between the age groups at sexual debut and the type of sexual debut against the alternative hypothesis that there was an association between the age groups at sexual debut and the type of sexual debut. The test rendered a p-value of 0.00 thus the null hypothesis was rejected at the 0.05 level of significance which is indicative of an association present between the age groups at sexual debut and the type of sexual debut. Next, the type of sexual engagement across age at sexual debut are investigated and compared for females and males.

Table 2.8 Distribution of reported type of sexual debut across age at sexual debut for females and males

		Age group at sexual debut					
		14 years or younger		15 to 16 years		17 to 18 years	
Gender		Count	%	Count	%	Count	%
Female	Voluntary	31	9.14	183	53.98	125	36.87
	Involuntary	24	27.59	46	52.87	17	19.54
	Total	55	13.91	229	53.76	142	33.33
Male	Voluntary	189	46.55	160	39.41	57	14.04
	Involuntary	80	53.69	53	35.57	16	10.74
	Total	269	48.47	213	38.38	73	13.15

Table 2.8 allows for a multitude of important comparisons by critically examining the reporting of coercion for younger females versus older females, younger males versus older males, females versus younger males and females versus older males.

Firstly, in comparing younger to older females, we see that 9.14% of all females who reported voluntary sexual debut were 14 years or younger, 53.98% were 15 to 16 years old and 36.87% were 17 to 18 years old. This tells us that the majority of females who engaged in consensual first sex were 15 to 16 years old.

For females who engaged in coerced first sex, 27.59% were 14 years or younger, 52.87% were 15 to 16 years old and 19.54% were 17 to 18 years old thus most females who were coerced into first sex were 15 to 16 years old.

In comparing younger and older males, we note that of all males that reported voluntary sexual debut, 46.55% were 14 years or younger, 39.41% were 15 to 16 years old and 14.04% were 17 to 18 years old. On the other hand, of the males who had reported that their sexual debut was coerced, 53.69% were 14 years or younger, 35.57% were 15 to 16 years old and 10.74% were 18 to 19 years old. It is interesting to see that for the males, reporting of involuntary sexual debut predominantly occurs in the 14 years or younger age group whereas for females it predominantly occurs in the 15 to 16 year old age group. Furthermore, for females, reporting of coercion starts moderately high in the early years (14 years or younger), peaks at 15 to 16 years old and reduces moderately at 17 to 18 years old whereas for males, reporting of coercion peaks in the early years (14 years or younger) and thereafter declines in subsequent years. At 17 to 18 years old, only a very small proportion of males reported coercion. Younger males (14 years or younger) are more likely to report coercion than females of any age group and older males (17 to 18 years old) are less likely to report coercion than females of any age group.

To detect whether there was an association between type of sexual debut and age (groups) at sexual debut for females and for males, Pearson's chi-square test of independence was used in each of the two cases, that is, for females and for males. Both cases tested the null hypothesis that there is no association between type of sexual engagement and the age group at sexual debut against the alternative hypothesis that there is an association between type of sexual engagement and the age group at sexual debut. For females, a p-value of 0.00 was obtained and a p-value of 0.29 was obtained for males. At the 0.05 level of significance the null hypothesis was rejected for females whereas the null hypothesis was accepted for males. This means that for females there is an association between type of sexual debut and age at sexual debut whereas the opposite is true for male respondents.

Table 2.9 Distribution of reported type of sexual debut across partner's age at sexual debut for females and males

Gender		Partners age at sexual debut					
		14 years or younger		15 to 16 years		17 years or more	
		Count	%	Count	%	Count	%
Female	Voluntary	6	1.82	48	14.55	276	83.64
	Involuntary	9	11.11	14	17.28	58	71.60
	Total	15	3.65	62	15.09	334	81.27
Male	Voluntary	151	40.05	156	41.38	70	18.57
	Involuntary	72	55.38	38	29.23	20	15.38
	Total	223	43.98	194	38.26	90	17.75

Table 2.9 records the age of the respondents partner and the type of sexual debut, for females and males. The vast majority of females who report voluntary sexual debut had engaged in first sex with a partner aged 17 or older. This is also the case with females who reported coercion. Males who reported voluntary sexual debut had partners who were predominantly 15 to 16 years old. Most males who reported involuntary sexual debut had partners aged 14 years or younger. We have already seen that most females who report involuntary sexual debut were 15 to 16 years at first sex. Now we see that females are most likely to be coerced by partners who are 17 years or older, so it is highly likely that coercion in females occurs by partners older than the respondent. Most males who report involuntary sexual debut were 14 years or younger and most males were reportedly coerced by partners in the same age category.

Pearson's chi-square test of independence was used to check whether there was an association between type of sexual debut and partners age for females and males. In both cases the null hypothesis was that there was no association between type of sexual debut and partners age group against the alternative hypothesis that there was an association between type of sexual debut and partners age group. A p-value of 0.00 was obtained for females and 0.01 for males. Thus, at the 0.05 level of significance, there is an association between type of sexual debut and partners age group for both females and males.

The subject of early sexual debut is an important public health concern and is therefore a crucial topic to understand. This chapter has introduced sexual debut among adolescents in the Birth to Twenty study. We have seen that by age 18 years, approximately half of all adolescents in the study had experienced sexual debut. Of all adolescents who engaged in sexual debut, approximately one quarter reported that it was coerced. We have also seen that the majority of coerced sexual debut occurred with partners who are either older or the same age as the respondent and this is true for both females and males. It is critical to understand the factors that affect these times to early sexual debut for both voluntary and involuntary first sex in an effort to contribute to research in this field which is used to design action plans to attempt to delay voluntary sexual debut and prevent sexual coercion in young people. Chapter 4 focuses on determining these risk factors.

Chapter 3

Methodology

3.1 Introduction to survival analysis

The Birth to Twenty study considers sexual behaviour outcomes which are observed longitudinally and give rise to what is termed survival data. In this chapter we discuss the statistical methods that deal with survival data.

Survival analysis comprises of several statistical tools and methods which are utilized in the study of time to event data. In pioneer studies, the prototypical event was death, thus accounting for the name given to such methods. Originally, the time until the event occurred was termed survival time, however it is now more broadly defined as *failure time*. In the current study, the event of interest is sexual debut and the time until the participant engages in first sex is the failure time.

Marubini & Valsecchi (1995) explain that the origin to what is today termed survival analysis, can be traced back to as early as 1662. The work of John Graunt, a London based English haberdasher, is generally considered to be the initial building blocks that provided a foundation for further research contributing to the development of the statistical study of human populations. This was initiated by his publication of the book titled *Natural and Political Observations upon the Bills of Mortality* (1662). The publication was so well received that he was granted admittance as a fellow of the Royal Society. Graunt's book reported on the births and deaths collected over some decades in the recordings of Lon-

don parishes. Deaths were then distributed into classes based on age, gender, time period and cause of death. His book provided a revolutionary stepping stone in scientific perspective: Death was viewed as an event for the very first time. Some time later, similar work was initiated in Poland by Edmund Halley, an English astronomer, which resulted in the formulation of the first life tables (Marubini & Valsecchi, 1995).

Smith et al. (1964) discuss that it was not until the second World War that researchers became more focussed on developing survival analysis. This development was primarily sparked by interest in evaluating the reliability of military equipment. Even after the war had ended, researchers were fast realizing the extent of the usefulness and applicability of these recently developed statistical methods and interest in research in such methods continued. As the usage of survival analysis became more popular in private industry, researchers began to further develop these methods. Since then survival analysis has undergone significant advancements over many years by a number of researchers in several fields. Survival analysis lends itself to a vast array of disciplines, where terminology differs from discipline to discipline. The following displays some terminology differences as mentioned by Fox (2006):

- *Survival Analysis*/*hazard models* in biostatistics and epidemiology (For example, clinical trial analysis),
- *Event history Analysis* in sociology,
- *Future time Analysis* in engineering/reliability analysis,
- *Duration Models* in economics/political science.

3.1.1 Characteristics of survival data

Survival data arise when the objective of the study is contingent on the time elapsed between entry into the observational study and some prespecified event, that is, the outcome variable is the time until the event occurs (Tsiatis & Zhang, 2005). The point of entry may be birth (as in life expectancy studies) or perhaps

the time at which a particular treatment is administered to patients with a certain disease (as is the case in many medical studies). The endpoint event may be death , disease relapse, disease remission (complete freedom of any signs of disease in an absolutely predefined sense) or any chosen event of interest. In practice, it is often the case that a subject may experience an event other than the event of interest which alters the probability of experiencing the event of interest. Such events are termed *competing risk events*. More on competing risks is explained in Section 3.4. Different methods for analysis are used in such cases and the reader is referred to Kleinbaum & Klein (2005), Lee & Wang (2003) and Marubini & Valsecchi (1995) for step by step methodologies.

Survival analysis is not only useful when there is an interest in studying the frequency of occurrence of an event, but also when there is an interest in characterizing the underlying distribution of the time processes to those occurrences. In addition, survival analysis allows for the comparison of time to event for different groups (this is typical in biostatistics and epidemiology; for example, treatment versus control in clinical trials) and researchers often employ survival analysis when there is an interest in modeling the association between time to event and other covariates (often called prognostic factors) (Tsiatis & Zhang, 2005).

Survival data is generally not symmetrically distributed. Upon constructing a histogram, it usually shows that the distribution of the survival times are *positively skewed*. Consequently, the researcher may not use conventional methods applicable to normally distributed data (Collett, 2003). To remedy this situation, the data may be transformed to give a more symmetric distribution, although, a preferred approach would be to rather find an alternative distributional model which fits the original data (Collett, 2003). However, the main characteristic that renders standard methods unsatisfactory is the presence of censored data.

Right censoring

In survival analysis studies, it is typical that survival data be collected for a finite/limited period of time. Consequently, some units in the study may not experience the stipulated event even by the end of the observational period. This gives rise to what is called *censored data* and is a distinctive characteristic of survival data. Marubini & Valsecchi (1995) provide an indepth understanding of censoring by explaining that censored data does not reflect the true survival time, instead, all we know is that the observed units' survival time is *at least* the recorded time. The incomplete data are referred to as *right censored* and the subjects providing the data are termed *withdrawn alive*.

There may be other restrictions apart from a limited time, dependent upon the nature of the experiment, that may also result in incomplete data. One such circumstance may be that subjects enrolled in a study are no longer willing to participate or are simply unable to do so for any particular reason. Subjects of this nature are referred to as *lost to follow-up* and provide right censored data. Suppose a subject enters the study at time t_0 and is either lost to follow-up or does not experience the event at time t_0 + T, then T is the right censored survival time. According to Marubini & Valsecchi (1995), survival data is best represented by a pair of variables for each subject (T_i, δ_i), where T_i is the time to the event and δ_i is an indicator of failure such that

$$\delta_i = \begin{cases} 1, & \text{if the subject experiences the event by time } \boldsymbol{T} \\ 0, & \text{if the survival time is censored.} \end{cases}$$

Now suppose that if the survival time for subject i is censored, then C_i is the censoring time. Thus, we only observe min(T_i, C_i).

Left censoring and interval censoring

We differentiate between two other less common types of censoring: *left censoring* and *interval censoring*. The former refers to the case where a subject experiences the event of interest before the formal start of the study. Interval censoring occurs when the exact survival time in unknown in the study, however, the interval of time during which the event occurred is known. In the

Birth to Twenty sexual debut survival analysis study, there are right censored observations and no left censored or interval censored observations.

Informative and non-informative censoring

Informative censoring refers to cases where censoring is related to any factors associated with the actual survival time. In contrast, non-informative censoring thus means that censoring is not related to any factors associated with the actual survival time. In other words, censoring is independent of the actual survival time. The methods used in this dissertation for the analysis of censored survival data pertain only to non-informative censoring where censoring is conducted for administrative purposes; an example is the censoring conducted on subjects that are *lost to follow-up*. The probability of being censored at time $t = T$ does not depend on the prognosis for failure at time $t = T$.

Independent censoring

According to Marubini & Valsecchi (1995) independent censoring means that the censoring mechanism is independent of the event process. In practical terms this would imply that the survival experience of censored subjects after censoring can be accurately estimated using the survival data of uncensored subjects. This is due to the fact that under the assumption of independent censoring, the censoring mechanism carries no prognostic information. In other words, a censored observation is no more or less at risk to fail as compared to other observations in the sample.

Truncation

Truncation is a variant of censoring but must not be confused with the censoring mechanism. Bagdonavicius et al. (2011) explain that truncation occurs when the incomplete nature of an observation is attributable to a systematic selection process which is inherent to the design of the study. We distinguish between left and right truncation. In left truncation, only subjects whose event time is greater than some truncated threshold will be observed. This threshold

need not necessarily be the same for all subjects. Conversely, in right truncation, only subjects with event times less than some truncated threshold will be included in the study.

3.1.2 Notations and concepts

In describing the distribution of survival times, the following three functions are of primary interest:

- Distribution function,
- Survivor (or survival) function,
- Hazard function.

Distribution function

The outcome variable is time to event and is denoted as the *survival time*. Typically, we refer to the endpoint event as a *failure*. Let T be a continuous non-negative random variable used to denote the actual survival time of a subject and t will be used to denote the values that T take on. Then T has a probability density function $f(t)$. The distribution function of T is the probability that the subject experiences the event of interest before some time t and is thus given by:

$$\begin{aligned} F(t) &= P(T < t) \\ &= \int_0^t f(u)du \end{aligned} \tag{3.1}$$

Equation (3.1) represents the probability that the survival time is some time less than t.

Survivor function

The *Survivor function* is denoted by *S(t)* and represents the probability that a subject survives for some time greater than *t*, where

$$\begin{aligned} S(t) &= P(\text{subject survives longer than t}) \\ &= P(T > t) \\ &= 1 - F(t). \end{aligned} \tag{3.2}$$

$S(t)$ is a non-increasing function of t and the basic assumption of survival analysis can be written in the following mathematical way:

$$\lim_{t \to \infty} S(t) = 0.$$

The survivor function (Equation 3.2) represents the probability that a subject has a survival time greater than t, or equivalently, that the subject is event-free for a time greater than t (Smith et al., 1964). The graphical presentation of $S(t)$ is called the *survival curve*.

Hazard function

The hazard function *h(t)* is used to measure the instantaneous failure rate of a subject at time t and is obtained from the probability that a subject fails at time $t + \delta t$ conditioned on the subject having been event-free up until time t. The hazard function is defined as

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \\ &= \lim_{\delta t \to 0} \left[\frac{P(t \leq T < t + \delta t)/\delta t}{P(T \geq t)} \right] \\ &= \frac{\lim_{\delta t \to 0} \{P(t \leq T < t + \delta t)/\delta t\}}{S(t)} \\ &= \frac{\lim_{\delta t \to 0} \{[F(t + \delta t) - F(t)]/\delta t\}}{S(t)} \\ &= \frac{F'(t)}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

It follows then that

$$h(t) = -\frac{d}{dt}\{\log S(t)\} \tag{3.3}$$

therefore,

$$S(t) = \exp\{-H(t)\} \tag{3.4}$$

where

$$H(t) = \int_0^t h(u)du. \tag{3.5}$$

$H(t)$ is called the *integrated* or *cumulative hazard*. From Equation (3.4), the cumulative hazard function may be determined from the survivor function since

$$H(t) = -\log S(t). \tag{3.6}$$

Parametric and nonparametric models

Usually in data analysis, it is customary to compute summary statistics to provide basic information regarding the distribution of the data. However, due to potential censoring and other characteristics inherent in survival data, summary statistics such as the mean and variance may no longer carry the desired statistical properties. For example, the summary statistics may not be unbiased. Other methods are thus needed to present the data. The survival experience is conveniently summarized through estimates of the hazard and survivor functions. *Parametric* or *nonparametric* methods may be used (Tsiatis & Zhang, 2005).

The usage of conventional parametric methods requires specific assumptions regarding the underlying distribution of the survival times (Collett, 2003). Popular parametric models include Weibull, Exponential (a special case of the Weibull), Log-normal, Log-logistic and the Generalized Gamma model. The reader is referred to Lee & Wang (2003) for a complete study on these models. Smith et al. (1964) explain that nonparametric methods gained popularity over parametric methods as it allows the analyst to be blind to the exact underlying distribution of the survival times. The advancement of nonparametric methods was stimulated by the difficulty in obtaining evidence to support an existing family of survival time distributions (Marubini & Valsecchi, 1995). When survival times

follow a known theoretical distribution, nonparametric methods will be less efficient than parametric methods and more efficient than parametric methods when the underlying distribution is unknown.

In particular, a widely employed nonparametric technique used to estimate the survivor function is the Kaplan Meier/Product Limit Estimator developed by Kaplan & Meier (1958). Most computer software packages use this method owing to its simplistic step idea. The Kaplan Meier estimator utilizes information from all available observations by treating any time point as a series of steps defined by the observed survival times and censored times (Smith et al., 1964).

A model that contains both parametric and nonparametric components is said to be semiparametric. A popular semiparametric model is the Cox proportional hazards model and is introduced in Section 3.3.

3.2 Proportional hazards regression models

3.2.1 Introduction to regression models

In most studies that give rise to survival data, it is often the case that supplementary information will be recorded on the subjects. This supplementary information is referred to as *prognostic factors*. Prognostic factors may also be termed *risk factors*, *covariates*, *concomitant variables* or *independent variables* (Lee & Wang, 2003).

An approach based on statistical modeling may be used to explore the relationship between the survival experience of the patient and the explanatory variables. Regression modeling is thus frequently used to explain the relationship between the outcome variable and the explanatory variables where the outcome variable in survival analysis is the survival time T or some function of the survival time. The popularity of this approach may be attributed to its simplistic nature that allows for easy model fitting and interpretation (Hosmer & Lemeshow, 1999).

In analysing survival data, interest is centered on the risk or hazard of the event occurring at any time after the study begins. Therefore, the hazard function is modeled directly (Collett, 2003).

According to Collett (2003), there are essentially two broad reasons for modeling survival data. Firstly, a clear aim of the modeling process is to work out which combination of the explanatory variables affect the form of the hazard function. Also, we can study the effect that a treatment has on the hazard function and the extent to which all the other explanatory variables affect the hazard function. The second reason focusses on evaluating an estimate of the hazard function itself for a subject.

In building regression models, we distinguish between three different types of regression models. These are:

- Proportional hazards regression models,
- Accelerated failure time models and
- Proportional odds models.

We focus on the first class of regression models listed above. In order to construct these regression models, one will need to estimate the parameters belonging to each probability distribution. The reader is referred to Appendix A.1 for detailed methods of computing these parameters.

3.2.2 The proportional hazards regression model

Let N be the number of subjects in the study. Each subject has observed vector (t_i, δ_i, $\mathbf{x}_i$). The hazard function $h_i(t)$ for failure time T for individual i with covariate vector $\mathbf{x}'_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is given by Marubini & Valsecchi (1995) as

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x_i}) \tag{3.7}$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$ is the vector of regression parameters.

Covariates are assumed to be constant in time; typical examples are sex, religiosity and biochemical features which are usually recorded at the onset of the study. Note however, that a persons' level of religiosity may not actually be constant over time but for purposes of analysis we may use the explanatory variable religiosity to be "constant" in time if the variable is assumed to remain the same once it has been measured, so that only one measurement on that variable is used per subject.

The hazard function in Equation 3.7 depends on both time and covariates, but through two separate factors: $h_0(t)$ is an arbitrary function of time only and is assumed to be the same for all subjects whereas the second quantity, $\exp(\boldsymbol{\beta}'\mathbf{x_i})$ is a function of the covariates but does not involve time. As mentioned earlier in this section, covariates are assumed to be constant in time. In other words, the covariates are *time-independent* (Kleinbaum & Klein, 2005).

Let the covariate vectors of any two individuals be $\mathbf{x}_1 = (x_{11}, x_{12}, ..., x_{1p})'$ and $\mathbf{x}_2 = (x_{21}, x_{22}, ..., x_{2p})'$. Making use of the hazard function in Equation (3.7), we find that the ratio of the hazards for these two individuals will simply be given by

$$\begin{aligned} \frac{h_1(t)}{h_2(t)} &= \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x_1})}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x_2})} \\ &= \exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)] \end{aligned} \tag{3.8}$$

which is called the *hazard ratio*.

Applying a logarithmic scale to the above equation yields

$$\ln(h_1(t)) - \ln(h_2(t)) = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2).$$

Notice, the above equation shows that the model assumes a constant difference between the logarithm of the hazards. A graphical representation which plots the hazard function against time studied simultaneously with the plot of the logarithm against time may be used to provide a clearer understanding

(Marubini & Valsecchi, 1995). In particular, consider two subjects with covariate vectors $\mathbf{x}_1 = \mathbf{x}$ and $\mathbf{x}_2 = 0$ then the hazard ratio is thus

$$\frac{h(t)}{h_0(t)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x})}{h_0(t)} = \exp(\boldsymbol{\beta}'\mathbf{x}).$$

This shows that $h_0(t)$ may be regarded as the hazard function of a subject with all covariates of value zero and it is thus for this reason that $h_0(t)$ is referred to as the *baseline hazard*.

Therefore, proportional hazards models are a class of models in which the effect of the covariates is to either increase or decrease the hazard function by a constant proportion relative to the baseline function $h_0(t)$.

The Cox proportional hazards model will be considered in detail in the section below.

3.3 Cox proportional hazards model

The Cox proportional hazards model is based on a modeling approach to the analysis of survival data in which a parametric form for the effects of the covariates are assumed although the model does allow for an unspecified baseline hazard function. It is thus for this reason that the Cox proportional hazards (PH) model is semiparametric (Smith et al., 1964). This popular model is widely used to explore the effects of several explanatory variables on survival time. In addition, it allows us to estimate the hazard (or risk) of death for a subject given their prognostic variables (Kleinbaum & Klein, 2005). The popularity of the model may be attributed to the fact that, even though the baseline hazard function is not specified, reasonably good estimates of regression coefficients, hazard ratios and adjusted survival curves may be calculated for a wide variety of data situations making the model fairly robust (Kleinbaum & Klein, 2005). Marubini & Valsecchi (1995) mention that it is important to note that the flexibility in the model as a tool for regressing prognosis on various factors lies in the nonparametric specification of the baseline hazard.

It is assumed that the survival time for each subject of the population has its own hazard function, $h_i(t)$, where

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x_i}). \tag{3.9}$$

$h_0(t)$ is an arbitrary and unspecified baseline hazard function, $\mathbf{x}_i$ is the vector of prognostic variables for the ith subject and $\boldsymbol{\beta}$ is the vector of unknown regression parameters associated with the explanatory variables and is assumed to be the same for all subjects (Kleinbaum & Klein, 2005).

3.3.1 Assumptions of the Cox proportional hazards model

The model requires that the hazards are proportional over time, that is, the hazard functions are multiplicatively related. The hazard ratio of any two subjects is assumed to be a time-independent constant over the survival time, thereby avoiding temporal biases from becoming influential on the endpoint (Collett, 2003).

3.3.2 The survivor function

The survivor function is given by

$$S_i(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)} \tag{3.10}$$

where

$$S_0(t) = \exp[-\int_0^t h_0(u)du]$$

is called the baseline survivor function (Collett, 2003).

3.3.3 Fitting the model

Fitting the model given in Equation (3.9) to an observed set of survival data entails estimating the regression coefficients $\boldsymbol{\beta}$ in the linear component of the model. $h_0(t)$, the baseline hazard function may also need to be estimated, however these two components of the model can be estimated separately (Collett,

2003). β is first estimated and the estimate obtained is then utilized to construct an estimate of the baseline hazard function. This is a vital result as it implies that in order to make inferences regarding the effects of the covariates on the relative hazard ($h_i(t)/h_0(t)$), an estimate of $h_0(t)$ is not needed (Collett, 2003). Therefore, methods of estimating $h_0(t)$ will be deferred until Section 3.3.6.

3.3.4 Method of maximum likelihood

The method of maximum likelihood is used to estimate the vector of regression coefficients (β). To carry out this method, we first obtain the *likelihood* of the sample data which is given by the joint probability of the observed data. In the proportional hazards model, this is a function of the observed survival times of the subjects and the unknown β parameters in the linear component of the model. The estimate of β will then be that vector of values which are most likely on the basis of the observed data. The *maximum likelihood estimates* are thus those values that maximize the likelihood function. Computationally, it is usually easier to maximize the logarithm of the likelihood function. Also, it is useful to mention that approximations to the variance of maximum likelihood estimates may be determined from the second derivatives of the log-likelihood function. The reader is recommended to see Appendix A.1.

Suppose that a study is conducted in which data are available for n subjects. Let there be r distinct failure times and $n - r$ right censored survival times. Here we will assume that only one subject experiences failure at any given time, that is, we shall ignore the possibility of *ties* in the data. Appropriate methods for dealing with ties will be dealt with in Section 3.3.5. Let $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ be the r ordered failure times so that $t_{(j)}$ is the jth ordered failure time. $R(t_{(j)})$ will denote the set of subjects who are at risk at time $t_{(j)}$ so that $R(t_{(j)})$ is the group of subjects who are still alive and uncensored at a time just prior to $t_{(j)}$. $R(t_{(j)})$ is called the *risk set*. Cox (1972) found that the relevant likelihood function for

the proportional hazards model, (Equation 3.9),

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x_i})$$

is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \tag{3.11}$$

where $\mathbf{x}_{(j)}$ is the vector of explanatory variables for the subject who fails at the jth ordered failure time, $t_{(j)}$. Collett (2003) explains that the summation in the denominator of the above equation includes values of $\exp(\boldsymbol{\beta}'\mathbf{x})$ over all subjects who are not at risk at time $t_{(j)}$, the product is then taken over all subjects for whom death has been recorded. Notice that subjects that have censored survival times do not contribute to the numerator of the likelihood function but they are included in the summation over the risk sets at failure times that occur before a censored time. Furthermore, the risk set at each failure time is determined by the ranking of the failure times, thus the likelihood function is only dependant upon the ranking of the failure times. As a result, it is not surprising that inferences about the effect of explanatory variables on the hazard function depend only on the rank order of the survival times (Collett, 2003).

Now suppose that there are n observed survival times which are denoted by $t_1, t_2, \ldots, t_n$. Let δ_i be an indictor variable that takes on the value zero if the ith survival time $t_i, i = 1, 2, \ldots, n$, is right censored, else δ_i will equal unity.

Therefore, the likelihood function in Equation (3.11) can be expressed as

$$\prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right)^{\delta_i}$$

where $R(t_i)$ is the risk set at time t_i. The corresponding log-likelihood is then

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{\beta}'\mathbf{x}_i - \log \sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) \right). \tag{3.12}$$

Section 3.3.4 provides a justification as to why Cox (1972) choose the likelihood function in Equation (3.11) and gives details on the structure of the likelihood function.

Likelihood function for the model

In constructing a likelihood function for the proportional hazards model, Cox (1972) reasons that no information can be contributed about the effect of explanatory variables during time intervals in which there are no failures. This is because the baseline hazard function has an arbitrary form and it may thus be conceivable that $h_0(t)$ is equal to zero in such intervals. Therefore, Cox (1972) considered the probability that the ith individual fails at some time $t_{(j)}$, conditional on the set $t_{(j)}$, where $t_{(j)}$ is a single set from r death times, $t_{(1)}, t_{(2)}, \ldots, t_{(j)}, \ldots, t_{(r)}$. If we let $\mathbf{x}_{(j)}$ denote the vector of covariates for the subject who fails at $t_{(j)}$ then this probability is

$$\text{P(subject with variables } \mathbf{x}_{(j)} \text{ fails at } t_j \mid \text{one failure at } t_{(j)}). \tag{3.13}$$

Using the theorem of conditional probability, Equation (3.13) becomes

$$\frac{\text{P(subject with variables } \mathbf{x}_{(j)} \text{ fails at } t_{(j)})}{\text{P(one failure at } t_{(j)})}. \tag{3.14}$$

Failure times are assumed to be independent, therefore the denominator in the above expression reduces to the summation of the probabilities of failure at time $t_{(j)}$ over all subjects who are at risk of failure at that time. If the subjects are indexed by l, with $R(t_{(j)})$ being the risk set at $t_{(j)}$, then expression (3.14) may be written as

$$\frac{\text{P(subject with variables } \mathbf{x}_{(j)} \text{ fails at } t_{(j)})}{\sum_{l \in R(t_{(j)})} \text{P(subject } l \text{ fails at } t_{(j)})}. \tag{3.15}$$

By replacing the time $t_{(j)}$ with the interval $(t_{(j)}, t_{(j)} + \delta t)$ and then dividing both the numerator and denominator by δt, we obtain

$$\frac{\text{P[subject with variables } \mathbf{x}_{(j)} \text{ fails in } (t_{(j)}, t_{(j)} + \delta t)]/\delta t}{\sum_{l \in R(t_{(j)})} \text{P[subject } l \text{ fails in } (t_{(j)}, t_{(j)} + \delta t)]/\delta t}.$$

If we now consider the limiting value of this expression as $\delta t \to 0$ then we will obtain the ratio of the probabilities in expression (3.15). This limit also turns out to be the ratio of the corresponding hazards of failure at time $t_{(j)}$, that is

$$\frac{\text{Hazard of death at time } t_{(j)} \text{ for subject with variables } \mathbf{x}_{(j)}}{\sum_{l \in R(t_{(j)})} [\text{Hazard of failure at time } t_{(j)} \text{ for subject } l]}.$$

In particular, if it is the ith subject who fails at time $t_{(j)}$, then the hazard function in the numerator of the above expression may be written as $h_i(t_{(j)})$. Likewise, since the denominator is the summation of the hazards of failure at time $t_{(j)}$ over all subjects who fall into the risk set, this may be expressed as the summation of the values $h_l(t_{(j)})$ over those subjects in the risk set at time $t_{(j)}$. Thus, the conditional probability in expression (3.13) becomes

$$\frac{h_i(t_{(j)})}{\sum_{l \in R(t_{(j)})} h_l(t_{(j)})}.$$

Upon substituting Equation (3.9) into the above expression, $h_0(t)$ cancels out and we now have

$$\frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_l)}.$$

To obtain the likelihood function in Equation (3.11) we take the product of these conditional probabilities over the r death times. This likelihood obtained is not actually a true full likelihood since it does not directly use the censored and uncensored survival times. Consequently, it is called a *partial likelihood function*. Note that standard results which are used in maximum likelihood estimation carry over without modification to maximum partial likelihood estimation (Collett, 2003).

Parameter estimates of the Cox partial likelihood function

The Cox partial likelihood given in Equation (3.11) is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_l)}.$$

Replacing l with i, merely so that the notation in the following derivation is not confused, the log likelihood function is then given by Marubini & Valsecchi (1995) as

$$
\begin{aligned}
l(\boldsymbol{\beta}, \mathbf{x}) &= \log L(\boldsymbol{\beta}) \\
&= \log \left\{ \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{(j)})}{\sum_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_i)} \right\} \\
&= \sum_{j=1}^{r} \log \left\{ \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{(j)})}{\sum_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_i)} \right\} \\
&= \sum_{j=1}^{r} \left[\log\{\exp(\boldsymbol{\beta}' \mathbf{x}_{(j)})\} - \log \left\{ \sum_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_i) \right\} \right] \\
&= \sum_{j=1}^{r} \left[\boldsymbol{\beta}' \mathbf{x}_{(j)} - \log \left\{ \sum_{i \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_i) \right\} \right] \\
&= \sum_{j=1}^{r} l_j
\end{aligned}
$$

where $l_j = \left\{ \boldsymbol{\beta}' \mathbf{x}_{(j)} - \log \sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_i) \right\}$ is the contribution to the log likelihood function for failure time $t_{(j)}$.

