Modelling a cluster randomised controlled trial to evaluate the impact of a cash incentivised prevention intervention to reduce teenage pregnancy among high school learners in rural KwaZulu-Natal, South Africa

A thesis presented to

The University of KwaZulu-Natal

in fulfilment of the requirement for the degree

of

Master of Science in Statistics

by

Mbali Charity Mlangeni

School of Mathematics, Statistics and Computer Science

2013

Abstract

A third of adolescent girls fall pregnant before the age of 20 in South Africa, and this happens despite contraceptives and condoms being free and mostly accessible.

This thesis aims to evaluate the impact of a cash conditional incentivised prevention intervention on teenage pregnancy through the use of appropriate statistical methods that take into account the clustering effect as well as the binary nature of the response variable. This thesis will focus on the analysis of interim data of the study which was collected at baseline and first follow up.

Fourteen schools were allocated to two study arms (intervention and control arm), totalling to 1412 teenage girls who were followed up annually for 3 years. Participants in the intervention arm received the conditional cash transfers while those in the control arm did not. The intervention arm comprised of 704 girls while the control arm had 708 girls at baseline. The null hypothesis for this study states that there is no difference in the rate of pregnancy across the study arms. Urine specimen were collected to test for pregnancy. Baseline data analysis revealed an overall pregnancy proportion of 3.47% with that of 2.84% and 4.10% respectively for the intervention and control arm. These findings together with all the other findings from the applied statistical methods yielded insignificant results thus, favoring the null hypothesis. Amongst the covariates used, age and grade were multi-collinear. In all the fitted models, the age variable was statistically significant ($p<0.01$) which is indicative that this variable plays an important role in pregnancy.

From this study, a total of 280 (approximately 21%) girls missed a follow up visit. No statistical analysis was done to account for the missing data as the study was analysed at an interim thus, there is a possibility that participants might miss a particular visit but return for another scheduled visit. Based on the outcome obtained from the interim data analysis, it is evident that teenage pregnancy occurs regardless of treatment arm thus, we conclude that cash conditional transfers does not exclusively prevent teenage girls from falling pregnant.

Declaration

I, Mbali C. Mlangeni, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis 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 re-written 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 thesis 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.

Mbali C. Mlangeni (Student)	Date
Prof. Henry Mwambi (Supervisor)	Date
Ms Anneke Grobler (Co-supervisor)	Date

Acknowledgements

Compiling this project has been a tough, challenging but very insightful journey, one that I could not have accomplished without the assistance and support of a number of people. Therefore, I want to thank my supervisor Professor H.G Mwambi and co-supervisor Ms Anneke Grobler for their dedication, support, guidance and expertise in helping me make this project a reality. I am most appreciative for the opportunity to have been supervised by such brilliant minds, I can never say enough- I hope "thank you" will suffice.

Many thanks to the Centre for the AIDS Programme of Research In South Africa (CAPRISA) for giving me the opportunity to use their world class data for the purposes of this thesis. I would also like to express my deepest gratitude to this organization for the fellowship that was awarded to me and for the invaluable experience I gained from working as a statistics fellow. In the same breath; I also want to express my sincere appreciation to Ms Anneke Grobler for mentoring me at CAPRISA during my fellowship. I owe a great deal of thanks to Mrs Nonhlanhla Yende-Zuma and the entire CAPRISA team for being so supportive and very influential. Thank you all very much, this project would not have been what it is today!

Special thanks to my late mother for her unconditional love and support, though she did not live long enough to see this project reach completion. Her prayers, wisdom and life lessons has brought me this far, I thank her for being my pillar of strength. To my wonderful sisters (Sinegugu and Nokubongwa), I thank you for believing in me. Thanks must also be expressed to Ms Aminata Lima and

all my friends and family for their remarkable support and love through both the difficult emotions and passions of making this project a success.

Last but not least I would like to express my heartfelt gratitude to the Almighty God for sustaining and providing me with the strength I needed to conquer all challenges thus far.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Globally, there are an estimated 1.2 billion children and adolescents aged between 10 and 19 (WHO, 2011). These individuals constitute about 18% of the world's population and the majority of them are females (WHO, 2011). The World Health Organization (WHO) estimates that 16 million girls, aged 15 to 19 give birth each year, accounting for roughly 11% of all child births worldwide (WHO, 2011). These appalling statistics are alleged to be primarily due to unprotected sex, no contraception use and a lack of sex education in teenagers.

In 2011, the World Health Organization reported that about 95% of teenage pregnancies occur mainly in the low and middle income countries (WHO, 2011). A comparison study showed that middle income countries have an average proportion of adolescent birth rates that is twice as high when compared to that of high income countries, while low income countries exhibit five times as high adolescent child births rates as that of high income countries (WHO, 2009). Klein (2005) estimated that 78.9% of teen births occurred outside of marriage. Over 50% of child births to teenage mothers in the world occur in sub-Saharan Africa (SSA), while Latin America and the Caribbean contribute about 18% and China has just about 2% of these child births to teenage mothers (WHO, 2009).

Several factors are understood to be associated with teenage pregnancy. These factors include; peer pressure, limited educational and employment prospects

(Mwaba, 2000), lack of knowledge and access to contraceptives and condoms as well as a lack of sex education (Panday et al., 2009; Richter et al, 2005). In South Africa, a third of adolescent girls fall pregnant before the age of 20 despite condoms and contraceptives being free and mostly accessible (Wood et al., 2006). Majority of South African teenage girls are fully aware of the facilities that offer these services and products free of charge, however, research shows that teenage girls fear accessing such facilities to get condoms or contraceptives as they claim defamation, insult and bad attitude from the health staff judging them of engaging in sexual activities at a young age (Wood et al., 2006). Pregnant teenage girls face a host of challenges as a result of falling pregnant, some of these setbacks include: incomplete education, unemployment, poverty, social embarrassment and numerous other social and emotional distress (Richter et al, 2005). Dropping out of school for most teenage girls is usually associated with pregnancy. When teenagers drop out of school, this endangers their future and wellbeing as a large number of them already live in poverty and cannot support themselves or their babies. Thus, it is difficult for a school dropout to acquire good employment without attaining adequate educational requirements for that particular occupation (Panday et al., 2009).

Regardless of trends showing a decline in the rate of teen births during the late 1990's and early 2000's, teenage pregnancy, however continues to be a global public health concern (Dangal, 2006). Teenage pregnancy is the greatest contributor to the gender gap in educational attainment, particularly at the secondary school level (Dangal, 2006). In order to turn the tide some sort of interventions need to be introduced. Cash conditional intervention studies such as the PROGRESA study which focused on reducing poverty, morbidity, high teenage pregnancy rates, school dropout rates, high infant mortality rates and unhealthy living conditions have been implemented across the globe and have proved to be a huge success in other parts of the world (de Janvry & Sadoulet, 2004). The success of such studies relies greatly on monitoring and evaluation. Recently, cash conditional transfers (CCT) have been introduced as a means of providing

intervention for many different phenomena, including the reduction of teenage pregnancy.

In 2007, 29 developing countries had some CCT programmes in place, with many other countries planning or piloting them (World Bank, 2009). The CCT programmes focus mainly on reducing poverty in poor households. In CCT, beneficiaries are awarded cash on condition that they meet a certain criterion, accomplish a particular task or behave in a certain way. The success of CCT is witnessed through the early CCT pilot studies such as Mexico's PROGRESA, Brazil's Bolsa Escola and Nicaragua's Red de Proteccion (World Bank, 2009). These programmes pioneered CCT schemes and became national programmes a few years later. CCTs have the potential to prevent or delay risky sexual behaviour among teenage girls and young women in SSA (Baird et al., 2009). In many interventions involving school children, the progress of CCTs has been observed mainly in school enrolment and attendance (Baird et al, 2009). With education suggested as a "social vaccine" to change sexual behaviour and prevent the spread of sexually transmitted infections (Jukes et al., 2008), it only makes sense that many of the CCT targeting young adults focus on school goers and are conducted at schools since majority of the participants of interest are highly accessible there.

Cluster randomised controlled trials (CRT) are highly recommended in CCTs or interventional studies that involve school children. This is due to their ability to prevent contamination. CRTs are study designs whereby subjects are not allocated to treatments individually, but as a group (Donner, 1998). The groups or schools in this case, are referred to as clusters and become the unit of randomisation. CRTs centred around school children usually concentrate on outcomes related to education, intervention and/or community education (Donner, 1998). According to Donner and Klar (2000) CRTs are increasingly being used in health services research and in primary care; however, the main concern is that the majority of these trials do not account appropriately for the clustering effect in their analyses. In individually randomised trials, the outcome for each participant is

assumed to be independent of the outcome of any other participant. In a CRT, participants in a cluster are assumed not to be independent, known as correlation. Participants in a cluster are more likely to have similar outcomes (Donner, 1998).

This thesis seeks to address the impact of a CCT prevention intervention aimed at reducing teenage pregnancy among high school learners. The thesis is structured as follows; Chapter 1 provides a brief review of the literature on teenage pregnancy, CCT programmes and CRT, while Chapters 2 looks at the data description and baseline characteristics. Chapters 3 and 4 looks at the different approaches suitable for analysing CRTs with a binary response variable. These methods include generalised linear models (Wedderburn & Nelder, 1972; McCullagh & Nelder, 1986), generalised linear mixed models (Breslow & Calyton, 1993) and multilevel modelling (Golstein, 2011). Chapter 5 concentrates on applying these methods to the data and Chapter 6 concludes the findings of this thesis.

1.1 Literature review

1.1.1 Teenage pregnancy

According to the World Health Organisation teenage pregnancy can be defined as a teenage girl, usually within the ages of 13 to 19, becoming pregnant (WHO, 2006). The term teenage pregnancy is widely used to mean unmarried adolescent girls who become pregnant (Klein, 2005). When referring to young people, often the terms "adolescent" and "teenager" are used. According to the Department of Health (DOH), these terms both describe the development stage between childhood and adulthood; they describe a time when individuals in this age group (13 to 19 year olds) start to experiment with adult behaviours (DOH, 2008). Some of these adult behaviours are good, such as part time work or helping around the house whilst others may be bad and not acceptable, such behaviours include; smoking, drinking alcohol and sexual relationships to name but a few. Nonetheless, the adolescent stage can be a time of profound change or extreme turmoil

depending on the behaviour of the adolescent (Panday et al., 2009).

Becoming a parent is a major event in a person's life and it becomes more important when it occurs early on (Klein, 2005) especially in a time where HIV and AIDS is recognised as the primary reproductive health concern for adolescents (Shaw et al., 2006). According to Shaw et al., teenage pregnancy remains a common social and public health concern worldwide, affecting nearly every society (Shaw et al., 2006). However, public health literature and family planning services treat pregnancy and HIV as distinct, even though they share a common predecessor of unprotected sex (Shaw et al., 2006). Research shows that pregnancy and lactation increases the susceptibility to HIV infection by induced immunological changes (Gray et al., 2005), thus teenage mothers, and in particular, pregnant teenagers represent an important target group for HIV prevention.

Antenatal data from the South African Department of Health shows that 12.9% of 15 to 19 year old pregnant teenagers are HIV positive (DOH, 2008), thus showing an association between pregnancy and HIV infection in South Africa (S.A.). The hazard of both events, infection and pregnancy, seem to be high in the teenage group. In survival analysis terms this presents a competing risk phenomenon which can be catastrophic if not abated. Therefore, HIV and pregnancy are the most critical threats to the health and overall wellbeing of teenagers in S.A., further making it imperative to understand teenage pregnancies and the pattern of high risk sexual activities that these adolescents are engaging in. According to Harrison (2008), from age 17 onwards, every second teenager who has been pregnant is infected with HIV. Consequently, the Millennium Development Goals (MDGs) defined by Heads of State in 2000 firmly placed maternal health on the international agenda by identifying it as the fifth of eight goals that the world must respond to decisively by 2015 (WHO, 2000).

The alarming teenage pregnancy rates have become a driving force for researchers to investigate this particular phenomenon. Although, the adolescent childbearing prevalence has continued to progressively decline over time or stay stagnant

throughout the world (Santelli et al., 2009), teenage pregnancy still remains very high and extreme in the poorest countries, such as those of SSA (Moultrie et al., 2007). One study estimated that 90% of the pregnancies occuring in teenagers are unintended/unwanted (Jewkes et al., 2006). In 2008, Brazil's 15-19 year old adolescent girls showed a fertility rate of 56 births per 1000, while the United States had 41.5 per 1000 (WHO, 2009). The overall fertility rates in countries such as Latin America, the Carribean as well as South-Eastern Asia have since declined substantively over the past two decades (United Nations, 2008). When teenage fertility in S.A. is compared with that of many middle-income countries, it appears to be lower than other African countries, but occurs more frequently out of wedlock than in other African countries (Manzini, 2001). According to Moultrie and McGrath, the mean age at first birth has not increased, and two-thirds of teenage pregnancies are unplanned and unwanted (Moultrie & McGrath, 2007).

A status of the youth survey conducted in 2003, showed that the median age of sexual debut is between the ages of 16 and 17, whilst national and international data show that fertility increases with age (Richter et al., 2005). Other studies conducted in Africa also confirm this relationship. Teenage pregnancy rates in Kenya double from 17% at ages 15 to 16 to 34% at ages 17 to 18 (Were, 2007). A small number of studies have conducted rural/urban comparisons of teenage pregnancy, this is due to the variations in how urban and rural areas are defined; thus making interpretation and comparability difficult. In addition, high pregnancy rates are observed in rural areas and in schools located in poorer neighbourhoods than in urban areas, this was established in the KwaZulu-Natal (KZN) transitions to adulthood study (Manzini, 2001). Hence, black and coloured South African adolescents are most affected as they reside mostly in rural areas. In S.A. teenage pregnancies are more prevalent in the Eastern Cape, Limpopo and KZN necessitating interventions in these areas (DOH, 2008). Despite pregnancy rates declining, the high pregnancy rate in teenagers is a serious problem in S.A. Pregnant teenagers face serious health, socio-economic and educational challenges

(Manzini, 2001).

The reason(s) why teenage girls become pregnant are difficult to categorise. Customs and traditions that lead to early marriage, lack of education and information about reproductive sexual health, lack of access to tools that prevent pregnancies, peer pressure to engage in sexual activity, incorrect use of contraception, sexual abuse, poverty, exposure to abuse and violence at home, low self esteem, low educational ambitions and goals are some of the factors that leads to teenage pregnancy (Panday et al, 2009; Klein, 2005; Imamura et al., 2007). Though, sexual behaviour in teenagers is influenced by many factors, most teenagers report "curiosity" and peer pressure as reasons for initiating in sexual activities.

Education has been suggested as a "social vaccine" for teenagers to change sexual behaviour and prevent the spread of HIV (Jukes et al., 2008). With the South African schooling system characterised by both high enrolment and high rates of repetition, dropout, late entry and re-entry means that a significant number of older learners, well past the onset of puberty, can be found in lower grades (Schindler, 2008). As a result, schools have had to accommodate traditionally high rates of teenage pregnancy. Studies have reported that over a third of girls below 19 years of age who had an early pregnancy were attending school in 1993 (Maharaj et al., 2000). A similar trend was seen in KZN in 2001 (Hallman & Grant, 2003). Imamura et al., (2007) suggests that when the relationship with schooling is weak, either through dislike of school, poor academic achievement or poor expectations of furthering education, girls are more likely to become pregnant (Cassell, 2002). Pregnancy may be associated with dropout but is often not the cause of dropout. Pregnancy and school dropout share many common social and economic conditions (Lloyd & Mensch, 1999), such as poverty and poor academic achievement (Cassell, 2002; Lloyd & Mensch, 1999). Using the KZN Transitions data, Hallman and Grant (2003) showed that poor school performance is a strong indication of the increased probability of falling pregnant in school as well as of dropping out of school at the time of pregnancy.

Some girls fall pregnant and drop out of school at the same time, but most girls fall pregnant after dropping out of school (Imamura et al., 2007) often due to poor academic performance resulting in a loss of interest in school (Manlove, 1998). However, dropping out of school not only increases risk for pregnancy, it also significantly increases risk for HIV. Poverty is both a cause and consequence of early childbearing and this is particularly observed in countries of SSA that experience high levels of poverty (Kirby, 2007). Since teenage pregnancy is mostly unplanned and often coincides with other changes such as schooling, it can result in negative outcomes for the teenage mother and for the child (Kirby, 2007; Cassel, 2002).

S.A. has a burden of both high risk sexual behaviour and substance use. About a third (31.8%) of adolescents report drinking in the past month and a quarter report binge drinking (Reddy et al., 2003). Several studies have reported that between 6 to 12% of adolescents have used drugs in their lifetime (Pluddemann et al., 2008). A significant proportion (13.3%) of sexually active learners in S.A. also report using alcohol or drugs before sex (Reddy et al., 2003). Data from a Cape Town study shows that when learners use drugs (methamphetamine) they are more likely to have anal, vaginal and oral sex as well as to be pregnant or responsible for a pregnancy (Pluddemann et al., 2008). Adolescents who take drugs are more likely to engage in casual sex (Palen et al., 2006). Studies have shown that girls are less likely to fall pregnant with someone they know or a boyfriend than with a casual partner (Jewkes et al., 2001).

