

Assessing Variance Components of Multilevel Models for Social Science Data: Application to Teenage Pregnancy Data

by

Marothi Peter Letsoalo



A Research Dissertation Submitted for the Degree of Master of Science in Statistics,
Department of Statistics and Operations Research, School of Mathematical and
Computer Sciences, Faculty of Science and Agriculture, University of Limpopo, South
Africa

Supervisor:

Prof. Yehenew G Kifle

Co-supervisors:


Prof. Christel Faes (Hasselt University, Belgium)

Prof. Maseka Lesaoana

March 2019

Declaration

I declare that this research project entitled “Assessing Variance Components of Multilevel Models for Social Science Data: Application to Teenage Pregnancy Data” is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references. Further, this work has not been submitted before for any other degree at any other institution locally and internationally.

Signature: 

Date: 04 March 2019

Copyright© 2019 University of Limpopo

All rights reserved

Dedication

I dedicate this work to my son, Leteba Letsoalo, I will forever love you and will always live to be a better father. A warm feeling of gratefulness to a very special woman who has been a mother to my son during the tough times of this work.

I also dedicate this work to my selfless student, Nkaselebale Molebohe, who has cared to learn through reading this work.

I further dedicate this to my parents (Letaba and Mogaladi), sister (Mantoa) and brothers (Letebele, Chueu and Matome), who have always supported me throughout this work. This work is also dedicated to my aunt, Itholeng Mantsha, who is at all times watching over me.

I am as well dedicating this work to my promoters, queens (Prof. Christel Faes, Prof. 'Maseka Lesaoana) and king (Prof. Yehenew G Kifle) of this work, this work would have not come to conclusion without your deliberate guidance.

Acknowledgements

Acknowledgment of the conscientious work done by my promoters Prof. Yehenew G Kifle, Prof. Christel Faes and Prof. 'Maseka Lesaoana. I also acknowledge the work and notes prepared by Prof. Molenberghs, G on both generalised linear model and generalised linear mixed model. I would also like to extend these acknowledgments to all authors whose works are referenced in this dissertation. Lastly, I acknowledge Africa Center in Kwa-Zulu Natal for enabling us to use their teenage pregnancy data to use in this research.

Acronyms

AIC	Akaike Information Criterion
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
CCDS	Cross-Classified Data Structure
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
HDS	Hierarchical Data Structure
HDSS	Health and Demographic Surveillance System
ICC	Intraclass Correlation Coefficient
INDEPTH	International Network for the Demographic Evaluation of Populations and their Health
KZN	KwaZulu-Natal
LRM	Logistic Regression Model
MLM	Multilevel Model
SAS	Statistical Analysis System
SPSS	Statistical Package for the Social sciences
STATA	Statistics and Data
STATSSA	Statistics South Africa
VC	Variance Component
VPC	Variance Partition Coefficient

Abstract

Most social and health science data are longitudinal and additionally multilevel in nature, which means that response data are grouped by attributes of some cluster. Ignoring the differences and similarities generated by these clusters results to misleading estimates, hence motivating for a need to assess *variance components* (VCs) using *multilevel models* (MLMs) or *generalised linear mixed models* (GLMMs). This study has explored and fitted teenage pregnancy census data that were gathered from 2011 to 2015 by the Africa Centre at Kwa-Zulu Natal, South Africa. The exploration of these data revealed a two level pure hierarchy data structure of teenage pregnancy status for some years nested within female teenagers. To fit these data, the effects that census year (*year*) and three female characteristics (namely age (*age*), number of household membership (*idhhms*), number of children before observation year (*nch*) have on teenage pregnancy were examined. Model building of this work, firstly, fitted a logit *generalised linear model* (GLM) under the assumption that teenage pregnancy measurements are independent between females and secondly, fitted a GLMM or MLM of female random effect. A better fit GLMM indicated, for an additional year on *year*, a 0.203 decrease on the log odds of teenage pregnancy while GLM suggested a 0.21 decrease and 0.557 increase for each additional year on *age* and *year*, respectively. A GLM with only *year* effect uncovered a fixed estimate which is higher, by 0.04, than that of a better fit GLMM. The inconsistency in the effect of *year* was caused by a significant female cluster variance of approximately 0.35 that was used to compute the VCs. Given the effect of *year*, the VCs suggested that 9.5% of the differences in teenage pregnancy lies between females while 0.095 similarities (scale from 0 to 1) are for the same female. It was also revealed that *year* does not vary within females. Apart from the small differences between observed estimates of the fitted GLM and GLMM, this work produced evidence that accounting for cluster effect improves accuracy of estimates.