In order to obtain the estimates $\hat{\boldsymbol{\beta}} = \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$, the Newton Raphson iterative procedure is used and is included in Appendix A.2.

3.3.5 Treatment of ties

In the proportional hazards model, the hazard function is assumed to be continuous so that tied survival times are not plausible (Collett, 2003). In practice however, recordings of survival times are usually done to the nearest day, month or year and so it is not unusual that tied survival times arise. Also, in addition to more than one failure at a given time, there may also occur more than one censored observation at a given time. The partial likelihood function may become really complicated in the presence of ties and must thus be modified to accommodate these tied survival times.

Collett (2003), in his discussion of Cox (1972) and Breslow & Crowley (1974),

proposed the following method:
Let $\mathbf{s}_j$ be the vector of sums of each of the p covariates for subjects who fail at the jth failure time, $t_{(j)}, j = 1, 2, \ldots, r$. $\mathbf{s}_{mj} = \sum_{k=1}^{d_j} x_{mjk}$ denotes the mth element of $\mathbf{s}_j$ if there are d_j failures at $t_{(j)}$ where x_{mjk} is the value of the mth explanatory variable, $m = 1, 2, \ldots, p$, for the kth of d_j subjects, $k = 1, 2, \ldots, d_j$, who fail at the jth time, $j = 1, 2, \ldots, r$.

According to Collett (2003), the simplest approximation to the likelihood function is the **Breslow approximate likelihood** which is given as

$$L_B(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_j)}{\left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\mathbf{x}_l)\right]^{d_j}}. \tag{3.16}$$

The d_j deaths at time $t_{(j)}$ are assumed to be distinct and to occur sequentially. Summing the probabilities of all possible sequences will then give Equation (3.16). This approximation is fairly easy to compute and useful when the number of tied survival times at any given time is not very large. In many statistical software packages designed to handle survival data, the Breslow approximate likelihood is usually the default method for dealing with ties (Collett, 2003).

The **Efron approximate likelihood** is given by

$$L_E(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_j)}{\prod_{k=1}^{d_j} \left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\mathbf{x}_l) - (k-1)d_j^{-1} \sum_{l \in D(t_{(j)})} \exp(\boldsymbol{\beta}'\mathbf{x}_l)\right]} \tag{3.17}$$

where $D(t_{(j)})$ is the set of all subjects who fail at time $t_{(j)}$. Collett (2003) comments that the Efron approximate likelihood gives a closer approximation than the Breslow approximate likelihood, although in practice, both often give similar results.

The approximation as suggested by Cox (1972) is given by

$$L_C(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp \boldsymbol{\beta}'\mathbf{s}_j}{\sum_{l \in R(t_{(j)}; d_j)} \exp(\boldsymbol{\beta}'\mathbf{s}_l)} \tag{3.18}$$

where $R(t_{(j)}; d_j)$ denotes a set of d_j subjects drawn from the risk set, $R(t_{(j)})$ at $t_{(j)}$. The summation in the denominator is over all possible sets of d_j subjects

sampled from the risk set. This is done without replacement. The approximation by Cox (1972) is based on a discrete time-scale so that tied observations are permissible. Consider the hazard function for a subject with vector of explanatory variables $\mathbf{x}_i$, then $h_i(t)$ is the probability of failure in the time interval $(t, t+1)$ given that the subject survived up to time t. Collett (2003) provides the following discrete version of the proportional hazards model:

$$\frac{h_i(t)}{1-h_i(t)} = \exp(\boldsymbol{\beta}'\mathbf{x}_i)\frac{h_0(t)}{1-h_0(t)}$$

for which the likelihood function is given in Equation (3.18). It should be noted that, in the limit as the width of these discrete time intervals reach zero, the model above tends to the original proportional hazards model (see Equation (3.9). In the absence of tied survival times, that is, $d_j = 1$ for each failure time, Equations (3.16), (3.17) and (3.18) all reduce to the partial likelihood function proposed by Cox (1972) (see Equation 3.11).

3.3.6 Estimating the hazard and survivor functions

Suppose that the linear component of a proportional hazards model contains p explanatory variables so that the estimated regression coefficients of these variables are $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. Then Collett (2003) gives the estimated hazard function for the ith of n subjects as

$$\hat{h}_i(t) = \hat{h}_0(t)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i) \tag{3.19}$$

where $i = 1, 2, \ldots, n$ and $\hat{h}_0(t)$ denotes the estimated baseline hazard function.

Kalbfleisch & Prentice (1973) proposed an estimate of the baseline hazard function using an approach based on the method of maximum likelihood as follows: Suppose that there are r distinct ordered failure times, $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Also, there are d_j failures and n_j subjects at risk at time $t_{(j)}$. The estimated baseline hazard function is thus

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\xi}_j \tag{3.20}$$

where $\hat{\xi}_j$ is the solution to the equation

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)}{1 - \hat{\xi}_j^{\exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l). \tag{3.21}$$

In the above equation, $D(t_{(j)})$ is the set of all d_j subjects who fail at $t_{(j)}, j = 1, 2, \ldots, r$. In the particular case when there are no tied failure times, that is, $d_j = 1$, the left hand side of Equation (3.21) will reduce to a single term. The equation can then be solved to give

$$\hat{\xi}_j = \left(1 - \frac{\exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} \right)^{\exp(-\hat{\boldsymbol{\beta}}' \mathbf{x}_{(j)})}.$$

However, when there are tied failure times, that is, when one or more d_j is greater than unity, Equation (3.21) becomes very complicated. This is due to the fact that the left hand side is now the summation of a series of fractions in which $\hat{\xi}_j$ occurs in the denominators, raised to different powers. Iterative methods are then used to determine a solution (Collett, 2003).

If we now assume that the hazard of failure is constant between adjacent failure times, then according to Collett (2003), the estimated baseline hazard function in this interval can be determined by dividing the estimated hazard in Equation (3.20) by the time interval, to produce a step function,

$$\hat{h}_0(t) = \frac{1 - \hat{\xi}_j}{t_{(j+1)} - t_{(j)}} \tag{3.22}$$

for $t_{(j)} \leq t < t_{(j+1)}, j = 1, 2, \ldots, r-1$, and $\hat{h}_0(t) = 0$ for $t < t_{(1)}$. The quantity $\hat{\xi}_j$ may be regarded as the estimated probability that a subject is event-free through the interval $(t_{(j)}, t_{(j+1)})$. The baseline survivor function can thus be estimated by

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j \tag{3.23}$$

for $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r-1$, which also is a step function. $\hat{S}_0(t)$ equals unity for $t < t_{(1)}$ and zero for $t \geq t_{(r)}$, however, when there are censored survival times greater than $t_{(r)}$ then $\hat{S}_0(t) = \hat{S}_0(t_{(r)})$ until the greatest censored

time. Beyond this time, $\hat{S}_0(t)$ is undefined.

From Equation (3.6) which states that the cumulative hazard function may be calculated from the survivor function, that is, $H_0(t) = -\log S_0(t)$, an estimate of the cumulative hazard function is thus

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{j=1}^{k} \log \hat{\xi}_j \tag{3.24}$$

for $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r-1$, and $\hat{H}_0(t) = 0$ for $t < t_{(1)}$. Now, estimates of $h_0(t), S_0(t)$ and $H_0(t)$ may be obtained for a subject with vector of covariates $\mathbf{x}_i$. In particular, the estimated hazard function is given by $\hat{h}_i(t) = \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)\hat{h}_0(t)$ (see Equation 3.19). Integrating both sides of this equation yields

$$\int_0^t \hat{h}_i(u)du = \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i) \int_0^t \hat{h}_0(u)du. \tag{3.25}$$

Therefore, the estimated cumulative hazard function for the ith subject will be

$$\hat{H}_i(t) = \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)\hat{H}_0(t). \tag{3.26}$$

Now, we multiply both sides of Equation (3.25) by -1 and exponentiate. Then, making use of Equation (3.4), which is given by $S(t) = \exp\{-H(t)\}$, we see that the estimated survivor function for the ith subject is given by

$$\hat{S}_i(t) = \{\hat{S}_0(t)\}^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)} \tag{3.27}$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$. Once the estimated survivor function has been determined, the integrated or cumulative hazard function is simply $-\log \hat{S}_i(t)$ (Marubini & Valsecchi, 1995).

The special case of no covariates

Let us consider the simple case where there are no covariates so that we will only deal with a single sample of failure times. Therefore, Equation (3.21) reduces to

$$\frac{d_j}{1 - \hat{\xi}_j} = n_j.$$

Rearranging the above equation to make $\hat{\xi}_j$ the subject of the formula gives

$$\hat{\xi}_j = \frac{n_j - d_j}{n_j}.$$

Therefore, the estimated baseline hazard at time $t_{(j)}$ is $1 - \hat{\xi}_j = \frac{d_j}{n_j}$. The corresponding estimated survivor function calculated from using Equation (3.23) is then

$$\begin{aligned} \hat{S}_0(t) &= \prod_{j=1}^{k} \hat{\xi}_j \\ &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \end{aligned}$$

which happens to be the Kaplan Meier estimate of the survivor function (Kaplan & Meier, 1958). This means that the estimated survivor function, that is, $\hat{S}_i(t) = \{\hat{S}_0(t)\}^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)}$, generalizes the Kaplan Meier estimate to the case where the hazard function depends on explanatory variables.

Some approximations to estimates of the baseline functions

As we have seen in Section 3.3.6, in the presence of tied survival times, $\hat{h}_0(t)$ is determined using iterative methods. However, Collett (2003) mentions that one way to avoid these iterative methods is to make use of an approximation to the summation of the left hand side of Equation (3.21). Recall,

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}{1 - \hat{\xi}_j^{\{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)\}}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l).$$

If we write the term $\hat{\xi}_j^{\{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)\}}$ as

$$\exp\{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) \log \hat{\xi}_j\}$$

and then taking the first two terms in the expansion of the exponent will give

$$\exp\{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) \log \hat{\xi}_j\} \approx 1 + \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) \log \hat{\xi}_j. \tag{3.28}$$

Writing $1 - \tilde{\xi}_j$ for the estimated baseline hazard obtained using the above approximation and substituting $1+\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)\log\hat{\xi}_j$ for $\hat{\xi}_j^{\{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)\}}$ in Equation (3.21) then gives

$$-\sum_{l\in D(t_{(j)})}\frac{1}{\log\tilde{\xi}_j}=\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l).$$

So then, since d_j is the number of deaths at $t_{(j)}$, we have

$$\frac{-d_j}{\log\tilde{\xi}_j}=\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l).$$

Therefore,

$$\tilde{\xi}_j=\exp\left(\frac{-d_j}{\sum_{l\in R(t_j)}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}\right). \tag{3.29}$$

Using Equation (3.23), the estimated baseline survivor function is then

$$\tilde{S}_0(t)=\prod_{j=1}^{k}\exp\left(\frac{-d_j}{\sum_{l\in R(t_j)}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}\right) \tag{3.30}$$

and the estimated baseline cumulative hazard function is thus

$$\tilde{H}_0(t)=-\log\tilde{S}_0(t)=\sum_{j=1}^{k}\left(\frac{-d_j}{\sum_{l\in R(t_j)}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}\right) \tag{3.31}$$

for $t_{(k)}\le t<t_{(k+1)}, k=1,2,\ldots,r-1$. The above estimate is referred to as the *Nelson-Aalen* or the *Breslow* estimate of $H_0(t)$. In the absence of covariates, $\tilde{S}_0(t)$ becomes

$$\tilde{S}_0(t)=\prod_{j=1}^{k}\exp\left(-\frac{d_j}{n_j}\right) \tag{3.32}$$

with corresponding $\tilde{H}_0(t)$ given by

$$\tilde{H}_0(t)=\sum_{j=1}^{k}\left(\frac{d_j}{n_j}\right),$$

since the number of subjects that are at risk at time $t_{(j)}$ is n_j (Collett, 2003).

Collett (2003) gives an additional approximation to the baseline functions by concentrating on the exponent of Equation (3.29), that is

$$\left(\frac{-d_j}{\sum_{l \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} \right),$$

and notes that unless there exists a large number of tied survival times at particular failure times, the above expression will usually be small. Considering the first two terms of this expression, and denoting this new approximation to ξ_j by ξ_j^* gives

$$\xi_j^* = 1 - \left(\frac{d_j}{\sum_{l \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} \right).$$

Under this approximation, the estimated baseline hazard function for the interval $(t_{(j)}, t_{(j+1)})$ is given by

$$h_0^*(t) = \frac{d_j}{(t_{(j+1)} - t_{(j)}) \sum_{l \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} \tag{3.33}$$

for $t_{(j)} \leq t < t_{(j+1)}, j = 1, 2, \ldots, r-1$. The corresponding estimate of the baseline survivor function is

$$S_0^*(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{\sum_{l \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} \right) \tag{3.34}$$

with estimated cumulative hazard function,

$$H_0^*(t) = -\log S_0^*(t) = -\sum_{j=1}^{k} \log \left(1 - \frac{d_j}{\sum_{l \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} \right). \tag{3.35}$$

Note that in the absence of covariates, the estimates will be the same as that given in Section 3.3.6.

In practice, using either $\tilde{S}_0(t)$ or $S_0^*(t)$ is preferred instead of using $\hat{S}_0(t)$. This is because $\hat{S}_0(t)$ requires computationally intense methods of evaluation. It is also noteworthy that when the number of tied survival times is low, all three estimates will be very similar.

3.3.7 Regression diagnostics

Regression diagnostics refers to checking the adequacy of a model once it has been fitted to an observed set of data. Generally in model checking, a visual inspection of the data is first conducted and reveals some key features in the data, however, this is not very useful when there are more than one or two explanatory variables. Moreover, the presence of censored observations complicates the situation even further and makes visual inspection of the data very complicated even in the simplest of situations.

Some key aspects in assessing model adequacy involve:

- Checking for the presence of outliers,
- Checking whether the correct functional form of the explanatory variables were used,
- Finding influential points,
- Checking whether the model violates the proportional hazards assumption,
- Checking for missing predictors.

It is rather difficult to identify influential points and outliers in survival data sets. Thus, we focus on the other aspects of model checking listed above.

A simple graphical check

A simple graphical check can reveal whether the data follow the proportional hazards assumptions or whether the assumptions are being violated. There are two graphical approaches that may be used.

A first approach is to check a plot of $-\ln(-\ln S(t))$ versus t for the various covariates (Kleinbaum & Klein, 2005). If the curves are more or less parallel then we would expect that the proportional hazards assumption holds, however, if the curves intersect, it means that there certainly is a violation of the proportional hazards assumption. Note that, as stated by Kleinbaum & Klein

(2005), the $-\ln(-\ln S(t))$ survival curve can be written as $\ln(-\ln S(t))$. This is because the $\ln(-\ln S(t))$ survival curve is actually a transformation of an estimated survival curve that is a result of taking the natural log of an estimated survival probability twice. Many statistical software packages including SAS and SPSS produce $\log(-\log\ S(t))$ survival curves.

A second graphical technique is to compare observed with predicted survivor curves. The observed survivor curves are derived for categories of the explanatory variable being assessed, without putting the variable in a proportional hazards model, whereas with the predicted survivor curve, the explanatory variable being assessed is included in a proportional hazards model. Now, if predicted and observed survivor curves are close, it is indicative that the proportional hazards assumption is reasonable (Collett, 2003).

Residuals for the Cox proportional hazards model

Many model checking procedures are based on residuals. Residuals may be calculated for each subject in the study and have the feature that their behaviour is known or at least approximately known when the model is adequate.

Collett (2003) proposes the usage of a number of residuals in connection with the Cox proportional hazards model. In particular we consider:

- Cox-Snell residuals,
- Martingale residuals,
- Schoenfeld residuals.

Cox-Snell Residuals: These are the most widely used residuals in the analysis of survival data and are so named as it is a particular case of the general definition of residuals given in a paper by Cox & Snell (1968).

The Cox-Snell residual for the ith subject is given by

$$r_{Ci} = \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_i)\hat{H}_0(t_i), \tag{3.36}$$

$i = 1, 2, \ldots, n$, where $\hat{H}_0(t_i)$ is the estimated baseline cumulative hazard function at time t_i. Note that r_{Ci} may also be written as

$$r_{Ci} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i). \tag{3.37}$$

This residual is derived in the following way:
Let T be the random variable which denotes the survival time of a subject and $S(t)$ is the corresponding survivor function then $Y = -\log S(T) \sim \exp(1)$ regardless of the form of $S(t)$.

According to a general result, if $f_X(x)$ denotes the probability density function of a random variable X, then the probability density function of a random variable $Y = g(X)$ is given by

$$f_Y(y) = f_X\{g^{-1}(y)\} \Big/ \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|$$

where $f_X\{g^{-1}(y)\}$ is the density of X expressed in terms of y. Making use of this transformation result, in our case with $Y = -\log S(T)$, we have

$$f_Y(y) = f_T\{S^{-1}(e^{-y})\} \Big/ \left|\frac{\mathrm{d}y}{\mathrm{d}t}\right| \tag{3.38}$$

where $f_T(t)$ is the probability density function of T. Next, since

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}\{-\log S(t)\}}{\mathrm{d}t} = \frac{f_T(t)}{S(t)}$$

and when we take the absolute value of this function, expressed in terms of y, the above becomes

$$\frac{f_T\{S^{-1}(e^{-y})\}}{S\{S^{-1}(e^{-y})\}} = \frac{f_T\{S^{-1}(e^{-y})\}}{e^{-y}}.$$

Finally, upon substituting the above in Equation (3.38) we obtain

$$f_Y(y) = e^{-y}$$

but this is the probability density function of an exponential distribution with mean equal 1, that is, $Y = -\log S(T) \sim \exp(1)$. Now, the argument put forward by Collett (2003) is as follows:

When the fitted model is adequate, the estimate that it will produce of the survivor function for the ith subject at time t_i, that is, the survival time of that subject, will be close to the corresponding true value $S_i(t_i)$. In other words, $\hat{S}_i(t_i)$ will be close to $S_i(t_i)$. Thus, it follows that $-\log \hat{S}_i(t_i), i = 1, 2, \ldots, n$, will behave as n observations from a Unit Exponential distribution. These estimates produced are then the Cox-Snell residuals.

If the observed survival time of a subject is right censored then the corresponding value of the Cox-Snell residual will also be right censored, however, the residual cannot be regarded on the same footing as those derived from uncensored observations (Collett, 2003). Thus, to account for censored observations, we must modify the Cox-Snell residuals.

When censoring occurs, we know that the actual survival time t_i is some time more than the observed censored time, which we may call t_i^*, that is, $t_i > t_i^*$. Thus, as defined earlier, the Cox-Snell residual for this subject evaluated at the censored survival time is

$$r_{Ci} = \hat{H}_i(t_i^*) = -\log \hat{S}_i(t_i^*). \tag{3.39}$$

Now, we have already established that if the fitted model is adequate, then the values r_{Ci} will follow an exponential distribution with a mean equal to unity. Collett (2003) then mentions that, since the cumulative hazard function of this distribution increases linearly with time, that is, the greater the survival time of a subject, the greater the value of the Cox-Snell residual, it follows that the residual for the ith subject at the true (unknown) survival time will be more than the residual evaluated at the observed censored survival time. It is thus on this basis that the *modified Cox-Snell* residuals are constructed to account for censoring. This is done by adding some positive constant Δ, which we call the *excess residual*. Therefore the modified residuals then have the form

$$r'_{Ci} = \begin{cases} r_{Ci}, & \text{for uncensored observations} \\ r_{Ci} + \Delta, & \text{for censored observations.} \end{cases}$$

Now all that remains is to determine a suitable value of Δ. By making use of the lack of memory property of a random variable following an exponential

distribution, Collett (2003) shows that since r_{Ci} follows a Unit Exponential distribution, the excess residual Δ also follows a Unit Exponential distribution. The proof of this has been omitted without loss of continuity. So now, the expected value of Δ is the value one and Collett (2003) suggests that the value of Δ itself is taken to be one, hence

$$r'_{Ci} = \begin{cases} r_{Ci}, & \text{for uncensored observations} \\ r_{Ci} + 1, & \text{for censored observations.} \end{cases} \tag{3.40}$$

An alternative way of representing the ith Cox-Snell residual is

$$r'_{Ci} = 1 - \delta_i + r_{Ci} \tag{3.41}$$

where δ_i is an event indicator that takes on the value zero if the observed survival time for the ith subject is censored and one if it is uncensored. Crowley & Hu (1977) argue that the addition of the value one to the Cox-Snell residual for a censored observation causes the residual to be far too inflated. They therefore suggested that the median value of the excess residual be used rather than the mean (expected value). Since $S(t) = e^{-t}$, the median, $t(50)$ is such that $e^{-t(50)} = 0.5$. Therefore, $t(50) = \log 2 = 0.693$. Thus a second version of the modified Cox-Snell residual is then

$$r'_{Ci} = \begin{cases} r_{Ci}, & \text{for uncensored observations} \\ r_{Ci} + 0.693, & \text{for censored observations.} \end{cases} \tag{3.42}$$

However, if the number of censored observations in the data set is not very large then the set of Cox-Snell residuals obtained from either method will yield approximately the same results.

It is noteworthy to mention that Cox-Snell residuals are not symmetrically distributed about zero and also they cannot be negative. Moreover, since the residuals follow a Unit Exponential distribution, it is expected that when the model is fitted adequately, residuals will have a skew distribution.

Martingale residuals: These residuals are defined as

$$r_{Mi} = \delta_i - r_{Ci}. \tag{3.43}$$

The name is based on the fact that these residuals are derived by using martingale theory. They may take on values in the interval $(-\infty, 1)$ and the residuals for censored survival times may be negative. Also, the summation of all the residuals is zero and in large samples these residuals are uncorrelated with each other and has an expected value of zero.

For the ith subject, δ_i can be thought of as the observed number of failures in the interval $(0, t_i)$ and the estimated cumulative hazard function may be viewed as the expected number of failures in the same interval. The difference between these two quantities gives the Martingale residuals in Equation (3.43). Collett (2003) goes on to explain that these residuals are particularly useful in examining the functional form of the relationship between survival and a covariate. A plot of r_{Mi} against a covariate should reveal the correct functional form for including the covariate in the model. This covariate may even be one that is not currently included in the model or one that the analyst wishes to check for non-linear effects.

Schoenfeld residuals: So far we have considered Cox-Snell residuals and Martingale residuals. These residuals have the disadvantage that they depend on the observed survival time and require the calculation of the cumulative hazard function. The residuals that were then proposed by Schoenfeld (1982) overcame both these disadvantages. The Schoenfeld residuals also differ from those previously considered because these residuals are calculated on each covariate for each subject as opposed to a single residual for each subject. Thus, Schoenfeld residuals will take on a set of values, one for each covariate and are a measure of the difference between the covariate for the ith subject and a weighted average of that covariate over the risk set at the corresponding subjects' failure time.

Including time-varying covariates in the model

A key underlying assumption of the Cox model is that the effects of the covariates on the hazard of the outcome does not change with time. If the hazard

ratio is found to change with time for some variable x, then we know that x interacts with time or some function of time. If this is the case then including a time interaction variable will yield a more suitable model to accommodate non-proportional hazard ratios. It follows then that, to test whether hazards are proportional, we can add a time interaction variable to the model and test its significance. If the variable is not significant then the proportional hazards assumption is not violated and the interaction variable can be dropped from the model, however, if the interaction variable is significant then it must remain in the model and the proportional hazards assumption is not satisfied. According to Allison (1995), using time interaction variables to validate the proportional hazards assumption is quite useful to the researcher as it provides both a test of the proportional hazards assumption and a fix to non-proportional hazards.

3.4 Survival analysis in the presence of competing risks

It is not uncommon for a participant in a survival analysis study to be at risk of more than one type of failure. The term competing risks refers to the situation where more than one type of failure can occur, and the observation of one type of failure hinders or precludes the observation of other types of failures (Dignam et al., 2012).

Figure 3.1 graphically depicts the Birth to Twenty competing risks model. The initial state is the event-free state. Participants who then engage in sexual debut may do so in one of two ways, either voluntarily or involuntarily. These separate causes to the event of interest are competing risks events as those participants who engage in voluntary sexual debut will never experience involuntary sexual debut and vice versa.

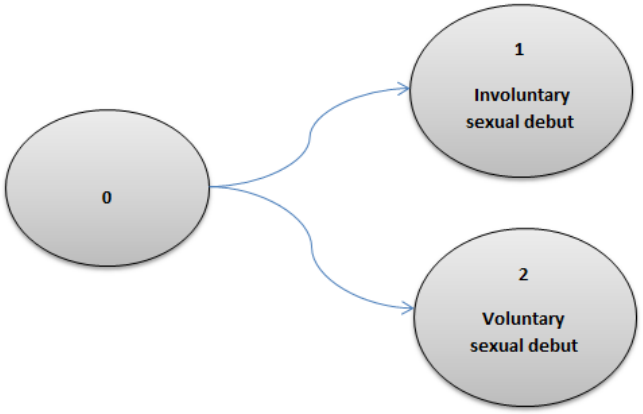

Figure 3.1: Birth to Twenty competing risks model

Statistical analysis and interpretation of competing risks data differ from standard survival analysis with only one cause of failure. Appropriate techniques must be applied to determine the correct estimate of the cumulative probability of each event in the presence of competing risk events (Dignam et al., 2012). The Cox proportional hazards model may be used for regression analysis, however the interpretation of the results become different compared to standard interpretation with a single event of interest (Putter et al., 2006). Recent techniques of analysing competing risks survival data involve two main quantities; namely, the *cause-specific hazard function* and the *cumulative incidence function*. The cause-specific hazard is defined as the instantaneous risk of failing from a specific type of event in the presence of competing events. First, a regression model for the cause-specific hazard was considered, however, the cause-specific hazard did not have a direct interpretation in terms of survival probabilities relevant to a specific cause of failure. Additionally, the effect of the prognostic factors on the cause-specific hazard function may be substantially different from the effect of the prognostic factors on the corresponding cumulative incidence function. The reader is referred to Fürstovà & Valenta (2011) for a full argument on the shortcomings of modeling the cause-specific hazard functions of competing risks. In lieu of these shortcomings, Fine & Gray (1999) proposed a direct regression modeling approach on the hazards of the cumulative incidence func-

tion of the competing events.

3.4.1 A proportional hazards model for the subdistribution of a competing risk

The methodology put forward by Fine & Gray (1999) makes use of a semiparametric proportional hazards model for the cumulative incidence function of the competing risk. In creating a model for the subdistribution, Fine & Gray (1999) intended to develop an equivalent to the Cox model for univariate survival analysis. In other words, they wanted to create a parsimonious semiparametric model for the subdistribution which has direct applicability to competing risks regression modeling. Under this methodology, the cumulative incidence function is also known as the subdistribution function or the marginal probability function and is so called to reflect that the cumulative probability of failing from the corresponding specific cause remains less than unity in the presence of competing risks (Fürstovà & Valenta, 2011).

Let $K \in (1, \ldots, k)$ be the cause of failure so that a participant can potentially fail from any one of k event types where k causes are assumed to be observed. $\mathbf{x}$ is a $p \times 1$ vector of covariates. Let T be the time until failure and C be the censoring time. Observations are represented by the pair (X, K) where $X = \min(T, C)$ and $K = 0$ for censored observations. The cumulative incidence function for failure from cause K before some time t, conditional on the covariates and in the presence of other competing risks is calculated as

$$\begin{aligned} I_K(t, \mathbf{x}) &= P(T \leq t, \text{cause} = K|\mathbf{x}) \\ &= \int_0^t S_0(u, \mathbf{x}) h_K(u, \mathbf{x}) du \end{aligned} \tag{3.44}$$

where

$$h_K(t, \mathbf{x}) = -\frac{d}{dt} \log\{1 - I_K(t, \mathbf{x})\} \tag{3.45}$$

is the subdistribution hazard and $S_0(t, \mathbf{x})$ is the probability of remaining event-free by time t, that is, it is the probability of not experiencing either event by

time t (Fine & Gray, 1999).

Applying the proportional hazards assumption to Equation (3.45), we obtain

$$h_K(t, \mathbf{x}) = h_{K,0}(t) \exp(\beta_K' \mathbf{x}) \tag{3.46}$$

where $h_{K,0}(t)$ is the baseline hazard function of the subdistribution for failure time *t* from cause *K*.

The partial likelihood function used in the model is a modification of the partial likelihood function proposed by Cox in the proportional hazards model (See Section 3.3.4) and is given as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} w_{jl} \exp(\boldsymbol{\beta}' \mathbf{x}_l)} \tag{3.47}$$

where r is the number of distinct failure times from failure type 1 and $\mathbf{x}_{(j)}$ is the vector of prognostic factors for the subject experiencing event type $K = 1$ at time $t_{(j)}$. The weights w_{jl} become active as soon as censoring occurs. Fine & Gray (1999) define the risk set $R(t_{(j)})$ as

$$R(t_{(j)}) = \{l : t_l \geq t \cup (t_l \leq t \cap K_l \neq 1)\}. \tag{3.48}$$

The risk set at any time point includes those who are still at risk of that type of event as well as those who have experienced a competing risk event prior to that time point. Those subjects still at risk of that type of failure contribute a weight of $w_{jl} = 1$ whereas subjects who have experienced a competing risk event prior to that time point contribute time-dependant weights to the partial likelihood function. These time-dependant weights are ≤ 1 and diminish over time (Fine & Gray, 1999). Estimation for this model and regression diagnostics are analogous to the Cox proportional hazards model which has been discussed in Section 3.3. A comprehensive discussion on the Fine and Gray proportional hazards model for the subdistribution of a competing risk may be found in Fine & Gray (1999).

Chapter 4

Birth to Twenty sexual debut study results

Standard survival analysis using the Cox proportional hazards model was applied to the Birth to Twenty data. Due to the presence of associations between the covariates in the model, univariate analyses were run and were stratified by gender. Results and regression diagnostics will be presented and interpreted for the Cox proportional hazards model. A more suitable and sophisticated approach to consider stems from the fact that the event of interest may occur from two separate causes, that is, sexual debut may occur voluntarily or involuntarily. These separate causes to the event of interest spark the consideration of a competing risks model which is explored in Section 4.2.

4.1 Cox proportional hazards model

The Pearson's chi-square test for independence was first run to assess the associations between the exposure measures. The p-values are tabled below and indicate the significance of the associations.