A concern raised in the South African society is that young women are deliberately conceiving in order to access the Child Support Grants (CSG). Research based on this concern found that 12.1% of pregnant teenagers had deliberately conceived so to access the CSG (Duflo, 2003). However, only 20% of teenage mothers are beneficiaries of these grants (Duflo, 2003). Older female relatives usually take care of the child and are often the beneficiaries of the CSG rather than the teenage mothers. During the period in which the CSG has been offered, rates of termination of pregnancy have increased, and as the CSG increased in

value, fertility rates have decreased (Duflo, 2003).

A number of HIV prevention interventions have been instituted in S.A. These include school-based sex education, peer education programmes, adolescent friendly clinics, mass media interventions and community level programmes (de Janvry & Sadoulet, 2004). The focus of these interventions has been to prevent HIV, but they could also impact teenage pregnancy because they aim to change sexual behaviour.

The following section will discuss and review some literature on the latest type of intervention known as CCT.

1.1.2 Conditional cash transfers (CCTs)

According to the World Bank, CCTs are programmes targeted at the poor with the objective of reducing poverty and vulnerability, by providing monetary (or sometimes in-kind assistance, such as food aid, shelter and support to livelihood recovery) transfers to households on the condition that they comply with some pre-defined requirements or stipulations (World Bank, 2009). Based on this definition, CCTs can be an important component of social protection policy and there is considerable evidence that they have improved the lives of poor people. There are increasingly perceived as being "a magic bullet in development" (World Bank, 2009).

CCTs are used around the world as an innovative way to deliver social assistance with two objectives: to provide poor households with a minimal income (reduce poverty in the very short run) and to invest in the human capital of future generations (reduce poverty in the long run). There is a large body of evidence supporting the success of CCTs throughout most of the developing world, particularly in relation to schooling (de Janvry & Sadoulet, 2004; Schultz, 2004). Currently, cash transfers are one of the most researched and evaluated forms of development intervention, however there are still gaps with regards to producing robust monitoring and evaluation techniques (World Bank, 2009).

CCT programmes were first initiated in the mid 90's from early pilots such as Mexico's PROGRESA, Brazil's Bolsa Escola and Nicaragua's Red de Protecc io?n, which advanced to national programmes a few years later (Baird et.al, 2009). These Latin American countries pioneered the current generation of CCT programmes and built in best-practice of high quality monitoring and evaluation which led to effective results. As of 2008, the popularity and success of these programmes led to at least 29 developing countries having some type of CCT programme in place (World Bank, 2009). Many of the CCT programmes implemented throughout most of the developing world have concentrated on increasing school enrolment and attendance (de Janvry & Sadoulet, 2004; Schultz, 2004). School attendance is not the only outcome one may want to affect or improve; sometimes the interest might be on other outcomes such as those related to sexual behaviour, drug taking or other risky behaviours (World Bank, 2009; Baird et al., 2010).

Significant positive impact on education indicators have been found to occur from CCT programmes. In Nicaragua, where primary school enrolment was low, the CCT programme increased overall enrolment by 13%, enrolment of children from the poorest households by 25% and regular primary school attendance by 20%. In Pakistan, a 2008 World Bank assessment showed that the Punjab Education Sector Reform Program increased enrolment rates for girls aged between 10 to 14 years by 11% from a baseline of 29% (Chaudhury, 2008), while the female secondary school assistance programme in Bangladesh increased the secondary school certificate pass rate for girls receiving the stipend from 39% in 2001 to nearly 63% in 2008 (World Bank, 2009). Participants in the Bolsa Fam?lia programme in Brazil are 63% less likely to drop out of school and 24% more likely to advance an additional year (Veras et al., 2007). Between 2003 and 2009, the World Bank reported that poverty in Brazil has fallen from 22% of the population to 7% and the income of poor Brazilians has grown seven times as much as the income of rich Brazilians (World Bank, 2009). The World Bank further announced the poverty reduction experience in Brazil as a dramatic success story

(World Bank, 2009). A study in Kenya finds that reducing the cost of schooling (by paying for uniforms) reduced dropout rates, teen marriage, and childbearing (Duflo et al., 2006). The success of CCT programmes has also been experienced in Malawi through a social cash transfer programme which targets households with children to go to school, this programme led to an increase in school enrolment of 5% among children aged 6 to 17, also yielding an increase of 4.2% from household with orphans (Schubert & Slater, 2006). In Malawi, the Social Cash Transfer Scheme has reduced the likelihood of female and child-headed households resorting to "risky behaviour" such as transactional sex, in order to survive (Schubert & Slater, 2006).

Robust evidence from various countries show that cash transfers have made quite an improvement in the access to health and education services, as measured by increases in school enrolment (particularly for girls) and use of health services (particularly preventative health, and health monitoring for children and pregnant women). Effects of such gains are usually larger in low income countries with lower baseline levels as compared to that of middle income countries. A social cash transfer scheme designed for adolescent girls in the Zomba district of south-eastern Malawi targeting current schoolgirls and recent dropouts to stay in or return to school demonstrated improved school attendance and decreased early marriage, teenage pregnancy, and self-reported sexual activity, and importantly decreased HIV infection rates among beneficiaries after just one year of the programme implementation (Baird,et al.,2009).

This programme also decreased participant's risk of HIV infection by 60%, compared to the control group (Baird et al., 2009). The Zomba CCT programme provides incentives in the form of school fees and cash transfers to programme beneficiaries. According to Baird et al. (2009), among programme beneficiaries who were out of school at baseline, the probability of getting married and becoming pregnant declined by more than 40 and 30%, respectively. In addition, the incidence of the onset of sexual activity was 38% lower among all programme beneficiaries than the control group (Baird et al., 2009). Overall, these results

suggest that CCT programmes not only serve as useful tools for improving school attendance but may also reduce sexual activity, teenage pregnancy, and early marriage. In relation to HIV treatment, the current evidence on the link between cash transfers and access to and use of anti-retroviral treatment (ART) is limited and requires further research. However, an important randomised controlled trial in rural Uganda found better HIV treatment adherence scores amongst programme participants than the control group. This led the researchers to conclude that "modest cash transfers of $5-8 per month to defray the costs of transportation may be an important strategy to reduce costs and improve treatment outcomes in rural, resource-limited treatment settings" (Jukes et al., 2008).

Many CCT programmes implemented target young people and are intended to improve the lives of children and adolescents. In an effort to improve the circumstances in which children from poor families start out life, the Mexican government developed an anti-poverty programme called PROGRESA, currently known as Oportunidades. The PROGRESA programme began in 1997 and it is a combination of traditional cash transfer programme with financial incentives for positive behaviour in health, education, and nutrition. Although, the PROGRESA programme has been successful in the many spheres that it covers, much of its success was observed through the educational sector, through a 6 and 9% increase for boys and girls respectively, in enrolment into secondary school (Fiszbein & Schady, 2009). Girls often dropped out before secondary school. This programme succeeded in keeping girls at school and those making the transition to secondary school increased by 15% (Fiszbein & Schady, 2009). Children in the PROGRESA programme also entered school at an earlier age and repeated fewer grades. PROGRESA had relatively little impact on school attendance rates, on achievement in standardised tests, or in bringing dropouts back to school (Fiszbein & Schady, 2009).

Unconditional cash transfers (UCT) (cash transfers that are administered to the poor or needy not based on any conditionalities) produce just as much positive results as CCTs. In S.A., school attendance rates are significantly higher in house-

holds receiving UCTs, including pensions (Devereux et al., 2006). UCTs have also resulted in significant positive impacts on school attendance in Ethiopia, where 15% of participants in the Productive Safety Nets Programme (PSNP) spent some of their unconditional transfer on education (Devereux et al., 2006). In Lesotho, studies have shown that those receiving a social pension are using their pension grants for buying uniforms, books and stationery for their grandchildren (Samson et al., 2007). The success of UCTs raises the possibility that cash alone might be enough, with no need for conditions. This view can be debated as conditionalities create costs for governments, in monitoring, and for recipients, in demonstrating compliance. CCT programmes have achieved considerable success by making a significant impact, however, it is not yet clear what role conditionalities have played in these achievements, since the success of some UCTs raises the possibility that cash alone might be enough and there may be no need for conditions (Ozer, 2009).

There is also an increasing volume of research into how cash transfers might support "graduation" from poverty for those of working age. Evidence from Bangladesh and Ethiopia suggests that transfers are unlikely to achieve graduation without complementary interventions (e.g. skills training or agricultural extension) to promote livelihoods. Cash transfers also have a proven role in supporting specific vulnerable groups such as people living with HIV and AIDS, orphans and vulnerable children. There is some evidence (from Zambia and Namibia) that the introduction of cash transfers into poor, remote areas can stimulate demand and local market development (Baird et al., 2009).

The primary function of most cash transfer programmes is the direct and immediate alleviation of poverty and reduction of vulnerability. These programmes can also have many other benefits, such as improved health, nutrition or education economic productivity and growth or empowerment (especially for women). However, incentive-based CCT programmes have been criticized for paying people to engage in behaviour that they should be expected to engage in without incentives. There are concerns about the sustainability of the desired behaviour once

payments are ended. Evidence from some incentive initiatives in both health and education suggest that many positive behavioural changes are not maintained after withdrawal of the cash incentives (World Bank, 2009).

In summary, CCT programmes have the potential to deliver a range of benefits. It not only reduces extreme poverty, but also provides effective support for broader human development objectives, including better nutrition, health and education outputs and outcomes, but high quality monitoring and evaluation is needed for optimal results. More research needs to be done on the area of CCTs and sexual behaviour. This thesis examines the impact of CCTs on teenage pregnancy.

Since the study is conducted in a school setting, cluster randomised controlled trials are reviewed below as it is the design used mostly in studies comprising of grouped data.

1.1.3 Cluster randomised Controlled Trials (CRTs)

CRTs are trials in which groups or clusters of individuals (subjects) rather than individuals themselves are randomised (Murray, 1998). The unit of randomisation might be school, community, worksite or hospital, etc.

CRTs were established as early as 1940 by Lindquist during a school based study where he wanted to use a design method that was able to avoid contamination (Lindquist, 1940). Since this was a school study, Linquist decided to randomise the classes to the treatment/intervention rather than to randomise individual learners to treatment, thus looking at the classes as groups (clusters). This idea proved to be appropriate and effective as it achieved Lindquist's primary goal; which was to avoid contamination. It was reasoned that this worked because many educational evaluations are within naturally occurring clusters (e.g. classroom or school) and therefore the only feasible approach of undertaking a CRT is to use cluster allocation. Lindquist's ideas were not initially well received (McNemar, 1940; Glass and Stanley, 1970; Oakley, 2000), with educational researchers still debating the need to account for the effects of clustering 40 years later. The

use of CRTs has increased all over the world and in almost all sectors of research, including health services, education, and many other fields. There are concerns that the majority of these trials do not account appropriately for the clustering in their analysis (Campbell, 1998). According to Donner and Klar (2000), CRTs may be the only feasible approach for some types of interventions.

CRTs have two key features, namely that clusters are often large and observations on individuals in the same cluster are usually correlated, thus special methods of design and analysis are needed for such trials (Donner & Klar, 2000). Individuals within a cluster (such as a school) are often more likely to respond in a similar manner, and thus can no longer be assumed to act independently. This lack of independence in turn leads to a loss of statistical power in comparison with an individually randomised trial. Hence, unlike individual randomised trials, CRTs do not assume that the observations on all the individuals in the trials are statistically independent; this assumption is violated the moment clusters are randomised.

Traditionally, analysis for CRTs has been focused at the cluster level; however, recent advances in statistics have led to the development of techniques which can incorporate individual level data analysis and within each approach, simple analyses such as t-tests or more complex approaches such as regression analyses may be undertaken (Donner & Klar, 2000). Both the cluster or individual level approach allows for the effect of the intervention to be tested; however, only complex analyses allow adjustment for potential covariates, such as baseline performance. A common mistake made in the analysis of CRT designs is to ignore the effect of clustering and analyse the data as if each treatment group were a simple random sample. If the clustering effect is ignored, many authors have highlighted that p-values will be artificially small, and confidence intervals will be over-narrow, increasing the chances of spuriously significant findings and misleading conclusions (Feng et al., 2001; Donner, 1998).

CRTs may be advantageous for the evaluation of specific interventions, but there

are substantial drawbacks to the use of this design. Firstly, compared with an individually randomised trial testing the same hypothesis, cluster randomisation requires a significantly larger sample size; because standard sample size calculations assume that outcomes between individuals are uncorrelated (Donner, 1998). CRTs are less costly and avoid contamination (Linquist, 1940). Although, individuals in the same cluster lack independence, the clusters themselves have an intracluster dependence. The intra-cluster correlation coefficient (ICC) is a statistical measure of the relatedness of clustered data. It accounts for the relatedness of clustered data by comparing the variance within clusters with the variance between clusters (Feng et al., 2001). Mathematically, the ICC is the between-cluster variability divided by the sum of the within-cluster and between-cluster variability and can be expressed as:

$$\rho = \frac{s_b^2}{s_b^2 + s_w^2} \tag{1.1}$$

where s_b^2 represents the variance between clusters, s_w^2 represents the variance within clusters and ρ represents the ICC. The value of the ICC ranges from 0 to 1, $\rho = 1$ if the within cluster variance s_w^2 is zero and $\rho = 0$ if the between cluster variance s_b^2 is zero. The calculation of ρ usually requires a pilot study. Standard sample size calculation for the CRT need to be inflated by a factor which is often referred to as the "design effect":

$$1 + (n - 1)\rho \tag{1.2}$$

where n is the average cluster size and ρ represents an estimate of the ICC. Both the ICC and the cluster size influence the calculation, as shown by the equation for the design effect above, therefore even small values of the ICC can have a substantial impact on power. A study of UK data sets relevant to implementation research showed that in primary care settings, the ICCs for process variables are much higher than those for outcome variables. ICCs for process variables were between 0.05-0.15, and ICCs for outcome variables were generally lower than 0.05 (Kjaergaard et al., 2001).

The analysis of CRTs must take into account the clustered nature of the data.

Standard statistical techniques are not appropriate, unless an aggregated analysis is performed at the level of the cluster, as standard techniques require data to be independent.

Although, CRTs are the most robust evaluative method, poorly conducted trials are susceptible to a number of factors that can bias their results (Donner & Klar, 2000). Methodological reviews of individually randomised trials have shown that rigorously conducted trials produce different effect estimates from those which are poorly conducted (Kjaergaard et al., 2001). CRTs are more complex to design and execute than individually randomised trials (Donner & Klar, 2000). The aim of this thesis is to apply appropriate analytical techniques to a CRT from a school based intervention study.

Chapter 2

Data description and baseline characteristics

2.1 A cash incentivised prevention intervention trial to reduce HIV infections in adolescents: RHIVA (CAPRISA 007) study

In September 2010, the Centre for the AIDS Programme of Research in South Africa (CAPRISA), together with Media in Education Trust (MiET Africa) and the Department of Education (DoE) in KwaZulu-Natal (KZN), S.A. began a CRT to evaluate the impact of a CCT intervention to reduce HIV infection among high-school learners in rural KZN. The study was given the name RHIVA (CAPRISA 007) which is an acronym for Reducing HIV in Adolescents. Ethics approval for the study was obtained from the University of KwaZulu-Natal Biomedical Research Ethics Committee (BREC). Study duration was 3 years and follow up data was collected at 2 annual study visits in the form of a structured questionnaires as well as biological specimen.

The RHIVA study was conducted in the Vulindlela district which is located about 150 km north-west of Durban in KwaZulu-Natal. The Vulindlela district is home to about 400 000 residents, with limited resources, infrastructure and employment

opportunities accounting for high levels of poverty and unemployment. Health services for the Vulindlela district residents are provided by seven public sector primary health care clinics with the closest referral hospitals approximately 30 kilometres away.

2.2 Study design

A spatial map of all 42 secondary (high) schools within the Vulindlela School Circuit boundaries was generated. These schools were then audited for study eligibility using a structured questionnaire created by MiET Africa. Based on the audit, 14 secondary schools out of the 42 secondary schools met all the selection criteria. The key selection criteria for the schools included school enrolment numbers in Grades 8 and 9 in 2009, matric pass rate in 2009, school enrolment and proximity to a primary health care clinic, and school infrastructure. Randomisation to the two study arms was then carried out on the 14 selected schools. The RHIVA study is powered at 80%, with alpha at 0.05, a 0.25 coefficient of variation between the clusters and a design effect of 1.6. A 50% reduction in HIV incidence rate in the intervention schools is expected with about 140 eligible female learners in each school to be included in each of the study arms or a total of approximately 1200 female learners per study arm.