Keywords: *Multilevel Model, Generalised Linear Mixed Model, Variance Components, Hierarchical Data Structure, Social Science Data, Teenage Pregnancy*

Table of Contents

Declaration	i
Dedications	ii
Acknowledgements	iii
Acronyms	iv
Abstract	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Problem Statement	2
1.2 Rationale	2
1.2.1 Aim	2
1.2.2 Objectives	2
1.2.3 Literature Review	3
1.3 Methodology and Analytical Procedure	4
1.3.1 Methodology	4
1.3.2 Analytical Procedures	4
1.4 Scientific Contribution	5
1.5 Ethical Consideration	6
2 Literature Review	7
2.1 History	8

2.2	Reasons for multilevel models	8
2.2.1	Theoretical reasons	8
2.2.2	Statistical reasons	9
2.3	Hierarchical and non-hierarchical data structures	10
2.3.1	Hierarchical data structures	10
2.3.2	Non-hierarchical data structures	11
2.4	Application of multilevel models to the study of teenage pregnancy	13
2.4.1	Nonlinear outcome data	13
2.4.2	Longitudinal outcome data	13
2.5	Evidence for the existence and application of multilevel models over the years . .	14
2.5.1	School effectiveness	14
2.5.2	Causality	16
2.5.3	Multivariate outcome data	17
2.5.4	Survey data	18
2.6	Evidence for application of multilevel models in african data	18
2.7	Software packages for multilevel modelling	19
3	Data Description and Exploration	20
3.1	Population and Data Structure	21
3.1.1	Population	21
3.1.2	Data structure	22
3.2	Variables of the dataset	24
3.2.1	Response variable	24
3.2.2	Covariates	24
3.3	Data exploration	27
3.3.1	Examining the mean and standard deviation of variables	27
3.3.2	Exploring the relationship of pregnancy status and the covariates	27
3.4	Summary	30
4	Prediction of the Probability of Teenage Pregnancy using a Logit GLM	31
4.1	Introduction	32
4.2	Methodology of generalised linear model for binary data	32
4.2.1	Definition of generalised linear models	32
4.2.2	Formulation of generalised linear models for binary data	33
4.2.3	Distribution of binary data	34

4.2.4	Maximum likelihood estimation of binary data	35
4.3	Specifying and fitting a logit generalised linear model	35
4.3.1	Model building	36
4.3.2	Model selection and interpretation of a better fit model	38
4.3.3	Prediction of the probability of teenage pregnancy	39
4.4	Logit generalised linear model building with census year as a factor covariate . .	41
4.4.1	Adding to the effect of different sub-populations to a null logit GLM . . .	41
4.4.2	Adding to the effect age of the female to a logit GLM with sub-population effect	42
4.4.3	Adding to the interaction effect of age of the female and each sub-population	43
4.5	Summary	44
5	Prediction of the Probability of Teenage Pregnancy using a Logit GLMM	45
5.1	Introduction	46
5.2	Methodology of generalised linear mixed model for binary data	46
5.2.1	Formulation of GLMM	47
5.2.2	GLMM for binary response data	47
5.2.3	Random part of a logit GLMM	48
5.2.4	Maximum likelihood parameter estimation of GLMM	49
5.2.5	Variance components estimation for binary response data GLMM	49
5.3	Specifying and fitting a Logit GLMM for teenage pregnancy	51
5.3.1	Null model	51
5.3.2	Adding census year fixed effect	52
5.3.3	Adding other main covariates effect	55
5.3.4	Variance components	56
5.3.5	Adding the random effect of census year	57
5.3.6	Model Selection	58
5.3.7	Interpretation of a better fit logit GLMM for teenage pregnancy	58
5.3.8	Probability predictions of teenage pregnancy	59
5.4	A Logit GLMM building for teenage pregnancy by sub-population	61
5.5	Summary	62
6	Model Extensions and More Complex Data Structure	64
6.1	Population and Data structure	65