Table 4.1 shows several significant associations between the exposure variables. Due to the presence of these associations, a univariate analysis is run. These significant associations could imply that the exposure variables may have direct and indirect effects on the time to sexual debut. To strictly investigate direct effects, the Cox proportional hazards model was run for each exposure variable

Table 4.1 P-values for Pearson's chi-square tests for independence

Significance of associations (p-value)								
	Race	Mat edu	SES	Reli	Height	Pub status	Foreplay	Oral sex
Race	.							
Mat edu	0.000	.						
SES	0.000	0.000	.					
Reli	0.215	0.162	0.068	.				
Height	0.000	0.088	0.063	0.609	.			
Pub status	0.082	0.044	0.016	0.044	0.000	.		
Foreplay	0.001	0.014	0.006	0.248	0.000	0.002	.	
Oral sex	0.000	0.604	0.016	0.067	0.717	0.589	0.000	.

separately. We focus on the effects of the exposure variables on the time to sexual debut for females and males separately. For significant variables, the cumulative incidence functions across the categories of that exposure variable are included in the results and a plot of the $\log(-\log\ S(t))$ versus the time to sexual debut is examined to verify the validity of the proportional hazards assumption. Thereafter, Cox-Snell residual plots are assessed to determine the goodness of fit of the Cox proportional hazards model to the Birth to Twenty data.

Table 4.2 presents the Cox proportional hazards model results. The table gives the hazard ratio along with the 95% confidence interval and the p-value for each of the categories of the covariates relative to the baseline category. Note that "$*$" is used to show which covariates are significant at the 0.10 level of significance and "$**$" is used to show covariates that are significant at the 0.05 level of significance. Furthermore, only covariates that are significantly different from each other are further investigated by examining the relevant graphical output.

From Table 4.2 we see that Coloured and Asian females have a significantly lower risk of engaging in sexual debut compared to Black adolescents with $\frac{h(t,\text{Coloured})}{h(t,\text{Black})} = 0.727$ (p-value $= 0.039$) and $\frac{h(t,\text{Asian})}{h(t,\text{Black})} = 0.199$ (p-value $= 0.005$). White female adolescents had a significantly higher risk of engaging in sexual debut compared to their Asian counterparts. The hazard ratio is calculated as

$$\frac{h(t,\text{White})}{h(t,\text{Asian})} = \frac{0.592\ h(t,\text{Black})}{0.199\ h(t,\text{Black})} = 2.975 \tag{4.1}$$

Table 4.2 Cox proportional hazards regression model results for females and males

		Female		Male	
		HR (95% CI)	p-value	HR (95% CI)	p-value
Race	Black (Baseline)				
	White	0.592 (0.294; 1.192)	0.142	0.388 (0.201; 0.750)	0.005**
	Coloured	0.725 (0.537; 0.984)	0.039**	0.542 (0.407; 0.721)	0.000**
	Asian	0.199 (0.064; 0.620)	0.005**	0.183 (0.068; 0.490)	0.001**
Maternal education	No formal/primary education (Baseline)				
	Secondary education	0.877 (0.662; 1.162)	0.361	1.120 (0.855; 1.468)	0.410
	Post-school training	0.725 (0.472; 1.113)	0.141	0.893 (0.607; 1.315)	0.567
Socioeconomic status	Low (Baseline)				
	Middle	1.044 (0.838; 1.300)	0.704	1.176 (0.962; 1.436)	0.113
	High	0.802 (0.625; 1.029)	0.082*	0.980 (0.797; 1.204)	0.844
Height	Stunted (Baseline)				
	Normal	1.279 (1.021; 1.603)	0.033**	1.256 (1.053; 1.498)	0.011**
	Tall	.	.	.	.
Pubertal status	Prepubertal (Baseline)				
	Early pubertal	1.296 (0.793; 2.120)	0.301	1.012 (0.807; 1.269)	0.917
	Late pubertal	1.844 (1.111; 3.061)	0.018**	1.223 (0.865; 1.730)	0.254
Foreplay	Did not engage (Baseline)				
	Engaged	6.377 (4.494; 9.050)	0.000**	6.220 (4.474; 8.647)	0.000**
Oral sex	Did not engage (Baseline)				
	Engaged	3.847 (3.181; 4.651)	0.000**	3.242 (2.743; 3.832)	0.000**
Religiosity	Not at all (Baseline)				
	Somewhat	1.291 (0.687; 2.428)	0.427	0.918 (0.656; 1.283)	0.615
	Very	1.082 (0.592; 1.978)	0.798	0.808 (0.602; 1.085)	0.157

with 95% confidence interval $(1.923; 4.594)$.

This means that White females are almost three times more likely to engage in sexual debut compared to Asian females. The hazard ratio for Coloured female adolescents compared to Asian female adolescents is calculated as

$$\frac{h(t, \text{Coloured})}{h(t, \text{Asian})} = \frac{0.727\ h(t, \text{Black})}{0.199\ h(t, \text{Black})} = 3.653 \tag{4.2}$$

with 95% confidence interval $(1.587; 8.391)$.

Therefore, Coloured females have 3.653 times the risk of engaging in sexual debut compared to Asian females.

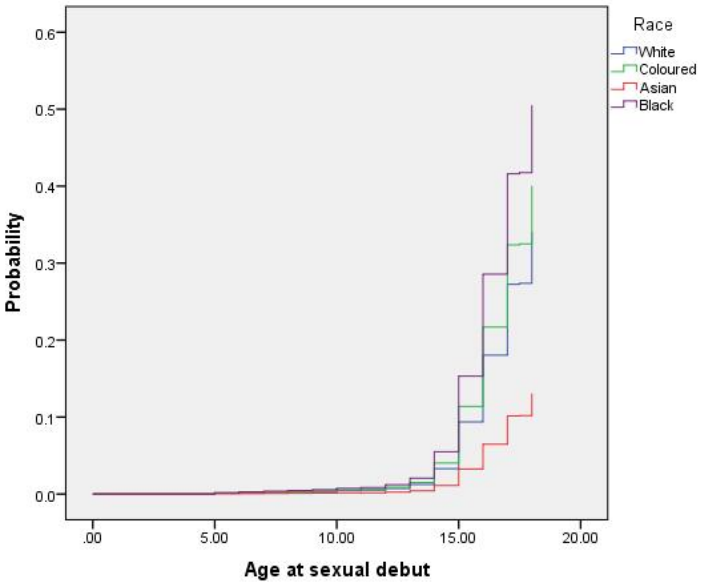

Figure 4.1: Cumulative incidence for race in female adolescents

These results are shown graphically for females in Figure 4.1 and in Figure 4.2 for males. For females, Black adolescents have a higher risk of engaging in sexual debut compared to Coloured females and Asian females, however, White adolescents do not have a significantly different risk of engaging in sexual debut compared to Black and Coloured adolescents. Asian females have the lowest risk of engaging in sexual debut compared to females in any of the other race groups. In males we see a similar risk grading to females based on race, however, here we also see that White male adolescents have a significantly lower risk of engaging in sexual debut compared to Black male adolescents (HR $= 0.388$, p-value $= 0.005$). Both White and Coloured males have a higher risk of

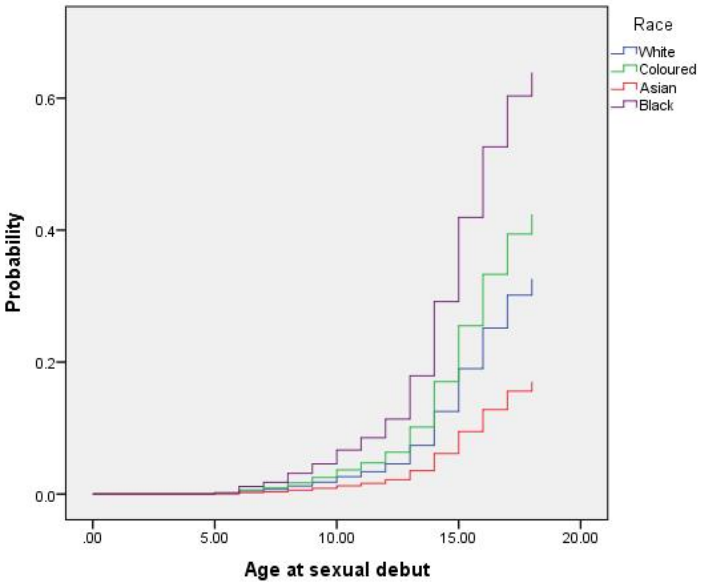

Figure 4.2: Cumulative incidence for race in male adolescents

engaging in sexual debut compared to Asian males. The relevant hazard ratios are calculated as

$$\frac{h(t, \text{White})}{h(t, \text{Asian})} = \frac{0.388\ h(t, \text{Black})}{0.183\ h(t, \text{Black})} = 2.120 \tag{4.3}$$

with 95% confidence interval $(1.531; 2.956)$ and

$$\frac{h(t, \text{Coloured})}{h(t, \text{Asian})} = \frac{0.542\ h(t, \text{Black})}{0.183\ h(t, \text{Black})} = 2.962 \tag{4.4}$$

with 95% confidence interval $(1.471; 5.985)$.

Black male adolescents had the highest risk of engaging in sexual debut followed by Coloured and White males who did not have a significantly different risk of engaging in sexual debut. Lastly, Asian males had the lowest risk of engaging in sexual debut (Figure 4.2).

We see significant differences in the risk to sexual debut between those adolescents whose mothers had secondary education and those whose mothers had

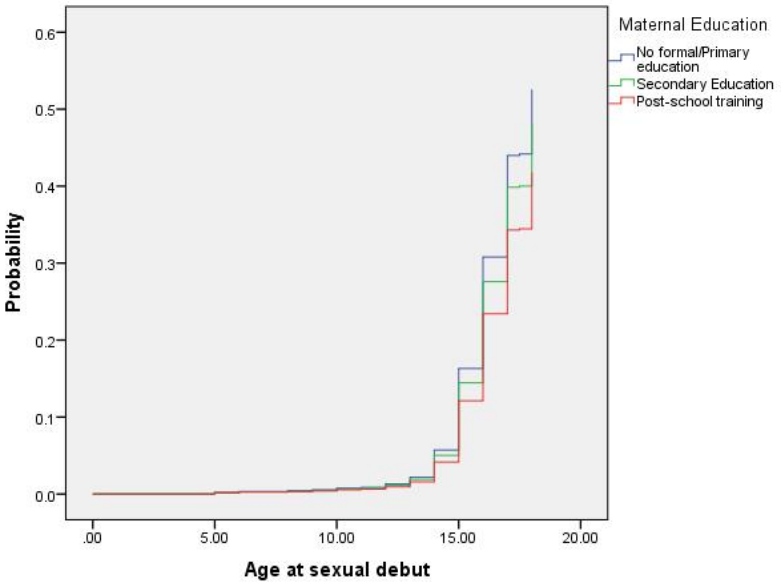

Figure 4.3: Cumulative incidence for maternal education in female adolescents

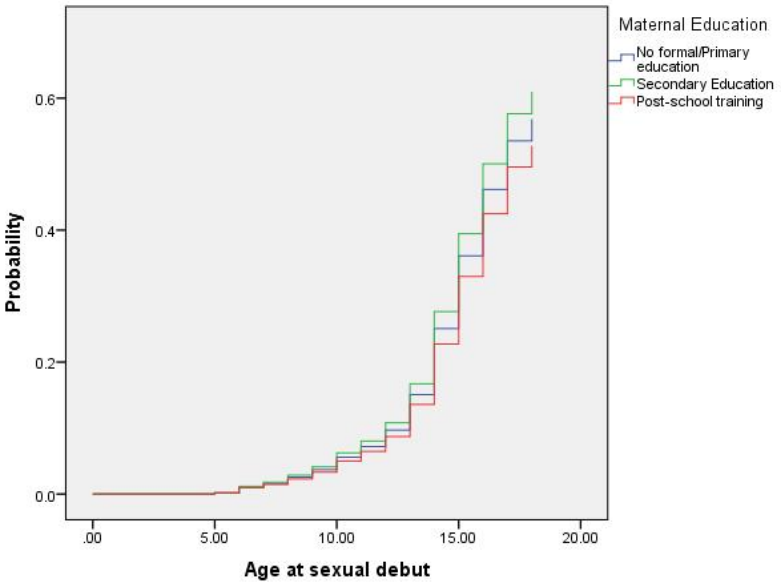

Figure 4.4: Cumulative incidence for maternal education in male adolescents

post-school training for both females and males. The hazard ratio for females with secondary level maternal education compared to those females with post-

school training maternal education is calculated as

$$\frac{h(t, \text{Secondary education})}{h(t, \text{Post-school training})} = \frac{0.877\ h(t, \text{No formal/primary})}{0.725\ h(t, \text{No formal/primary})} = 1.210 \qquad (4.5)$$

with 95% confidence interval $(1.044; 1.403)$.

The hazard ratio for males with secondary level maternal education compared to those males with post-school training maternal education is calculated as

$$\frac{h(t, \text{Secondary education})}{h(t, \text{Post-school training})} = \frac{1.120\ h(t, \text{No formal/primary})}{0.893\ h(t, \text{No formal/primary})} = 1.254 \qquad (4.6)$$

with 95% confidence interval $(1.116; 1.409)$.

Figure 4.3 and Figure 4.4 show these results for females and males respectively. Females and males with mothers who have post-school training have a lower risk of engaging in sexual debut compared to those with mothers who have secondary education. Note that the risk of those females and males with mothers who have no formal/primary school education was not found to be significantly different from either of the other two categories.

It was found that for both females and males, participants with a high socioeconomic status had a lower hazard of engaging in sexual debut compared to participants with a middle level socioeconomic status. The relevant hazard ratio for females is calculated as

$$\frac{h(t, \text{High})}{h(t, \text{Middle})} = \frac{0.802\ h(t, \text{Low})}{1.044\ h(t, \text{Low})} = 0.768 \qquad (4.7)$$

with 95% confidence interval $(0.746; 0.792)$.

The relevant hazard ratio for males is calculated as

$$\frac{h(t, \text{High})}{h(t, \text{Middle})} = \frac{0.980\ h(t, \text{Low})}{1.176\ h(t, \text{Low})} = 0.834 \qquad (4.8)$$

with 95% confidence interval $(0.829; 0.838)$.

Additionally, for female adolescents it was found that those participants with a high socioeconomic status had a significantly lower hazard of engaging in sexual debut compared to those participants with a low socioeconomic status (HR $= 1.247$, p-value $= 0.082$). Thus, for females, a high socioeconomic status acted as a protective factor against engaging in sexual debut. The results are shown in Figure 4.5 and Figure 4.6 for females and males respectively. Furthermore, adolescents with a middle socioeconomic status do not have a significantly different hazard of engaging in sexual debut compared to adolescents with a low socioeconomic status for both females and males.

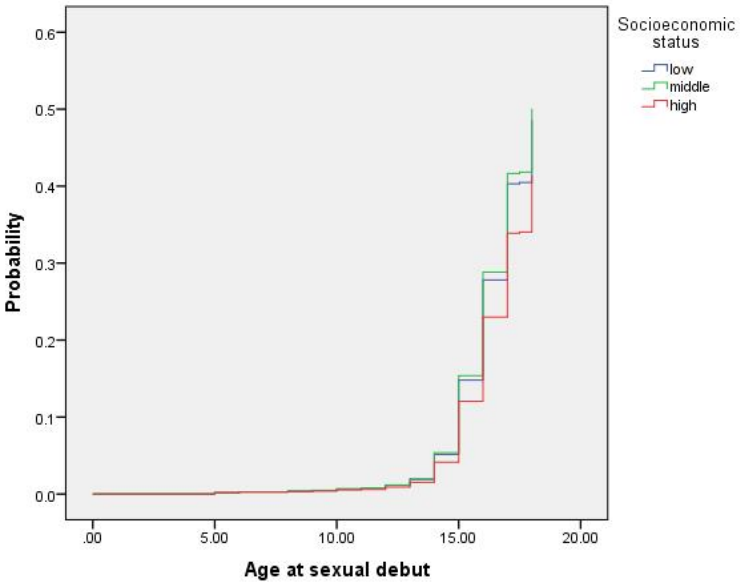

Figure 4.5: Cumulative incidence for socioeconomic status in female adolescents

Height had been divided into three categories; namely, stunted, normal and tall for age. A frequency count of tall adolescents showed that there were only four tall females and four tall males. Therefore, we cannot readily interpret hazard ratios for tall adolescents due to insufficient data. We will thus investigate differences in stunted and normal height females and males on the time to sexual

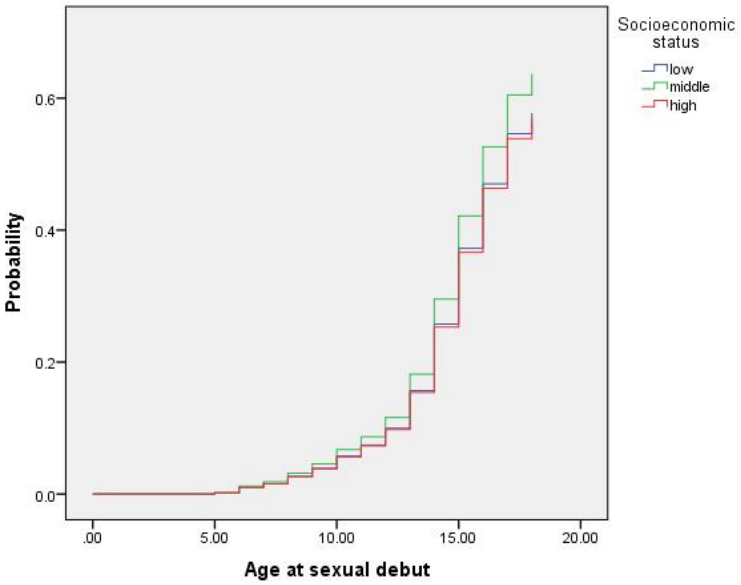

Figure 4.6: Cumulative incidence for socioeconomic status in male adolescents

debut. Normal height females had a significantly higher hazard of engaging in sexual debut compared to stunted females (HR = 1.279, p-value $= 0.033$). Additionally, normal height males had a significantly higher hazard of engaging in sexual debut compared to stunted males (HR = 1.256, p-value $= 0.011$). The cumulative incidence curves show these results in Figure 4.7 and Figure 4.8 respectively.

Next we consider whether the pubertal status of the adolescents had an affect on the hazard of sexual debut. Figure 4.9 shows that females who were in the late pubertal stage of development had a higher hazard of engaging in sexual debut compared to prepubertal females (HR $= 1.844$, p-value$= 0.018$). Additionally, females in the late pubertal stage also had a significantly higher risk of engaging in sexual debut compared to females in the early pubertal stage and the relevant hazard ratio is calculated as

$$\frac{h(t, \text{Late pubertal})}{h(t, \text{Early pubertal})} = \frac{1.844\ h(t, \text{Prepubertal})}{1.296\ h(t, \text{Prepubertal})} = 1.423 \tag{4.9}$$

with 95% confidence interval $(1.401; 1.443)$.

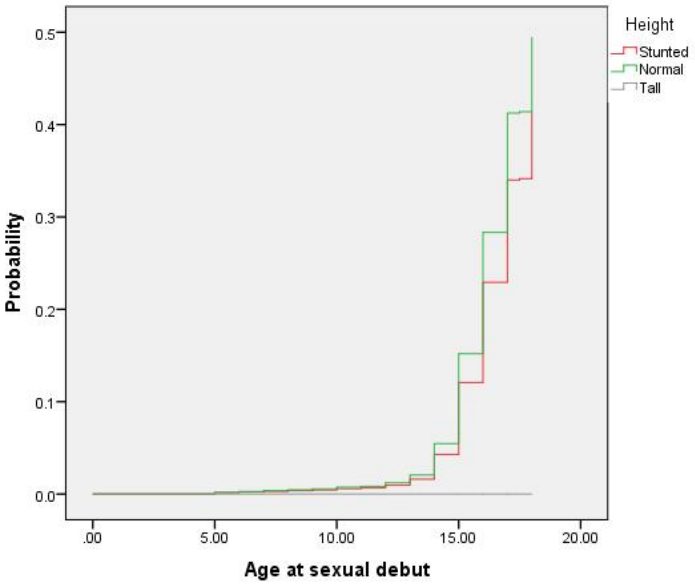

Figure 4.7: Cumulative incidence for height in female adolescents

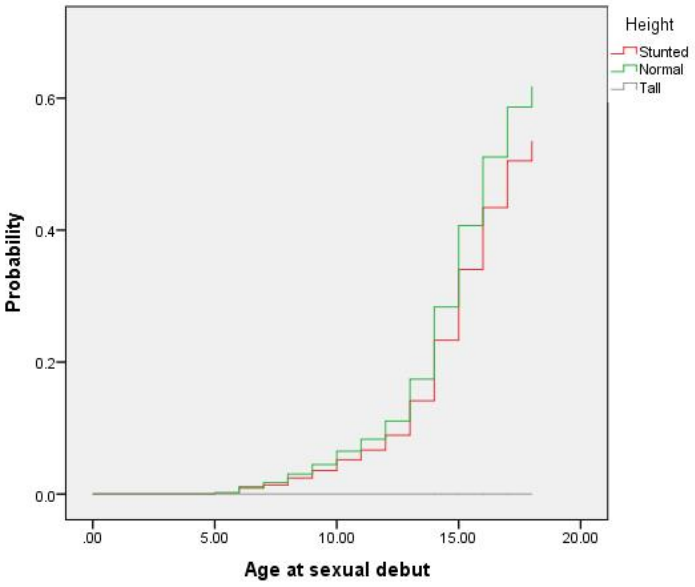

Figure 4.8: Cumulative incidence for height in male adolescents

Thus, females in the late pubertal stage of development have the highest risk

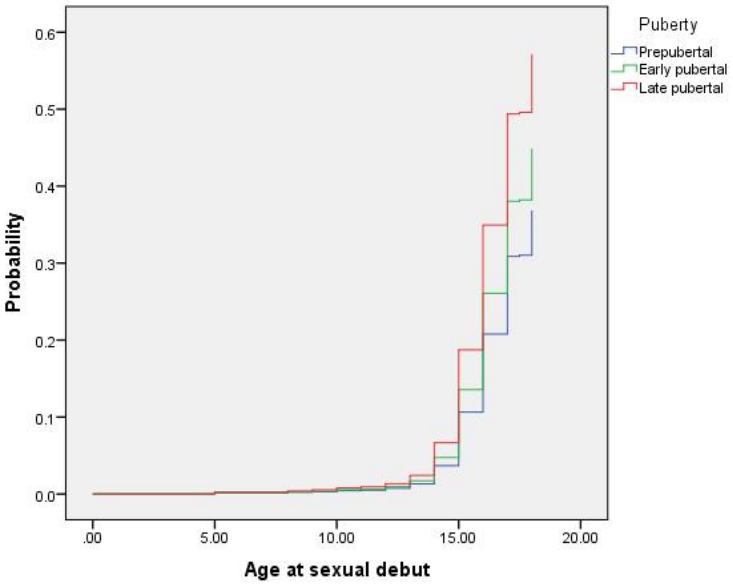

Figure 4.9: Cumulative incidence for pubertal status in female adolescents

of engaging in sexual debut.

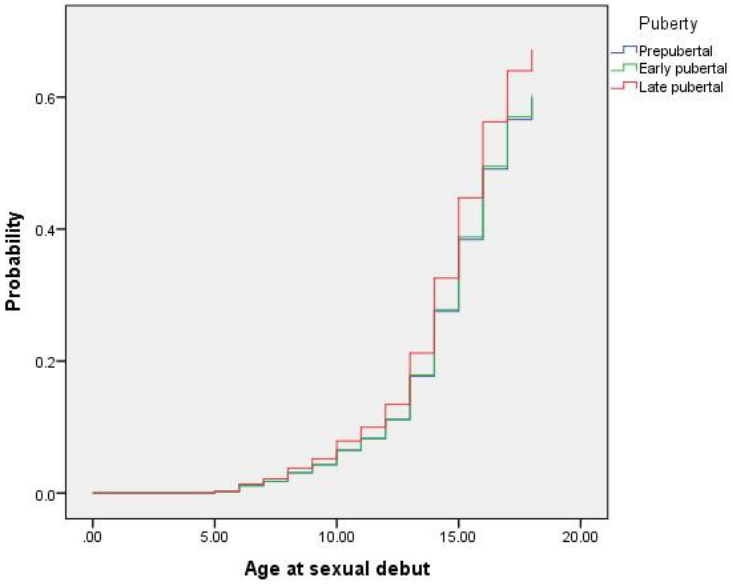

Figure 4.10: Cumulative incidence for pubertal status in male adolescents

In males, we note that adolescents in the late pubertal development stage have a significantly higher hazard of engaging in sexual debut compared to adolescents in the early pubertal stage of development. The hazard ratio is calculated as

$$\frac{h(t, \text{Late pubertal})}{h(t, \text{Early pubertal})} = \frac{1.223\ h(t, \text{Prepubertal})}{1.012\ h(t, \text{Prepubertal})} = 1.209 \tag{4.10}$$

with 95% confidence interval $(1.072; 1.362)$.

This result is shown graphically in Figure 4.10. Note that prepubertal males were not found to have a significantly different hazard of engaging in sexual debut compared to males in either of the other two pubertal stages.

Figure 4.11 shows that females who engaged in foreplay had a significantly higher hazard of engaging in sexual debut compared to those who did not engage in foreplay (HR = 6.377, p-value = 0.000).

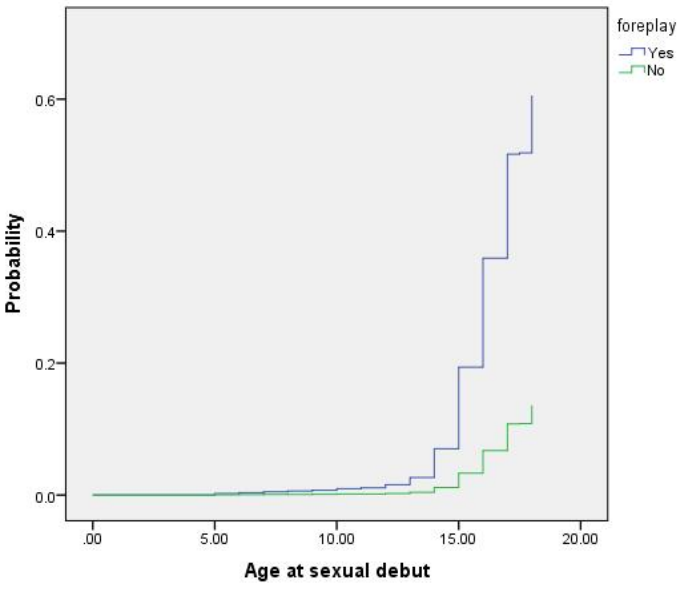

Figure 4.11: Cumulative incidence for foreplay in female adolescents

Similarly, males who engaged in foreplay had a significantly higher risk of en-

gaging in sexual debut compared to males who who did not engage in foreplay (HR = 6.220, p-value = 0.000). The cumulative incidence functions for the role of foreplay on sexual debut in males is shown in Figure 4.12.

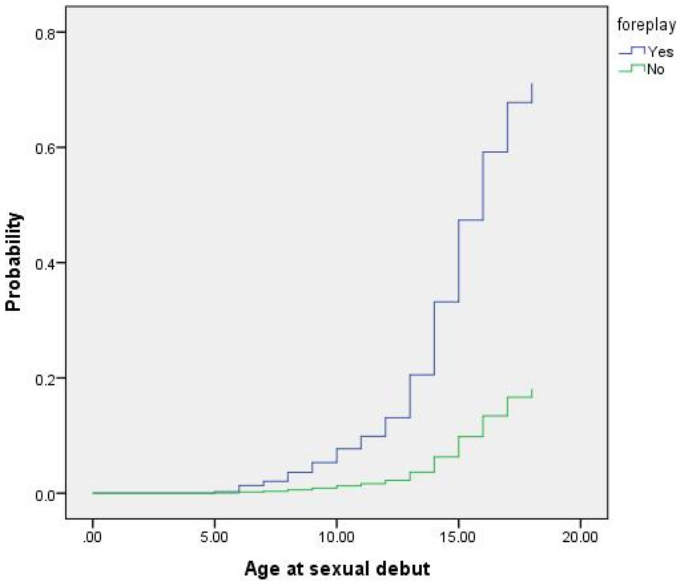

Figure 4.12: Cumulative incidence for foreplay in male adolescents

Both females and males who engaged in oral sex had a higher hazard of engaging in sexual debut. Females who engaged in oral sex were 3.847 (p-value $= 0.000$) times likely to engage in sexual debut compared to females who did not engage in oral sex and males who engaged in oral sex were 3.242 times likely to engage in sexual debut compared to males who did not engage in oral sex. Figure 4.13 shows these results for females and Figure 4.14 shows the results for males.

The level of religiosity of the adolescents were measured and categorized as not at all, somewhat or very religious. However, as mentioned previously, exploratory analysis of the data revealed 24% missing data for religiosity. Albeit this concern, religiosity was included in the analysis due to the significance it possibly held to the time to sexual debut according to the literature. Therefore,

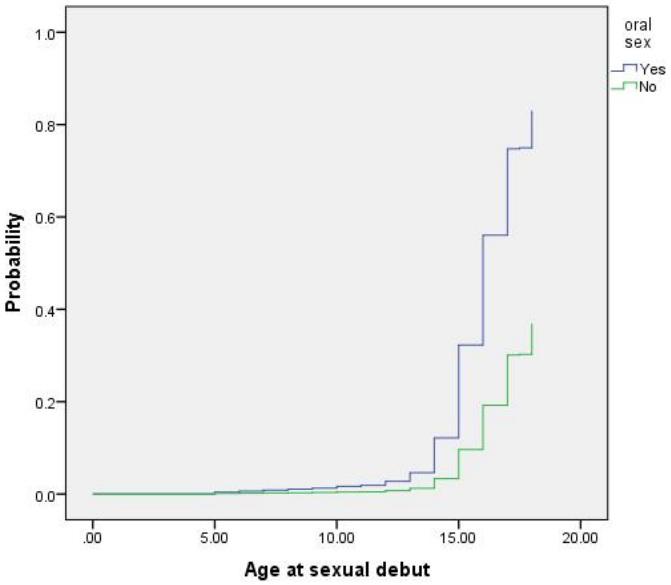

Figure 4.13: Cumulative incidence for oral sex in female adolescents

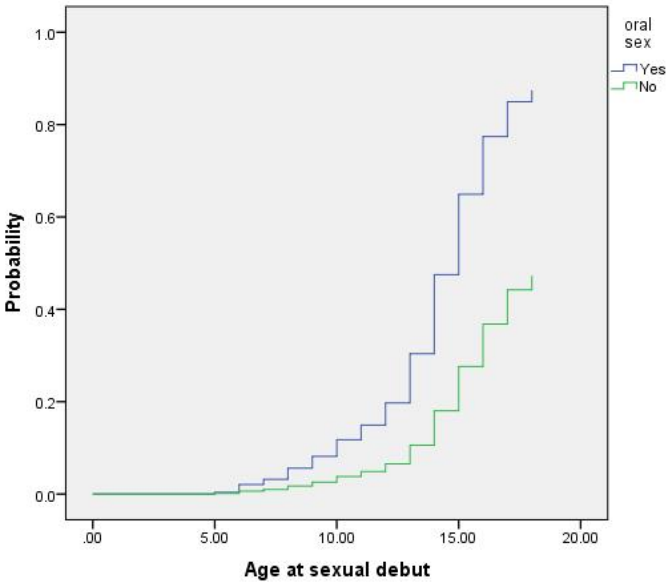

Figure 4.14: Cumulative incidence for oral sex in male adolescents

we proceed to analyse the role of religiosity on time to sexual debut in the Birth to Twenty cohort but do so with caution. We note that both females and males

who were reportedly very religious had a significantly lower hazard of engaging in sexual debut compared to their counterparts who were somewhat religious. The relevant hazard ratio for females is calculated as

$$\frac{h(t, \text{Very})}{h(t, \text{Somewhat})} = \frac{0.728\ h(t, \text{Not at all})}{0.865\ h(t, \text{Not at all})} = 0.842 \tag{4.11}$$

with 95% confidence interval $(0.819; 0.864)$.