2.2.1 Randomisation

The schools were the unit of randomisation and were randomised to either the intervention or control arm in a 1:1 ratio. Each study arm consisted of seven schools or rather clusters. Once the schools were assigned to treatment arms, the enrolled cohort in each school was locked and no participants were allowed migration from control schools to intervention schools or vice versa as this might later bias the results and bring about confusion due to the "contamination" effect between schools which are the primary study units. Although, this is a CRT with schools representing the clusters, data was collected at an individual level, grouped by schools, and analysed by study arm.

2.2.2 Inclusion criteria

Study participants had to meet all of the below mentioned criteria in order to be considered eligible for inclusion in the study:

- Male or female learner in Grade 9 or 10 in one of the 14 selected schools
- Willing and able to provide informed consent and/or assent to participate in the study
- Willing to provide locator data for home visits if necessary
- Not planning to move to another school or relocate in the next 36 months
- Willing to be finger-printed to verify identity for study procedure purposes
- Willing to complete all study procedures

2.2.3 Exclusion criteria

Potential study participants were excluded from the study based on the following exclusion criteria:

- Refusal by the learner and/or parent or legal guardian to participate in the study
- Unable to provide necessary informed consent
- Cognitively challenged learners

2.2.4 Recruitment and enrolment

General information sessions were held for all volunteers in the selected schools and grades as part of the study recruitment process. These sessions were aimed at providing volunteers with information on HIV/AIDS risk, prevention and treatment options as well as information on the RHIVA study; which included a brief report on study design, study procedures, frequency of study procedures and duration of the study. Similar sessions were held for parents and community

members in order to bring awareness about the proposed study and to make them mindful of the role they would have to play should their children become potential participants.

Volunteers agreeing to enrol into the study after attending the information sessions were enrolled into the study once all the necessary consents were provided. Most of the study participants were minors, below the age of 18 which in South Africa is regarded as the age of maturity, thus their parents/guardians also signed the informed consent forms. Procedures for participants who do not have parents or guardians were followed.

During enrolment, study participants were also assigned a 6 digit unique number known as the patient identification number (PID). The primary function of the PID was to provide confidentially by making participants anonymous by way of being identified by the PID instead of their own personal details.

2.3 Data collection

Data was collected from study participants in the form of structured questionnaires and biological specimens at every study visit. The analysis presented in this thesis is an interim analysis. A dipstick pregnancy test was performed on the urine sample of female participants to test if these participants were pregnant or not. As part of the intervention, study participants who were identified as being pregnant were referred to a local antenatal care clinic. Follow-up assessments were undertaken about 12 months apart. During follow-up visits, data was collected the same way using a combination of self-reported data on a standard questionnaire and biological specimens.

2.4 Study objectives

2.4.1 Primary endpoint

The primary endpoint for the RHIVA study is to compare the incidence of HIV infection between the study arms. This will be assessed by measuring HIV status in all learners at baseline (pre-intervention), and twice thereafter (12 months apart). The date of HIV infection will be estimated as being the midpoint between the last negative HIV test result and the first positive HIV test result. Learners who test HIV positive at baseline will not contribute to the incidence rate calculation as they would have already reached the endpoint.

2.4.2 Secondary endpoints

Secondary endpoints included comparing pregnancy and substance use rates between the arms. The standardised, structured questionnaires were used to assess self-reported rates of:

- Condom use
- Primary and secondary abstinence
- Age of sexual debut
- Number of concurrent sex partners
- Frequency of partner change
- Medical male circumcision
- Rates of anal sex
- Sexually-transmitted infection
- Intergenerational sexual coupling
- Non-barrier method contraceptive use

This thesis focuses on evaluating the impact of the CCT on one of the secondary study endpoints, namely pregnancy. Thus, pregnancy is the primary objective in this thesis and it will be analysed at baseline and first follow up visit.

2.5 The RHIVA study programme

This study intervened using a combination prevention approach that is an incentivised structural and behavioural intervention. The foundation of the intervention is an enhanced essential package complemented with a sustainable livelihood component that culminates with voluntary uptake of HIV testing in the context of pre- and post-test counselling. The essential package comprise of a seven-module package which includes a life skills programme that is delivered to all selected schools, regardless of study arm.

All enrolled participants were encouraged to participate in a programme known as Sustainable Livelihood Programme (SLP). The SLP is also known as My Life! My Future! Programme and it is an extra-mural activity facilitated by peer/youth workers. This programme comprised of weekly one hour sessions. The overall goal of the programme was to give learners a positive view of self and a greater sense of future (through increased engagement in relevant life skills). The programme was also a combination of theory and practical instruction which included:

- financial management skills training for learners
- business plan development
- gardening to enhance food security
- conducting a community needs and resource audit
- health and well-being (sexual reproductive health, HIV/AIDS, substance abuse, mental health, nutrition)
- entrepreneurship to improve learner self-esteem
- gain a more positive view of the future

- be given the ability to identify new opportunities, and the commitment and drive to pursue them

The aim of My life! My future! programme is to provide basic entrepreneurial skills to study participants. All of the above mentioned were quality assured by life orientation (LO) educators, MiET Africa's training coordinators and identified KwaZulu-Natal DoE officials. Participating learners were furnished with a list of clinics they could access services for voluntary counselling and testing (VCT) of HIV. This list had detailed information on clinic hours and schools closest to each clinic. Learners were encouraged to self-initiate the HIV testing as this was anticipated to lead to a decrease in high-risk sexual behaviours from these learners.

2.5.1 What will be incentivised

Learners in the intervention arm only were awarded cash incentives based on the following:

- Academic Performance- The aim of incentivising learners on this task was to support learners to improve their academic performance, school attendance, and improve their self-esteem while helping them to complete schooling. In order to receive cash for this incentive, learners had to pass their June and December examinations with an average mark of at least 50%. Educators from the selected schools were required to design, administer and mark the June and December examination papers. The cash payment was made in instalments of R150 twice a year to learners who achieved at least 50% in each of the June and December examinations.

- Sustainable livelihoods- Payment for participation in the programme was linked to learner attendance; completing a portfolio which includes a community audit report, business plan and evidence of having implemented a project. In order to receive the incentive, learners had to:

    i. attend 80% of the My Life! My Future! SLP sessions

ii. identify a project and develop a business plan that includes a community audit in terms of needs, existing resources and opportunities, and implement the identified project

iii. complete portfolio showing evidence that the project was implemented.

For the My Life! My Future! programme, attendance was measured and recorded by taking a register at each session. The completed portfolio was assessed (MiET Africa provided an assessment guide) and signed off by a peer educator. The total amount of the SLP incentive was R400 and payment was issued as follows: A quarterly payment of R50 (in total R200) for attendance; an annual payment of R200 for the completion of a portfolio which include: R50 for completing the community audit, and identifying a project; R50 for developing a business plan; R100 for implementing the project. Both arms participated in the My Life! My Future! Program but only intervention arm learners received cash incentives.

- Voluntary uptake of HIV testing services- Learners had to provide to their educators a receipt from the clinic that specifies that the learner had tested for HIV. A once off payment of R200 for voluntarily utilising a VCT service to establish HIV status was paid to qualifying learners within the quarter that they were tested. This payment was made once a year only during the course of the study.

2.5.2 Statistical methods appropriate for the outcome of interest: Teenage Pregnancy

Pregnancy is the outcome of interest in this thesis. This response variable is binary in nature (a female learner is either pregnant or not pregnant at follow up time). Our study population is only female learners enrolled in the RHIVA study since male learners cannot experience the outcome of interest. Since we are analysing interim data at only two time points, the date in which a participant fell pregnant will be estimated between the last negative pregnancy test result and

the first positive pregnancy test result. This will be done using the midpoint rule. A SAS code was prepared before the analysis to specify and check the possible pregnancy status that might occur, that is to identify whether a participant fell pregnant more than once during the interim, terminated a pregnancy or had a pregnancy that lasted well over the normal pregnancy period of nine months. Fortunately, no such case was identified in this data. Appropriate statistical methods that take clustering and correlation into account will be utilised to analyse this data and to evaluate the effectiveness of administering a cash incentive.

2.6 Baseline characteristics

A total of 2675 male and female learners were eligibly enrolled into the RHIVA study at baseline. Of the 2675 enrolled participants, 1423 were females and 1252 were male participants. All enrolled learners were in Grade 9 and 10.
The enrolled female participants had a mean age of 16.1 with individual ages ranging from 12 to 24 years, while 20 (2.84%) and 29 (4.10%) learners, respectively from the intervention and control arm were pregnant at baseline. The overall mean age of pregnant girls at baseline was 17.0 and individual ages of the pregnant girls ranged from 14 to 21 years.

Out of the 49 pregnant females at baseline, 34 (69.4%) of them were in grade 10 whilst the remaining 15 (30.4%) were in grade 9. Baseline characteristics are shown in Table 2.1. Only 1412 of the female learners performed the dipstick pregnancy test, the remaining 11 refused to do the dipstick pregnancy test but agreed to complete the structured questionnaire and to run other laboratory tests. This led to 704 (49.9%) female participants in the intervention arm and 708 (50.1%) in the control arm; see Figure 2.1.

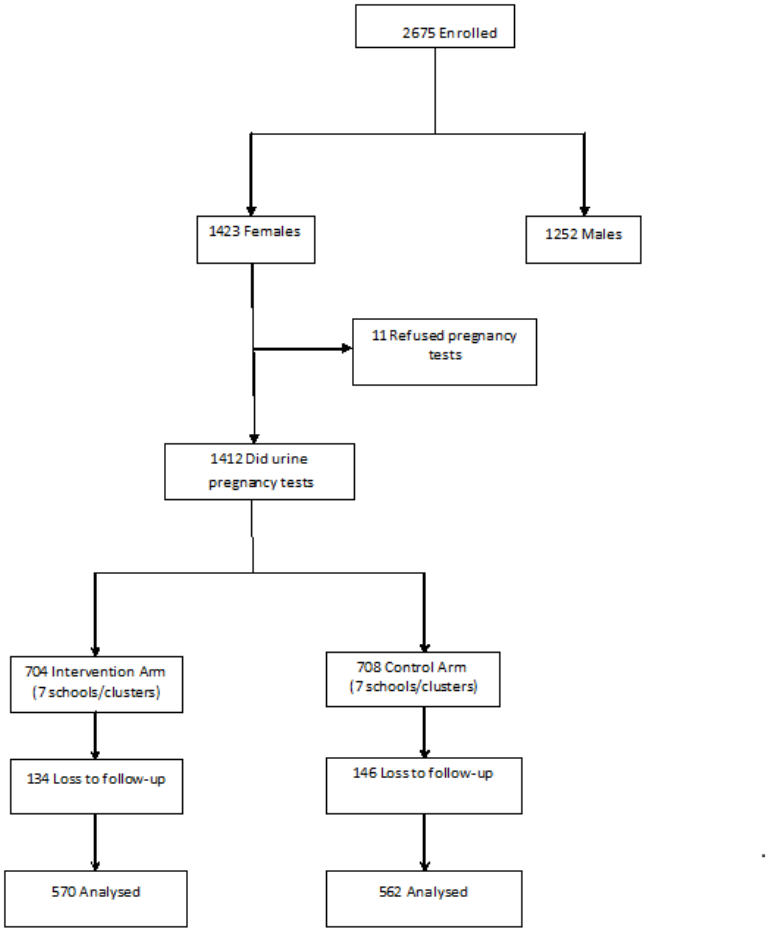

Figure 2.1: RHIVA study CONSORT diagram

Table 2.1: Baseline characteristics of female learners in the RHIVA study

Characteristic	Intervention Arm (n=704)	Control Arm (n=708)
Mean age (SD)	15.9 (1.6)	16.4 (2.0)
Grade distribution		
Grade 9 (%)	359 (51.80)	334 (48.20)
Grade 10 (%)	343 (48.04)	371 (51.96)
Pregnancy		
Proportions (%)	20 (2.84%)	29 (4.10%)

SD: Standard deviation

Chapter 3

Generalised Linear Models (GLM)

3.1 Introduction

The Generalised Linear Model (GLM) is an extension of the general linear model. In its simplest form, the GLM specifies a linear relation between the response variable and the predictor variables. General linear models assume that the observations and residuals are independent, the mean of the observations is a linear function of the explanatory variables and that observations are normally distributed with a constant variance. Although, the general linear model provides a useful framework, it is inappropriate in cases where the data modelled is not normally distributed (e.g. binary or count data) as well as in cases where the variance of the response variable depends on the mean. Hence, the GLM was developed to extend the general linear model to address both these issues where the general linear model framework is limited (McCullagh & Nelder, 1989).

GLMs were first introduced and formulated by Nelder and Wedderburn in 1972 as a way of unifying various statistical models; including linear regression, logistic regression and Poisson regression. Nelder and Wedderburn (1972) also proposed an iteratively reweighted least squares method for maximum likelihood estimation (MLE) of the model parameter estimation under GLM. To date MLEs remain

popular and are the default method of estimating model parameters in any statistical computing software.

GLMs were further developed by McCullagh and Nelder in 1989. These models generalise the classical linear models based on the normal distribution. This generalisation has two aspects in addition to the linear regression part of the classical model; these models can involve a variety of distributions which are selected from a special family of exponential dispersion models. The other aspect involves transformation of the mean through a link function. GLMs can have an extension to include repeated measures where the experimental units are measured over time including the case of correlated observation taken from the same cluster such as a household, class within a school and other similar structures. In situations where the data is longitudinal and/or clustered, as is the case with the RHIVA data, a common approach to use when analysing such data in order to estimate population averaged effects is the method of Generalised Estimating Equations (GEE) by Liang and Zeger (1986).

The concept of GEE was described by Liang and Zeger in 1986 as an extension of the GLM; and were specifically extended to accommodate correlated and/or clustered data. GEEs belong to the class of marginal (or population-averaged) models and are applicable to both discrete and continuous data. The GEE approach requires no distributional assumptions, but requires the regression model for the mean response and the working correlation structure of the longitudinal or clustered data to be specified. The correlation structure is described by a variance-covariance matrix. Marginal models are an appropriate way of dealing with correlated GLM-type observations. Studies show that the application of GEE requires the number of clusters to be 20 or more per arm (Donner et al., 2000). In the RHIVA study the total number of clusters is 14, which is below the recommended number required for the application of GEE analysis.

Donner et al (2000) advised that using the GEE approach to analyse less than 30 clusters has a high probability of yielding small p-values, narrow confidence

intervals and small standard errors, which might be misleading. In this thesis, we rely mostly on the generalised linear mixed model (GLMM) which falls among the class of models known as mixed effects models as extensions of the GLM in order to model and analyse the RHIVA data. Such models are also known as multilevel models in social science based on the formulation by Goldstein (2011). Since the methods that will be used are an extension of the GLM, this chapter will look at the structure of the GLM, define and explain the Exponential family of distributions on which the GLMs are based. It will also focus on the MLE as a method of estimating parameters. Model fitting and model diagnostics will be explored as part of assessing the goodness of fit of a GLM via the deviance.

3.2 The structure of a GLM

Consider, the classical linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i \tag{3.1}$$

The response Y_i $(i = 1, 2, ..., n)$ is modelled as a linear function of explanatory variables X_{ij} $(j = 1, 2, ..., p)$ plus an error term ε_i $(i = 1, 2, ..., n)$, n is the total number of individuals whilst p is the number of variables included in the model. We assume that Y_i has a normal distribution with mean μ_i and variance σ^2. Typically we assume,

$$Y_i \sim N(\mu_i, \sigma^2)$$

For the model in equation (3.1) above, we assume that the error terms ε_i are independent and identically distributed such that

$$E(\varepsilon_i) = 0$$

and

$$Var(\varepsilon_i) = \sigma^2$$

Typically we assume

$$\varepsilon_i \sim N(0, \sigma^2) \tag{3.2}$$

In contrast, GLMs consist of three components namely;

i. A random component which identifies the response variable (Y), which specifies or assumes a probability distribution for the response variable

ii. A systematic component which consists of a set of explanatory variables and some linear functions of them. The functional dependence is in the form of a linear function of the form $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$ known as the linear predictor. The linear predictor is denoted by η_i where

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

iii. A link function which specifies the relationship between the mean or expected value of the random component (i.e. $E(Y_i)$) and the systematic component. The link function describes how the mean $E(Y_i) = \mu_i$ depends on the linear predictor. This relationship can equivalently be expressed as:

$$g(\mu_i) = \eta_i$$

where $g(\cdot)$ is a monotone, differentiable function.

Various link functions are commonly used depending on the assumed distribution of the dependent variable (Y). Table 3.1 shows the various link functions.

Table 3.1: Various link functions

Distribution	Link name	Link Function
Normal	Identity	μ
Binomial	Logit	$\ln(\frac{\mu}{1-\mu})$
Poisson	Log	$\ln(\mu)$
Gamma	Inverse	μ^{-1}

For every assumed link function there is a corresponding variance function which describes how the variance of a particular distribution depends on the mean. The variance mean relationship is expressed as

$$Var(Y_i) = \phi(V(\mu))$$

where $V(.)$ is called the mean variance function and ϕ represents the dispersion parameter which is always a constant. Thus, the variance is principally a product of two components

i. The factor ϕ

ii. The variance function $V(\mu)$

Table 3.2 shows the various variance functions:

Table 3.2: Various variance functions

Distribution	Variance function	Dispersion ϕ
Normal	$V(\mu) = 1$	σ^2
Binomial	$V(\mu) = \mu(1-\mu)$	1
Poisson	$V(\mu) = \mu$	1
Gamma	$V(\mu) = \mu^2$	$\frac{1}{a}$

If we were to consider a Normal general linear model with $\varepsilon_i \sim N(\mu_i, \sigma^2)$ as a special case of the GLM then we would have a linear predictor

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

with link function

$$g(\mu_i) = \mu_i$$

and variance function

$$V(\mu_i) = 1$$

Since this thesis explores data with a binary outcome variable, $Y_i \in (0, 1)$, the Bernoulli distribution was considered as a basis for the modelling. That is we assume that

$$Y_i \sim Bernoulli(1, p_i)$$

We wish to model probability $p_i = P(Y_i = 1)$. Usually,

$$E(Y_i) = p_i$$

and

$$var(Y_i) = p_i(1 - p_i)$$

So the variance function is

$$V(\mu_i) = \mu_i(1 - \mu_i)$$

where $\mu_i = p_i$ and the link function is given by

$$g(\mu_i) = logit(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right)$$

Sometimes the response variable itself can be transformed to improve linearity and homogeneity of the variance so that a general linear model can be applied. However, doing the transformation has its own drawbacks because when the response variable is transformed, the transformation may not be defined on the sample space. This may also be disadvantageous since transformation may not simultaneously improve linearity and homogeneity of general linear models. For example, applying a log transform is considered a remedy for a variance that is increasing with the mean but does not always remedy the problem.

3.3 The Exponential Family

A distribution is a member of the exponential family if its probability mass function (if discrete) or its density function (if continuous) has the following form:

$$f(y_i, \theta_i, \psi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi)\right\} \tag{3.3}$$

where θ_i is the location parameter which is a function of the mean response and ϕ is the scale parameter, while $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. If a GLM uses a canonical link function then $g(\mu_i) = \theta_i$. The canonical link is a special link function that is structurally inherent in every distribution that belongs to the exponential family as stated in equation (3.3). The exponential family of distributions has good statistical properties; hence GLMs are naturally constructed from

them. The numerical algorithm iterated weighted least squares (IWLS), is used for parameter estimating purposes. The exponential family includes most distributions such as the Normal, Exponential, Gamma, Beta, Chi-square, Bernoulli and Poisson distribution.

For members of the exponential family, a special relationship exists between the mean and variance. If Y_i has a distribution in the exponential family then its mean and variance can be expressed as below

$$E(Y_i) = \mu_i = b'(\theta_i)$$

$$Var(Y_i) = \sigma_i^2 = a_i(\phi)b''(\theta_i)$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. Thus in the earlier statement about the mean-variance relationship it follows that $V(\mu_i) = b''(\theta_i)$. In this thesis the response variable is binary and the Bernoulli distribution belongs to the exponential family.

3.4 Maximum Likelihood Estimation (MLE)

The idea behind MLE is to provide estimates for a model's parameters. The estimated parameters maximise the likelihood of the sample data. In general, MLEs do not have a closed form for GLMs and therefore one has to rely on approximation methods such as the Newton-Raphson or Fisher scoring to find MLEs. From a statistical point of view, the method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties when compared to other methods such as the method of least square estimation. The MLE estimation method is versatile and applies to most models and to different types of data. In our case we primarily use the MLE method to estimate the regression parameters β in the GLM (Molenberghs & Verbeke, 2005) and this is achieved by assuming that the observations are independent. The estimation of the dispersion parameter ϕ may also become necessary, particularly if it differed from one implying the case of over-dispersion

($\phi > 1$) or under-dispersion ($\phi < 1$) in order to correctly determine standard errors for parameter estimates. Following this assumption, the joint density of the sample observations y_i, given parameters θ and ϕ, is defined by the product of the density stated in equation (3.3) over the individual observations as expressed below:

$$f_{y_1,y_2,...,y_n}(y_1, y_2, ..., y_n; \theta, \phi) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi)\right\} \tag{3.4}$$

The joint probability density function may be expressed as a function of θ_i and ϕ_i given the observations y_i, this function is then called the likelihood. The MLE procedure begins by specifying the likelihood joint probability density or joint probability mass function in which the data are taken as given and the parameters are estimated.

For any GLM, the likelihood function depends on β only through η_i. Therefore, we wish to obtain estimates of (β, ϕ) that will maximise the likelihood function above. This is more convenient to obtain when working with the log-likelihood rather than the likelihood since the values that maximise the likelihood are the same values that maximise the log likelihood. The log-likelihood for the exponential family model is expressed as

$$\log L(\theta, \phi; y_1, y_2, ..., y_n) = \sum_{i=1}^{n} \left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi)\right\} \tag{3.5}$$

The score equation is obtained from equating the first-order derivatives of the log-likelihood to zero. This is given by

$$S(\beta) = \sum_i \frac{\partial\theta_i}{\partial\beta}[y_i - b'(\theta_i)] = 0 \tag{3.6}$$

Since $\mu_i = b'(\theta_i)$ and $v_i = v(\mu_i) = b''(\theta_i)$, then we have that

$$\frac{\partial\mu_i}{\partial\beta} = b''(\theta_i)\frac{\partial\theta_i}{\partial\beta} = v_i\frac{\partial\theta_i}{\partial\beta}$$

Under suitable regularity, the MLE is a solution to the score equation re-expressed below as

$$S(\beta) = \sum_i \frac{\partial\mu_i}{\partial\beta} v_i^{-1}(y_i - \mu_i) = 0 \tag{3.7}$$

Generally, the solution to the above score equations are solved or optimised iteratively by using numerical algorithms such as iteratively (re-)weighted least squares (IRLS), Newton-Raphson or Fisher scoring. The IRLS algorithm as an approach of optimising MLE will be discussed briefly in what will follow. Most of the algorithms (IWLS, IRLS, Fisher scoring) are based on the fundamental Newton-Raphson method which is based on successive approximations to the solution using Taylor's theorem to approximate the equation. These algorithms start with a reasonable guess to the initial solution of the equation $\hat{\beta}^{(0)}$ and keep on updating the solution until the iterative algorithm converges to the solution of β. ML estimators have the following properties. They are

i. Consistent

ii. Asymptotically normal

iii. Asymptotically efficient

iv. Asymptotically achieve the Cramer Rao Lower Bound (CRLB). The property states that if $\hat{\theta}$ is ML estimate of the parameter θ and $g(\theta)$ is a function of θ then the ML estimate of $g(\theta)$ is $g(\hat{\theta})$. The asymptotic variance of ML estimators is found by using the Fisher information of the parameter θ which is defined as

$$I(\theta) = -E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right] \tag{3.8}$$

hence the asymptotic variance of $\hat{\theta}$ is given by $I^{-1}(\theta)$

3.5 Numerical Algorithm

3.5.1 Iteratively Reweighted Least Squares (IRLS)

IRLS is a technique used in MLE as an efficient optimisation methods to ensure convergence to the required maximum. Unlike the Fisher scoring and Newton Raphson algorithm, the IRLS algorithm does not require any initial guess for the parameter vector of interest β, but instead, it requires initial guesses for the fitted

values of $\hat{\mu}_i$; which are much easier to implement. The IRLS numerical algorithm is advantageous as it lessens the influence of outliers in an otherwise normally distributed data set by solving functions of the form:

$$argmin_{\beta} \sum_{i=1}^{N} \omega_i(\beta) \mid y_i - f_i(\beta) \mid^2 \tag{3.9}$$

Each step involves a weighted least square problem of the form

$$\beta^{(t+1)} = Argmin_{\beta} \sum_{i=1}^{N} \omega_i(\beta^{(t)}) \mid y_i - f_i(\beta) \mid^2 \tag{3.10}$$

IRLS are not the only methods used to make the iterative algorithm of the score equation converge as stated above, Newton-Raphson or Fisher scoring can also be used.

3.6 Model Selection and Diagnostics

Selecting an optimal model in statistics can be quite a challenge since a number of models can fit the same data set and yield good results. According to Lindsey (1997), a good model is one that is simple, interpretable and fits the data reasonably well. Model selection can sometimes be a trade off between model fit and complexity of the model. When selecting the best model, a comparison between models can be made using the Aikake Information Criterion (AIC), Bayesian Information Criterion (BIC) and the F-tests. With AIC and BIC, the model giving the smallest value is regarded as the best model.

After fitting any model, it is essential to do a model check in order to select a model that best describes the data. The deviance is defined as a measure of the fit of the model to the data, it uses the log-likelihood and it is twice the difference between the log-likelihood of the model of interest and the saturated model. Since this difference is a measure of the model of interest from a perfectly fitting model, it is therefore called the deviance (McCullagh & Nelder, 1989; Hardin & Hilbe, 2001). A graphical approach is commonly used to plot the deviance residual against predicted values and look for outliers, constant variance, patterns and

normality. Research suggests that the goal should be to find the simplest model (preferably one that has fewest parameters) that has the smallest deviance (which in turn reproduces the data). The deviance, D, is given by:

$$D(y;\hat{\mu}) = 2\Big\{ l(y;y) - l(\hat{\mu};y) \Big\} \tag{3.11}$$

where $l(y;y)$ is the log-likelihood under the maximum achievable (saturated) model and $l(\hat{\mu};y)$ is the log-likelihood under the current model. When fitting a particular model, we usually seek values of the parameters that minimise the deviance. We aim to minimise $D(y;\hat{\mu})$ by maximising $l(\hat{\mu};y)$. Thus the values of the parameters that minimise the deviance are the same as the values of the parameters that maximise the likelihood (Hardin & Hilbe, 2001).

Chapter 4

Generalised Linear Mixed Models (GLMMs)

4.1 Introduction

Hierarchical linear models, nested models, mixed models, random coefficient, random effects models, random parameter models, split-plot designs or multilevel linear models (MLM) are all names that describe the same advanced regression technique (Raudenbush et al., 2002). The hierarchies in MLM are known as levels (Goldstein, 2011). These levels are generally made up of grouped units. The units of analysis are usually individuals (at a lower level) who are nested within contextual/aggregate units (at a higher level). A model may have more than one level, but 2-level models are the most common. A schooling system presents a clear example of a hierarchical structure with pupils clustered or nested within schools, while the schools themselves may be nested within education authorities or boards. In the schooling example, we have a 2-level structure where the pupils are the level 1 unit while the schools are the level 2 units. Thus in these type of models it is important to clearly understand at which level is the effect of interest going to be analysed.

The basic concept behind MLMs is similar to that of ordinary least squares (OLS) regression but complex in that it is used to analyse variance in the outcome vari-

ables when the predictor variables are at varying hierarchical levels (Merlo, 2005). MLMs are generalisations of linear models (in particular, linear regression), but can also extend to non-linear models. MLMs are particularly appropriate for research designs where the data for participants is organised at more than one level (i.e. nested, hierarchical or clustered data) (Gelman and Hill; 2006).

In MLMs, characteristics or processes occurring at a higher level of analysis influence characteristics or processes at a lower level. MLMs simultaneously investigate relationships within and between hierarchical levels of grouped data, thereby efficiently accounting for variance among variables at different levels. MLMs provide an alternative type of analysis for univariate or multivariate analysis of clustered data. They can be used to adjust scores on the dependent variable for covariates (i.e. individual differences) before testing treatment differences. These models are also able to analyse experiments without the assumptions of homogeneity of regression slopes that is required by analysis of covariance (ANCOVA).

Generalised linear mixed models (GLMMs) are an extension of the GLM. GLMMs were developed in 1993 by Breslow and Clayton (Breslow & Clayton, 1993). The extension involves random or subject-specific effects to be present in the linear predictor (McColloch & Searle, 2001) in addition to the usual fixed effects of regression analysis. In other words, the linear predictor of a GLMM includes both fixed and random effects. The inclusion of random effects in the linear predictor reflects the idea that there is natural heterogeneity across subjects in (some of) their regression coefficients which is possibly the case with the RHIVA study analysed in this thesis.

The extension also allows the GLM univariate data to be obtained in the context of clustered measurements (Molenberghs & Verbeke, 2005) as is the case with the RHIVA study (which is a randomised controlled clustered trial). The inclusion of random effects assist in determining the correlation structure between observations in the same cluster whilst also taking into account the heterogeneity among clusters due to unobserved characteristics. Random effects are usually assumed

to have a normal distribution.

GLMMs are the most frequently used random effects models for discrete outcomes from cross-sectional and longitudinal (i.e. responses that are collected over time) data types whereby the aim is to evaluate how subject or cluster specific effects change over time as well as the variables that influence the change. Longitudinal data is widely used for at least three reasons

i. To increase the sensitivity by making within subject comparison.

ii. To study changes over time, and

iii. To utilise subjects efficiently.

For clustered data such as the one in this thesis, the aim is mainly to capture the within cluster correlation. GLMMs are a combination of two statistical frameworks, namely the linear mixed model (which incorporates random effects) and GLM (which handle non-normal data by using the link function and exponential family distribution; which has been described in detail in Chapter 3). Linear mixed models are a special case of random effects models (McCulloch & Searle, 2001). However, linear mixed models require the observations or response variable to be continuous and normally distributed in order to be applied as a method of analysis (Jiang, 2007). The word "generalised" refers to the non-normal distributions for the response variable while "mixed" refers to random effects present in the model in addition to fixed effects. GLMMs are also known as mixed effects models, because they contain both fixed and random effects which aid in explaining the outcome.

Fixed effects are those factors in which the levels in the experiment represent all the levels about which inference is to be made, whilst random effects are those factors in which the levels are considered to be random samples from a larger population of levels. In short, the fixed effects determine a model for the mean of the response variance while random effects determine a model for the variance covariance matrix. SAS statistical software has a specific procedure, `PROC`

`GLIMMIX`, which is often used to specify the relationship between the response of Y and the levels of the random effects. For this thesis, the randomised schools as well as individuals in the schools are the random effects in the model. It should be noted that the schools are nested within study arms, thus we have a 2-level model, but in general, one could have a 3-level model.

The GLMM for a binary response will be used here as an illustration of a random effects model because of its relevance to the RHIVA data. GLMMs will be applied as a method of analysis as they account for correlation amongst the clusters and individuals within clusters. In contrast to this, there are other approaches available for modelling correlation especially in clustered data; some of these approaches include the GEEs (which was briefly explained in the previous chapter). Although the GEE approach might seem appealing to utilise in the analysis of binary (discrete) data because of its computational simplicity compared to the maximum likelihood based approaches, the disadvantage of this approach is that it is not fully likelihood based and in case where there are missing observations, they are assumed to be missing completely at random (MCAR) under GEEs. However, the fact that GEEs allow for empirical based standard errors gives reliable confidence intervals to avoid inflated chance of committing type I error. But, with fewer clusters in the study these standard errors can be over-estimated (Donner & Klar, 2000; Bland, 2010). However, under GEEs it not quite clear how one can allow for multilevel effects as in GLMMs or MLMs. We cannot use the GEE approach to model the RHIVA dataset as this study only has 14 clusters in total.

GLMMs seem to be the best approach to use that will account for correlation in the current problem of study. Unlike GEEs, GLMMs do not require a minimum number of clusters or groups present per arm. In addition to this, there is one difference between GLMMs and GEEs; GLMMs model the data on an individual level whereas GEEs model data on the population level. GLMMs are more robust in cases where there are missing data as well as in cases where there are unbalanced clusters. Another advantage is that GLMMs can estimate variances

at different levels when dealing with nested or multileveled data types.

In the sections to follow, we define the model formulation of the GLMM, explore the consequences of adding random factors, briefly discuss the numerical measures of the strength of a relationship between two random variables under the section on covariance and correlation as well as introduce some correlation structures. Afterwards, we shall discuss the three approaches toward maximum likelihood estimation and end the chapter by discussing model selection.

4.2 The Model

The linear mixed model (LMM) is generally defined by

$$Y = X\beta + Zu + \epsilon \tag{4.1}$$

where Y is an N×1 vector of observations, β is a p×1 vector of unknown constants, u is a q×1 vector of unknown effects of random variables, ϵ is an N×1 vector of unknown residual effects, X is a vector of known matrix of order N×p that relate elements of β to elements of Y and Z is a vector of known matrix of the form N×q which relate elements of u to elements of Y. Thus, the above equation of a LMM has three components; namely the fixed component ($X\beta$), the random component (Zu) and the error components (ϵ). Comparing the above equation with that of a GLM, the major distinction between the two is the inclusion of the random component (Zu) and that the expected response $E(Y)$ is directly equated to the linear predictor through the identity link.

Including random effects in the model is useful as it explains the excess variability in the dependent variable that is not accounted for by the measured covariates. The elements in β are considered to be fixed effects while the elements in u are the random effects from populations of random effects with some variance-covariance structure. Both β and u may be partitioned into one or more variables depending on the situation.