For males, the relevant hazard ratio is calculated as

$$\frac{h(t, \text{Very})}{h(t, \text{Somewhat})} = \frac{0.808\ h(t, \text{Not at all})}{0.918\ h(t, \text{Not at all})} = 0.880 \tag{4.12}$$

with 95% confidence interval $(0.846; 0.918)$.

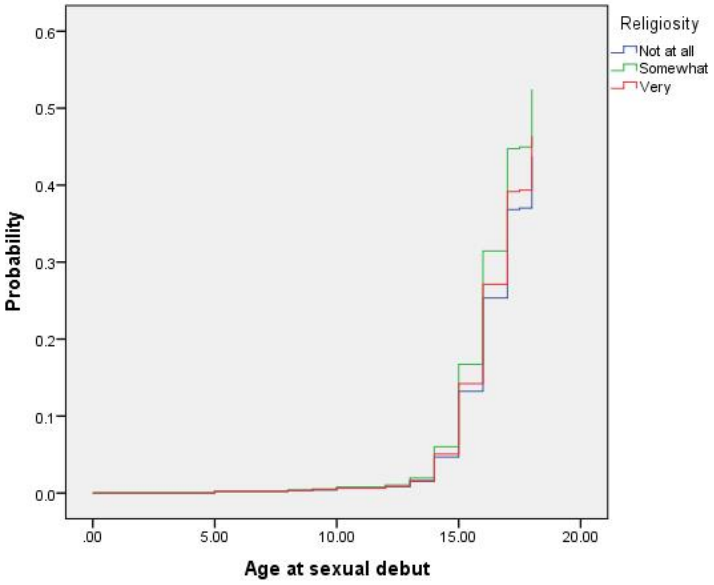

Figure 4.15: Cumulative incidence for religiosity in female adolescents

The results for females can be seen graphically in Figure 4.15 and the results for males can be seen in Figure 4.16. Across both categories of gender we note that those participants who were not at all religious were not found to have a significantly higher or lower risk of engaging in sexual debut compared to the

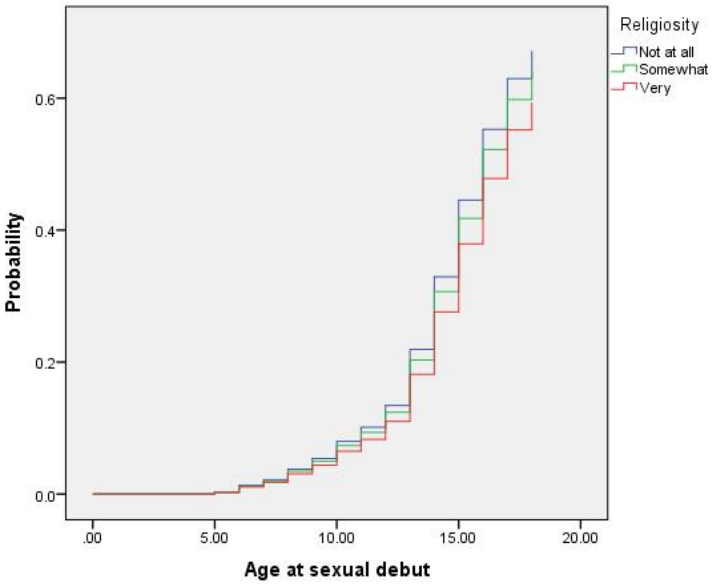

Figure 4.16: Cumulative incidence for religiosity in male adolescents

other categories of religiosity.

Finally, a plot of the $\log(-\log\ S(t))$ versus the survival time was checked for all significant variables to assess whether the proportional hazards assumption was maintained. Intersection of the curves are indicative of a violation of the proportional hazards assumption. If the proportional hazards assumption is maintained then we expect to see parallel curves across the categories of the relevant covariate. Figure B.1 to Figure B.16 in Appendix B.1 show these results. Note that the curves did not show any cases of a violation in the proportional hazards assumption.

A plot of the Cox-Snell residuals versus the cumulative hazard of the residuals was investigated for each of the significant variables to assess the fit for each model. If the models are fitted adequately, we expect the plot to resemble that of a unit exponential distribution, that is, the fitted model is adequate if the residual plot is a straight line through the origin with a slope of 1. Figure B.17 to Figure B.32 in Appendix B.2 show these results. For ease of visual inspection,

a plot of a Unit Exponential curve has been imposed on the residual plots so as to serve as a reference line. Note that all the curves are substantially close to the unit distribution curve. There are only two small deviations and these are noted where the curves are presented in Appendix B.2. These deviations do not appear to be substantial, thus the curves show no cause for concern in modeling the Birth to Twenty data using the Cox proportional hazards model.

4.1.1 Conclusion

Asian adolescents were found to have the lowest risk of engaging in sexual debut. This result was common across both strata of gender. For females, Black adolescents demonstrated a higher risk of engagement in sexual debut relative to Coloured and Asian adolescents while White adolescents did not show any evidence of differing hazards relative to Black and Coloured adolescents. In males, Black adolescents had the highest hazard of engagement in sexual debut, followed by Coloured and White males who did not have a significantly different hazard of engaging in sexual debut. A post-school training level of maternal education was associated with lower levels of sexual debut for all adolescents. A high socioeconomic status acted as a protective factor for female adolescents in delaying first sex. For males, there was less of a significant distinction in the association between the risk of sexual debut and the socioeconomic status. Males with a high socioeconomic status were found to be at a significantly lower risk of engaging in sexual debut in comparison to males with a middle level socioeconomic. Normal height females and males were associated with a higher hazard of early sexual debut relative to those who were classified as stunted for age, however, tall adolescents could not be included in the analysis due to insufficient data. For females, those in the late pubertal stage were associated with the highest risk of engaging in sexual debut whereas for males it was found that those in the late pubertal stage had a higher risk of engaging in sexual debut compared to those in the early pubertal stage, but not those in the prepubertal stage of development. Across both strata of gender, engagement in foreplay and oral sex acted as a significant risk factor for early sexual debut.

Graphical checks include $\log(-\log\ S(t))$ versus the survival time plots which showed no violation in the proportional hazards assumption. Additionally, Cox-Snell residual plots confirmed an adequate fit of the Cox proportional hazards model to the Birth to Twenty data.

4.2 Competing risks regression model

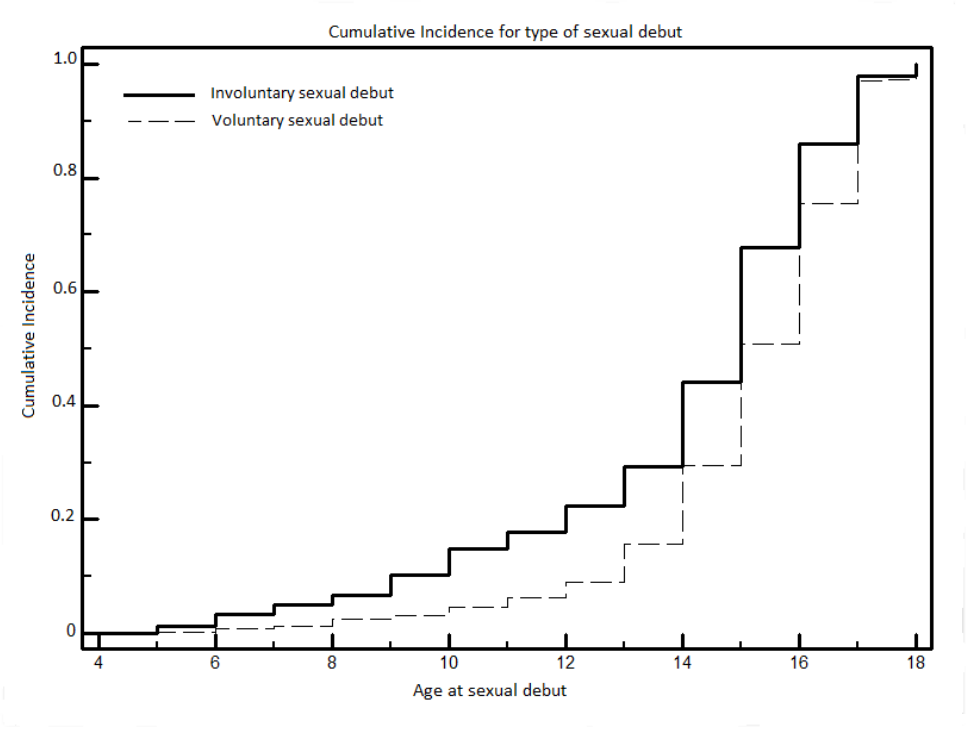

Figure 4.17: Cumulative incidence for type of sexual debut

Sexual debut can occur either voluntarily or involuntarily. We wish to determine whether the type of sexual debut is associated with the time to sexual debut, in which case, sexual debut will need to be modeled by type of sexual debut. Figure 4.17 shows the cumulative incidence for respondents who reported voluntary sexual debut and those who reported involuntary sexual debut. The curves show that those adolescents who had been coerced into sexual debut had reached the event of interest earlier compared to those who had engaged in voluntary sexual debut. These differing curves tell us that adolescents from the

two groups may have different risk factors which affect their survival curves. A log-rank test was employed to test whether there was a significant difference in the cumulative incidence for voluntary sexual debut versus the cumulative incidence for involuntary sexual debut. The test showed a p-value less than 0.0001. This provided sufficient evidence to conclude that the cumulative incidence of voluntary sexual debut and the cumulative incidence of involuntary sexual debut differ. The use of a competing risks regression model where voluntary sexual debut and involuntary sexual debut are competing events is thus justified.

The competing risks regression model was run separately for females and males. Race, maternal education, socioeconomic status, religiosity, height, pubertal status, foreplay and oral sex were individually modeled with type of sexual debut as the dependent variable due to the strong associations between the covariates. Table 4.3 presents the results of the competing risks regression model for females. The table gives the hazard ratios along with the 95% confidence intervals and the p-values for each of the covariates relative to the associated baseline category. Note that "$*$" was used to show covariates that are significant at the 0.10 level of significance and "$**$" was used to show covariates that are significant at the 0.05 level of significance. Only significant covariates are further investigated by examining the relevant graphical output. Additionally, to assess the validity of the proportional subhazards assumption, a formal test is conducted to detect whether time-varying covariates are significant in the model. We expect that time interaction variables are not significant in the model. If it is significant at the 0.05 level then the proportional subhazards assumption will be violated and if that is the case then the competing risks model must be re-interpreted taking the time-varying effect of the covariate into account. Testing the proportional subhazards assumption in each of the 16 models for females rendered no violations.

The model including race is the first to be assessed and tested. Race has four categories. In Table 4.3, White, Coloured and Asian females are compared to Black females which is the baseline category. Voluntary sexual debut dif-

Table 4.3 Competing risks regression model results for female adolescents

		Voluntary		Involuntary	
		HR (95% CI)	p-value	HR (95% CI)	p-value
Race	Black (Baseline)				
	White	0.591 (0.268; 1.305)	0.193	0.661 (0.165; 2.639)	0.557
	Coloured	0.748 (0.544; 1.030)	0.075*	0.721 (0.371; 1.400)	0.334
	Asian	0.182 (0.045; 0.728)	0.016**	0.341 (0.049; 2.378)	0.278
Maternal education	No formal/primary education (Baseline)				
	Secondary education	1.010 (0.732; 1.392)	0.954	0.659 (0.390; 1.112)	0.118
	Post-school training	0.928 (0.585; 1.472)	0.751	0.386 (0.143; 1.038)	0.059*
Socioeconomic status	Low (Baseline)				
	Middle	1.026 (0.807; 1.303)	0.835	1.037 (0.668; 1.607)	0.873
	High	0.921 (0.708; 1.197)	0.536	0.520 (0.289; 0.934)	0.029**
Height	Stunted (Baseline)				
	Normal	1.703 (1.309; 2.215)	0.000**	0.618 (0.413; 0.923)	0.019**
	Tall	.	.	.	.
Pubertal status	Prepubertal (Baseline)				
	Early pubertal	1.210 (0.731; 2.002)	0.458	1.517 (0.479; 4.802)	0.478
	Late pubertal	1.729 (1.028; 2.908)	0.039**	1.787 (0.544; 5.866)	0.338
Foreplay	Did not engage (Baseline)				
	Engaged	8.475 (5.405; 13.333)	0.000**	2.639 (1.529; 4.545)	0.000**
Oral sex	Did not engage (Baseline)				
	Engaged	3.846 (3.155; 4.695)	0.000**	2.141 (1.431; 3.208)	0.000**
Religiosity	Not at all (Baseline)				
	Somewhat	1.338 (0.712; 2.516)	0.366	0.968 (0.216; 4.344)	0.967
	Very	1.055 (0.579; 1.924)	0.861	1.147 (0.279; 4.726)	0.849

fers for Black and Coloured female adolescents at the 0.10 level of significance (HR $= 1.337$, p-value $= 0.075$) and voluntary sexual debut differs for Black and Asian adolescents at the 0.05 level of significance (HR $= 5.495$, p-value $= 0.016$). A significant difference in voluntary sexual debut was also detected between Coloured and Asian female adolescents. The hazard ratio is calculated as

$$\frac{h(t, \text{Coloured})}{h(t, \text{Asian})} = \frac{0.748\ h(t, \text{Black})}{0.182\ h(t, \text{Black})} = 4.110 \tag{4.13}$$

with 95% confidence interval $(1.415; 12.089)$.

Therefore, Coloured females have more than four times the risk of engaging in voluntary sexual debut compared to Asian females.

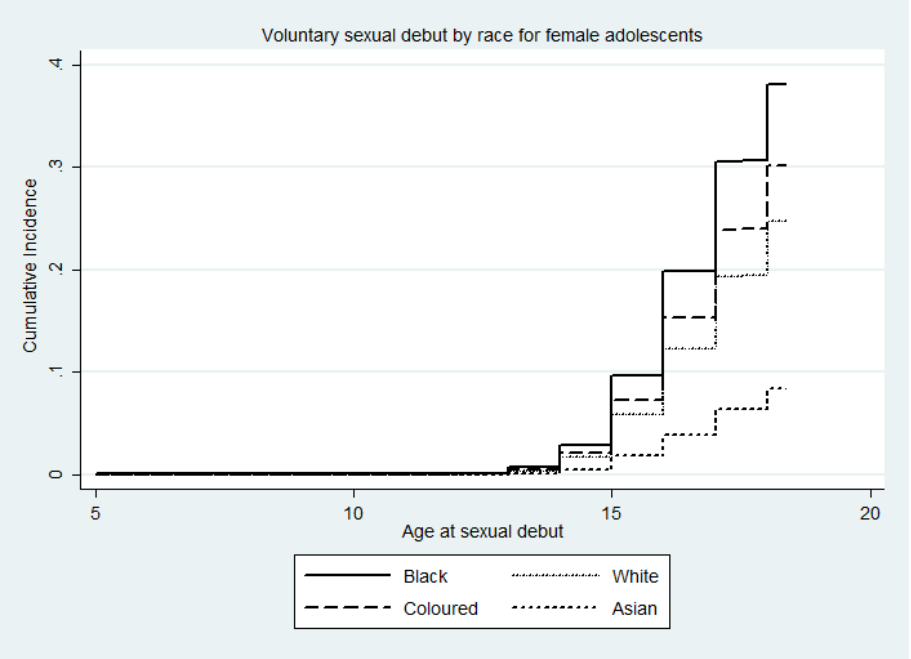

Figure 4.18: Cumulative incidence for voluntary sexual debut by race for female adolescents

The results for females are shown in Figure 4.18. Black female adolescents have a higher risk of engaging in sexual debut compared to Coloured females and Asian females. Asian females have a lower risk of engaging in voluntary sexual debut compared to Black and Coloured females. Due to these significant differences detected among race, race is said to be a significant variable in affecting the hazard of voluntary sexual debut in females. Thus, the next step is to test whether these effects of race interpreted above comply with the underlying assumption of the model, that is, we test whether the subhazards are proportional.

Table 4.4 shows the results when we include time interaction variables in the model. We note that the p-values are greater than 0.05, therefore there is no

Table 4.4 Test of proportional subhazards assumption for race in voluntary sexual debut for females

	Variable interacted with time	Voluntary (p-value)
Race	Black (Baseline)	
	White	0.790
	Coloured	0.260
	Asian	0.824

evidence to suggest a violation in the proportional subhazards assumption.

Race did not play a significant role in time to involuntary sexual debut and thus the relative hazards are not further investigated.

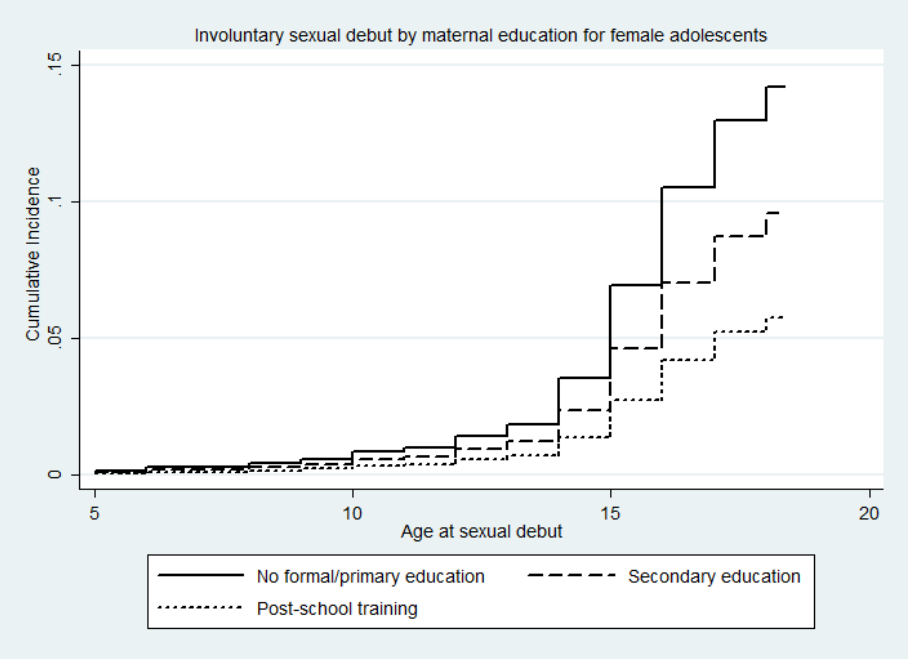

Figure 4.19: Cumulative incidence for involuntary sexual debut by maternal education for female adolescents

Also, the level of maternal education for female adolescents did not play a significant role in affecting the risk of voluntary sexual debut. However, females who had mothers with post-school training were at a significantly lower risk of being coerced into sexual debut as compared to those individuals who had mothers with no formal/primary school education (HR $= 0.386$, p-value $= 0.059$) and those who had mothers with secondary level maternal education. The hazard

ratio of the latter comparison is given by

$$\frac{h(t, \text{Secondary education})}{h(t, \text{Post-school training})} = \frac{0.659\ h(t, \text{No formal/primary})}{0.386\ h(t, \text{No formal/primary})} = 1.707 \qquad (4.14)$$

with 95% confidence interval $(1.071; 2.727)$.

Figure 4.19 shows that post-school maternal education acted as a protective factor against coerced sexual debut.

Table 4.5 Test of proportional subhazards assumption for maternal education in involuntary sexual debut for females

	Variable interacted with time	Involuntary (p-value)
Maternal education	No formal/primary (Baseline)	
	Secondary	0.878
	Post-school	0.426

According to the p-values found in Table 4.5, there is no evidence to suggest a time-varying effect of race on the subhazard functions of involuntary sexual debut. Thus, the proportional subhazards competing risks model assumption is not violated.

Socioeconomic status did not show any significant influence on voluntary sexual debut in females. However, socioeconomic status was found to affect involuntary sexual debut. Female adolescents with a high socioeconomic status were significantly less likely to have been coerced into sexual debut compared to female adolescents with a low socioeconomic status (HR $= 0.520$, p-value $= 0.029$). Additionally, significant differences were also detected between females with a middle socioeconomic status and those with a high socioeconomic status. The relevant hazard ratio is given as

$$\frac{h(t, \text{Middle})}{h(t, \text{High})} = \frac{1.037\ h(t, \text{Low})}{0.520\ h(t, \text{Low})} = 1.994 \qquad (4.15)$$

with 95% confidence interval $(1.721; 2.311)$.

Thus, females with a middle level socioeconomic status were almost twice as likely to be coerced into sexual debut as compared to females who had a high socioeconomic status.

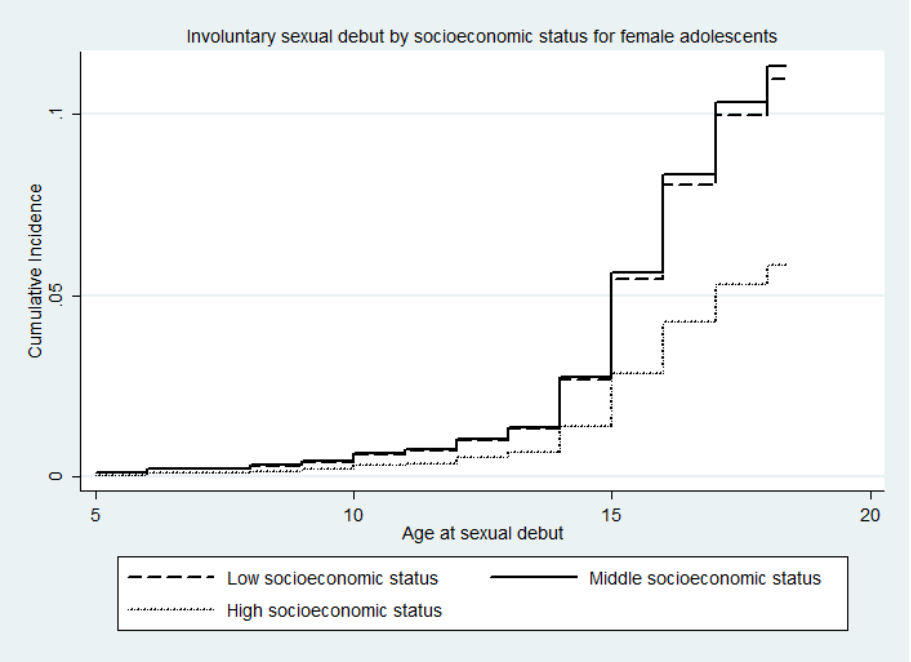

Figure 4.20: Cumulative incidence for involuntary sexual debut by socioeconomic status for female adolescents

Figure 4.20 shows graphically that females who come from families with low and middle socioeconomic statuses are at a significantly higher risk of being coerced than females who come from families with a high socioeconomic status and thus a high socioeconomic status acts as a protective factor against coerced first sex.

Table 4.6 Test of proportional subhazards assumption for socioeconomic status in involuntary sexual debut for females

	Variable interacted with time	Involuntary (p-value)
Socioeconomic status	Low (Baseline)	
	Middle	0.163
	High	0.422

Table 4.20 shows the results when time interaction variables are included in the model. Since the p-values are less than 0.05 there is no evidence to suggest

that the proportional subhazards assumption is violated.

Height plays a significant role in both voluntary and involuntary sexual debut in female adolescents. Females with normal height had a significantly higher risk of engaging in voluntary sexual debut compared to those with a stunted height (HR $= 1.703$, p-value $= 0.000$). The opposite is the case for involuntary sexual debut where females with stunted height had a significantly higher risk compared to those with normal height (HR $= 1.618$, p-value $= 0.019$). Hazard ratios involving tall females could not be accurately calculated from the existing data since there were only four $(0.348\%$ of all females) tall females and of those tall females none had reported voluntary or involuntary sexual debut. Figure 4.21 and Figure 4.22 show these results graphically. Normal height females are more likely to engage in voluntary sexual debut relative to stunted height females whereas the opposite is true for coerced first sex.

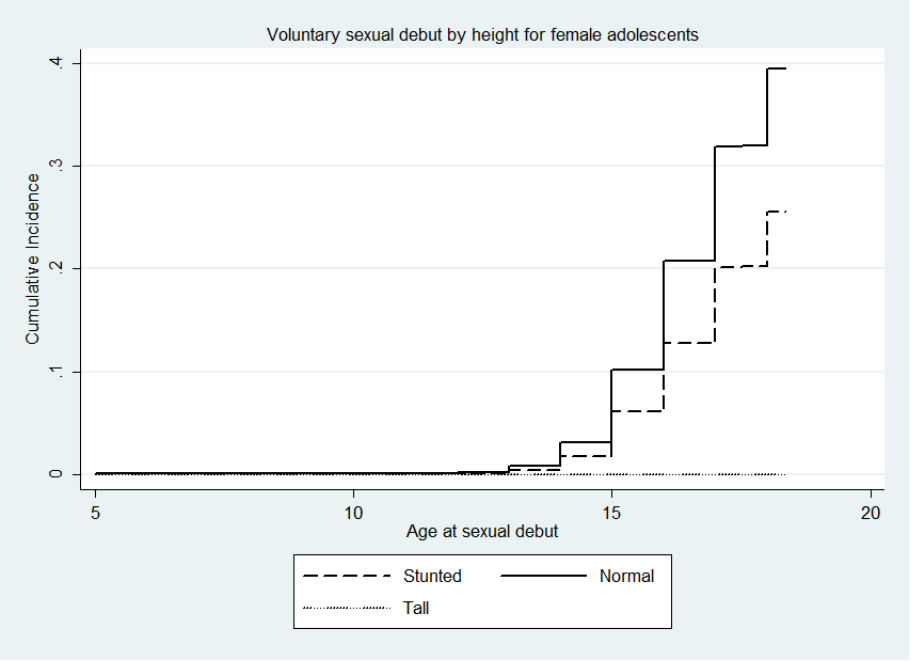

Figure 4.21: Cumulative incidence for voluntary sexual debut by height for female adolescents

Table 4.7 gives the results of the proportional subhazards assumption test for both voluntary and involuntary sexual debut. The p-values for variables in-

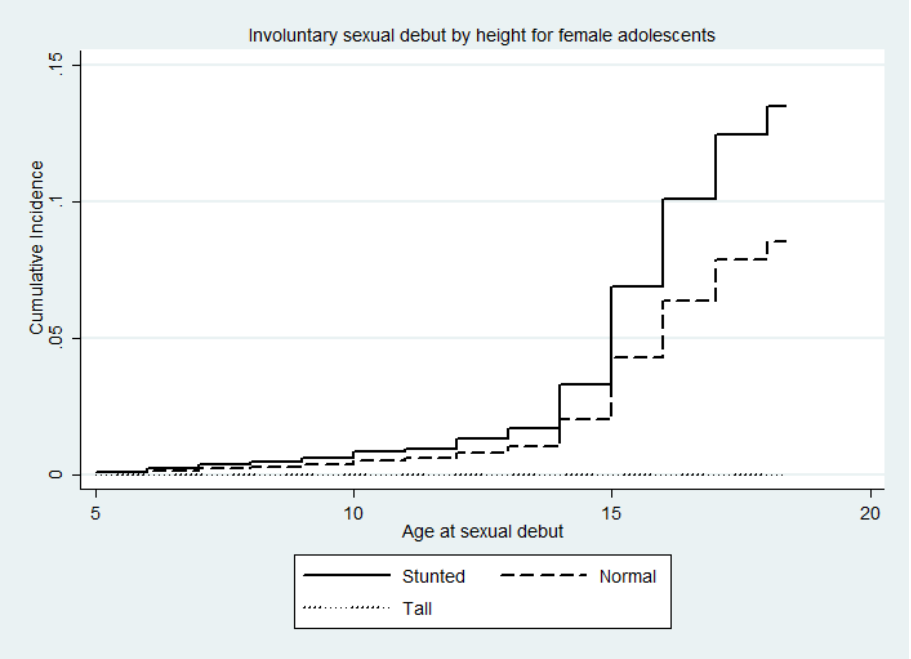

Figure 4.22: Cumulative incidence for involuntary sexual debut by height for female adolescents

Table 4.7 Test of proportional subhazards assumption for height in voluntary and involuntary sexual debut for females

	Variable interacted with time	Voluntary (p-value)	Involuntary (p-value)
Height	Stunted (Baseline)		
	Normal	0.127	0.785
	Tall	.	.

teracted with time are larger than 0.05 thus indicating that the proportional subhazards assumption is not violated for height in voluntary and involuntary sexual debut. We cannot comment on the proportionality of the hazard of tall females relative to the other height categories due to insufficient data.

Females who were at a late pubertal development stage showed a higher risk of engaging in voluntary sexual debut compared to those who were in the prepubertal stage of development (HR $= 1.729$, p-value $= 0.039$) and those who were in the early pubertal stage. The latter hazard ratio is given by

$$\frac{h(t, \text{Late})}{h(t, \text{Early})} = \frac{1.729\ h(t, \text{Prepubertal})}{1.210\ h(t, \text{Prepubertal})} = 1.429 \tag{4.16}$$

with 95% confidence interval $(1.406; 1.453)$.

For voluntary sexual debut, female adolescents who were in the late pubertal stage had 1.429 of the hazard of those who were in the early pubertal stage.

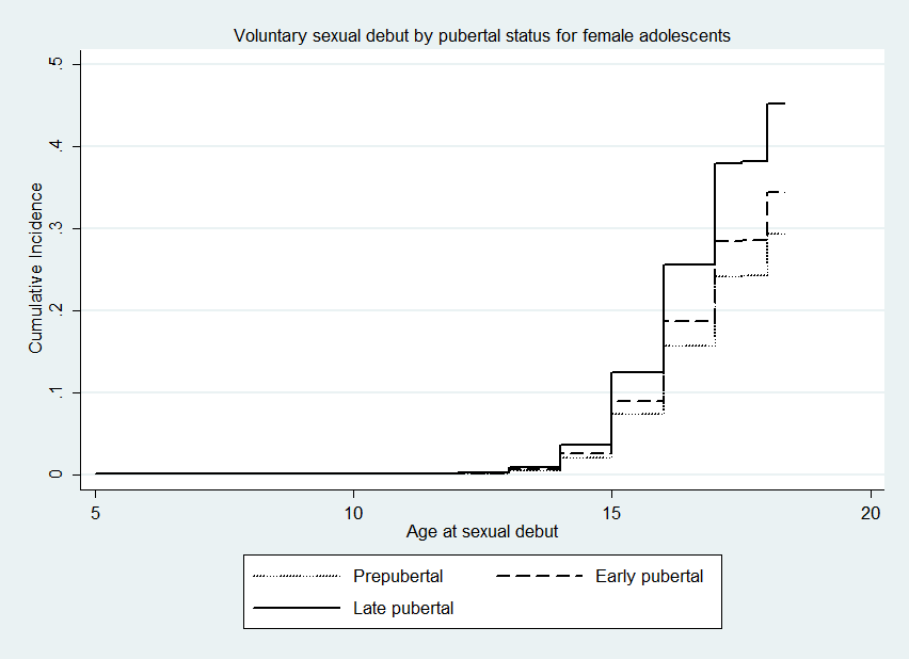

Figure 4.23: Cumulative incidence for voluntary sexual debut by pubertal status for female adolescents

Figure 4.23 shows that females in the late pubertal development stage had the highest risk of engaging in voluntary sexual debut. Intuitively, this is expected.