Let Y_{ij} denote the j^{th} observation in the i^{th} cluster, $(i = 1, ..., n)$ and $(j =$

$1, ..., n_j$). X_{ij} denote a p×1 vector of known covariates, and let Z_{ij} denote a q×1 vector of random effects. The same methods can be applied for repeated measures. In this case the i^{th} cluster would be replaced with the i^{th} individual. The elements of $Y_i = (y_{i1}, ..., y_{in_i})'$ are conditionally on the random effects u_i, assumed to be independent random variables from a simple exponential family expressed as:

$$f_i(y_{ij}, u_i, \beta, \phi) = exp\left\{ \frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi) \right\} \tag{4.2}$$

The conditional mean (μ_{ij}) of Y_{ij} is modelled through a linear predictor which contains fixed effects parameters contained in a vector β as well as subject specific parameters u_i. The parameters of the random effects u_i are assumed to be distributed with density function $f(u_i \mid \alpha)$, where α denotes the unknown parameters in the density function. Therefore, we assume that the random effects u_i are $N \sim (\mathbf{0}, \mathbf{G})$. The GLMM model can be expressed as $g(\mu_{ij}) = X'_{ij}\beta + Z'_{ij}u_i$, where $g(\cdot)$ is the link function relating the mean of Y_{ij} to the linear predictor. Now, this model specification is similar in properties to that of a GLM, except that the current model includes the random effects component. The properties of a link function are standard irrespective of the model fitted. The model specification for the exponential family equation as well as that of the link function are made conditional on the value u_i.

Suppose that given the random effects u_i, binary response $y_1, ..., y_n$ are conditionally independent Bernoulli. Moreover, with $p_{ij} = P(Y_{ij} = 1 \mid u_i)$, one has

$$logit(p_{ij}) = x'_{ij}\beta + z'_{ij}u_i \tag{4.3}$$

where x_{ij} and z_{ij} are as in the definition of GLMM. The above equation is a special case of a GLMM, in which the conditional exponential family is Bernoulli. Thus, the link function is $g(\mu) = logit(\mu)$ as with the case of any standard Bernoulli distribution.

4.3 The consequences of having random effects

To better appreciate the inclusion of random effects in a GLMM, one needs to understand the consequences of such effects by studying the first two moments of the marginal distribution of y_{ij}. With linear mixed models, the marginal mean of y_{ij} coincide with the conditional mean given that $E(u_i) = 0$. However, this property is not necessarily true in GLMMs. The aspects of the marginal distribution of y_{ij}; namely the mean, variance and co-variances are derived (McCulloch & Searle, 2001) below. The marginal mean of y_{ij} is defined as;

$$\begin{aligned} E(y_{ij}) &= E\{E[y_{ij}|\mathbf{u_i}]\} && (4.4)\\ &= E[\mu_{ij}] && (4.5)\\ &= E[g^{-1}(\mathbf{x'_{ij}}\beta + \mathbf{z'_{ij}u_i})] && (4.6) \end{aligned}$$

The function g^{-1} above is non-linear and therefore there is no direct link between the conditional and the marginal model as is the case with LMM for normal responses. The marginal variance of y_{ij} is expressed as:

$$\begin{aligned} Var(y_{ij}) &= Var(E[y_{ij} \mid \mathbf{u_i}]) + \mathbf{E}[\mathbf{var}(\mathbf{y_{ij}} \mid \mathbf{u_i})] && (4.7)\\ &= var(g^{-1}[\mathbf{x_{ij}^{'}}\beta + \mathbf{z_{ij}^{'}u}]) + \mathbf{E}[\phi\mathbf{a_i v}(\mu_{\mathbf{ij}})(\mathbf{g^{-1}}[\mathbf{x'_{ij}}\beta + \mathbf{z_{ij}^{'}u}])] && (4.8) \end{aligned}$$

In contrast, the induced marginal variance in a linear mixed model is generally reduced to $var(\mathbf{Y_i}) = \mathbf{Z_i G Z'_i} + \mathbf{R}$, where $\mathbf{R} = \sigma^{\mathbf{2}}\mathbf{I_i}$, with I_i denoting the identity matrix of order n_i. Again, the above equation for the variance of the GLMM cannot be specified without making specific assumptions about the function $g(\cdot)$ and/or the conditional distribution of y_{ij}. The marginal covariance of y_{ij} is derived as follows

$$\begin{aligned} cov(y_{ij}, y_{ik}) &= cov(E[y_{ij} \mid \mathbf{u}], \mathbf{E}[\mathbf{y_{ik}} \mid \mathbf{u}] + \mathbf{E}[\mathbf{cov}(\mathbf{y_{ij}}, \mathbf{y_{ik}} \mid \mathbf{u})]) && (4.9)\\ &= cov(\mu_{ij}, \mu_{ik}) + E[0] && (4.10)\\ &= cov(g^{-1}[\mathbf{x_{ij}^{'}}\beta + \mathbf{z_{ij}^{'}u}], \mathbf{g^{-1}}[\mathbf{x_{ik}^{'}}\beta + \mathbf{z_{ik}^{'}u}]) && (4.11) \end{aligned}$$

In a linear mixed model the above equation of the covariance is reduced to $cov(y_i, y_j) = \mathbf{Z_i G Z_j^{'}}$ and to $cov(y_{ij}, y_{ik}) = \sigma_u^2$ if the model has only one random

random effect. For example, in the case of longitudinal data a random intercept model is described as the simplest case of a mixed effects model.

4.3.1 Random intercepts model

A random intercepts model is a model in which the intercepts are allowed to vary between individuals or clusters (Molenberghs & Verbeke, 2005). It is the simplest case of a mixed effects model. According to Molenberghs and Verbeke (2005), conditional mean of the dependent variable for each individual observation are predicted among other things by the intercept that varies across groups. Random intercept models assume that slopes are fixed. These models also provide information about intraclass correlations, which are helpful in determining whether multilevel models are required in the first place.

4.3.2 Random slopes model

A random slopes model is a model in which slopes are allowed to vary between individuals (Molenberghs & Verbeke, 2005). This means that the slopes are different across groups. Random slopes models assume that the intercepts are fixed.

4.3.3 Random intercepts and slopes model

A model that includes both random intercepts and random slopes is known as a random intercept and slope model. Such a model is likely to be the most realistic type of model, even though it is also the most complex. In a random intercept and slope model, both the intercepts and slopes are allowed to vary across groups.

4.4 Covariance structures

Covariance structures are commonly used in clustered, repeated and/or longitudinal data. Repeated measurements data refers to data that is generated by repeatedly observing a response or outcome from the same individual or experimental

unit over time (Crowder & Hand, 1990). Generally, in longitudinal data, measurements on the same observational unit are correlated. The same phenomenon can occur for clustered data, whereby clusters play the role of observational units, and repetition of measurement occurring within subjects in a cluster. In both cases, the usual model assumption of independent errors may be violated. A model that can incorporate this lack of independence is needed. There is nothing peculiar about repeated measures and longitudinal mixed models except for the distinct covariance structure of the observed data. For the model

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{u_i} + \epsilon_\mathbf{i} \tag{4.12}$$

$$\begin{pmatrix} u_i \\ \epsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{G_i} & 0 \\ 0 & \mathbf{R_i} \end{pmatrix} \right]$$

The term covariance structure is used to describe how the matrices $\mathbf{G_i}$ and $\mathbf{R_i}$ are constrained in the (Normal case of the) general linear mixed model:

$$Y_i \sim N(X_i\beta, v_i) \tag{4.13}$$

where $v_i = \mathbf{Z_i}\mathbf{G}\mathbf{Z_i'} + \mathbf{R_i}$ is the variance of Y_i and v_i represents the i^{th} diagonal element of the variance-covariance matrix. The covariance structure is not the primary interest of analysis but it is essential for valid inference. The covariance structure should be specified. The four commonly used covariance structures (Compound symmetry, Toeplitz, Autoregressive and Unstructured) are illustrated below using a 4x4 variance-covariance matrix. It is important to separate or distinguish between the correlation structure of the elements in u_i and the correlation structure of repeated measures. This is because marginally one is interested in the correlation between the Y_{ij}'s while in conditional models such as random intercept and slope models one is more concerned with covariance between the random effects.

4.4.1 Compound symmetry (CS)

This structure is also known as the exchangeable working correlation. It assumes a constant correlation between all pairs of measurements within a subject, regardless of the time interval between the measurements. Consequently, the Compound Symmetry (CS) structure assumes constant variance and constant covariance correlation. According to Horton and Lipsitz (1999), the exchangeable structure is appropriate for datasets that have clustered observations. A 4x4 exchangeable correlation structure matrix is shown below:

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

The CS structure has homogeneous variances. A common drawback with using CS is that close observations are expected to be more correlated than observations that are far apart.

4.4.2 The Autoregressive Regressive structure of order one: AR(1)

This structure assumes that measurements closer to each other in time are more correlated than measurements that are further away from each other. In addition, the autoregressive (AR(1)) assumes that the variance of any measurements is constant, regardless of when the measurement or observation was made. The AR(1) has the following 4x4 structure:

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

AR(1) structures require that measurements be equally spaced and that the variance decays exponentially. The AR(1) covariance structure also has homogeneous variances and correlations that decline with time or distance.

4.4.3 Toeplitz (Toep)

Similar to the AR(1), the Toeplitz (Toep) covariance structure assumes that all observations of the same distance have the same correlation but the structure is not assumed to decay exponentially as is the case with the AR(1). The Toep covariance structure has the resulting covariance matrix:

$$\sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

4.4.4 Unstructured (UN)

The unstructured (UN) covariance structure does not assume any particular pattern about the variance and covariance between measurements; therefore it permits all the variance and covariance of a particular matrix to be different. While this structure might seem as the most suitable to fit, its only pitfall is that it requires the most number of parameters to estimate and can cause computational difficulties such as non-convergence. The unstructured covariance matrix has the following form:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}$$

One should note that for the UN $\sigma_{ij} = \sigma_{ji}$. There are other covariance structures available but the above are the most frequently used.

4.5 Parameter estimation for a GLMM

Estimating the parameters of a statistical model is a key step in most statistical analyses. For a GLMM, this can be done in one of two ways, either by using the Bayesian approach or the MLE approach. According to Molenberghs and Verbeke (2005), the Bayesian approach requires one to specify the priors of β, G and ϕ through their density functions of $f(\beta), f(G)$ and $f(\phi)$ respectively. For the purposes of this thesis, we will use the method of MLE to estimate the parameters.

The MLE is considered the primary method of estimation and it is frequently used in many modern statistical tools, including GLMM estimation. According to Bolker et al. (2008), to find ML estimates for a GLMM, one must integrate likelihoods over all possible values of the random effects. Consequently, for GLMM this calculation is at best slow and at worst computationally unfeasible, especially for large numbers of random effects. Thus, the maximum likelihood techniques are hindered by the integration over the q-dimensional vector of random effects, meaning, the likelihood of a GLMM may be compromised by the high-dimensional integrals that cannot be solved analytically. As a result, statisticians have proposed various ways to approximate the likelihood to estimate the parameters for a GLMM.

In the sections to follow, we will describe the likelihood function and look at some of the numerical approximation techniques that can be used in the MLE to approximate the integrand.

4.6 Maximum Likelihood Estimation (MLE)

Let $P(Y_{ij}|u_i)$ represent the conditional probability for any form of the response Y_{ij} for a given subject i in cluster j. To avoid any difficulties with the conditional probability, we omit conditioning on the covariates x_{ij}. Now, let Y_i denote the vector of responses from subject i. The probability of any response pattern Y_i (of size n_i), conditional on u_i, is equal to the product of the probabilities of the level-1 responses, thus the likelihood function for the unknown parameters β, α and ϕ with $y = (y_1^{'}, ..., y_N^{'})$ then becomes:

$$L(\beta, \alpha, \phi, y) = \prod_{i=1}^{N} f(y_i \mid \alpha, \beta, \phi) \tag{4.14}$$

$$= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij} \mid u_i, \beta, \phi) f(u_i \mid \alpha) du_i \tag{4.15}$$

where the integral is with respect to the q-dimensional vector u_i. When both the data and the random effects are normally distributed as in the case of the linear mixed model, the integral can be worked out analytically and closed-form expressions exist for the maximum likelihood estimator of β and the best linear unbiased predictor (BLUP) for u_i. However, for general GLMMs approximations to the likelihood or numerical integration techniques are required to maximise the above equation with respect to the unknown parameters. Therefore, in order to solve the likelihood solution, integration over the random effects distribution must be numerically done as estimation tends to be much more complicated than in models for continuous normally distributed outcomes whose solution can be detailed in closed form.

Several approximating techniques that are used to evaluate the integral over the random-effects distribution are based on first or second order of the Taylor expansions. These approaches include pseudo- and penalized quasi-likelihood (PQL), Laplace approximations and Gauss-Hermite quadrature (GHQ), as well

as Markov chain Monte Carlo (MCMC) algorithms. When applying these approaches, one must distinguish between standard MLE (which estimates the standard deviations of the random effects assuming that the fixed-effect estimates are precisely correct) and restricted maximum likelihood (REML) estimation (which is a variant that averages over some of the uncertainty in the fixed-effect parameters). The above mentioned numerical approximating techniques for the integrand excluding the MCMC will be discussed in detail in the sub-sections that follow below.

4.6.1 Laplace Approximation

Laplace approximation is a well known method to approximate integrals where the exact likelihood is difficult to evaluate. If the marginal distribution of the data in a mixed model is expressed as:

$$\begin{aligned} P(y) &= \int P(y|u,\beta,\alpha)P(u|\theta^*)du && (4.16)\\ &= \int \exp(\log P(y|u,\beta,\alpha) + \log P(u|\theta^*))du && (4.17)\\ &= \int \exp c_\imath f(u,\beta,\alpha)du && (4.18) \end{aligned}$$

where θ^* is the vector of the G-side parameter, the constant $c_\imath$ is large, then the Laplace approximation of the integral can be illustrated as follows:

$$L(\beta,\alpha,u,y) = \Big(\frac{2\pi}{c_\imath}\Big)^{\frac{n_\gamma}{2}} \mid -f''(y,\beta,\alpha,u) \mid^{\frac{-1}{2e_{c_\imath} f(y,\beta,\alpha,u)}} \tag{4.19}$$

where n_γ is the number of elements in γ and f'' is the second derivative matrix

$$f''(y,\beta,\alpha,\hat{u}) = \Big(\frac{\partial^2 f(y,\beta,\alpha,u)}{\partial\gamma\partial\gamma}\Big) \mid_{\hat{\gamma}} \tag{4.20}$$

and $\hat{\gamma}$ satisfies the first order condition

$$\frac{\partial f(y,\beta,\alpha,u)}{\partial\gamma} = 0 \tag{4.21}$$

The objective function of the Laplace parameter estimation is to optimise $-2logL(\beta,\alpha,\hat{u},y)$. An advantage of the Laplace approximation is that it yields accurate results and is said to be computationally fast as opposed to other methods suitable for approximating the integrand.

If you have longitudinal or clustered data with n independent subjects or clusters, then the vector of observations can be written as $y = (y_1', ..., y_n')$ where y_i is an $n_i \times 1$ vector of observations for subject (or cluster) $i(i = 1, ..., n)$. Assuming conditional independence such that

$$P(y_i \mid u_i) = \prod_{j=1}^{n_i} P_j(y_{ij} \mid u_i), \tag{4.22}$$

the marginal distribution of the data can be expressed as

$$\begin{aligned} P(y) &= \prod_{i=1}^{n} P(y_i) \\ &= \prod_{i=1}^{n} \int P(y_i \mid u_i) P(u_i) du_i \\ &= \prod_{i=1}^{n} \int \exp n_i f(y_i, \beta, \alpha, u_i) du_i \end{aligned} \tag{4.23}$$

where

$$n_i f(y_i, \beta, \alpha, u_i) = \log P(y_i \mid u_i) P(u_i) \tag{4.24}$$

$$= \sum_{j=1}^{n_i} \log P(y_{ij}|u_i) + n_i \log P(u_i) \tag{4.25}$$

when the number of observations within a cluster, n_i is large, then the Laplace approximation to the ith individual's marginal probability density function can be written as:

$$P(y_i \mid \beta, \alpha) = \int \exp n_i f(y_i, \beta, \alpha, u_i) du_i \tag{4.26}$$

$$= \frac{(2\pi)^{\frac{n_u}{2}}}{\mid -n_i f''(y_i, \beta, \alpha, \hat{u_i}) \mid^{\frac{-1}{2}}} \exp n_i f(y_i, \beta, \alpha, u_i) \tag{4.27}$$

The parameter n_{u_i} is the common dimension of the random effects, u_i. The Laplace approximation to the marginal log likelihood of clustered data is defined as:

$$\log \left\{ L(\beta, \alpha, \hat{u_i}, y) \right\} = \sum_{j=1}^{m} \left\{ n_i f(y_i, \beta, \alpha, \hat{u_i}) + \frac{n_{u_i}}{2} \log 2\pi - \frac{1}{2} \log \mid -n_i f''(\beta, \alpha, \hat{u_i}) \mid \right\} \tag{4.28}$$

and it serves as the objective function in SAS, `PROC GLIMMIX`. It is therefore important to note that the Laplace approximation implemented in the `GLIMMIX` procedure differs from that in Wolfinger (1993) and Pinheiro and Bates (1995) in important respects. This difference is as a result of prior assumptions, Wolfinger (1993) assumed a flat prior for β and expanded the integrand around β and u, in turn leaving only the covariance parameters for the overall optimisation.