Table 4.8 Test of proportional subhazards assumption for pubertal status in voluntary sexual debut for females

	Variable interacted with time	Voluntary (p-value)
Pubertal status	Prepubertal (Baseline)	
	Early pubertal	0.364
	Late pubertal	0.148

Table 4.8 shows the results of the proportional subhazards assumption test for pubertal status in time to voluntary sexual debut. The p-values show that the assumption has not been violated.

Note that pubertal status did not play a significant role in the time to involuntary sexual debut.

Foreplay plays a highly significant role in both voluntary and involuntary sexual debut. Figure 4.24 shows that females who engaged in foreplay were at a significantly higher risk of engaging in voluntary sexual debut as compared to females who had not engaged in foreplay (HR $= 8.475$, p-value $= 0.000$). Figure 4.25 shows that females who engaged in foreplay were more likely to have been coerced into sexual debut as compared to females who had not engaged in foreplay (HR $= 2.639$, p-value $= 0.000$).

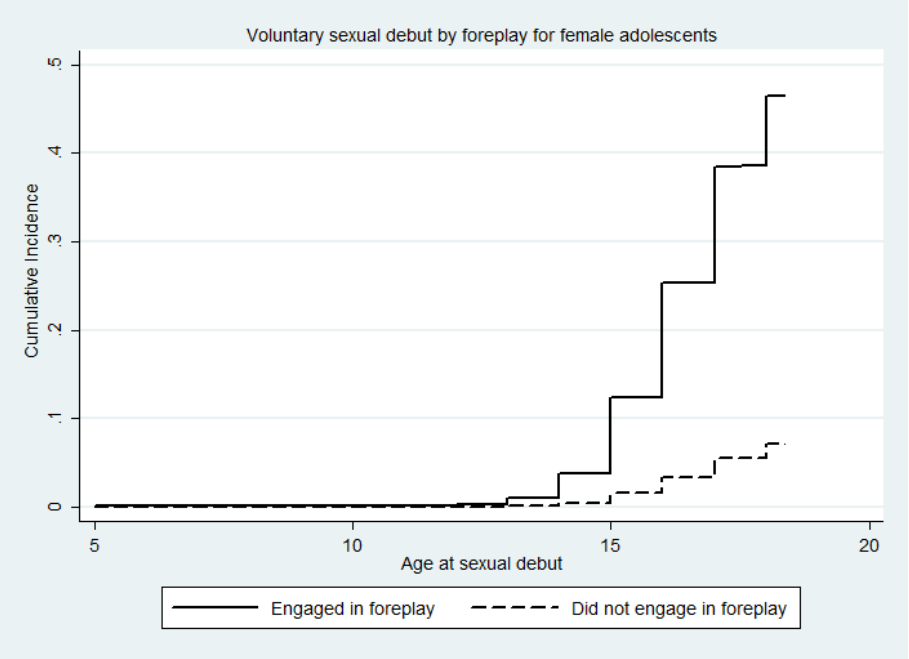

Figure 4.24: Cumulative incidence for voluntary sexual debut by foreplay for female adolescents

Table 4.9 Test of proportional subhazards assumption for foreplay in voluntary and involuntary sexual debut for females

	Variable interacted with time	Voluntary (p-value)	Involuntary (p-value)
Foreplay	Did not engage (Baseline) Engaged	0.100	0.533

Table 4.9 records the results of the proportional subhazards assumption test for

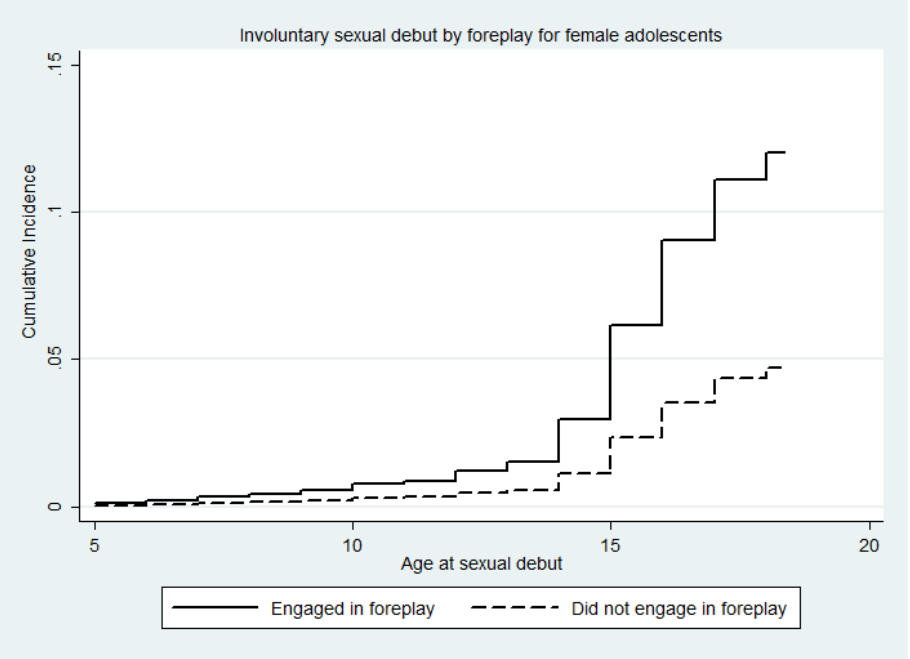

Figure 4.25: Cumulative incidence for involuntary sexual debut by foreplay for female adolescents

foreplay for both voluntary and involuntary sexual debut. Since the p-values are larger than 0.05, we accept that the effects of foreplay on voluntary and involuntary sexual debut are constant through time. Thus, the proportional subhazards assumption is not violated.

Female adolescents who engaged in oral sex had a significantly higher risk of engaging in voluntary sexual debut than female adolescents who had not engaged in oral sex (HR $= 3.846$, p-value $= 0.000$) (Figure 4.26). It was also found that females who engaged in oral sex were more likely to have been coerced than females who did not engage in oral sex (HR $= 2.141$, p-value $= 0.000$) (Figure 4.27).

Table 4.10 Test of proportional subhazards assumption for oral sex in voluntary and involuntary sexual debut for females

	Variable interacted with time	Voluntary (p-value)	Involuntary (p-value)
Oral sex	Did not engage (Baseline) Engaged	0.222	0.078

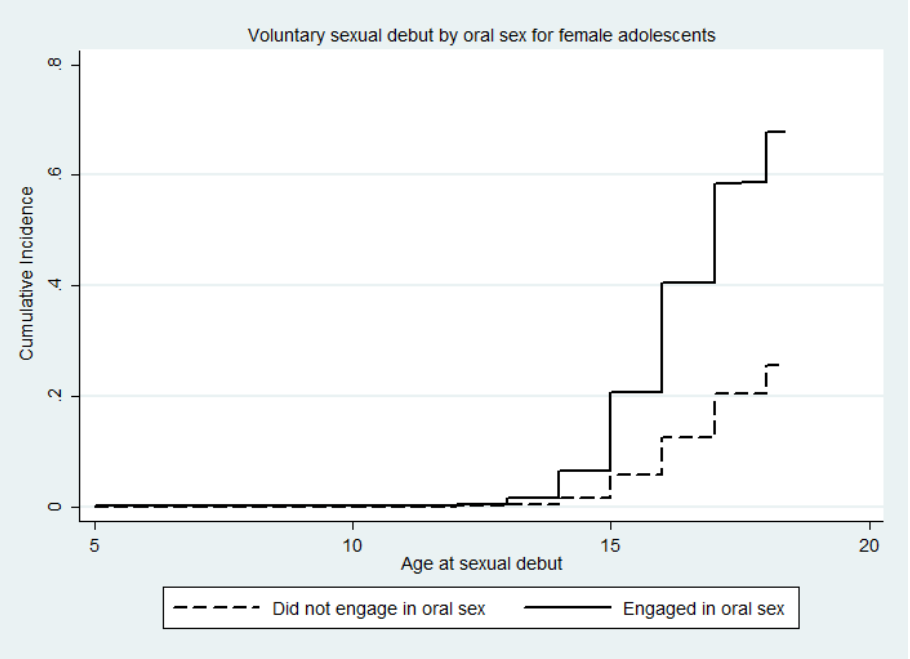

Figure 4.26: Cumulative incidence for voluntary sexual debut by oral sex for female adolescents

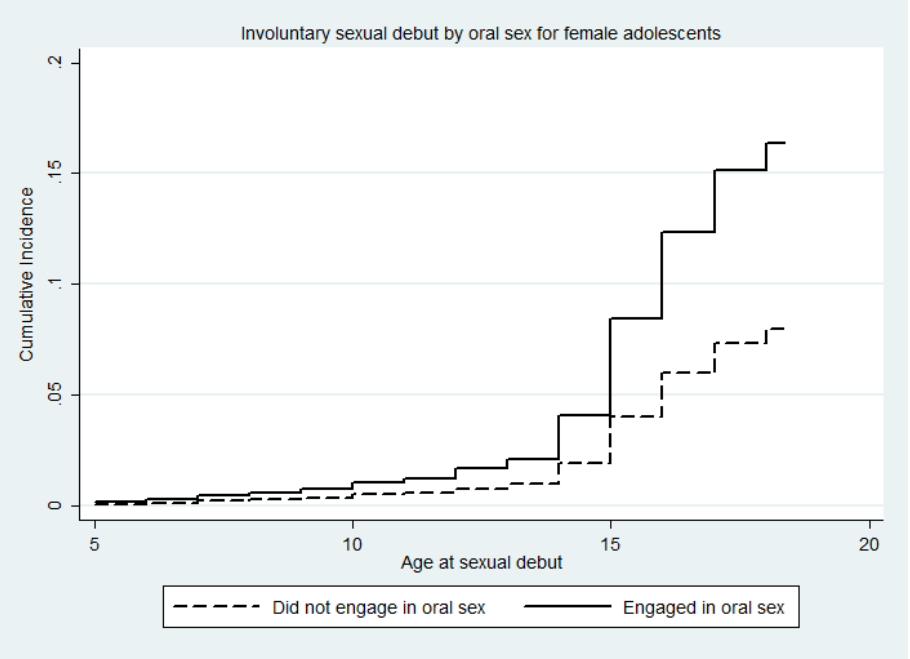

Figure 4.27: Cumulative incidence for involuntary sexual debut by oral sex for female adolescents

The proportional subhazards assumption was tested for oral sex for both voluntary and involuntary sexual debut and the results are given in Table 4.10. Time interaction variables are not significant in the model since the relevant p-values are larger than 0.05. Therefore, there is no evidence to suggest a violation in the proportional subhazards assumption.

As previously mentioned, religiosity had 24% missing data and it was not possible to impute the values. We proceed to interpret the results of the analysis including religiosity, however, we do so with caution. Figure 4.28 shows that female adolescents who were very religious had a higher risk of being coerced into first sex compared to those who were somewhat religious. The hazard ratio is given by

$$\frac{h(t, \text{Very})}{h(t, \text{Somewhat})} = \frac{1.147\ h(t, \text{Not at all})}{0.968\ h(t, \text{Not at all})} = 1.185 \tag{4.17}$$

with 95% confidence interval $(1.088; 1.291)$.

This means that female adolescents who were very religious were 1.185 times at risk of being coerced into first sex compared to female adolescents who were somewhat religious. Note that females who were not at all religious did not have a significantly different risk of engaging in involuntary sexual debut compared to the other categories.

Table 4.11 Test of proportional subhazards assumption for religiosity in involuntary sexual debut for females

	Variable interacted with time	Involuntary (p-value)
Religiosity	Not at all (Baseline)	
	Somewhat	0.838
	Very	0.534

The p-values in Table 4.11 indicate that the effect of religiosity on the hazard of involuntary sexual debut does not vary with time, thus the proportional subhazards assumption is not violated.

Next, we investigate the risk factors associated with voluntary and involuntary

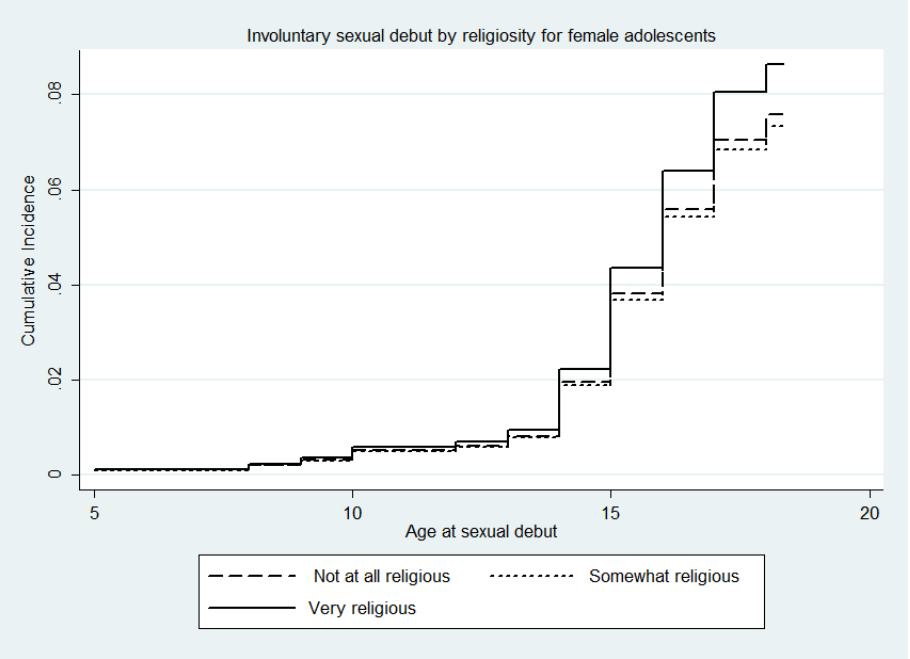

Figure 4.28: Cumulative incidence for involuntary sexual debut by religiosity for female adolescents

sexual debut in male adolescents. Table 4.12 presents the proportional subhazards model results for the competing risks regression in male adolescents. However, unlike the case for females, violations in the proportional subhazards assumption were detected. In particular, in 5 of the 16 models explored, the time interaction variables included in the models were in fact found to be significant. As previously mentioned, the inclusion of time interaction variables in the model is both a useful test of the proportional subhazards assumption and a remedy in the case of non-proportional hazards. The time interaction variables account for the time-varying effect of the hazard functions on the time to either voluntary or involuntary sexual debut. Similar to the analysis for females, only significant risk factors are explored graphically.

For the models including race for both voluntary and involuntary sexual debut, violations in the proportional subhazards assumption were found. Thus, a non-proportional subhazards model is fitted. First we assess the model including race for time to voluntary sexual debut. Table 4.13 shows that at the 0.05 level of significance, the variable Asian*time is significant in the model (p-value $=$

Table 4.12 Competing risks regression model results for male adolescents

		Voluntary		Involuntary	
		HR (95% CI)	p-value	HR (95% CI)	p-value
Race	Black (Baseline)				
	White	0.622 (0.339; 1.142)	0.126	.	.
	Coloured	0.604 (0.443; 0.824)	0.001**	0.587 (0.342; 1.008)	0.054*
	Asian	0.218 (0.075; 0.639)	0.005**	0.196 (0.028; 1.378)	0.101
Maternal education	No formal/primary education (Baseline)				
	Secondary education	1.403 (1.000; 1.967)	0.050**	0.778 (0.502; 1.204)	0.259
	Post-school training	1.433 (0.921; 2.229)	0.110	0.341 (0.149; 0.779)	0.011**
Socioeconomic status	Low (Baseline)				
	Middle	1.259 (1.0001; 1.585)	0.049**	0.941 (0.651; 1.360)	0.747
	High	1.175 (0.939; 1.472)	0.159	0.686 (0.417; 0.975)	0.038**
Height	Stunted (Baseline)				
	Normal	1.346 (1.100; 1.647)	0.004**	0.964 (0.703; 1.322)	0.819
	Tall	.	.	.	.
Pubertal status	Prepubertal (Baseline)				
	Early pubertal	1.003 (0.772; 1.304)	0.981	1.002 (0.666; 1.509)	0.991
	Late pubertal	1.082 (0.720; 1.628)	0.704	1.240 (0.664; 2.314)	0.500
Foreplay	Did not engage (Baseline)				
	Engaged	5.988 (4.000; 9.000)	0.000**	4.082 (2.326; 7.194)	0.000**
Oral sex	Did not engage (Baseline)				
	Engaged	2.770 (2.304; 3.333)	0.000**	2.193 (1.613; 2.976)	0.000**
Religiosity	Not at all (Baseline)				
	Somewhat	0.814 (0.559; 1.185)	0.283	1.217 (0.642; 2.307)	0.547
	Very	0.787 (0.568; 1.090)	0.150	0.958 (0.536; 1.711)	0.885

0.018), thus time interaction variables are included in the model. The hazard function is modeled by

$$h(t, \text{race}) = h_0(t) \exp(\beta_1{}^* w + \beta_2{}^* c + \beta_3{}^* a + \beta_4{}^* w^* t + \beta_5{}^* c^* t + \beta_6{}^* a^* t) \qquad (4.18)$$

where $h_0(t)$ is the baseline hazard function (which is the hazard function for Black male adolescents), w, c and a are the indicator variables for White, Coloured and Asian respectively, β_1, β_2 and β_3 are the associated regression coefficients of the covariates, β_4, β_5 and β_6 are the regression coefficients of the time inter-

Table 4.13 Non-proportional hazards regression model results for race in voluntary sexual debut for males

Variable	Coefficient ($\hat{\beta}$)	(95% CI)	p-value
Black (Baseline)			
White	-5.544	$(-12.254; 1.166)$	0.105
Coloured	-3.613	$(-7.561; 0.334)$	0.073^{*}
Asian	-18.712	$(-33.816; -3.608)$	0.015^{**}
White*time	0.343	$(-0.091; 0.776)$	0.121
Coloured*time	0.211	$(-0.048; 0.470)$	0.110
Asian*time	1.091	$(0.188; 1.993)$	0.018

action covariates and t is the time to voluntary sexual debut.

Next we examine the regression results to help understand the relative hazards. Note that coefficients are merely average values. Thus, on average, Black male adolescents have a significantly higher hazard of engaging in voluntary sexual debut compared to Coloured and Asian male adolescents (p-value $= 0.073$ and p-value $= 0.015$ respectively). Additionally, Asian males have the lowest hazard of engaging in voluntary sexual debut. In the case of a non-proportional hazards model, the hazard functions fluctuate with time . Thus, in order to calculate hazard ratios at a particular time, one would have to plug-in the relevant time point into the model.

From Equation (4.18), it follows that

$$\begin{aligned} h(w,b) &= \exp(\beta_1 + \beta_4{}^{*}t) \\ h(c,b) &= \exp(\beta_2 + \beta_5{}^{*}t) \\ h(a,b) &= \exp(\beta_3 + \beta_6{}^{*}t) \end{aligned}$$

Given the regression coefficients from Table 4.13 and substituting $t = 12$ in the above equations yield

$$\begin{aligned} h(w,b) &= \exp(\beta_1 + \beta_4{}^{*}t) \\ &= \exp(-5.544 + 0.343{}^{*}12) \\ &= 0.240 \end{aligned} \tag{4.19}$$

$$\begin{aligned} h(c,b) &= \exp(\beta_2 + \beta_5{}^{*}t) \\ &= \exp(-3.613 + 0.211^{*}12) \\ &= 0.339 \end{aligned} \tag{4.20}$$

$$\begin{aligned} h(a,b) &= \exp(\beta_3 + \beta_6{}^{*}t) \\ &= \exp(-18.712 + 1.091^{*}12) \\ &= 0.004 \end{aligned} \tag{4.21}$$

Similarly, we can calculate hazard ratios for these male adolescents at age 13 up to age 18 years. The results are seen in Table 4.14. We don't consider ages below 12 years as we know that all sexual debut prior to 12 years old was regarded as involuntary sexual debut.

Table 4.14 Hazard ratios by time for race in voluntary sexual debut

	Hazard Ratio (Baseline = Black)		
Age	White	Coloured	Asian
12 years	0.240	0.339	0.004
13 years	0.338	0.419	0.011
14 years	0.476	0.517	0.032
15 years	0.671	0.639	0.096
16 years	0.946	0.789	0.285
17 years	1.332	0.974	0.848
18 years	1.878	1.203	2.524

At 12 years old, Coloured adolescents have 0.339 times the hazard of engaging in voluntary sexual debut relative to Black adolescents whereas at 15 years old, the hazard ratio is 0.639. This shows that the effects are becoming smaller with time until some time between 17 and 18 years where Coloured and Black males have the same hazard function. After this intermediate time, the effects become larger with time since Coloured adolescents have 1.203 the hazard of engaging in voluntary sexual debut compared to Black adolescents at 18 years old. Asian males have 0.004 times the hazard of engaging in voluntary sexual debut compared to Black males at 12 years old. At each year following this, the effects are becoming smaller with time until some age between 17 and 18 years. Note how rapidly the hazard ratio is becoming larger. At 18 years old, Asian males have 2.525 times the hazard of engaging in voluntary sexual debut compared to Black

males. The high hazard ratios at 18 years for Coloured and Asian adolescents relative to Black adolescents show that Black adolescents engaged in sexual debut at much earlier ages and Coloured and Asian adolescents engage in sexual debut at relatively later ages. Additionally, with regard to White males, recall that this group of adolescents were not found to have a significantly different hazard of engaging in voluntary sexual debut compared to Black male adolescents.

Table 4.15 Non-proportional hazards regression model results for race in involuntary sexual debut for males

Variable	Coefficient ($\hat{\beta}$)	(95% CI)	p-value
Black (Baseline)			
White	.	.	.
Coloured	-1.451	$(-4.739; 1.837)$	0.387
Asian	-5.250	$(-7.452; -3.049)$	0.000^{**}
White*time	.	.	.
Coloured*time	0.067	$(-0.164; 4.298)$	0.571
Asian*time	0.253	$(0.149; 0.356)$	0.000^{**}

Table 4.15 depicts the results for the model including race for time to involuntary sexual debut. Note that the hazard ratios involving White males could not be determined because there were no White males who had reported coercion. At the 0.05 level of significance we note that the covariate Asian*time is highly significant in the model (p-value $= 0.000$), thus time interaction variables are necessary in the model. The hazard function is modeled by

$$h(t, \text{race}) = h_0(t) \exp(\beta_1{}^* w + \beta_2{}^* c + \beta_3{}^* a + \beta_4{}^* w^* t + \beta_5{}^* c^* t + \beta_6{}^* a^* t) \qquad (4.22)$$

where $h_0(t)$ is the baseline hazard function (which is once again the hazard function for Black male adolescents), w, c and a are the indicator variables for White, Coloured and Asian respectively, β_1, β_2 and β_3 are the associated regression coefficients of the covariates, β_4, β_5 and β_6 are the regression coefficients of the time interaction covariates and t is now the time to involuntary sexual debut.

The results in Table 4.15 show that on average Asian males have the lowest hazard of being coerced into first sex. There is no significant difference between

the hazards for Black and Coloured adolescents.

Using Equation (4.22) we calculate hazard ratios from age 5 up to 18 years. Here we must include the earlier years because the exploratory analysis in Chapter 2 revealed that a substantial proportion of males reported sexual debut prior to 12 years old, which is regarded as involuntary sexual debut.

Table 4.16 Hazard ratios by time for race in involuntary sexual debut

	Hazard Ratio (Baseline = Black)	
Age at involuntary sexual debut	Coloured	Asian
5 years	0.328	0.019
6 years	0.350	0.024
7 years	0.375	0.031
8 years	0.401	0.040
9 years	0.428	0.051
10 years	0.458	0.066
11 years	0.490	0.085
12 years	0.524	0.109
13 years	0.560	0.141
14 years	0.599	0.181
15 years	0.640	0.233
16 years	0.685	0.301
17 years	0.732	0.387
18 years	0.783	0.499

Asian male adolescents have 0.019 times the hazard of being coerced into sexual debut at the age of 5 years. This hazard ratio steadily increases and by the age of 18 years, Asian males have approximately half the hazard of coercion relative to Black males. With regard to Coloured male adolescents, we know that there is no significant difference relative to Black adolescents.

Next we test whether the level of maternal education affects the time to voluntary and involuntary sexual debut. Figure 4.29 shows that males who had mothers with a secondary education were found to have a significantly higher risk of engaging in voluntary sexual debut as compared to males who had mothers with no formal/primary school education (HR $= 1.403$, p-value $= 0.005$). The data did not provide further evidence to suggest that the level of maternal education significantly affected the risk of male adolescents engaging in sexual debut.

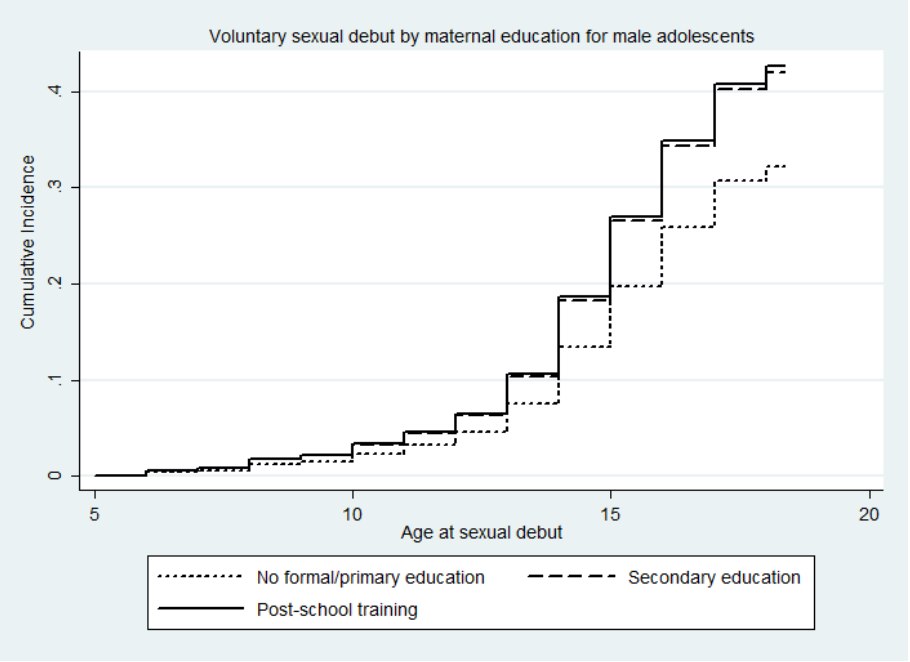

Figure 4.29: Cumulative incidence for voluntary sexual debut by maternal education for male adolescents

Males who had mothers with post-school training had a significantly lower risk of being coerced into first sex compared to males who had mothers with no formal/primary education (HR $= 0.341$, p-value $= 0.011$). Additionally, males whose mothers had post-school training also showed a much lower risk of being coerced into first sex relative to males whose mothers had secondary education. The hazard ratio is given by

$$\frac{h(t, \text{Post-school training})}{h(t, \text{Secondary education})} = \frac{0.341\ h(t, \text{No formal/primary education})}{0.778\ h(t, \text{No formal/primary education})} = 0.438 \tag{4.23}$$

with 95% confidence interval $(0.297; 0.647)$.

Male adolescents with post-school training level maternal education had 0.438 of the hazard of male adolescents with secondary level maternal education. Figure 4.30 shows that male adolescents who had mothers with post-school training had the lowest risk of being coerced into sexual debut.

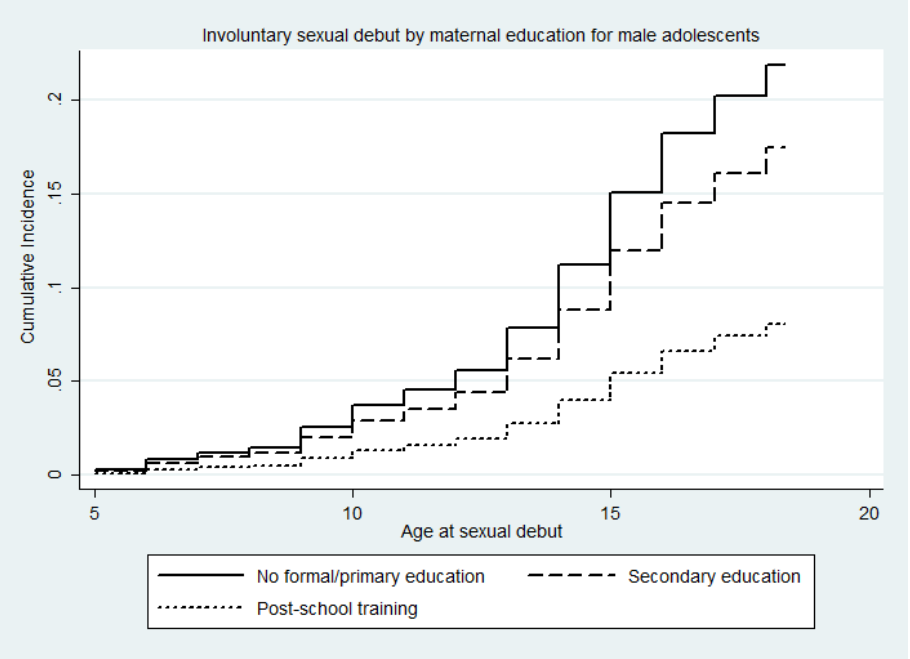

Figure 4.30: Cumulative incidence for involuntary sexual debut by maternal education for male adolescents

Table 4.17 Test of proportional subhazards assumption for maternal education in voluntary and involuntary sexual debut for males

	Variable interacted with time	Voluntary (p-value)	Involuntary (p-value)
Maternal education	No formal/primary (Baseline)		
	Secondary	0.560	0.633
	Post-school	0.602	0.083

Table 4.17 gives the results of the proportional hazards test for maternal education for voluntary and involuntary sexual debut. The p-values are greater than 0.05 indicating no evidence to suggest that the proportional subhazards assumption has been violated.