4.6.2 Gauss-Hermite quadrature

The Gauss-Hermite quadrature is often used to evaluate and maximise the likelihood for random component probit models (Rabe-Hesketh et al, 2008).The Gauss-Hermite quadrature is a standard approach for evaluating the marginal likelihood numerically particularly in limited and discrete dependent variable models with normally distributed random effects whose marginal likelihood generally do not have a closed form. Gauss-Hermite quadratures are not as good as the adaptive Gauss-Hermite quadrature (Pinheiro & Bates, 1995). An advantage of adaptive Gauss-Hermite quadratures over Gauss-Hermite quadrature is that they largely overcome biased estimates due to cluster sizes and/or intra-class correlations being large.

The adaptive quadrature does this by using the same weights and nodes as Gauss-Hermite quadrature, but to increase efficiency it centres the nodes with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. This dramatically reduces the number of quadrature points needed to approximate the integrals effectively. Although additional computing time is needed to compute the mode and curvature for each unique cluster, fewer quadrature points are needed to obtain the same degree of accuracy thus, the number of unique clusters is the biggest factor in determining the extra amount of time adaptive quadrature requires. It is highly efficient for multinomial response data with categorical covariates, but may be slower for continuous covariates with large data sets, as the mode and curvature must be calculated for every cluster for every iteration. Sometimes the computing time is a prob-

lem. Liu and Pierce (1994) recommend using ordinary quadrature to get starting values with few quadrature points, and then use the adaptive version to improve accuracy. The Gauss-Hermite quadrature approximates the integral by:

$$\int h(s)c(s)ds \approx \sum_{q=1}^{Q} w_q h(s_q) \tag{4.29}$$

From the expression above, s_q represents the nodes which are the solutions to the Q^{th} order Hermite polynomial and w_q is the corresponding weights. If one wishes to use the adaptive Gaussian quadrature rule, then the nodes of the Gaussian quadrature would be shifted such that the integrand is sampled in an appropriate region. According to Molenberghs and Verbeke, (2005), if the nodes were shifted or rescaled, then the integral to be approximated, together with new quadrature points and corresponding weights will be given respectively as expressed below:

$$\int h(s)c(s)ds \approx \sum_{q=1}^{Q} w_q^+ h(s_q^+) \tag{4.30}$$

where the above represents the equation from which the integral will be approximated. Below is the quadrature point

$$s_q^+ = \hat{s} + \left[-\frac{\partial^2}{\partial s^2} \ln[h(s)c(s)]|_{s=\hat{s}} \right]^{\frac{-1}{2}} s_q \tag{4.31}$$

and the corresponding weights

$$w_q^+ = \hat{s} + \left[-\frac{\partial^2}{\partial s^2} \ln[h(s)c(s)]|_{s=\hat{s}} \right]^{\frac{-1}{2}} \frac{c(s_q^+)}{c(s_q)} w_q \tag{4.32}$$

Liu and Pierce, (1994) have shown that when the equation of the integral above is applied with only one node, this is equivalent to approximating the integrand using the Laplace approximation. Adaptive Gaussian quadratures have an advantage over usual Gaussian quadratures, adaptive Gaussian quadratures require less quadrature points and have high accuracy compared to the Gaussian quadrature.

4.6.3 Penalized quasi-likelihood (PQL)

The penalized quasi-likelihood (PQL) was popularised by Breslow and Clayton (1993) and is related to the work on semi parametric regression which was done by

Green (1987). Since the integrated likelihood equation in the MLE section above does not have a closed form expression, Breslow and Clayton (1993) therefore proposed estimation of the regression coefficients β using the PQL method by applying the Laplace approximation to the integrated log likelihood function. According to Breslow and Lin (1995), the PQL likelihood can be written as:

$$\ell_p(\beta, \theta) = \sum_{i=1}^{n} \left(\widetilde{\ell_i} - \frac{\widetilde{u_i^2}}{2\theta} \right) \tag{4.33}$$

where $\widetilde{u_i}$ satifies $\widetilde{u_i} = \theta \frac{\partial \ell_i(\beta,\theta)}{\partial u_i}|_{u_i=\widetilde{u_i}}$ and

$$\widetilde{\ell_i} = \ell_i(\beta, \widetilde{u_i}) = \sum_{j=1}^{m} \frac{a_{ij}}{\phi} \left\{ y_{ij}\widetilde{\eta_{ij}} - c(\widetilde{\eta_{ij}}) \right\} + k(y_{ij}, \phi) \tag{4.34}$$

where $\widetilde{\eta_{ij}} = x_{ij}^T\beta + \widetilde{u}_i$

Breslow and Lin (1995) assumed that θ is known and thus derived the asymptotic bias and variance of the PQL estimator of the regression coefficient $\hat{\beta}$ in grouped randomised trial settings.

4.7 Model selection

Statistical model selection is essential as it helps one to choose the simplest model that provides the best fit to the data. The idea of model selection is commonly based on the model parsimony principle, thus models should be kept as simple as possible. Model selection compares fits of candidate models. One can select these models either by using hypothesis tests (i.e. testing simpler nested models against more complex models) (Stephens et al., 2005) or by using information theoretic approaches, which use measures of expected predictive power to rank models or average their predictions (Burnham & Anderson, 2002).

Bayesian methods have the same general scope as frequentist or information-theoretic approaches, but differ in their philosophical underpinnings as well as in the specific procedures used (Bolker et al., 2008). The likelihood ratio (LR) test is used to compare two nested models. It determines the contribution of a single

(random or fixed) factor by comparing the fit (measured as the deviance, i.e. -2 times the log-likelihood ratio) for models with and without the factor, namely nested models (Bolker et al., 2008). LR tests can assess the significance of particular factors or, equivalently, choose the better of a pair of nested models. However, researchers have criticized model selection via such pair-wise comparisons as an abuse of hypothesis testing (Burnham & Anderson, 2002). The test statistic for LR test compares the maximised log-likelihoods for the full and reduced models respectively and is defined below as:

$$-2\ln\lambda_N = -2\left[\frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})}\right] \tag{4.35}$$

Where $\hat{\theta}_{ML,0}$ and $\hat{\theta}_{ML}$ are the MLEs which maximise the maximum likelihood functions of the reduced and full models respectively. Alternatively, the information criterion (IC) can be used to select the best model. However, ICs are not formal testing procedures, as they only provide rules of thumb to discriminate between several statistical models (Verbeke & Molenberghs, 2000). ICs also provide a natural basis for averaging parameter estimates and predictions across models, this provides better estimates as well as confidence intervals that correctly account for model uncertainty. Some of the commonly used information criterions include; the Akaike (AIC), Hannan and Quinn (HQIC), Schwarz (BIC) and Bozdogan (CAIC) Information Criteria. According to Bolker et al. (2008), the AIC and related information criteria use deviance as a measure of fit, adding a term to penalize more complex models (i.e. greater numbers of parameters).

Table 4.1 illustrates some of the commonly used information criterion.

Table 4.1: Commonly used information criterion

Criterion	Definition of $\gamma(\theta)$
Akaike (AIC)	θ
Schwarz (BIC)	$\theta \frac{\ln n^*}{2}$
Hannan and Quinn (HQIC)	$\theta \ln(\ln n^*)$
Bozdogan (CAIC)	$\theta \frac{(\ln n^* + 1)}{2}$

n^* is equal to the total number of observations

Apart from using the ICs and/or the LR, one can also use the approximate Wald statistic to test the hypothesis about fixed and random effects. The additive Guassian quadrature method is applied to the RHIVA data to estimate parameters by maximising an approximation to the likelihood integrated over the random effects.

Chapter 5

Applications

5.1 Introduction

In this chapter, we apply appropriate statistical methods to model teenage pregnancy by arm in order to evaluate the impact of the CCT. We include the following variables; age, grade, site (which represents the 14 schools), arm and pregnancy as our response variable of interest. The response variable is binary, meaning a study participant is either pregnant or not. We also emphasise that the analysis at this stage is interim thus the findings here may differ from the findings when the study is complete. Statistically, our model is:

$$logit[Prob(Y_{ij} = 1)] = \beta_0 + \beta' X_{ij} \tag{5.1}$$

where Y_{ij} is the response variable indicating whether a participant is pregnant or not and β is a vector of unknown parameters. If a participant is pregnant, the response variable Y_{ij} is '1' or if the participant is not pregnant at follow up $Y_{ij} = 0$. To establish the time a participant either became pregnant or gave birth during the study, the midpoint rule was applied. This rule assumes that a pregnancy start or stop date occurs half way through the study between the follow up collection date and the last visit date prior to follow up. The midpoint rule with regards to this data is explained in Section 2.5.2. Equation 5.1 can be

expressed as

$$logit[Prob(Y_{ij} = 1)] = \beta_0 + \beta_1 Arm_{ij} + \beta_2 Age_{ij} + \beta_3 Grade_{ij} \qquad (5.2)$$

where i=*subject within a cluster* and j = *cluster*. Arm_{ij} represents the randomisation arm. Age_{ij} represents the age of participant. $Grade_{ij}$ represents the grade of participant where 0 = grade 9; 1 = grade 10.

For the purposes of this thesis, three SAS procedures will be utilised to fit these models, from simple through to multivariate. The reason for fitting simple models and then an adjusted model is to check for any confounding variable(s) that might be present. The software used in the analysis is SAS version 9.2 (SAS Institute Inc., Cary, NC, USA). `PROC SURVEYLOGISTIC, PROC GLIMMIX` and `PROC NLMIXED` will be used to analyse the RHIVA data. The primary focus of this analysis is on comparing pregnancy outcome by arm, whilst also taking into account other covariates such as age, grade and school.

The aim of doing this is to establish which of the three procedures will be able to best handle a cluster randomised dataset with a binary response, whilst also capturing the best covariance structure of individuals within a cluster. Comparison between the three SAS procedures will be made based on the procedures statements, and the ability of the procedure to handle such data. Critical focus will also be based on the procedure's ability to describe, analyse and accurately interpret the results of the data from the fitted model without any bias. For each procedure applied, comparisons will be made by study arms and an ideal model will be selected based on the model that produces the smallest information criterion.

In the sections to follow, each of the above mentioned SAS procedures will be applied and function(s) of each statement used will be explained. The variable school is denoted by SAS variable `site`, treatment arm is denoted by `arm`, the response variable pregnancy is denoted by `newpreg` (which is 0 if the learner is not pregnant and 1 if the learner is pregnant), the age of the learner is `age` and the grade is `CURRENT_GRADE`. Output from the fitted model will be tabled with a

clear explanation and interpretation of the results. Lastly, comparisons among the procedures will be made in order to select the optimal procedure to use in future for the analysis of data of this kind. This will be achieved by first looking at what each SAS code for each procedure does best, we also look at the results of each procedure on parameter estimation, convergence criterion, best information criterion, odds ratios as well as covariance matrices. Every model in each procedure will have the same covariates namely; age, grade, school, arm and pregnancy status as the response variable. These covariates are all discrete, except for the age variable which is continuous. According to van Walvaren and Hart (2008), categorising a continuous variable can cause information loss and artificially make other variables appear to be associated with the outcome whereas in actual fact it is not (Taylor & Yu, 2002).

According to Donner and Klar (2000), a t-test is suitable for analysing CRTs, though for the scope of this thesis a t-test will not be used. We set a null hypothesis that there is no significant difference in the rate of teenage pregnancy across the two study arms along with an alternative hypothesis which states that there is a significant difference in the rate of teenage pregnancy across the two study arms.

5.2 PROC SURVEYLOGISTIC

The `SURVEYLOGISTIC` procedure is similar to the `LOGISTIC` procedure and other regression procedures in the SAS system. This procedure flexible and suitable for fitting GLMs, theory for a GLM can be obtained in Chapter 3 of this thesis. The one feature that sets `PROC SURVEYLOGISTIC` apart from `PROC LOGISTIC` is its ability to incorporate the sample design information into the analysis, including designs with stratification, clustering, and unequal weighting. `PROC SURVEYLOGISTIC` fits linear logistic regression models for discrete response data by the method of maximum likelihood. The MLE is carried out with either the Fisher scoring algorithm or the Newton-Raphson algorithm. One can optionally

specify starting values for the parameter estimates. An option to specify the explanatory variables for which odds ratio estimates are desired is available. The odds ratio estimates are displayed along with parameter estimates. Variances of the regression parameters and odds ratios are computed by using either the Taylor series (linearisation) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs.

Below is a series of `SURVEYLOGISTIC` statements used to analyse the RHIVA dataset. We start by fitting simple models to see the impact of each covariate and then finally run a multivariate model with all the covariates to evaluate the significance of the adjusted model. Using `PROC SURVEYLOGISTIC`, the SAS code is expressed below as:

```
proc surveylogistic data=preg total=enrollment;
class site arm;
model newpreg (ref='0') = arm age /link=logit COVB EXPB;
cluster site;
run;
```

The `TOTAL` statement specifies the total number of subjects per cluster that contribute to the pregnancy rate calculation. Specifying enrolment means it includes the total number of female subjects per cluster excluding subjects who were pregnant at baseline. The `CLASS` statement always precedes the `MODEL` statement and it specifies categorical variables. The `CLASS` statement tells SAS that site and arm are categorical variables.

The `MODEL` statement gives the response variable and the explanatory effects, including covariates, main effects, interactions, and nested effects. On the left hand side of the `MODEL` statement, we specify the dependent variable and all independent variable(s) are on the right hand side of the equal sign. In the `MODEL` statement, we specified the link function. The option EXPB asks SAS to display exponentiated estimates (i.e., the odds ratios) while COVB displays the

covariance matrix of the parameter estimates. If the explanatory variables are omitted from the `MODEL` statement, the procedure fits an intercept-only model. The `CLUSTER` statement comprises of categorical variables that define the clusters in the sample. This statement is responsible for specifying the primary sampling unit (PSU) to account for design effects of clustering. The variable in the `CLUSTER` statement can either be character or numeric.

Parameter estimates obtained from running the SAS code of proc surveylogistic above are given with a brief interpretation in Table 5.1.

Table 5.1: Parameter estimates for model 1 of proc surveylogistic

Parameter	Estimate	Standard Error	DF	Wald Chi-Square	Pr <Chi-Square
Intercept	-5.6271	1.0642	1	27.9561	<0.0001
Arm	0.0463	0.0985	1	0.2210	0.6383
Age	0.1833	0.0625	1	8.6108	0.0033

In this analysis, the Fisher's Scoring was used as the optimisation technique. The estimate of age is significant with a p-value of 0.0033. This indicates that age plays an important role in pregnancy. Age was fitted as a continuous variable instead of dealing with it as a categorical variable. This result implies that increasing age is associated with higher pregnancy rates. There is no statistically significant difference between the two arms.

The surveylogistic procedure calculates odds ratio estimates as shown in Table 5.2.

Table 5.2: Odds ratio estimates for model 1 of proc surveylogistic

Effect	Point estimate	95% Wald Chi-Square confidence interval
Arm A vs B	1.097	0.746 - 1.614
Age	1.201	1.063 - 1.358

The odds ratio estimates given in Table 5.2 shows that the pregnancy rates in

the two arms are similar. For each 1 year increase in age, there is a 20% increase in the odds of pregnancy, meaning that younger participants are less likely to fall pregnant. The covariance structure of the fitted model is expressed in Table 5.3.

Table 5.3: Covariance matrix for model 1 of proc surveylogistic

Estimate	Intercept	Arm	Age
Intercept	1.1326	-0.0617	-0.0663
Arm	-0.0617	0.0097	0.0038
Age	-0.0663	0.0038	0.0039

The covariance matrix shows the degree to which two variables change together or co-vary. This covariance matrix takes into account that the design is that of a cluster randomised controlled trail. The output in Table 5.3 shows that the regression parameter for arm and age is 0.0038. This value is positive indicating that the arm and age variables vary in the same direction relative to their expected values.

Subsequently, we fit the same model as above but we replace the age covariate with the grade variable.

```
proc surveylogistic data=preg total=enrollment;
class site CURRENT_GRADE (PARAM=REF REF='Grade 10');
model newpreg (ref='0') = arm CURRENT_GRADE (PARAM=REF REF='Grade 10')
/link=logit COVB;
cluster site;
run;
```

Everything else remains the same except for the `CLASS` and `MODEL` statement which now incorporates the grade variable. Since the grade variable is categorical, we specify its reference category which we chose to be grade 10. Table 5.4 shows the parameter estimates obtained from applying the SAS code above. Grade is statistically significant with a p-value of 0.0410, meaning that pregnancy

Table 5.4: Parameter estimates for model 2 of proc surveylogistic

Parameter	Estimate	Standard Error	DF	Wald Chi-Square	Pr <Chi-Square
Intercept	-2.4331	0.1135	1	459.2638	<0.0001
Arm A vs B	0.04386	0.0940	1	0.0017	0.9672
Grade 9 vs 10	-0.4914	0.2405	1	4.1756	0.0410

is associated with grade.