Table 4.18 Non-proportional hazards regression model results for socioeconomic status in voluntary sexual debut for males

Variable	Coefficient ($\hat{\beta}$)	(95% CI)	p-value
Low (Baseline)			
Middle	-0.051	$(-1.452; 1.351)$	0.943
High	-1.946	$(-3.536; -0.356)$	0.016^{**}
Middle*time	0.020	$(-0.078; 0.117)$	0.690
High*time	0.146	$(0.038; 0.254)$	0.008^{**}

Table 4.18 displays the results for the model including socioeconomic status for time to voluntary sexual debut. At the 0.05 level of significance, the variable High*time (p-value $= 0.008$) is significant so a non-proportional hazards model is fitted. The hazard function is modeled by

$$h(t, \text{ses}) = h_0(t) \exp(\beta_1{}^* m + \beta_2{}^* h + \beta_3{}^* m^* t + \beta_4{}^* h^* t) \tag{4.24}$$

where $h_0(t)$ is the baseline hazard function (which is the hazard function for male adolescents with a low socioeconomic status), m and h are the indicator variables for males with a middle socioeconomic status and those with a high socioeconomic status respectively, β_1 and β_2 are the associated regression coefficients of the covariates, β_3 and β_4 are the regression coefficients of the time interaction covariates and t is the time to voluntary sexual debut.

On average, male adolescents with a high socioeconomic status have the lowest risk of engaging in voluntary sexual debut whereas those with a low and middle level socioeconomic status do not have a significantly different hazard of engaging in voluntary sexual debut.

Table 4.19 Hazard ratios by time for socioeconomic status in voluntary sexual debut

	Hazard Ratio (Baseline = Low)	
Age at voluntary sexual debut	Middle	High
12 years	1.208	0.824
13 years	1.232	0.953
14 years	1.257	1.103
15 years	1.283	1.276
16 years	1.309	1.477
17 years	1.335	1.709
18 years	1.362	1.978

Table 4.19 shows that male adolescents with a high socioeconomic status had a lower risk of engaging in voluntary sexual debut at ages 12 and 13 years old relative to males with a low socioeconomic status. The hazard of engaging in voluntary sexual debut for these two classes of socioeconomic status is equivalent at some time between 13 and 14 years old. Thereafter, at 14 years old we see an opposite effect of socioeconomic status where males with a high socioeconomic status have a higher hazard of engaging in voluntary sexual debut

compared to those with a low socioeconomic status (HR $= 1.103$). These effects become larger with time and at 18 years we see that male adolescents with a high socioeconomic status have almost twice the hazard of engaging in voluntary sexual debut compared to those with a low socioeconomic status.

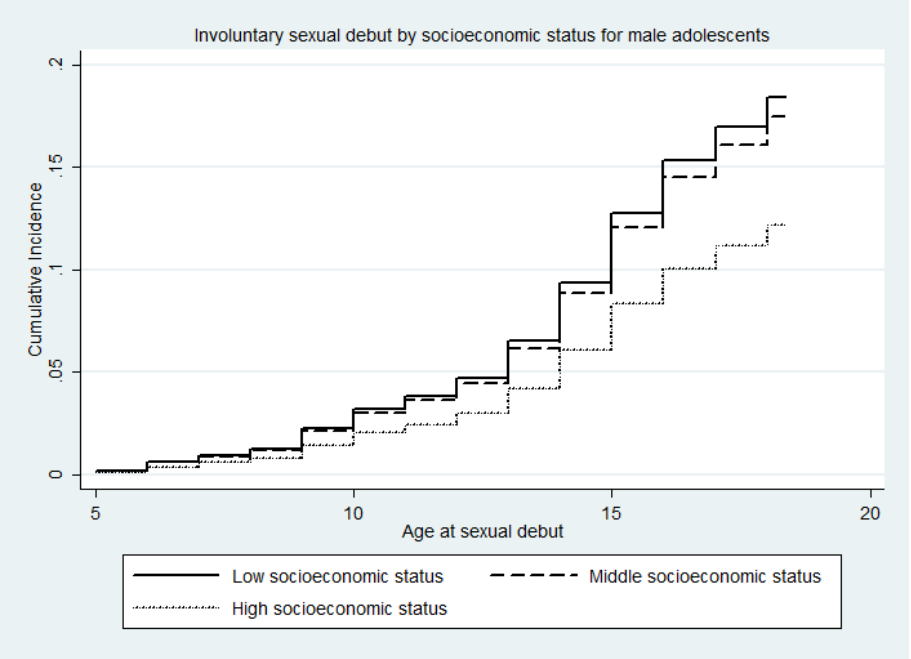

Figure 4.31: Cumulative incidence for involuntary sexual debut by socioeconomic status for male adolescents

For involuntary sexual debut, a significant difference was only detected between males with a high socioeconomic status and males with a low socioeconomic status where those who possessed a high socioeconomic status were less likely to be coerced relative to those with a low socioeconomic status (HR $= 0.638$, p value $= 0.038$). The result is shown graphically in Figure 4.31.

Table 4.20 Test of proportional subhazards assumption for socioeconomic status in involuntary sexual debut for males

	Variable interacted with time	Involuntary (p-value)
Socioeconomic status	Low (Baseline)	
	Middle	0.457
	High	0.052

The results of testing the proportional subhazards assumption for socioeconomic status in coerced sexual debut for male adolescents is shown in Ta-

ble 4.20. Since the p-values are smaller than 0.05, there is no evidence to suggest that the proportional subhazards assumption is not satisfied.

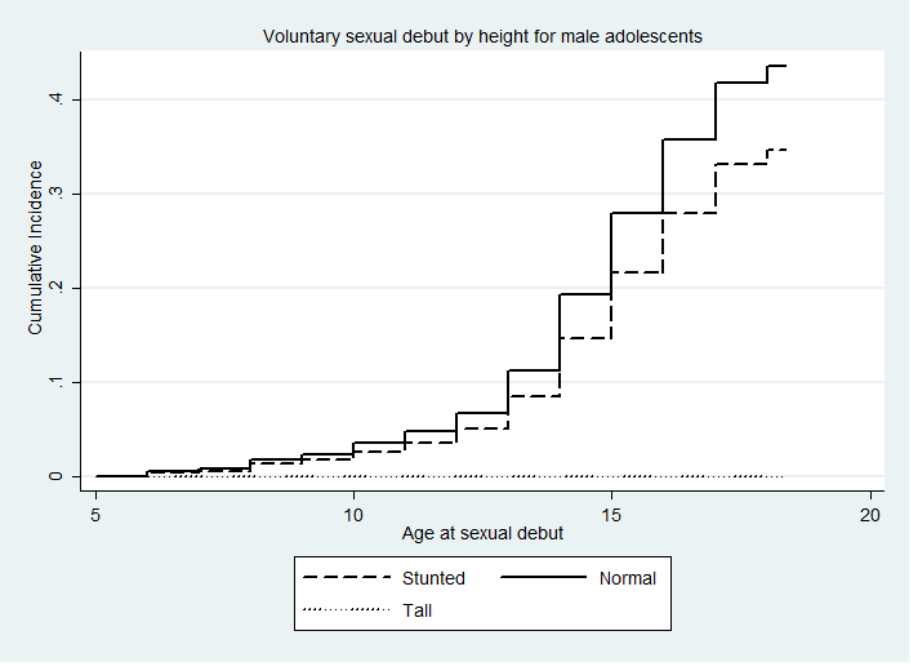

Figure 4.32: Cumulative incidence for voluntary sexual debut by height for male adolescents

Male adolescents who were classified as normal height were 1.346 times likely to engage in voluntary sexual debut compared to males whose height was classified as stunted (p-value $= 0.004$). Hazard ratios involving tall males could not be determined as there were no tall males who reported voluntary or involuntary sexual debut. Figure 4.32 shows that males who were of normal height were more likely to engage in sexual debut than males who had stunted height. The data did not provide sufficient evidence to suggest that the risk of coerced sexual debut was affected by whether a male respondent had normal or stunted height.

The proportional subhazards assumption for height for voluntary sexual debut was tested and the results are given in Table 4.21. The p-value is larger than 0.05 which indicates that the proportional hazards assumption is not violated however, we interpret this with caution since we did not include tall males in

Table 4.21 Test of proportional subhazards assumption for height in voluntary sexual debut for males

	Variable interacted with time	Voluntary (p-value)
Height	Stunted (Baseline)	
	Normal	0.324
	Tall	.

our testing due to insufficient data.

Table 4.22 displays the results for the model including foreplay for time to voluntary sexual debut. At the 0.05 level of significance, the variable Engaged*time (p-value $= 0.003$) is significant thus indicating the possible usage of a non-proportional hazards model. This means that the effects of the hazards vary with time, however, note that the main analysis shows that on average, the hazard of those males who engaged in foreplay is not significantly different relative to those who did not engage in foreplay (p-value $= 0.224$). Therefore, even though the effects vary with time, there is no evidence to suggest that the effects differ from each other. It follows then that foreplay is not considered as a significant predictor in the time to voluntary sexual debut analysis.

Table 4.22 Non-proportional hazards regression model results for foreplay in voluntary sexual debut for males

Variable	Coefficient ($\hat{\beta}$)	(95% CI)	p-value
Did not Engage (Baseline)			
Engaged	-1.210	$(-3.161; 0.741)$	0.224
Engaged*time	0.218	$(0.072; 0.364)$	0.003^{**}

Table 4.22 shows the results of the model including foreplay for involuntary sexual debut. The time interaction variable Engaged*time is highly significant in the model at the 0.05 level (p-value $= 0.011$). The hazard function is modeled by

$$h(t, \text{Foreplay}) = h_0(t) \exp(\beta_1{}^* e + \beta_2{}^* e^* t) \tag{4.25}$$

where $h_0(t)$ is the baseline hazard function (which is the hazard function of those male adolescents who did not engage in foreplay), e is the indicator variable for those males who engaged in foreplay, β_1 is the associated regression coefficient of the covariate, β_2 is the regression coefficient of the time interac-

tion covariate and t is the time to involuntary sexual debut.

Table 4.23 Non-proportional hazards regression model results for foreplay in involuntary sexual debut for males

Variable	Coefficient ($\hat{\beta}$)	(95% CI)	p-value
Did not Engage (Baseline)			
Engaged	3.986	$(1.904; 6.068)$	0.000^{**}
Engaged*time	-0.185	$(-0.326; -0.043)$	0.011^{**}

The results in Table 4.23 show that on average, males who engaged in foreplay had a higher hazard of being coerced into first sex compared to males who did not engage in foreplay (p-value $= 0.000$). Substituting $t = 12, 13, \ldots, 18$ into the model, we obtain hazard ratios for those who had engaged in foreplay relative to those who had not engaged in foreplay from age 12 up to 18 years old. These are shown in Table 4.24.

Table 4.24 Hazard ratios by time for foreplay in involuntary sexual debut

	Hazard Ratio (Baseline = Did not engage)
Age at involuntary sexual debut	Engaged
12 years	5.847
13 years	4.860
14 years	4.039
15 years	3.357
16 years	2.790
17 years	2.319
18 years	1.927

Table 4.24 shows that for ages 12 to 18 years old, male adolescents who engaged in foreplay had a higher hazard of being coerced into first sex compared to those who did not engage in foreplay. However, we notice that these effects become substantially smaller with time. At age 12, males who engaged in foreplay had 5.847 times the hazard of being coerced into sexual debut relative to males who had not engaged in foreplay, whereas at age 18 years we see that this hazard ratio has diminished to 1.927. This means that at 18 years old, males who have engaged in foreplay are almost twice as likely to have been coerced into sexual debut relative to males who have not engaged in foreplay.

Males who engaged in oral sex had 2.770 times the hazard of engaging in voluntary sexual debut compared to male adolescents who did not engage in oral sex (p-value $= 0.000$). It was also found that males who engaged in oral sex were 2.193 times more likely to have been coerced into sexual debut than males who did not engage in oral sex. Figure 4.33 and Figure 4.34 show the associated cumulative incidence functions for voluntary and involuntary sexual debut.

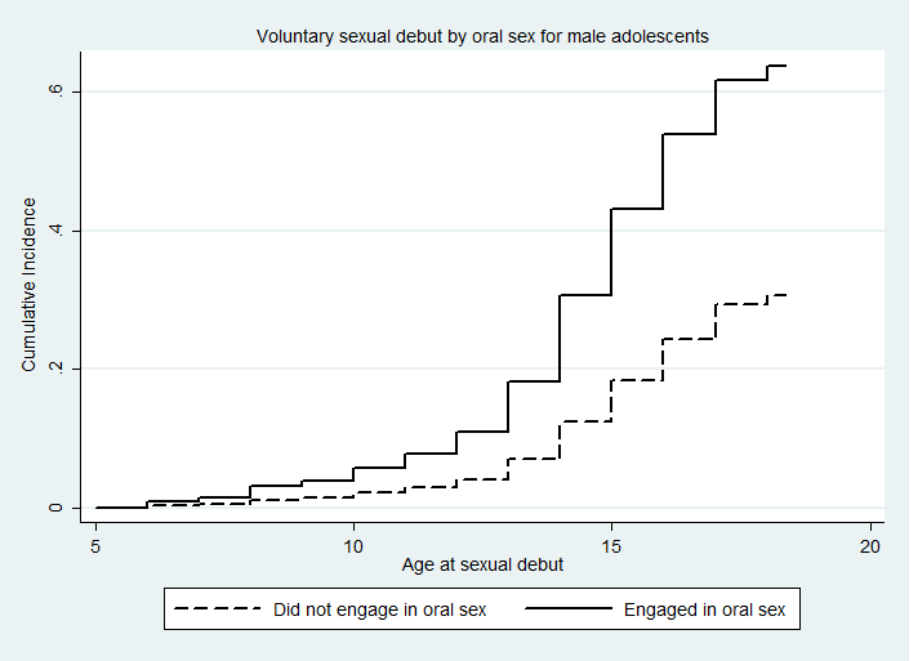

Figure 4.33: Cumulative incidence for voluntary sexual debut by oral sex for male adolescents

Table 4.25 Test of proportional subhazards assumption for oral sex in voluntary and involuntary sexual debut for males

	Variable interacted with time	Voluntary (p-value)	Involuntary (p-value)
Oral sex	Did not engage (Baseline)		
	Engaged	0.410	0.156

Table 4.25 shows the proportional subhazards assumption test results for oral sex. For both voluntary and involuntary sexual debut we see that time dependant oral sex variables are rejected at the 0.05 level, thus indicating that the proportional subhazards assumption has not been violated.

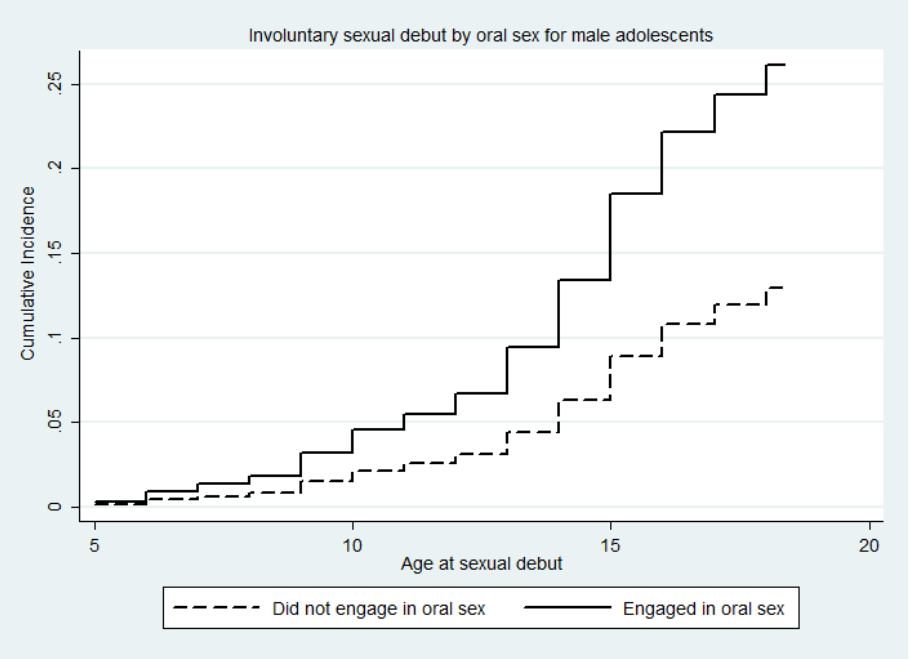

Figure 4.34: Cumulative incidence for involuntary sexual debut by oral sex for male adolescents

4.2.1 Conclusion

Black adolescents have a higher risk of engaging in voluntary sexual debut compared to Coloured and Asian adolescents for females. White adolescents have no significantly different risk of engaging in voluntary sexual debut compared to the other three race groups. For males, the effect of race on the hazard of voluntary sexual debut varies with time. In particular, Asian males usually had a lower risk of engaging in voluntary sexual debut compared to Black, Coloured and White males. Black males usually had a higher risk of engaging in voluntary sexual debut relative to Colured and Asian males with these effects initially becoming smaller with time. By age 18 years the effects become larger with time and Asian and Coloured males have a higher risk of engaging in voluntary sexual debut. The risk of White males are indifferent to the risk of Black and Coloured males in the time to voluntary sexual debut. Race had no effect on the hazard of involuntary sexual debut for female adolescents whereas the effect of race on the hazard of involuntary sexual debut varied with time for male adolescents. Asian male adolescents consistently had a lower

risk of coerced first sex relative to Black and Coloured males. Hazard ratios involving the effect of White males in the time to coerced first sex could not be determined due to insufficient data. The level of maternal education played no significant role in affecting the hazard of consensual sexual debut in females. In males however, those who had mothers with no formal/primary school education had a lower hazard of engaging in voluntary sexual debut compared to those whose mothers had a secondary school education. Post-school maternal education acted as a protective factor against coerced first sex of adolescents for both females and males. Socioeconomic status did not affect the hazard of voluntary sexual debut for female adolescents however it had a time varying effect on the hazard of voluntary sexual debut for male adolescents. A high socioeconomic status acted as a protective factor against involuntary sexual debut for female adolescents. For male adolescents, those with a high socioeconomic status had a lower risk of being coerced into first sex compared to those with a low socioeconomic status and those with a middle socioeconomic status did not have a significantly different hazard of being coerced relative to the other two categories. Normal height adolescents have a higher risk of engaging in voluntary sexual debut compared to adolescents with stunted height across both strata of gender. Females with stunted height had a higher risk of being coerced into first sex relative to those with normal height while height did not play a significant role in affecting involuntary sexual debut in males. No conclusions could be reached for tall adolescents due to insufficient data. Females with a late pubertal status had the highest risk of engaging in voluntary sexual debut whereas pubertal status did not significantly affect the hazard of voluntary sexual debut in males. Additionally, pubertal status played no role in affecting the hazard of involuntary sexual debut in both females and males. In females adolescents, it was found that those who engaged in foreplay were at a significantly higher risk of engaging in voluntary sexual debut. Surprisingly, foreplay was not found to play a significant role in voluntary sexual debut in males. Engagement in foreplay for both female and male adolescents was associated with higher levels of coerced first sex relative to those who did not engage in foreplay. In particular, for male adolescents, engagement in foreplay had a time varying effect on the hazard of coerced sexual debut. At all ages in the study, male adolescents who

engaged in foreplay had a higher hazard of being coerced into first sex relative to those who did not engage in foreplay with these effects becoming smaller with time. Engagement in oral sex was significantly associated with a higher hazard of engaging in voluntary sexual debut for both females and males. The results were similar for involuntary sexual debut although the risk was more evenly distributed between those who engaged in oral sex and those who did not compared to voluntary sexual debut. The level of religiosity was not found to be a significant variable in affecting the hazard of voluntary sexual debut for both females and males. Females who were reportedly very religious had a higher hazard of being coerced relative to those who were somewhat religious whereas religiosity played no role on the hazard of involuntary sexual debut for male adolescents.

Chapter 5

Conclusion

Risks arising from early sexual debut in adolescents are of particular importance as this group of the population represents the calibre of the next generation of adults in South Africa. Their sexual behaviour today will influence the overall social well-being and health status of the adults of tomorrow. The Birth to Twenty sexual debut survival analysis is concentrated around understanding the factors that are associated with early sexual debut in a South African context.

Two methods for analyzing time to sexual debut for adolescents were investigated. The first approach used standard survival analysis by employing the popular Cox proportional hazards regression model. Next, a more appropriate approach than standard survival analysis was considered to analyse the Birth to Twenty sexual debut data, namely the Fine & Gray (1999) method of analysing competing risks data. This stemmed from identifying that the event of interest can occur from two separate causes and the occurrence from one cause made it impossible for the event to occur from the other cause. A log-rank test showed inequality of the survival curves for voluntary and involuntary sexual debut which provided justification to use a competing risks model. The competing risks regression model results showed that the risk factors for time to voluntary and involuntary sexual debut differed which reiterates the need of a competing risks model.

The results and methods used in this study contribute to the topic of Modeling

Survival Data but more importantly contributes to the research around early sexual debut in a South African context. In a country where teenage pregnancies and HIV are a significant concern and fairly little research has been conducted in the area of early sexual debut, research is vital in understanding the predictors of sexual debut among adolescents. This research can be used to provide insight when designing strategies and action plans (such as sexual education programmes and workshops) in an attempt to educate adolescents in making informed and safe decisions about their sexual behaviour so as to prevent adolescents from compromising their health and social statuses. Furthermore, the study also identifies risk factors for time to coerced sexual debut. Very few studies in a South African context focus on coerced sexual debut however interest in this area is growing due to its association with adverse social and health implications (Agardh et al., 2011).

The first possible limitation to consider in this study is the question of the representativeness of the data. If we consider when the adolescents in the cohort were 16 years old then according to Statistics South Africa (2006) White people constituted 9.2% of the population whereas only 3.3% of the respondents in the Birth to Twenty study are White. If we consider current race statistics then the proportion of Coloured and Asian respondents in the study are not representative of the South African population since 12.8% of the respondents in the study are Coloured and 2.7% are Asian whereas according to Statistics South Africa (2014) these proportions are 8.8% and 8.4% respectively. The usage of univariate survival analysis techniques are also a limitation in the study as it allows only to investigate direct effects of the variables on the time to sexual debut and it does not allow for investigation of how variables affect each other in the analysis of time to sexual debut. Additionally, making the assumption that all sexual behaviour occurring in respondents at ages below 12 years is involuntary must also be included as a limitation in the study. Another possible limitation of the current study is that the majority of the variables recorded in the Birth to Twenty study are not directly subject to intervention. Demographic and anthropometric measures are inherent to an individual. Additionally, social factors including maternal education, socioeconomic status and religiosity

are also not subject to direct intervention. In other words, efforts cannot be dedicated to changing these factors however understanding the effects of these predictors on time to voluntary and involuntary sexual debut can assist in providing direction in terms of which groups of individuals to target with strategies and action plans in an effort to delay first sex.

Future research in South African studies should focus on including factors around formal sex education. In line with several similar studies conducted outside South Africa, formal sex education was listed among the most influential predictors of first sex and is a factor that is directly subject to intervention (Mueller et al., 2008). The effects are worth investigating in a South African context. Currently, only very few studies in South Africa have included formal sex education in analysing time to sexual debut where it is known whether the adolescents were exposed to sexual education before or after sexual debut. Future research should also consider conducting analysis segmented by age group to investigate whether the risk factors of time to voluntary and involuntary sexual debut vary for adolescents belonging to different age groups.

.

Appendices

Appendix A

Some generalized linear models concepts

A.1 Maximum likelihood estimation

This section of the appendix gives a summary of the results on maximum likelihood estimation that are relevant to survival analysis. The results presented apply equally to inferences based on a partial likelihood function, and so can be used for estimation in the Cox regression model and the competing risks model described in Section 3.3 and Section 3.4 respectively. A full treatment of the theory of maximum likelihood estimation and likelihood ratio testing is given by Cox & Hinkley (1974). The main source used in the following is Collett (2003).

A.1.1 Inference about a single unknown parameter

Suppose that the likelihood of n observed survival times $t_1, t_2, \ldots, t_n$ is a function of a single unknown parameter β, and denoted $L(\beta)$. The *maximum likelihood estimate* of β is then the value $\hat{\beta}$ for which this function is a maximum. In almost all applications, it is more convenient to work with the natural logarithm of the likelihood function, $\log L(\beta)$. The value $\hat{\beta}$, which maximizes the log-likelihood, is the same value that maximizes the likelihood function itself, and is generally found using differential calculus.

Specifically, $\hat{\beta}$ is the value of β for which the derivative of $\log L(\beta)$, with respect

to β, is equal to zero. In other words, $\hat{\beta}$ is such that

$$\left.\frac{\mathrm{d}\log L(\beta)}{\mathrm{d}\beta}\right|_{\hat{\beta}} = 0$$

The first derivative of $\log L(\beta)$ with respect to β is known as the *efficient score* for β, and is denoted $u(\beta)$. Therefore,

$$u(\beta) = \frac{\mathrm{d}\log L(\beta)}{\mathrm{d}\beta}$$

and so the maximum likelihood estimate of β, $\hat{\beta}$, satisfies the equation:

$$u(\hat{\beta}) = 0$$

The asymptotic variance of the maximum likelihood estimate of β can be found from

$$\left(-E\left\{\frac{\mathrm{d}^2\log L(\beta)}{\mathrm{d}\beta^2}\right\}\right)^{-1} \tag{A.1}$$

or from the equivalent formula,

$$\left(E\left\{\frac{\mathrm{d}\log L(\beta)}{\mathrm{d}\beta}\right\}^2\right)^{-1}$$

The variance calculated from either of these expressions can be regarded as the approximate variance of $\hat{\beta}$, although it is usually more straightforward to use expression (A.1). When the expected value of the derivative in expression (A.1) is difficult to obtain, a further approximation to the variance of $\hat{\beta}$ is then given by

$$\mathrm{var}(\hat{\beta}) \approx -\left\{\frac{\mathrm{d}^2\log L(\beta)}{\mathrm{d}\beta^2}\right\}^{-1}\Bigg|_{\hat{\beta}} \tag{A.2}$$

The second derivative of the log-likelihood function is sometimes known as the *Hessian*, and the quantity

$$-E\left\{\frac{\mathrm{d}^2\log L(\beta)}{\mathrm{d}\beta^2}\right\}$$

is called the *information function*. Since the information function is formed from the expected value of the second derivative of $\log L(\beta)$, it is sometimes

called the *expected information function*. In contrast, the negative second derivative of the log-likelihood function itself is called the *observed information function*. This latter quantity will be denoted $i(\beta)$, so that

$$i(\beta) = -\left\{\frac{\mathrm{d}^2 \log L(\beta)}{\mathrm{d}\beta^2}\right\}$$

The reciprocal of this function, evaluated at $\hat{\beta}$, is then the approximate variance of $\hat{\beta}$ given in Equation (A.2), that is,

$$\mathrm{var}(\hat{\beta}) \approx \frac{1}{i(\hat{\beta})}$$

The standard error of $\hat{\beta}$, that is, the square root of the estimated variance of $\hat{\beta}$, is found from

$$\mathrm{se}(\hat{\beta}) = \frac{1}{\sqrt{i(\hat{\beta})}}$$

This standard error can be used to construct the confidence intervals for β.

In order to test the null hypothesis that $\beta = 0$, three alternative test statistics can be used. The *likelihood ratio test statistic* is the difference between the values of $-2 \log L(\hat{\beta})$ and $-2 \log L(0)$.

The *Wald test* is based on the statistic $\hat{\beta}^2 i(\hat{\beta})$.
The *score test statistic* is $\{u(o)\}^2 / i(0)$. Each of these statistics has an asymptotic chi-squared distribution with 1 degree of freedom, under the null hypothesis that $\beta = 0$. Note that the Wald statistic is equivalent to the statistic

$$\frac{\hat{\beta}}{\mathrm{se}(\hat{\beta})}$$

which has an asymptotic standard normal distribution.

A.1.2 Inference about a vector of unknown parameters

The main source used in the following section is Collett (2003). The results in Section A.1.1 can be extended to the situation where n observations are used to

estimate the values of p unknown parameters, $\beta_1, \beta_2, \ldots, \beta_p$. These parameters can be assembled into a p-component vector, $\boldsymbol{\beta}$, and the corresponding likelihood function is $L(\boldsymbol{\beta})$. The maximum likelihood estimates of the p unknown parameters are the values $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$, which maximize $L(\boldsymbol{\beta})$. They are therefore found by solving the p equations

$$\left.\frac{\mathrm{d} \log L(\boldsymbol{\beta})}{\mathrm{d}\beta_j}\right|_{\hat{\boldsymbol{\beta}}} = 0$$

for $j = 1, 2, \ldots, p$, simultaneously.

The efficient score for β_j is

$$u(\beta_j) = \frac{\mathrm{d} \log L(\boldsymbol{\beta})}{\mathrm{d}\beta_j}$$

$j = 1, 2, \ldots, p$ and these quantities can be assembled to give a p-component vector of efficient scores, denoted $\mathbf{u}(\boldsymbol{\beta})$. The vector of maximum likelihood estimates is therefore such that

$$\mathbf{u}(\hat{\beta}) = \mathbf{0}$$

where $\mathbf{0}$ is the $p \times 1$ vector of zeroes.

Now let $\mathbf{H}(\boldsymbol{\beta})$ be the $p \times p$ matrix of second partial derivatives of the log-likelihood function, $\log L(\hat{\boldsymbol{\beta}})$. The (j, k)th element of $\mathbf{H}(\boldsymbol{\beta})$ is then

$$\frac{\partial^2 \log L(\hat{\boldsymbol{\beta}})}{\partial\beta_j \partial\beta_k}$$

for $j = 1, 2, \ldots, p$, $k = 1, 2, \ldots, p$ and $\mathbf{H}(\boldsymbol{\beta})$ is called the *Hessian matrix*. The matrix

$$\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta})$$

is called the *observed information matrix*. The (j, k)th element of the corresponding *expected information matrix* is

$$-E\left\{\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}\right\}$$

The variance-covariance matrix of the p maximum likelihood estimates, var$(\hat{\beta})$, can then be approximated by the inverse of the observed information matrix, evaluated at $\hat{\beta}$, so that

$$\text{var}(\hat{\beta}) \approx \mathbf{I}^{-1}(\hat{\beta})$$

The square root of the (j, j)th element of this matrix can be taken to be the standard error of $\hat{\beta}_j, j = 1, 2, \ldots, p$.

The test statistics given in Section A.1.1 can be generalized to the multiparameter situation. Consider the test of the null hypothesis, that is, $\beta_1, \beta_2, \ldots, \beta_p = 0$. The likelihood ratio test statistic is the value of

$$2\{\log L(\hat{\beta}) - \log L(\mathbf{0})\}$$

and the Wald test is based on

$$\hat{\beta}'\mathbf{I}(\hat{\beta})\hat{\beta}$$

and the score test statistic is

$$\mathbf{u}'(\mathbf{0})\mathbf{I}^{-1}(\mathbf{0})\mathbf{u}(\mathbf{0})$$

Each of these statistics has a chi-squared distribution with p degrees of freedom, under the null hypothesis.

In comparing alternative models, interest centers on the hypothesis that some of the β-parameters in a model are equal to zero. To test this hypothesis, the likelihood ratio test is the most suitable, and so we only consider this procedure here. Suppose that a model contains $p + q$ parameters, $\beta_1, \beta_2, \ldots, \beta_p, \ldots, \beta_{p+q}$, is to be compared with a model that only contains the p parameters $\beta_1, \beta_2, \ldots, \beta_p$. This amounts to testing the null hypothesis that the q parameters, $\beta_{p+1}, \ldots, \beta_{p+q}$, in the model with $p+q$ unknown parameters are all equal to zero. Let $\hat{\boldsymbol{\beta}}_1$ denote the vector of estimates under the model with $p + q$ parameters and $\hat{\boldsymbol{\beta}}_2$ that for the model with just p parameters. The likelihood ratio test of the null hypothesis

$$H_0 = \beta_{p+1}, \beta_{p+2}, \ldots, \beta_{p+q} = 0$$

in the model with $p + q$ parameters is then based on the statistic

$$2\{\log L(\hat{\beta}_1) - \log L(\hat{\beta}_2)\}$$

which has a chi-squared distribution with

$$\begin{aligned} \text{Degrees of freedom} &= p + q - p \\ &= q \end{aligned}$$

A.2 The Newton-Raphson procedure

Let $\mathbf{u}(\boldsymbol{\beta})$ be the $p \times 1$ vector of first derivatives of the log-likelihood function in Equation (3.12) with respect to the $\boldsymbol{\beta}$ parameters. This result is referred to as the *vector of efficient scores*. Let $\mathbf{I}(\boldsymbol{\beta})$ be the $p \times p$ matrix of negative second derivatives of the log-likelihood function, where the (j, k)th element of $\mathbf{I}(\boldsymbol{\beta})$ is given by

$$-\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_j \beta_k}$$

$\mathbf{I}(\boldsymbol{\beta})$ is called the *observed information matrix* (Collett, 2003).

The Newton-Raphson method gives an estimate of the vector of $\boldsymbol{\beta}$-parameters at the $(t + 1)$th cycle of the iterative procedure, $\hat{\boldsymbol{\beta}}_{t+1}$, as

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_t)\mathbf{u}(\hat{\boldsymbol{\beta}}_t)$$

for $t = 0, 1, 2, \ldots$, where $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_t)$ is the inverse of the information matrix and $\mathbf{u}(\hat{\boldsymbol{\beta}}_t)$ is the vector of efficient scores, both of these quantities are evaluated at $\hat{\boldsymbol{\beta}}_t$. To start the procedure, take $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$. When there are relatively small changes in the log-likelihood, the process may be terminated. Once the iterative procedure has converged, the variance-covariance matrix of the parameter estimates may be approximated at $\hat{\boldsymbol{\beta}}$, using the inverse of the information matrix, that is, $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$. To obtain the standard errors of the estimated parameter estimates, we merely take the square root of the diagonal elements of the $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$ matrix (Collett, 2003).

Appendix B

Cox proportional hazards model regression diagnostics

B.1 Log minus log (survival time) versus survival time plots

Figure B.1 to Figure B.16 show parallel curves across strata for each of the covariates. The curves show no intersection across strata and thus the proportionality assumption of the Cox proportional hazards model has not been found to be violated. There are a few cases where the curves of categories of a covariate either superimpose each other or they are very close to each other and these cases are noted.

Figure B.1 and Figure B.2 show the curves for the race models for females and males respectively. As seen, the curves appear to be approximately parallel.

Figure B.3 and Figure B.4 show the curves for the maternal education models for female and male adolescents respectively. The curves are fairly close across both strata of gender, however, no intersection occurs and thus there is no evidence to suggest a violation of the proportionality assumption.

Figure B.5 and Figure B.6 show the curves for the pubertal status models for females and males respectively. For females, the curves are approximately parallel over time. For males, we note that the curve for adolescents in the prepubertal stage has superimposed the curve for those who belong to the early

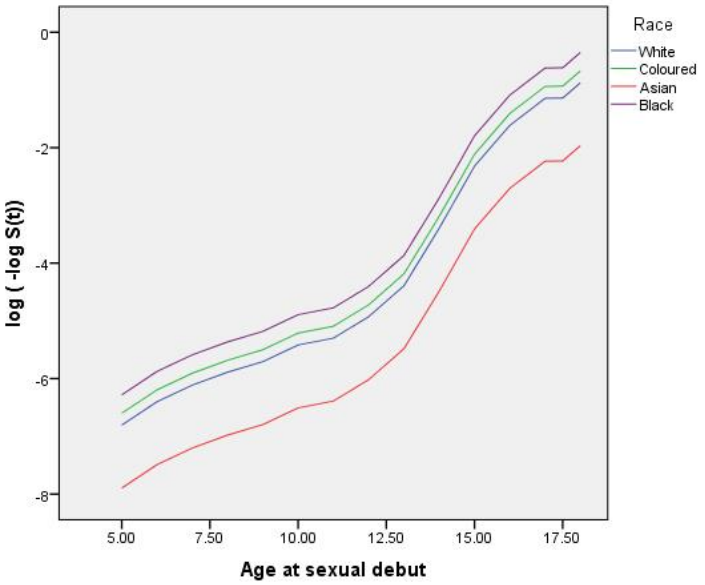

Figure B.1: $\log(-\log\ S(t))$ versus survival time for race in female adolescents

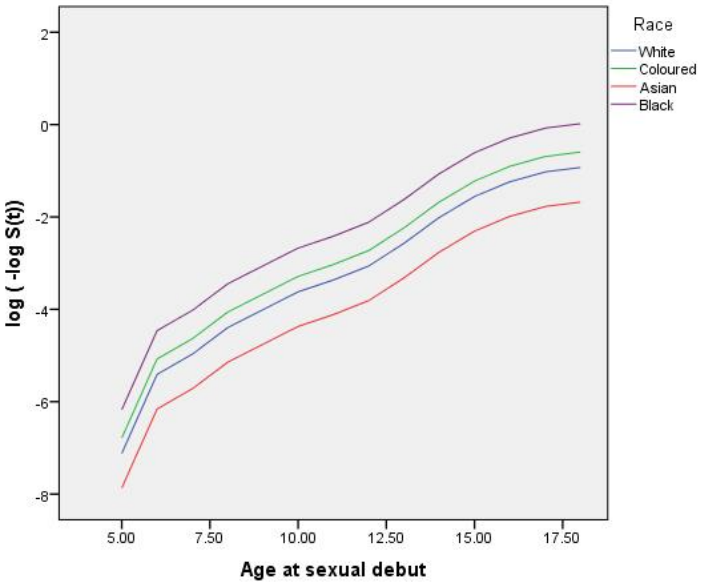

Figure B.2: $\log(-\log\ S(t))$ versus survival time for race in male adolescents

pubertal stage of development. Similar curves does not imply intersection of curves. Therefore, for both females and males there is no evidence to suggest a

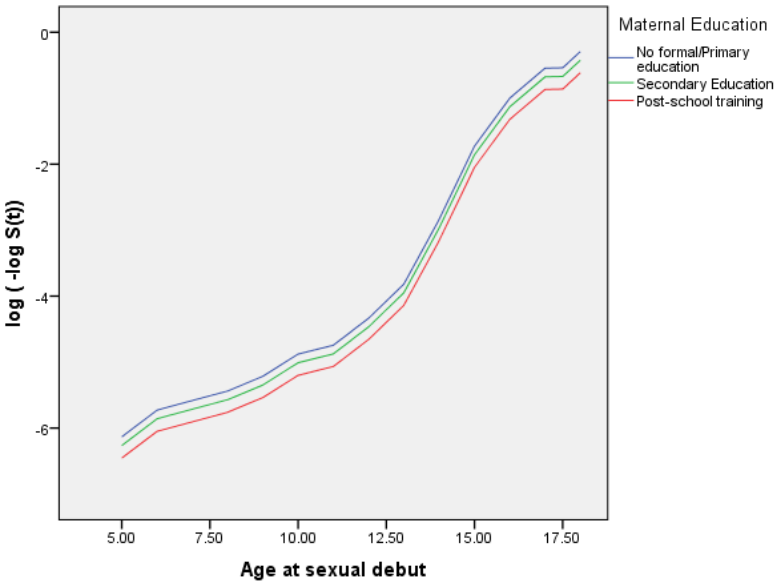

Figure B.3: $\log(-\log\ S(t))$ versus survival time for maternal education in female adolescents

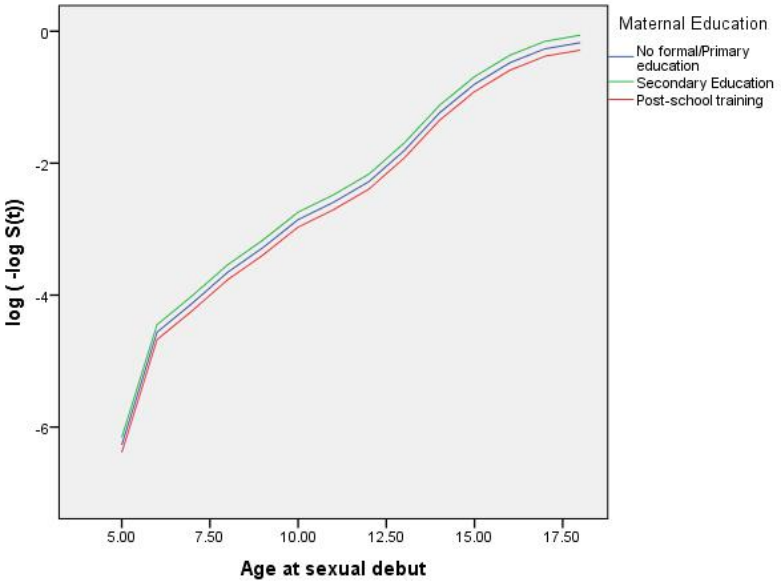

Figure B.4: $\log(-\log\ S(t))$ versus survival time for maternal education in male adolescents

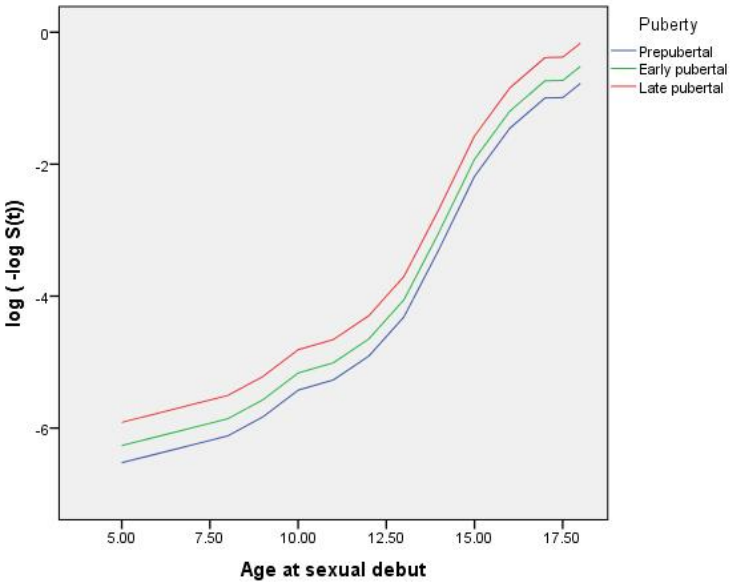

Figure B.5: $\log(-\log\ S(t))$ versus survival time for puberty in female adolescents

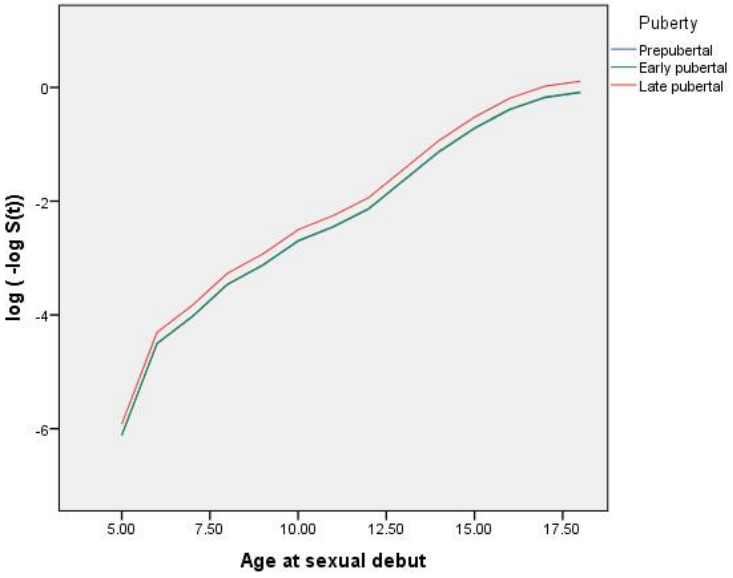

Figure B.6: $\log(-\log\ S(t))$ versus survival time for puberty in male adolescents

violation of the proportional hazards assumption.

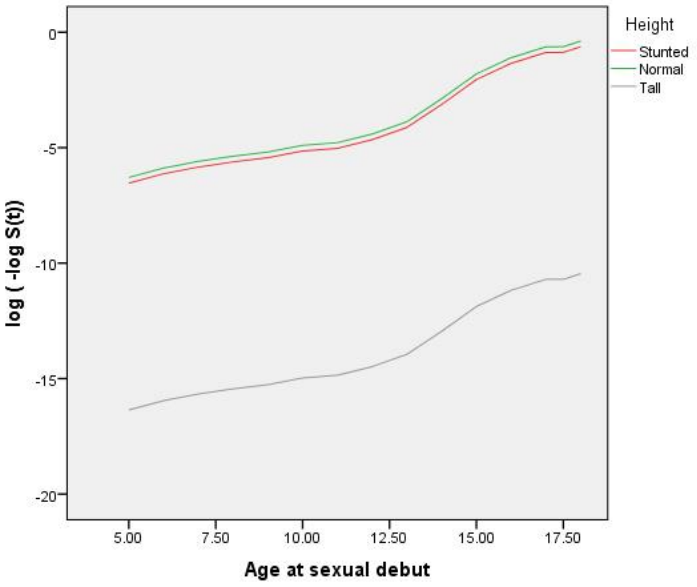

Figure B.7: $\log(-\log\ S(t))$ versus survival time for height in female adolescents

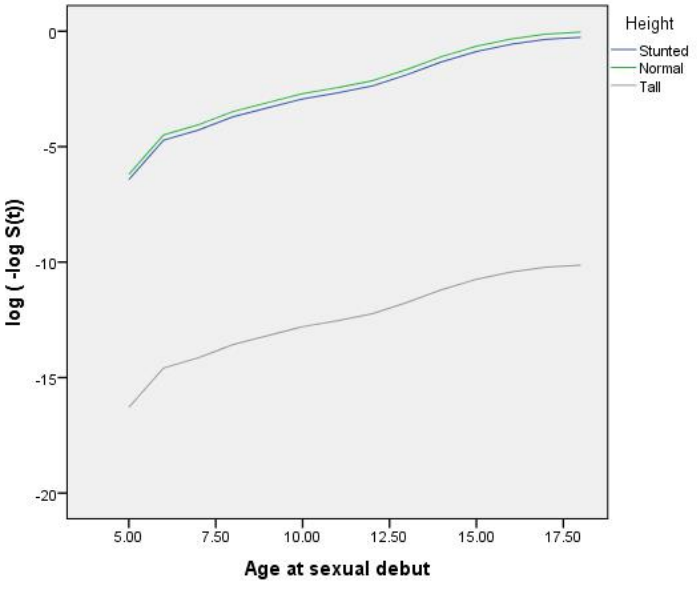

Figure B.8: $\log(-\log\ S(t))$ versus survival time for height in male adolescents

Figure B.7 and Figure B.8 show the curves for the height models for females and males respectively. For both females and males we cannot readily inter-

pret the hazard ratios involving tall adolescents due to insufficient data in the study. The remaining curves are approximately parallel and thus the proportional hazards assumption continues to hold true.

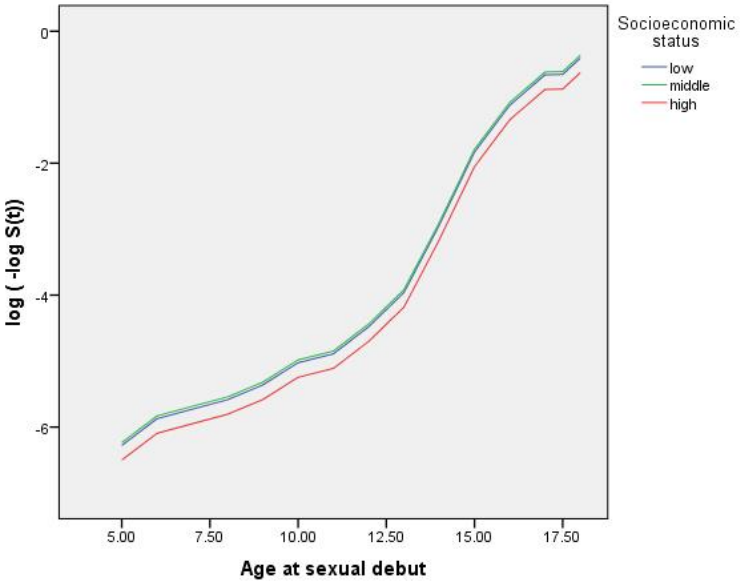

Figure B.9: $\log(-\log\ S(t))$ versus survival time for socioeconomic status in female adolescents

The curves for the socioeconomic status models can be seen in Figure B.9 and Figure B.10 for females and males respectively. For females, it appears that the curves for adolescents with a low socioeconomic status and adolescents with a middle socioeconomic status are almost identical. Additionally, for males we note that the curve for adolescents with a low socioeconomic status has been superimposed by the curve belonging to adolescents who possess a high socioeconomic status. No curves are found to intersect for both females and males, thus the proportional hazards assumption has not been violated.

Figure B.11 and Figure B.12 show the curves for the foreplay models and Figure B.13 and Figure B.14 show the curves for the oral sex models. The curves are shown for females and males respectively. The curves in each of the models are parallel and thus the proportional hazards assumption is not violated.

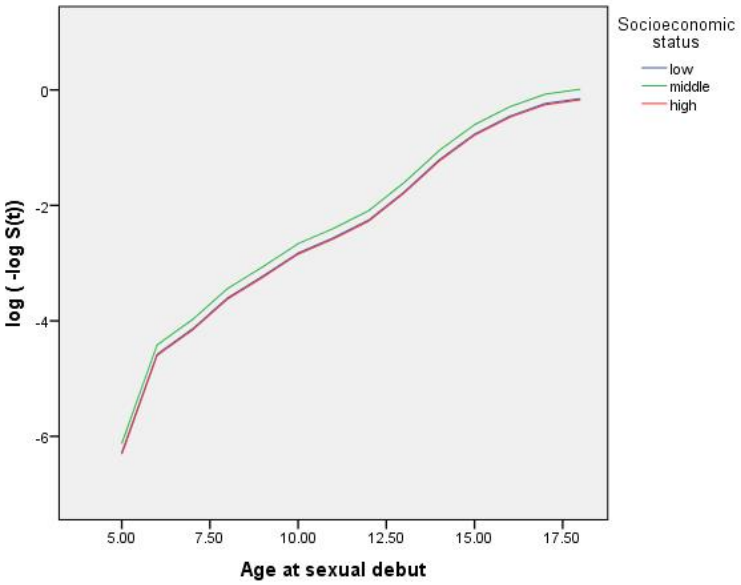

Figure B.10: $\log(-\log\ S(t))$ versus survival time for socioeconomic status in male adolescents

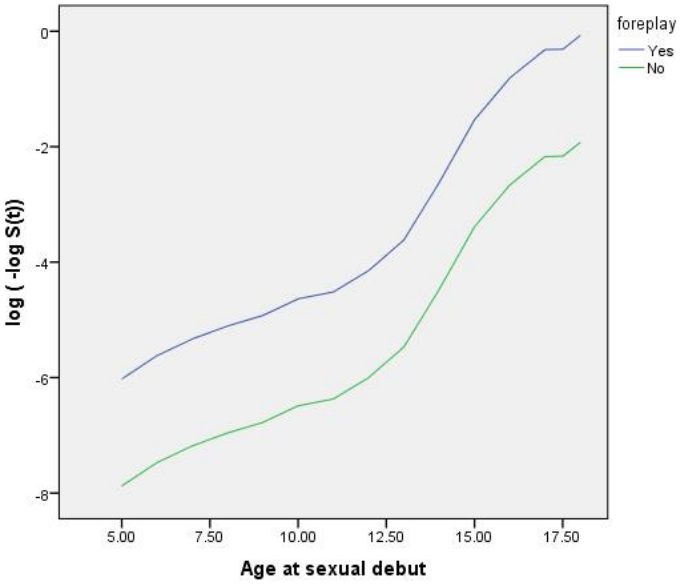

Figure B.11: $\log(-\log\ S(t))$ versus survival time for foreplay in female adolescents

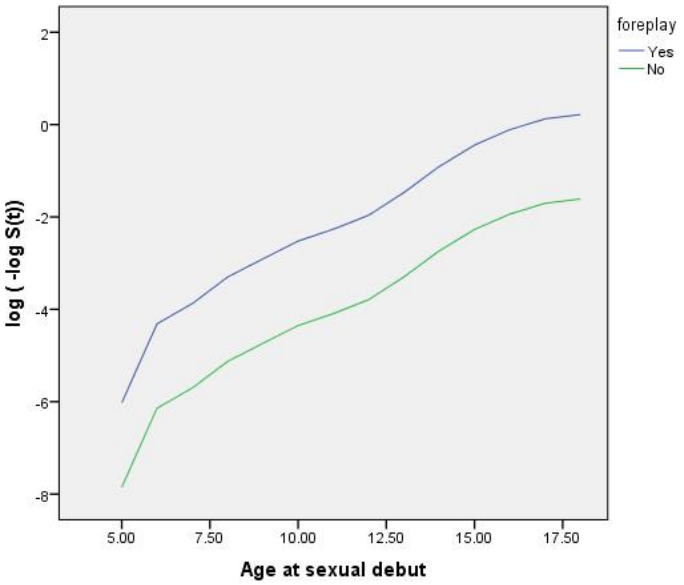

Figure B.12: $\log(-\log\ S(t))$ versus survival time for foreplay in male adolescents

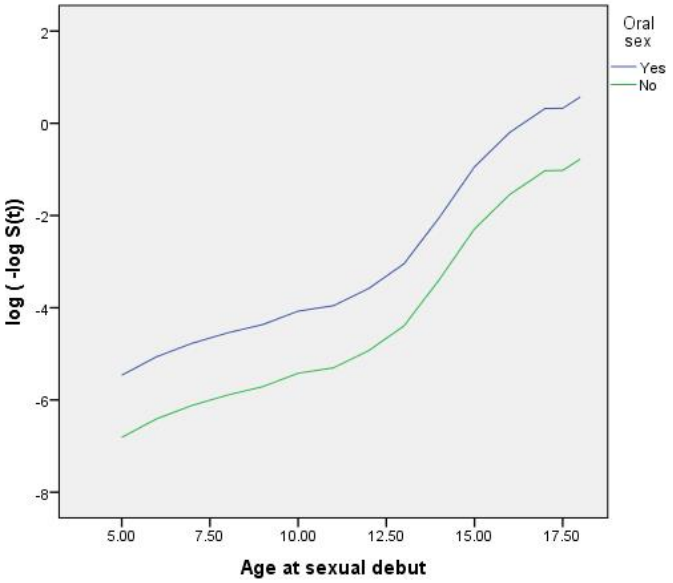

Figure B.13: $\log(-\log\ S(t))$ versus survival time for oral sex in female adolescents

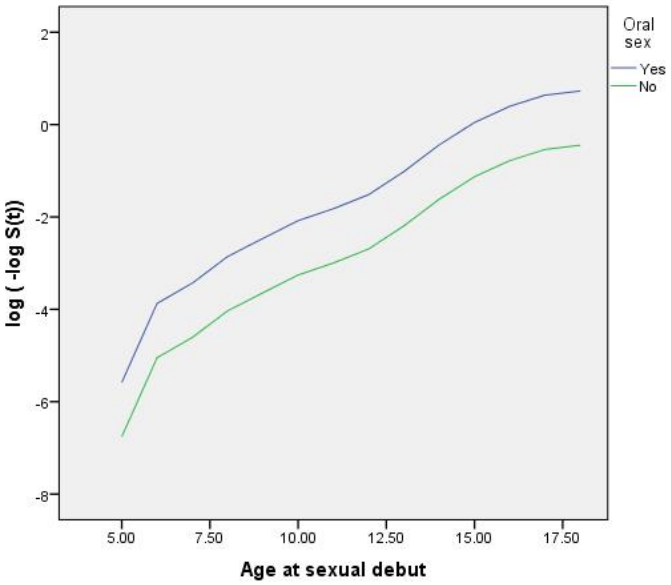

Figure B.14: $\log(-\log\ S(t))$ versus survival time for oral sex in male adolescents

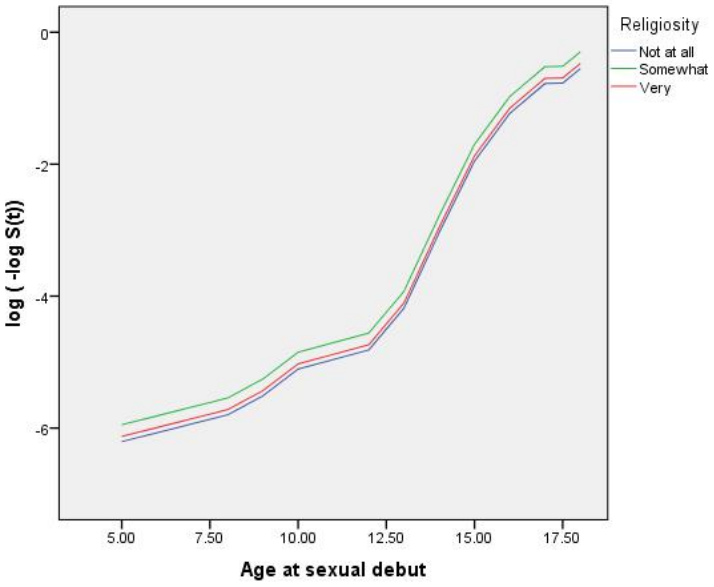

Figure B.15: $\log(-\log\ S(t))$ versus survival time for religiosity in female adolescents

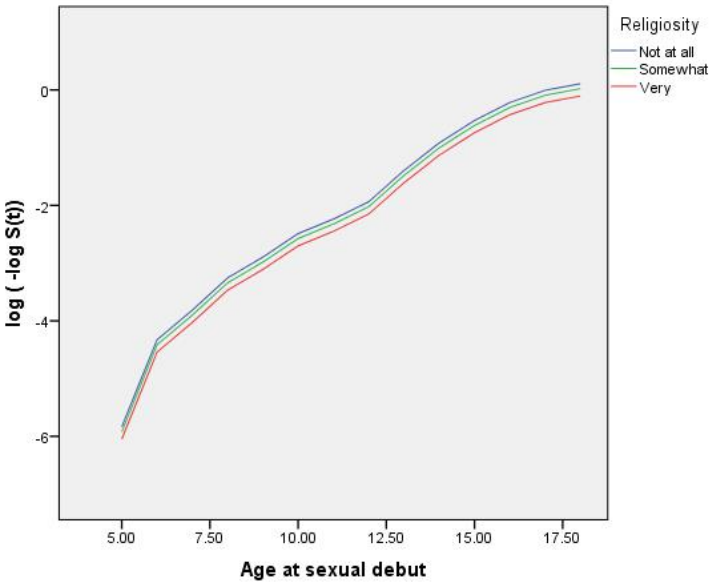

Figure B.16: $\log(-\log\ S(t))$ versus survival time for religiosity in male adolescents

Figure B.15 and Figure B.16 show the curves for the religiosity models for females and males respectively. For female adolescents we note that curves for those who are classified as not at all religious and very religious are fairly similar to each other. However, no intersection of curves across strata has occurred. For males, the curves are approximately parallel. Therefore, the models do not violate the proportional hazards assumption.

B.2 Cox-Snell Residuals

Figure B.17 to Figure B.32 graph the Cox-Snell residuals versus the cumulative hazard of the residuals for each of the models. All plots are fairly close to that of the Unit Exponential distribution curve. Only two models showed deviations greater than 0.20 units from the Unit Exponential curve, namely pubertal status and religiosity in females. Both residual plots show a maximum deviation of 0.25 units from the Unit Exponential distribution. These deviations are isolated events and appear to be outliers in the analysis. It is thus no cause for

concern. Overall, all models show an adequate fit based on the residual plots.