The odds ratio obtained from using the above SAS code is given in Table 5.5. A point estimate of 1.008 of the odds ratio observed between the study arms indicates that there is no statistical difference between the arms. This means that the likelihood of getting pregnant is the same regardless of the study arm one is enrolled in. The grade variable shows that learners in grade 9 are 38.8% less likely to fall pregnant compared to learners in grade 10.

Table 5.5: Odds ratio estimates for model 2 of proc surveylogistic

Effect	Point estimate	95% Wald Chi-Square confidence interval
Arm A vs B	1.008	0.746 - 1.614
Grade 9 vs 10	0.612	1.063 - 1.358

The covariance matrix of the modelled parameters is shown in Table 5.6. Table 5.6 shows a negative covariance of -0.0124 between arm and grade. A variation of 0.0578 is observed within the grade variable.

Table 5.6: Estimated covariance matrix for model 2 of proc surveylogistic

Estimate	Intercept	Arm	Grade
Intercept	0.01289	0.005771	-0.01737
Arm	0.005771	0.008828	-0.0124
Grade	-0.01737	-0.0124	0.057824

The SAS code for generating a full model consisting of all the covariates is ex-

pressed below:

```
proc surveylogistic data=preg total=enrollment;
class site arm CURRENT_GRADE (PARAM=REF REF='Grade 10');
model newpreg (ref='0')=arm age CURRENT_GRADE/link=logit COVB EXPB;
cluster site;
run;
```

The SAS code above is similar to the SAS code fitted previously using proc surveylogistic, the difference is in the `CLASS` and `MODEL` statements. The `CLASS` statement above tells SAS that site, arm and grade are all categorical variables. This statement also allows us to state reference categories of the referenced variables. The `MODEL` statement indicates the model we want to fit taking into account all the other statements. We want SAS to fit pregnancy as the outcome with study arm, age and grade as predictor variables. The fitted model yields the following parameter estimate expressed in Table 5.8.

Table 5.7: Parameter estimates for model 3 of proc surveylogistic

Parameter	Estimate	Standard Error	DF	Wald Chi-Square	Pr <Chi-Square
Intercept	-5.1208	1.0035	1	26.0409	<0.0001
Arm A vs B	0.0427	0.0968	1	0.1942	0.6594
Age	0.1580	0.0588	1	7.2282	0.0072
Grade 9 vs 10	-0.2090	0.2138	1	0.9560	0.3282

Age is statistically significant with a p-value of 0.0072 and grade appears to be insignificant with a p-value of 0.3282. In the simple models fitted both age and grade were significant. We perform a multi-collinearity test to see if these two variables are correlated. The Pearson correlation coefficient between the age and grade variable is 0.49335. This correlation coefficient indicates a positive correlation between the age and grade variables implying that these two variables possess an element of multi-collinearity, meaning that age explains grade and vice versa. The mean age by grade is 16.6. Since the age and grade variables have

multi-collinearity, we can remove either age or grade in the final model and end up with results as those from Table 5.1 or Table 5.4 above. For simplicity we will remove the grade variable in the model and therefore fit a final model with arm and age which yields similar results as those of Table 5.1, Table 5.2 and Table 5.3. Table 5.8 below shows the covariance matrix of the `surveylogistic` procedure modelling all the covariates.

Table 5.8: Estimated covariance matrix for model 3 of proc surveylogistic

Effect	Intercept	Arm	Age	Grade
Intercept	1.006977	-0.04038	-0.05863	-0.02023
Arm	-0.04038	0.009367	0.002697	-0.00929
Age	-0.05863	0.002697	0.003452	0.00031
Grade	-0.02023	-0.00929	0.00031	0.045704

Table 5.8 reflects a small but positive variation of 0.00031 between age and grade. A negative variation of -0.00929 is observed between arm and grade, while arm and age has a small positive variation of 0.002697.

5.3 PROC GLIMMIX

The `GLIMMIX` procedure fits statistical models to data with correlations or non constant variability. This procedure is also applicable for fitting models where the response variable is not necessarily normally distributed; such models are known as generalised linear mixed models (GLMM). More information on GLMM can be found in Chapter 4 of this thesis. The GLMMs, like linear mixed models, assume normal (Gaussian) random effects. Conditional on these random effects, data can have any distribution in the exponential family which comprises of many elementary discrete and continuous distributions. The response variable for this thesis is binary thus, discrete and has a binomial distribution; other examples of distributions with discrete outcome include the Poisson, and negative binomial distributions. `PROC GLIMMIX` allows one to incorporate both fixed and random effects in modelling repeated measures problems or longitudinal data. Applying

the GLIMMIX procedure, we fit a model shown below:

```
proc glimmix data=preg;
class site arm ;
model newpreg=arm age /dist=bin link=logit COVB cl;
random site;
estimate "A vs B" arm 1 -1/exp;
estimate "AGE" age 1 / exp;
run;
```

In the SAS GLIMMIX procedure above, the CLASS statement tells SAS to consider the variables site and arm as categorical variables which means that these variables are discrete. The MODEL statement specifies the model that is being fitted with the response variable on the left of the equal sign. As previously, we are modelling teenage pregnancy by arm and age but this time using the GLIMMIX procedure. The option statement in the MODEL statement is represented by backward slash and we have specified the type of distribution, the link function as a logit function and requested the covariance matrix as well as the confidence intervals of the model. The function of the RANDOM statement is to specify variables that have random effects. If the RANDOM statement is not specified, the GLIMMIX procedure fits generalised linear models since the random effect is what sets GLM and GLMM apart. The ESTIMATE statement yields estimates of specified contrasts. The option statement EXP used in the ESTIMATE statement tells SAS to generate exponentiated estimates. From the SAS GLIMMIX procedure above, we get the output as documented in Table 5.9.

Table 5.9: Solution for fixed effects of model 1 of proc glimmix

Effect	Estimate	Standard error	DF	t Value	Pr>$\|t\|$
Intercept	-5.6733	1.0650	12	-5.33	0.0002
Arm A	0.09266	0.2485	1087	0.37	0.7093
Age	0.1833	0.06257	1087	2.93	0.0035

The variable age is statistically significant with an estimate of effect of 0.09266

and a p-value of 0.0035, this p-value is almost the same as that of 0.0033 which gave an estimate of effect value 0.1833 as observed in Table 5.1, even though different SAS procedures were used to fit the same model. Again these results indicate that the pregnancy rate increases with age. The regression coefficient of the study arms was not statistically significant meaning that the rate of pregnancy is not statistically different across the two arms. The estimates of the effects in Table 5.9 yielded 0.09266 and 0.1833. We exponentiate these values to obtain 1.0971 and 1.2012 for arm and age respectively. The exponentiated values are regarded as the odds ratio estimates. The covariance matrix for the fixed effects is given in Table 5.10.

Table 5.10: Estimated covariance matrix of model 1 of proc glimmix

Effect	Intercept	Arm	Age
Intercept	1.1343	-0.07483	-0.06575
Arm	-0.07483	0.06176	0.002667
Age	-0.06575	0.002667	0.003915

A variance of 0.0618 is observed for arm, while the variation for arm and age is 0.00267.

Next, we use `PROC GLIMMIX` to fit the same model but with the grade variable in place of age. The SAS code for this model is expressed below:

```
proc glimmix data=preg;
class site arm CURRENT_GRADE;
model newpreg=arm CURRENT_GRADE /dist=bin link=logit
COVB cl; random site;
estimate "A vs B" arm 1 -1/exp;
estimate "grade9 vs grade10" CURRENT_GRADE -1 1/exp;
run;
```

This SAS code is similar to that of `PROC GLIMMIX` above. Since the grade variable is a categorical variable, we have included it in the `CLASS` statement and also

specified the reference grade in the `ESTIMATE` statement. Results from Table 5.11 give solutions for fixed effect of the fitted model using `PROC GLIMMIX`.

Table 5.11: Solution for fixed effects of model 2 of proc glimmix

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	-2.9283	0.2296	12	-12.75	<0.0001
Arm A vs B	0.007727	0.2444	1087	0.03	0.9748
Grade 9 vs 10	0.4914	0.2484	1087	1.98	0.0482

For the arm variable we can conclude that arm is not significant with p-value of 0.9748, this means that no difference in pregnancy is observed between the two study arms, which has been the case with the previous models and procedures fitted. The grade variable yielded a p-value of 0.0482, which indicates that grade is statistically significant.

This means that grade 9 pupils are less likely to be pregnant. When comparing the p-value of grade to that in Table 5.4, again we can see that these results are similar even though different procedures have been used. The covariance matrix of the model fitted above is displayed in Table 5.12.

Table 5.12: Estimated covariance matrix of model 2 of proc glimmix

Effect	Intercept	Arm	Grade
Intercept	0.05272	-0.03129	-0.03758
Arm	-0.03129	0.05975	0.002384
Grade	-0.03758	0.002384	0.06173

The covariance matrix in Table 5.12 indicates a variation of 0.00238 between the arm and grade variable. A variance of 0.0617 and 0.0598 was obtained respectively, for the grade and arm variables.

In the SAS code to follow, we fit a model with all the covariates using `PROC GLIMMIX`. The SAS code for this model is:

```
proc glimmix data=preg;
class site arm CURRENT_GRADE;
model newpreg=arm age CURRENT_GRADE /dist=bin link=logit
COVB cl;
random site;
estimate "A vs B" arm 1 -1/exp;
estimate "AGE" age 1 /exp;
estimate "grade9 vs grade10" CURRENT_GRADE -1 1/exp;
run;
```

The statements of the above code do not differ a lot from the previous `proc glimmix`. The one difference is the fact that it includes all the covariates and yields the results as per Table 5.13.

Table 5.13: Solution for fixed effects of model 3 of proc glimmix

Effect	Estimate	Standard error	DF	t Value	Pr>|t|
Intercept	-5.3724	1.1462	12	-4.69	0.0005
Arm A vs B	0.08530	0.2485	1087	0.34	0.7315
Age	0.1580	0.07175	1087	2.20	0.0279
Grade 9 vs 10	0.2090	0.2826	1087	0.74	0.4596

The results of the multivariate model in Table 5.13 are not consistent with that of the models discussed above, the inconsistency is with the grade variable which appears to be insignificant in the multivariate model whereas it was statistically significant in the pevious model (Table 5.11). This inconsistency has been observed previously in the `PROC SURVEYLOGISTIC` procedure. The age and grade variables appear to be statistically significant when fitted separately however, grade becomes statistically insignificant when fitted in the same model as age. These results both show that age and grade have multi-colinearity, meaning that one variable has the ability to explain the other, hence we can have a model which excludes either variable.

A good practice is to choose a model that excludes the grade variable rather than that of age, since the age variable is a strong predictor of pregnancy. Hence, this leads to the model and the results given in Table 5.9 and Table 5.10 above. Table 5.14 shows the covariance matrix generated from fitting a multivariate model using `proc glimmix`.

Table 5.14: Estimated covariance matrix of model 3 of proc glimmix

Effect	Intercept	Arm	Age	Grade
Intercept	1.3137	-0.07792	-0.08055	0.1120
Arm	-0.07792	0.06177	0.002952	-0.00271
Age	-0.08055	0.002952	0.005147	-0.00958
Grade	0.1120	-0.00271	-0.00958	0.07985

A variation of 0.00295 is observed between age and arm, while a variance of of 0.0618 and 0.00515 are observed from the arm and age variable respectively. The grade variable has a variance of 0.0799.

5.4 PROC NLMIXED

The `NLMIXED` procedure fits nonlinear mixed models; models with both fixed and random effects. This procedure allows the random effects to enter the model non-linearly. Such a procedure is used to fit data of a MLM which is incorporated in the GLMM. Theory for GLMM can be obtained in Chapter 4 of this thesis. `PROC NLMIXED` fits such models by maximising an approximation to the likelihood integrated over the random effects using different integral approximations available, such as adaptive Gaussian quadrature and a first-order Taylor series approximation. The adaptive Gaussian quadrature is the commonly used integral approximation technique while the quasi-Newton algorithm is the default optimisation technique.

`PROC NLMIXED` is frequently applied in the analysis of pharmacokinetics and over dispersed binomial data. This procedure allows one to fit not only data that is

normally distributed, but also binomial, Poisson or a general distribution that you code using SAS programming statements. PROC NLMIXED assumes that the input data set is clustered. This procedure can be quite tricky to program thus we only fit a multivariate model with age and arm, using the output from PROC SURVEYLOGISTIC as initial parameters. Below is a SAS code used to model the RHIVA dataset using the NLMIXED procedure.

```
proc nlmixed data=preg;
parms beta0=2 beta1=0.09266 beta2=0.1833 s2u=0.5;
eta = beta0 + beta1*arm+ beta2*age +u;
expeta = exp(eta);
p = expeta/(1+expeta);
model newpreg ~ binary(p);
random u ~ normal(0, s2u) subject=site;
predict eta out=eta;
estimate beta1 ;
estimate beta2 ;
run;
```

The PROC NLMIXED statement was fitted using SAS. The PARMS statement defines the parameters used, generally the parameters fitted require priors or starting values. The starting values are often obtained by fitting a simpler model or from other similar studies. The next three statements construct the variable p to correspond to the p_{ij}, while the MODEL statement defines the conditional distribution of the variable of interest to be binomial. The RANDOM statement defines U and SITE to be the random effect. The PREDICT statement constructs predictions for each observation in the input data set, thus for this data, the predictions of n_i are output to a SAS data set named ETA. The ESTIMATE statement yields reciprocals of the betas. From running the above application procedure, we obtain the following results as shown in Table 5.15.

Table 5.15: Parameter estimates of proc nlmixed

Parameter	Estimate	Standard error	DF	t Value	Pr <—t—	Lower	Upper
Intercept (beta0)	-3.1754	0.5140	13	-6.18	<0.0001	-4.2858	-2.0650
Arm (beta1)	-0.08103	0.1400	13	-0.58	0.5727	-0.3835	0.2215
Age (beta2)	0.2860	0.1400	13	2.04	0.0620	-0.01653	0.5885
s2u	-111E-14	01256	13	-0.00	1.000	-0.2714	0.2714

Table 5.15 lists the maximum likelihood estimates of the parameters and their approximate standard errors computed using the Hessian matrix.
Approximate t-values and Wald-type confidence limits are also provided, with degrees of freedom equal to the number of subjects minus the number of random effects. Again with this procedure we reach the same conclusion as with other procedures applied previously. The estimates for arm and age are -0.08103 and 0.2860, respectively.

5.5 Choosing between the methods

The SAS procedures applied above support the analyses of CRTs. Although, these procedures yield similar results they are different and also similar in several ways. The most common similarity is that they all have an added advantage over `PROC LOGISTIC`.

The `SURVEYLOGISTIC` and `GLIMMIX` procedures can specify the reference categories but the same cannot be done in the `NLMIXED` procedure. The SAS code for `PROC NLMIXED` is quite complicated and can be tricky to program.

The methodology that these procedures use to generate the results is different, `PROC GLIMMIX` gives approximate ML estimates as opposed to `PROC NLMIXED`. The `PROC GLIMMIX` uses pseudolikelihood estimation, which is similar to PQL, while `PROC NLMIXED` uses numerical integration for ML estimation. Another advantage of `PROC GLIMMIX` over `PROC NLMIXED`, is that it allows greater flexibility in the

types of models that can be estimated and the number of random effects that can be specified, that is, `GLIMMIX` can fit complex models that accommodate serial correlation in addition to random effects.

All these procedures are a valid tool in most cluster analysis data sets as they are able to accommodate or assume the study design and take into consideration the random effect, except for `PROC SURVEYLOGISTIC` which does not accommodate the random effects. When analysing such data, the ideal method to use is a choice between `PROC SURVEYLOGISTIC` and `PROC GLIMMIX`, thus we choose `PROC SURVEYLOGISTIC` since it is simple, efficient and straight-forward.

5.6 Missing data

From the interim data analysed using the SAS procedures above, we found that 280 out of 1412 enrolled participants missed a follow up visit. The reasons for missing a visit are unknown but, it is common to assume that it is because participants had reached an endpoint of interest, which in this case is falling pregnant. Missing data can pose a risk of bias, depending on the reasons why data are missing. Reasons for missing data are commonly classified as; missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). It is impossible to distinguish between MAR and MNAR using observed data. Therefore, biases caused by data that are MNAR can be addressed only by sensitivity analyses examining the effect of different assumptions about the missing data mechanism. In this thesis, the analysis was likelihood based thus, we assume that the missing data was MAR. However, the scope of missing data can be investigated as an extension for future work.