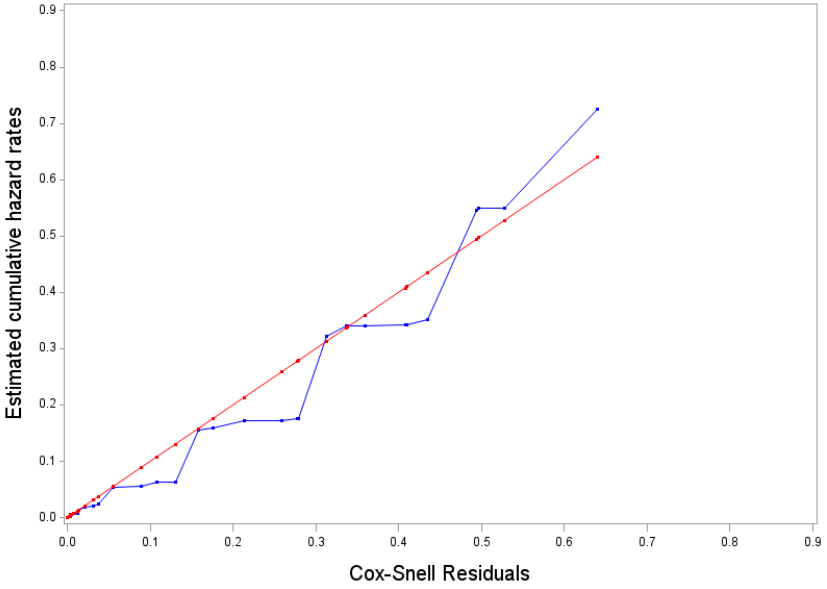

Figure B.17: Cox-Snell residuals for race in female adolescents

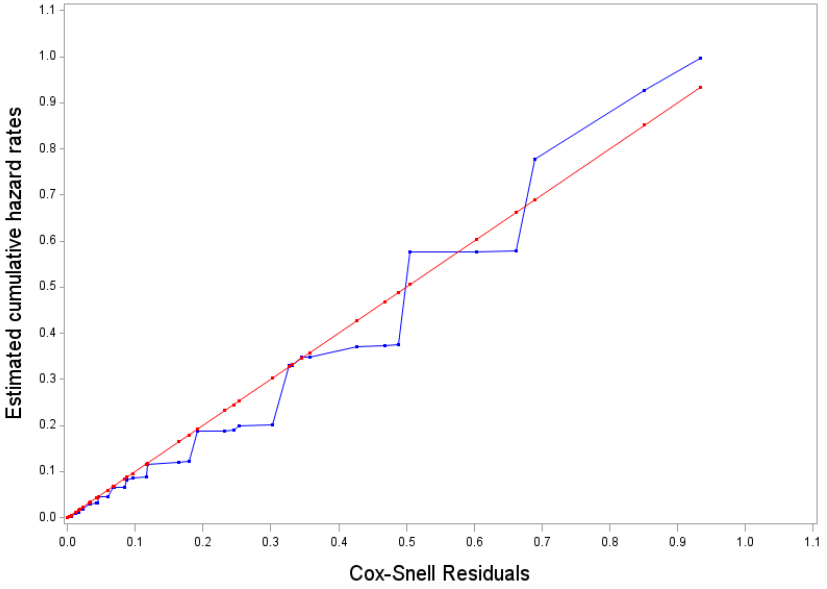

Figure B.18: Cox-Snell residuals for race in male adolescents

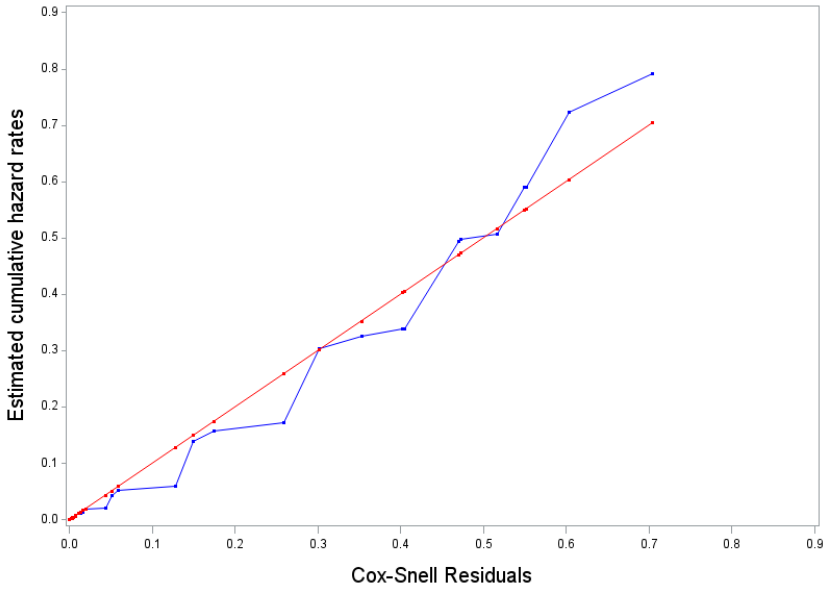

Figure B.19: Cox-Snell residuals for maternal education in female adolescents

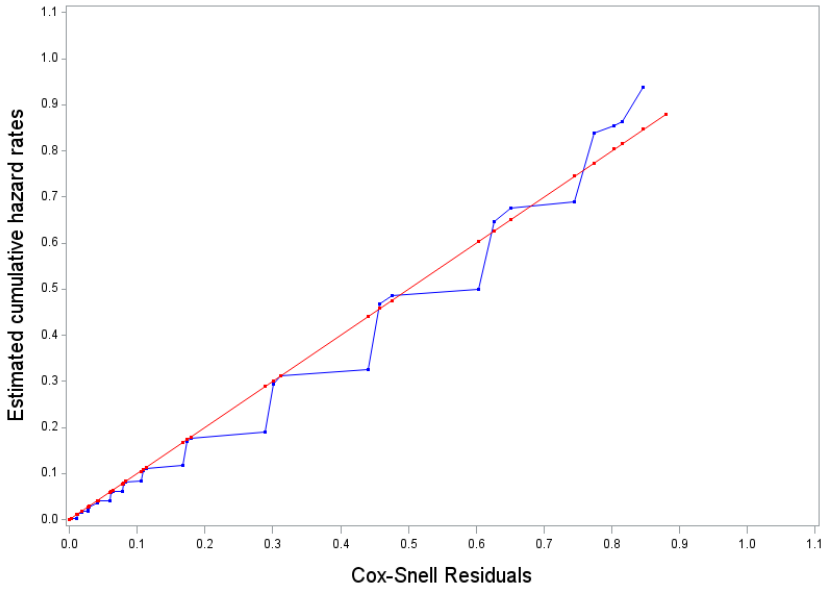

Figure B.20: Cox-Snell residuals for maternal education in male adolescents

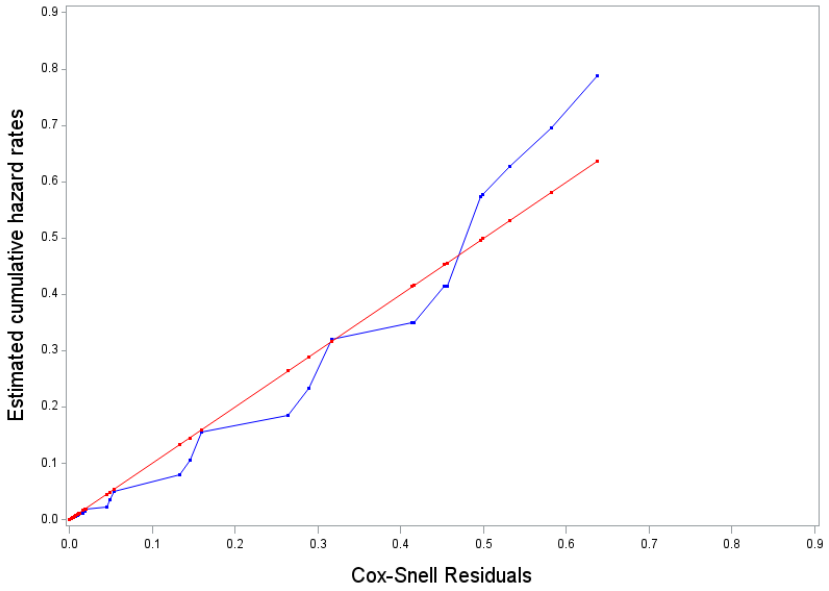

Figure B.21: Cox-Snell residuals for socioeconomic status in female adolescents

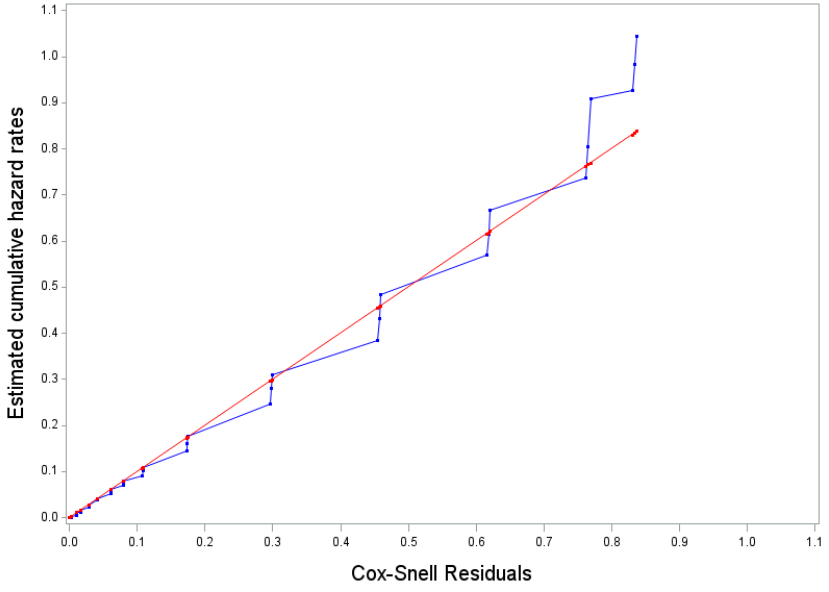

Figure B.22: Cox-Snell residuals for socioeconomic status in male adolescents

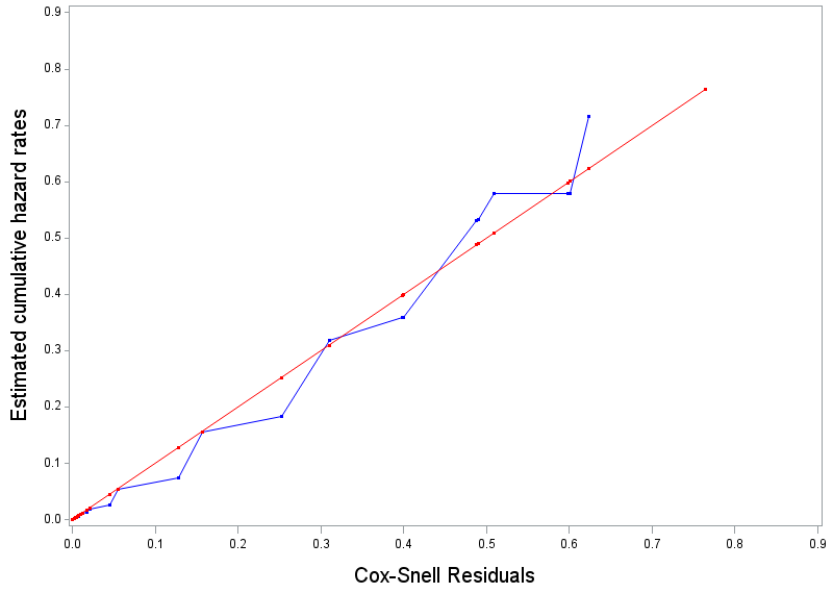

Figure B.23: Cox-Snell residuals for height in female adolescents

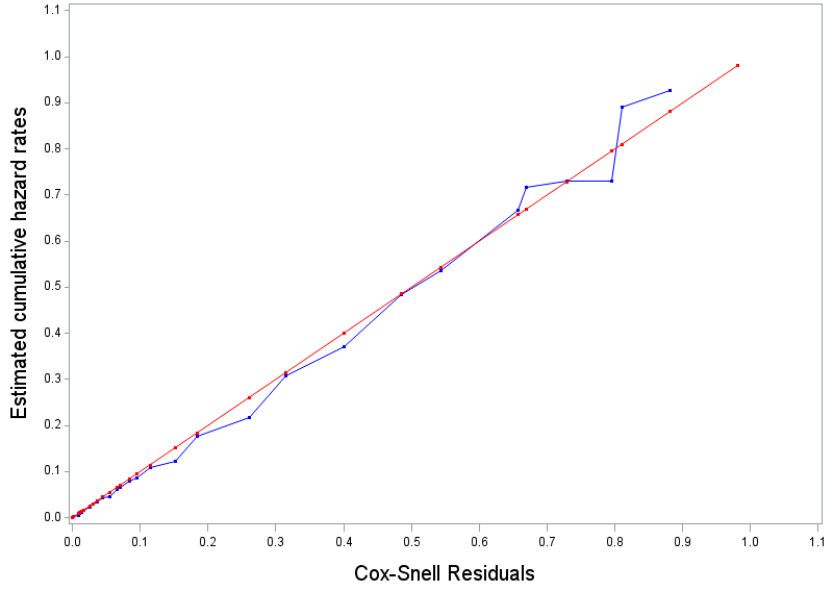

Figure B.24: Cox-Snell residuals for height in male adolescents

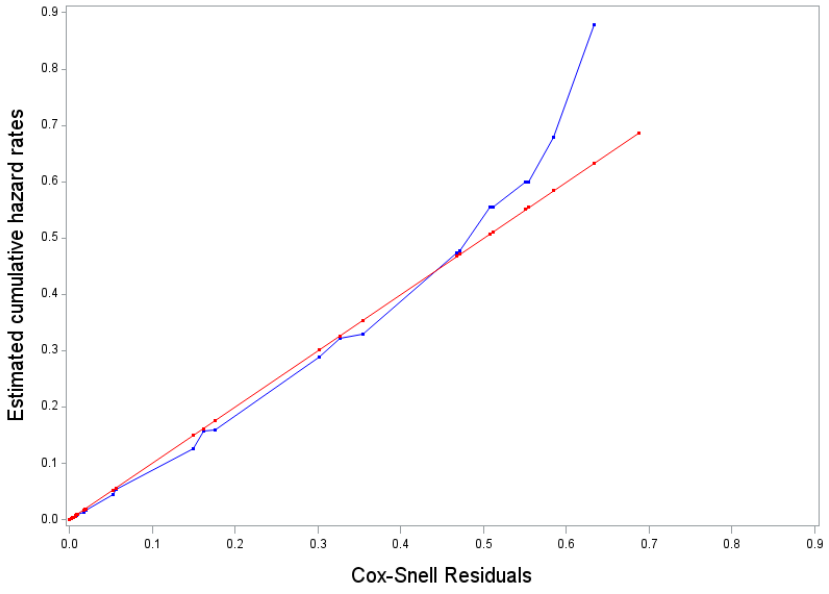

Figure B.25: Cox-Snell residuals for pubertal status in female adolescents

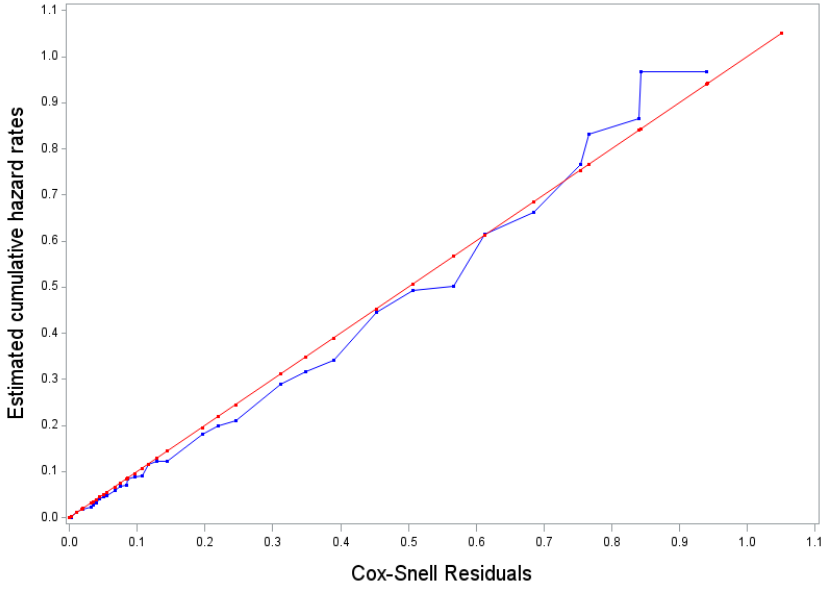

Figure B.26: Cox-Snell residuals for pubertal status in male adolescents

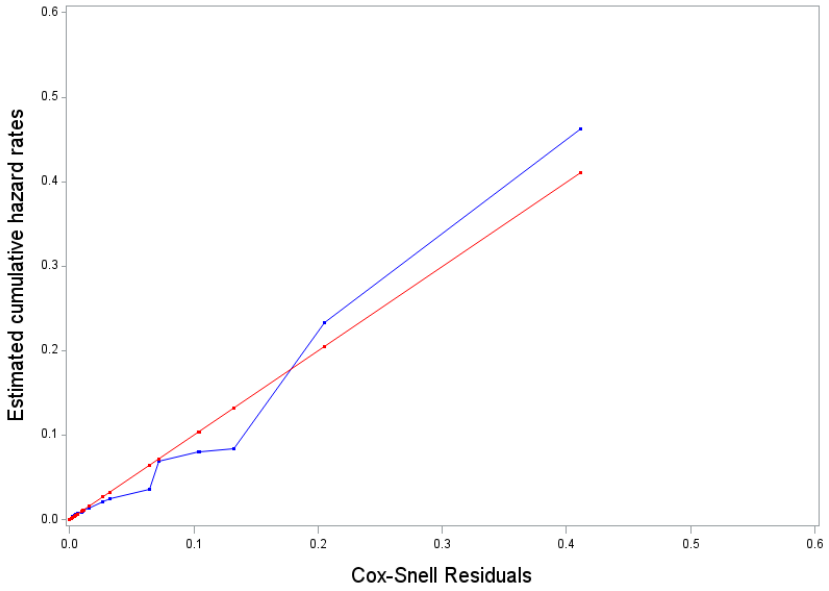

Figure B.27: Cox-Snell residuals for foreplay in female adolescents

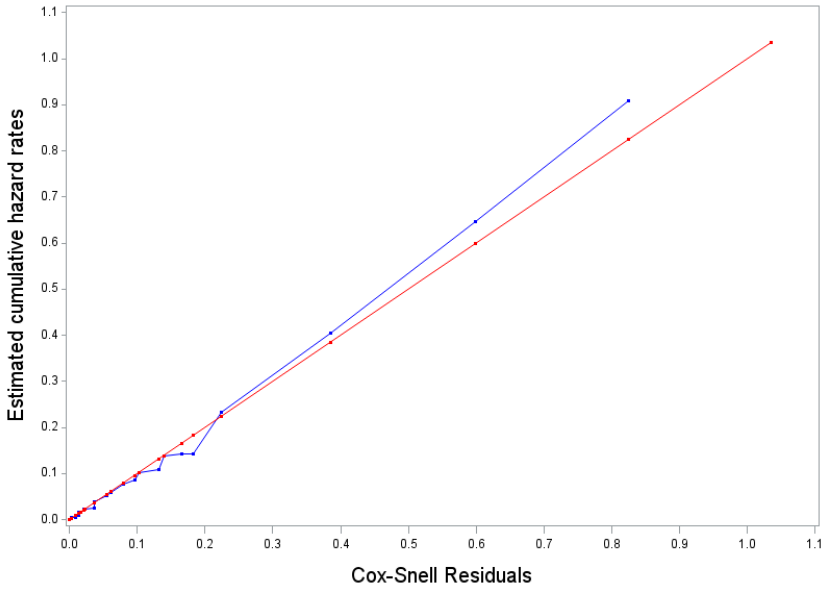

Figure B.28: Cox-Snell residuals for foreplay in male adolescents

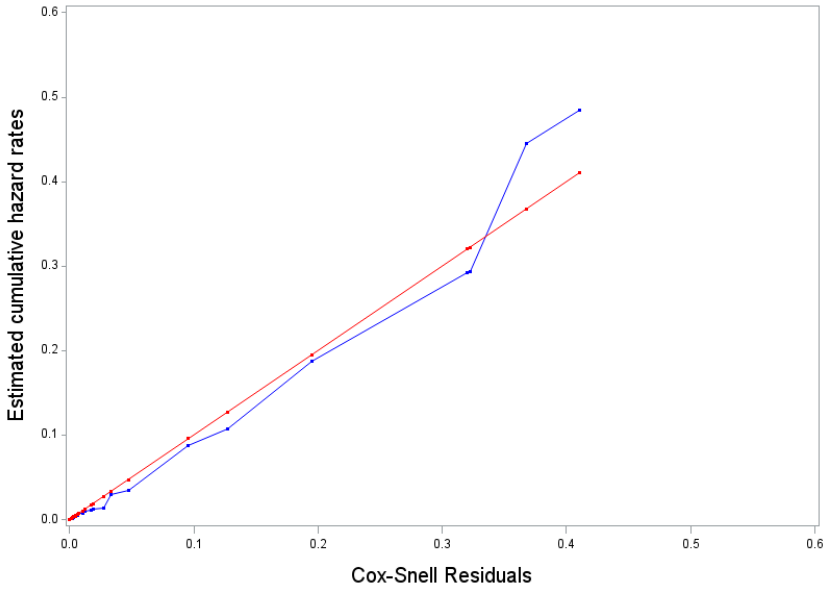

Figure B.29: Cox-Snell residuals for oral sex in female adolescents

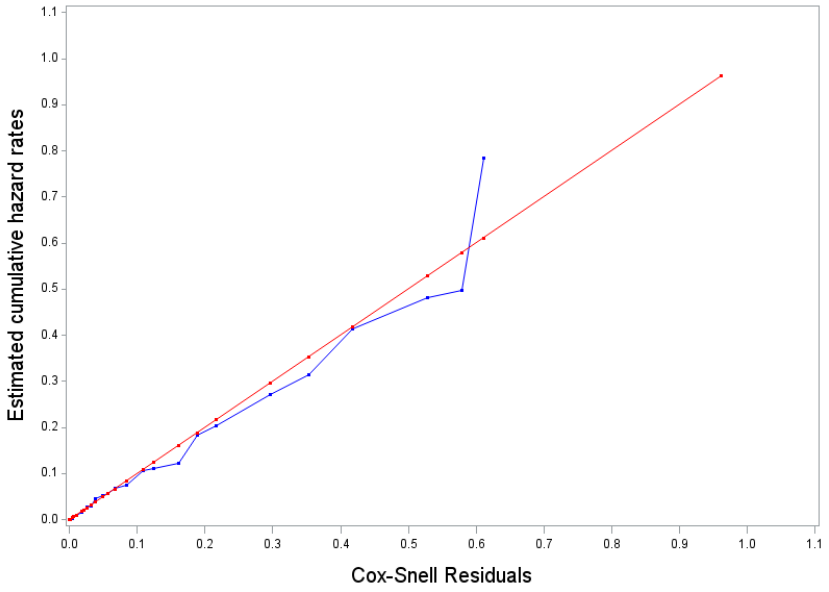

Figure B.30: Cox-Snell residuals for oral sex in male adolescents

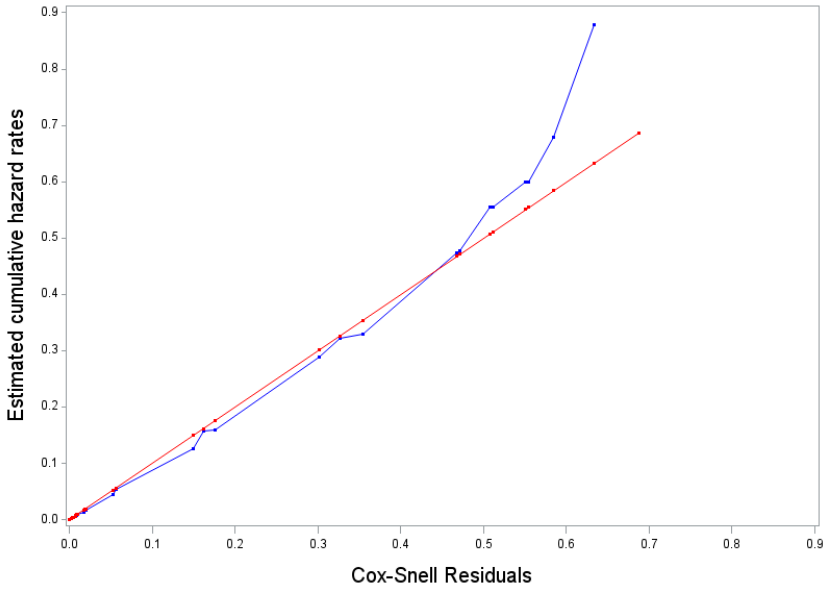

Figure B.31: Cox-Snell residuals for religiosity in female adolescents

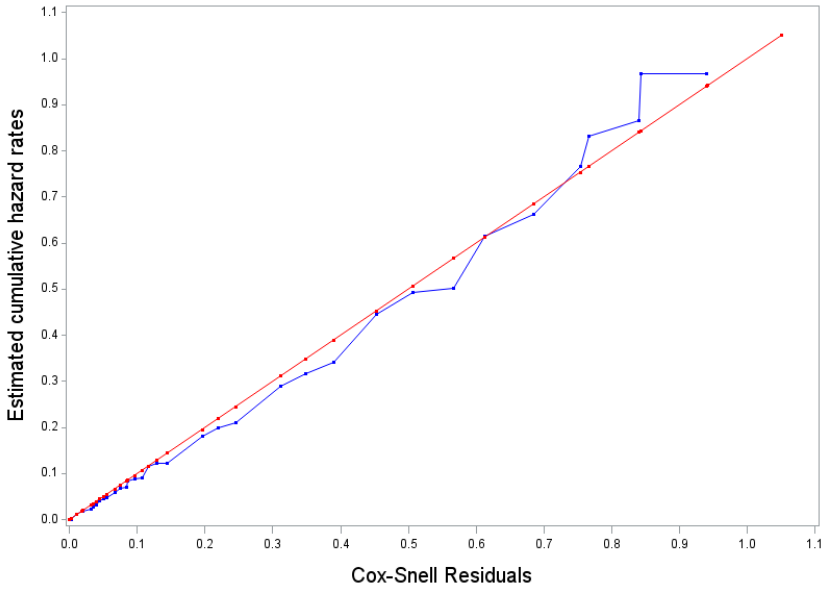

Figure B.32: Cox-Snell residuals for religiosity in male adolescents

Bibliography

Agardh, A., Odberg-Pettersson, K., & Ostergren, P.-O. (2011). Experience of sexual coercion and risky sexual behaviour among Ugandan university students. *BMC Public Health*, 11(527). Retrieved 30 July 2015 from http://www.biomedcentral.com/1471-2458/11/527.

Allison, P. D. (1995). *Survival Analysis Using SAS: A Practical Guide*. SAS Publishing.

Bagdonavicius, V., Kruopis, J., & Nikulin, M. S. (2011). *Non-parametric Tests for Censored Data*. ISTE Ltd and John Wiley and Sons.

Berry, L. & Hall, K. (2009). HIV & AIDS and STI. National Strategic Plan. Age at sexual debut.

Breslow, N. E. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, 2, 437–453.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall, 2nd edition.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

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

Crowley, J. & Hu, M. (1977). Covariance analysis of the heart transplant survival data. *Journal of the American Statistical Association*, 72, 27–36.

Dawson, D. A. (1986). The effects of sex education on adolescent bahavior. *Fam Plann Perspect*, 18(5), 162–170.

de Onis, M., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., & Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, 85(9), 660–667.

Dignam, J. J., Zhang, Q., & Kocherginsky, M. N. (2012). The Use and Interpretation of Competing Risks Regression Models. *American Association for Cancer Research*, 18, 2301–2308.

Fine, J. P. & Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446), 496–509.

Fox, J. (2006). Introduction to Survival Analysis. Retrieved 5 September 2013 from http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf.

Fürstovà, J. & Valenta, Z. (2011). Statistical Analysis of Competing Risks: Overall Survival in a Group of Chronic Myeloid Leukemia Patients. *European Journal for Biomedical Informatics*, 7(1), 2–10.

Harrison, A., Cleland, J., Gouws, E., & Frohlich, J. (2005). Early sexual debut among young men in rural South Africa: heightened vulnerability to sexual risk? *Sex Transm Infect*, 81, 259–261.

Hogan, D. & Kitagawa, E. (1985). The impact of social status, family structure and neighbourhood on fertility of black adolescents. *American Journal of Sociology*, 90(4), 825–855.

Hosmer, D. W. & Lemeshow, S. (1999). *Applied Survival Analysis*. John Wiley and Sons.

IBM Corp (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY:IBM Corp.

Joint United Nations Programme on HIV/AIDS (2013). Global report.

Kalbfleisch, J. D. & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60, 267–278.

Kaplan, E. L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Kleinbaum, D. G. & Klein, M. (2005). *Survival Analysis*. Springer, 2nd edition.

Lammers, C., Ireland, M., Resnick, M., & Blum, R. (2000). Influences on Adolescents' Decision to Postpone Onset of Sexual Intercourse: A Survival Analysis of Virginity Among Youths Aged 13 to 18 Years. *Journal of Adolescent Health*, 26, 42–48.

Lee, E. T. & Wang, J. W. (2003). *Statistical Methods for Surival Data Analysis*. John Wiley and Sons, 3rd edition.

Manlove, J., Romano, P. A., & Ikramullah, E. (2004). Not Yet: Programs to Delay Sex Among Teens.

Marsiglio, W. & Mott, F. L. (1986). The impact of sex education on sexual activity, contraceptive use and premarital pregnancy among American teenagers. *Fam Plann Perspect*, 18(4), 151–161.

Marubini, E. & Valsecchi, M. G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley and Sons.

McGrath, N., Nyirenda, M., Hosegood, V., & Newell, M.-L. (2009). Age of first sex in rural South Africa. *Sexually Transmitted Infections*, 85, i49–i55.

Mueller, T. E., Gavin, L. E., & Kulkarni, A. (2008). The Association Between Sex Education and Youth's Engagement in Sexual Intercourse, Age at First Intercourse, and Birth Control Use at First Sex. *Journal of Adolescent Health*, 42, 89–96.

Muula, A. S. (2008). HIV Infection and AIDS Among Young Women in South Africa. *Croatian Medical Journal*, 49(3), 423–425.

Nnadozie, R. C. (2013). Access to basic services in post-apartheid South Africa: What has changed? Measuring on a relative basis. *The African Statistical Journal*, 16, 81–103.

Nnko, S., Boerma, J. T., Urassa, M., Mwaluko, J., & Zaba, B. (2004). Secretive females or swaggering males? An assessment of the quality of sexual partnerships reporting in rural Tanzania. *Social Science amd Medicine*, 59(2), 299–310.

Pettifor, A. E., Rees, H. V., Kleinscmidt, I., Steffenson, A. E., MacPhail, C., Hlongwa-Madikizela, L., Vermaak, K., & Padian, N. S. (2005). Young peoples's sexual health in South Africa: HIV prevalence and sexual behaviors from a nationally representative household survey. *AIDS*, 23(19), 1525–1534.

Putter, H., Fiocco, M., & Geskus, R. B. (2006). Tutorial in Biostatistics: Competing Risks and Multi-State Events. *Statistics in Medicine*, 26(11), 2389–2430.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., & Udry, J. R. (1997). Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *JAMA*, 278(10), 823–832.

Richter, L., Norris, S., Pettifor, J., Yach, D., & Cameron, N. (2007). Cohort Profile: Mandela's children: The 1990 birth to twenty study in South Africa. *International Journal of Epidemiology*, 36, 504–511.

Richter, L. M., Panday, S., Emmett, T., Makiwane, M., du Toit, R., Brookes, H., Potgieter, C., Altman, M., & Makhura, M. (2005). *The Status of the Youth Report* 2003. Technical report, Human Sciences Research Council.

Rogol, A. D., Clark, P. A., & Roemmich, J. N. (2000). Growth and pubertal development in children and adolescents: effects of diet and physical activity. *American Journal of Clinical Nutrition*, 72(2), 521s–528s.

Sandfort, T. G. M., Orr, M., & Santelli, J. (2008). Long-Term Health Correlates of Timing of Sexual Debut: Results From a National US Study. *American Journal of Public Health*, 98(1), 155–161.

SAS Institute Inc (2002-2004). SAS 9.3. Cary, NC: SAS Institute.

Schoenfeld, D. A. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239–241.

Smith, T., Smith, B., & Ryan, M. A. K. (1964). Survival Analysis Using Cox Proportional Hazards Modeling For Single And Multiple Event Time Data. *Illinois Journal of Mathematics*, 8, 248–251.

StataCorp (2013). Stata Statistical Software: Release 13. College Station, TX:StataCorp LP.

Statistics South Africa (2006). *Mid-year Population Estimates*. Pretoria: Statistics South Africa.

Statistics South Africa (2014). *Mid-year Population Estimates*. Pretoria: Statistics South Africa.

Tsiatis, A. & Zhang, D. (2005). Analysis of Survival Data. Retrieved 16 January 2014 from http://www4.stat.ncsu.edu/ dzhang2/st745/chap1.pdf.

United Nations Publication (2011). Opportunity in Crisis: Preventing HIV from early adolescence to young adulthood.

Zaba, B., Pisani, E., Slaymaker, E., & Boerma, T. (2004). Age at first sex: Understanding recent trends in African demograhpic surveys. *Sexually Transmitted Infections*, 80, ii28–ii35.