Chapter 6

Discussion

The issue of teenage pregnancy poses a huge challenge worldwide and in a time where HIV and AIDS is recognised as the primary reproductive health concern for adolescents, teenage pregnancy remains a common social and public health concern (Dangal, 2006) and it is also the greatest contributor to the gender gap in educational attainment, particularly at the secondary level (Eloundou-Enyegue, 2004). A third of adolescent girls in South Africa become pregnant before the age of 20, despite contraception being free and mostly accessible (Wood et al., 2006). Though, teenage pregnancy can be prevented by condom use or taking contraceptives, teenagers feel afraid to access facilities that provide these services. Hence, more awareness and interventions need to be introduced. In this thesis, the data analysed reveals that teenage pregnancy is still a problem which needs to be tackled smartly and effectively so that it can be reduced.

Logistic regression models could not be used to analyse data from this thesis as they cannot handle clustered data. Thus, other appropriate methods that account for clustering were used and yielded the same results even though they differed. Surveylogistic regression, glimmix and nlmixed models were used as the standard approach to analyse the relationship between the binary dependent variable and a set of explanatory variables by incorporating the sample design information, including clustering. These procedures were similar in that we were able to obtain the parameter estimates and the covariance estimates. Glimmix can cope

with many variations of canonical and hybrid generalised linear models, but does not explicitly incorporate clustering effects (effectively reducing the sample size), unlike surveylogistic which has a special feature in SAS where one can state the cluster. The nlmixed procedure requires starting values and like glimmix, it does not explicitly incorporate clustering effects as it already assumes that the data is from a clustered design. Nonetheless, all the methods applied in the RHIVA data yielded the same results which in turn yielded the same conclusion.

The results obtained in the Chapter 5 show that the age and grade variable play a role in teenage pregnancy, regardless of the method applied. In the simple models fitted, age and grade were statistically significant but the grade variable was not significant in the multivariate model, the age variable remained significant throughout. This is a result of multi-colinearity, meaning that either (age or grade) variable was able to explain the other. The study arms showed no significant difference since the pregnancy rates obtained per study arm were 7.1 and 7.2 per 100 person years. This means that we cannot conclude that the intervention tested reduced teenage pregnancy. A few factors are considered to have influenced the results obtained. These include the duration of the study which was rather short. It is possible that a CCT over a longer time period could lead to changes in behaviour and reduction in pregnancies. The amount of money awarded (as a cash incentive that is too small might not be able to influence behaviour) and the fact that the conditions in RHIVA did not directly target pregnancy but HIV prevention could also have lead to lack of effect in reducing pregnancies.

We found no difference in pregnancy rate by study arms. There was no evidence that the intervention lowers pregnancy rates. However, a similar study carried out in Malawi was a success (Baird et al, 2009). Results from the Zomba Cash Transfer Programme (ZCTP), a CCT carried out in Malawi which targeted current schoolgirls and recent dropouts to stay or remain in school, showed a decline by more than 30% in schoolgirls becoming pregnant (Baird et al, 2009). Intervention for the ZCTP focused on school enrolment and attendance, though this programme can also affect sexual behaviour.

The most distinctive difference between the RHIVA and ZCTP study programme is the conditionality or tasks put in place for the learner to qualify for the cash transfers. For the ZCTP programme, participants were awarded the incentives if they attended school for at least 75% of the days her school was in session in the previous month. The conditions for the RHIVA study on the other hand was that participants undergo VCT testing, pass their June and December examination, attend the My Life! My Future! programme as well as do a community project which was aimed at increasing the participant's entrepreneurial and business skills.

Nonetheless, teenage pregnancy still continues to occur in this cohort and addressing it is a battle that requires the active involvement of all stakeholders. These stakeholders include other government departments, key organisations in the non-governmental sector; the research community, the religious sector, community leaders and more importantly, parents and the learners themselves.

Many girls who were pregnant at baseline were not included in the follow up assessment. This alone is proof of the impact pregnancy has on education. Missed visits or loss to follow up is often a problem in any study, thus it is essential to perform an analysis that take missing data into account. This thesis can be extended as possible future work to apply the methodology of handling missing data.

The analyses presented here have some limitations. These include

- The small number of clusters which reduce the power of the study and limit the applicable statistical methods that can be used (such as GEE).
- It is crucial to mention that the analysis performed in this thesis was that of interim data therefore conclusions made are not final since the study was still in progress.
- There was only one timepoint where pregnancy was measured. If this was measured repeatedly, the methods can be expanded to analyse repeated

measures that are clustered within school.

- This study did not directly target pregnancy but HIV prevention
- The analysis for missing data was done using crude methods. These could be expanded in future work.

References

Baird, S., Chirwa, E., McIntosh, C., Ozler, B. (2009) The short-term impacts of a schooling conditional cash transfer programme on the sexual behavior of young women. Health economics, 19, 55-68.

Baird, S., McIntosh, C., Ozler, B. (2010) Cash or Condition? Evidence from a randomised Cash Transfer Program, Technical Report. UC San Diego February.

Bland, J.M. (2010) Analysis of a cluster-randomised trial in education. Effective Education, 2, 165-180.

Bolker, B.M. (2008) Generalised linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution, 24, 127-135.

Breslow, N.E. & Clayton, D.G. (1993) Approximate inference in generalised linear mixed models. Journal of the American Statistical Association, 88, 9-25.

Breslow, N.E. & Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika, 82, 81-91.

Burnham, K.P. & Anderson, D.R. (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag.

Cassell, C. (2002) Let It Shine: Promoting School Success, Life Aspirations to

Prevent School-Age Parenthood. Washington, DC: SIECUS

Chaudhury, N. (2008) Income transfers and female schooling: The impact of the female school stipend programme on public school enrolments in Punjab. Washington: World Bank.

Crowder, M.J. & Hand, D.J. (1990) Analysis of Repeated Measures. London: Chapman and Hall/CRC Press.

Dangal, G. (2006) Teenage pregnancy: complexities and challenges. Journal of the Nepal Medical Association, 45, 262-272.

de Janvry, A. & Sadoulet, E. (2004) Conditional cash transfer programmes: are they really magic bullets? ARE Update, 7, 10.

Department of Health. (2008) Department of Health Annual Report 2007/2008 (Report No. 1-85). Pretoria: Department of Health.

Devereux, S., Sabates-Wheeler, R., Slater, R., Tefera, M., Brown, T. & Teshome, A. (2006) Ethiopia's Productive Safety Net Programme (PSNP): Trends in PSNP transfers within targeted households. IDS/INDAK

Donner, A. (1998) Some aspects of the design and analysis of cluster randomisation trials. Application Statistics, 47, 95-113.

Donner, A. & Klar, N. (2000) Design and analysis of cluster randomised trials in health research. London: Arnold.

Duflo, E. (2003) Grandmothers and Granddaughters: Old-Age Pensions and Intrahousehold Allocation in South Africa, World Bank Economic Review, 17, 1-25.

Feng, Z., Diehr, P., Peterson, A. & McLerran, D. (2001) Selected statistical issues in group randomised trials. Annual Review of Public Health, 22, 167-187.

Fiszbein, A. & Schady N. (2009) Conditional Cash Transfers: Reducing Present and Future Poverty. The World Bank: Washington DC.

Gelman, A. & Hill, J. (2006) Data Analysis Using Regression and Multilevel/ Hierarchical Models. Cambridge University Press.

Glass, G.V. & Stanley, J.C. (1970) Statistical Methods in Education and Psychology. Englewood Cliffs: Prentice-Hall.

Goldstein, H. (2011) Multilevel Statistical Models 4th Edition. United Kingdom: Wiley.

Gray, R., Li, X., Kigozi, G., Serwadda, D., Brahmbhatt, H., Wabwire-Mangen, F., Nalugoda, F., Kiddugavu, M., Sewankambo, N. & Quinn, T. (2005) Increased risk of incident HIV during pregnancy in Rakai, Uganda: a prospective study. The Lancet, 366, 1182-1188.

Green, P.J. (1987) Penalized Likelihood for General Semi-Parametric Regression Models. International Statistical Review, 55, 245-259.

Hallman, K. & Grant, M. (2003). Disadvantages and youth schooling, work and childbearing in South Africa. Paper presented at the Annual Meeting of the Population Association of America, Minneapolis

Hardin, J.W. & Hilbe, J.M. (2001) Generalised Linear Models and Extensions: Stata Press.

Harrison, D. (2008) Three ways to reduce teen pregnancy in South Africa. Paper

prepared for the HSRC Youth Policy Initiative Roundtable 5: Teenage Pregnancy. Reserve Bank, Pretoria.

Horton, N.J. & Lipsitz, S.R. (1999) Review of software to fit generalised estimating equations (GEE) regression models. The American Statistician 55, 242-254.

Imamura, M., Tucker, J., Hannaford, P., da Silva, M.O., Astin, M., Wyness, L. et al. (2007). Factors associated with teenage pregnancy in the European Union countries: a systematic review. European Journal of Public Health: 17, 630-639.

Jiang, J. (2007) Linear and Generalised Linear Mixed Models and Their Applications. New York: Springer.

Jewkes, R., Vundule, C., Maforah, F.& Jordaan, E. (2001) Relationship dynamics and teenage pregnancy in South Africa. Social Science and Medicine, 52, 733-744.

Jewkes, R. (2006) Response to Pettifor et al. "Young people's sexual health in South Africa": HIV prevalence and sexual behaviour from a nationally representative household survey. AIDS, 20, 949-958.

Jukes, M., Simmons S. & Bundy D. (2008) Education and Vulnerability: the role of schools in protecting young women and girls from HIV in Southern Africa. AIDS, 22, S41-S46.

Kjaergaard, L.L., Villumsen J. & Cloud C. (2001) Reported methodologic quality and discrepancies between large and small randomised trials in meta-analyses. Annals of Internal Medicine, 135, 982-989.

Klein, J.D & the Committee on adolescent (2005) Adolescent Pregnancy: Current trends and issues. American Academy of Paediatrics, 116, 281-286

Kirby, D. (2007) Emerging Answers 2007: Research Findings on Programs to Reduce Teen Pregnancy and Sexually Transmitted Diseases. Washington, DC: The National Campaign to Prevent Unplanned Pregnancy.

Liang, K.Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalised linear models. Biometrika, 73, 13-22.

Lindsey, K.L. (1997) Applying Generalised Linear Models. New York: Springer-Verlag.

Liu, Q. & Pierce, D. (1994) A note on Gauss-Hermite quadrature. Biometrika, 81, 624-629.

Lindquist, E.F. (1940) Statistical Analysis in Educational Research. Houghton Mifflin Company, Boston.

Lloyd, C.B. & Mensch, B.S. (1999) Implications of formal schooling for girls, transition to adulthood in developing countries. In C. H. Bledsoe, J. B. Casterline, J. A. Johnson-Kuhn & J. G. Haaga (Eds.), Critical Perspectives on Schooling and Fertility in the Developing World. Washington, DC: National Academy Press.

Maharaj, P., Kaufman, C. & Richter, L. (2000) Children's Schooling in South Africa: Transitions and Tensions in Households and Communities. (CSDS working paper No. 3. Durban: University of Natal, Center for Social and Development Studies.

Manlove, J. (1998) The influence of high school dropout and school disengagement on the risk of school-age pregnancy. Journal of Research on Adolescence, 8, 187-220.

Manzini, N. (2001) Sexual initiation and childbearing among adolescent girls in KwaZulu Natal, South Africa. Reproductive Health Matters, 9, 44-52.

McCullagh, P. & Nelder, J.A. (1989) Generalised Linear Models, 2nd ed. Chapman and Hall, London.

McCulloch, C.E. & Searle, S.R. (2001) Generalised, Linear, and Mixed Models. New York: Wiley.

McNemar, Q. (1940) Book review of Lindquist E.F. Statistical analysis in educational research. Psychological Bulletin, 37,746-748.

Merlo, J., Chaix, B., Yang, M., Lynch, J. & Rastam, L. (2005). A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. Journal of Epidemiol Community Health, 59, 443-449.

Molenberghs, G. & Verbeke, G. (2005) Models for Discrete Longitudinal Data. Springer Science and Business Media, Inc.

Moultrie, T.A. & McGrath, N. (2007) Teenage fertility rates falling in South Africa. South African Medical Journal, 97, 442-443.

Murray, D.M. (1998) Design and analysis of group randomised trials. New York: Oxford University Press.

Mwaba, K., (2000) Perceptions of teenage pregnancy among South African adolescents, Health S.A. Gesondheid, 5, 30-35.

Nelder, J. A. & Wedderburn, R. W. M. (1972) Generalised linear models. Journal of the Royal Statistical Society, 135, 370-384.

Oakley, A. (2000) Experiments in Knowing. Gender and Method in the Social Sciences. New York: The New Press.

Ozer, E.J. (2009) Effects of a Conditional Cash Transfer Program on Children's Behaviour Problems. American Journal of Pediatrics, 123, 630.

Palen, L., Smith, E. A., Flisher, A. J., Caldwell, L. L. & Mpofu, E. (2006) Substance use and sexual risk behaviour among South African eighth grade students. Journal of Adolescent Health, 39, 761-763.

Panday, S., Makiwane, M., Ranchod, C. & Letsoalo, T. (2009) Teenage pregnancy in South Africa- with a specific focus on school-going learners. Child, Youth, Family and Social Development, Human Sciences Research Council. Pretoria: Department of Basic Education.

Pinheiro, J.C & Bates, D.M. (1995) Approximations to the Log-likelihood function in Nonlinear Mixed-Effects Models. Journal of Computational and Graphical Statistics, 4, 12-35

Pluddemann, A., Flisher, A. J., Mathews, C., Carney, T. & Lombard, C. (2008) Adolescent methamphetamine use and sexual risk behaviour in secondary school students in Cape Town, South Africa. Drug and Alcohol Review, 27, 687-692.

Rabe-Hesketh, S. & Skrondal, A. (2008) Generalised linear mixed-effects models in Longitudinal Data Analysis: A handbook of modern statistical methods. London: Chapman and Hall.

Raudenbush, S.W. & Bryk, A.S. (2002) Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage

Reddy, S.P., Panday, S., Swart, D., Jinabhai, C.C., Amosun, S.L., James, S. et al. (2003) Umthenthe Uhlaba Usamila- The South African National Youth Risk Behaviour Survey 2002. Cape Town: Medical Research Council.

Ritcher, M.S. & Mlambo, G.T. (2005) Perceptions of rural teenagers on teenage pregnancy, Health S.A Gesondheid, 10, 61-69.

Santelli, J.S., Carter, M., Orr, M. & Dittus, P. (2009) Trends in sexual risk behaviours, by nonsexual risk behaviour involvement, U.S. high school students, 1991-2007. Journal of Adolescent Health, 44, 372-379.

Schindler, J. (2008) Public schooling. In A Kraak and K Press (Eds.) Human Resource Development Review 2008: Education, Employment and Skills in South Africa 2008. Cape Town: Human Sciences Research Council.

Schubert, B. & Slater, R. (2006) Social cash transfers in low-income African countries: conditional or unconditional? Development Policy Review, 24, 571-578.

Schultz, T.P. (2004) School subsidies for the poor: evaluating the Mexican PROGRESA Poverty Program. Journal of Development Economics 74, 199-250.

Shaw, M., Lawlor, D.A. & Najman, J.M. (2006) Teenage children of teenage mothers; psychological, behavioural and health outcomes from an Australian prospective longitudinal study. Social Science and Medicine, 62, 2526-2539.

Stephens, P.A. et al. (2005). Information theory and hypothesis testing: a call for pluralism. Journal of Application in Ecology, 42, 4-12.

Verbeke, G. & Molenberghs, G. (2000) Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

Taylor, J.M.G. & Yu, M.G. (2002) Bias and efficiency loss due to categorizing an explanatory variable. Journal of Multivariate Analysis, 83, 248263.

United Nations. (2000) The Millennium Development Goals Report. New York: United Nations.

van Walraven, C. & Hart , R. G. (2008) Leave 'em Alone- Why Continuous Variables Should Be analysed as Such. Neuroepidemiology, 30, 138-139.

Veras, S.F., Perez, R.R. & Guerreiro, O.R. (2007) Evaluating the Impact of Brazil's Bolsa Famlia: Cash Transfer Programmes in Comparative Perspective IPC Evaluation Note No. 1. International Poverty Centre: Brasilia.

Were, M. (2007) Determinants of teenage pregnancies: The case of Busia District in Kenya. Economics and Human Biology, 5, 322-339.

Wolfinger, R. & O'Connell, M. (1993) Generalised linear mixed models: A pseudo-likelihood approach. Journal of statistical computation and simulation, 48, 233-243.

Wood, K. and Jewkes, R., (2006) Blood Blockages and Scolding Nurses: Barriers to Adolescent Contraceptive Use in South Africa. Reproductive Health Matters, 14, 109-118.

World Bank. (2009) Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank Publications: Washington, DC.

World Health Organisation. (2006) Adolescent Friendly Health Services. An Agenda for Change. Geneva: World Health Organisation.

World Health Organization. (2009). Child and Adolescent Health and Development Progress Report 2009: Highlights, World Health Organisation: Geneva.

World Health Organization. (2011) World Health Organization Guidelines on Preventing Early Pregnancy and Poor Reproductive Outcomes among Adolescents in Developing Countries. World Health Organization: Geneva.