6.2	Specifying and fitting a Logit GLMM for teenage pregnancy with household cluster effect	67
6.3	Model building of a two level logit GLMM for teenage pregnancy	68
6.3.1	Model building for three level logit GLMM	69
6.4	Summary	72
7	Summary, Conclusions and Recommendations	73
7.1	Summary	74
7.1.1	Chapters 1 and 2	74
7.1.2	Chapters 3	74
7.1.3	Chapter 4 and 5	74
7.2	Conclusions and Recommendations	77
	Appendices	83
A	Fitting Logit GLMs	84
A.1	Logit generalised linear models stata codes in chapter 4	84
A.1.1	Load a two level teenage pregnancy dataset	84
A.1.2	Fitting and Building Logit generalised linear model (GLM)	84
B	Fitting Logit GLMMs	85
B.1	Logit generalised linear mixed models stata codes in chapter 5	85
B.1.1	Load a two level teenage pregnancy dataset	85
B.1.2	Null generalised linear mixed model	85
B.1.3	Adding the census year fixed effect	85
B.1.4	Building a logit GLMM that adds, to Model (5.3.2), other main covariates	85
B.2	Logit generalised linear mixed models stata codes in chapter 6	86
B.2.1	Load a three level teenage pregnancy dataset	86
B.2.2	Building a two level logit GLMM with a household cluster	86
B.2.3	Building a three level logit GLMM with female cluster and household cluster	86

List of Figures

2.1	Classification diagram for two and three level pure hierarchical data structures. . .	10
2.2	A classification diagram for cross-classified individuals within families and organ- isations.	11
2.3	A two-level classification diagram for MMDS of students within institutions . . .	12
2.4	A three-level classification diagram for multiple membership dataset of yearly average mark for students within institutions	12
3.1	Classification diagram for a two level pregnancy dataset (a).	23
3.2	Classification diagram for a two level pregnancy dataset (b).	23
3.3	Histogram for the frequency of number of occasion measurements across females.	25
3.4	The relationship proportion of teenage pregnancy and <i>age</i>	28
3.5	The relationship proportion of teenage pregnancy and <i>idhhms</i>	28
3.6	The relationship proportion of teenage pregnancy and <i>nch</i>	29
4.1	The relationship of probabilities and proportion of teenage pregnancy with fe- male's ages.	40
4.2	The relationship of probabilities and proportion of teenage pregnancy with census year.	40
5.1	Caterpillar plot of random intercept by rank with 95% confidence interval for a null GLMM.	53
5.2	Caterpillar plot of random intercept by rank with 95% confidence interval for Model 5.3.2.	55
5.3	Cluster specific and population averaged probability of teenage pregnancy by census year using estimated random intercepts.	60
5.4	Comparison of the population averaged probability of teenage pregnancy by cen- sus year using the estimated and simulated random intercepts.	60

6.1 Classification diagram for the three level dataset. 65

6.2 The relationship of probabilities of teenage pregnancy with *year*, *age*, *idhhms*
and *nch*. 67

List of Tables

3.1	Population of female teenagers for each <i>spft</i> , 2011 - 2015.	21
3.2	Number of observed females per pattern of subpopulations.	22
3.3	The number of unique identifiers and units within the levels.	23
3.4	The number of observations for each pregnancy status category by year.	24
3.5	The number of observations for each female age category within <i>spft</i> and across <i>year</i>	25
3.6	The number of females for each household membership category within <i>spft</i> and across <i>year</i>	26
3.7	The number of pregnancies the female teenager had before year of observation within <i>spft</i> and across <i>year</i>	26
3.8	Descriptive summary of teenage pregnancy status and all four covariates.	27
4.1	Estimates of model building for a logit GLM of pregnancy status.	38
4.2	Akaike's information criterion for each logit GLM of pregnancy status.	38
4.3	Likelihood estimates of a better fit logit GLM of pregnancy status.	39
4.4	Likelihood estimates of Model (4.4.1).	42
4.5	Likelihood estimates of Model (4.4.2).	42
4.6	Likelihood estimates of Model (4.4.3).	43
5.1	Maximum likelihood estimates for null mixed logit model with female random effect.	52
5.2	Female random intercept estimates for group of females.	53
5.3	Maximum likelihood estimates for Model (5.3.2).	54
5.4	Estimates of model building for a logit GLMM of pregnancy status.	56
5.5	ICC estimates for each model.	57
5.6	Maximum likelihood estimates for null mixed logit model with female random effect.	57

5.7	Akaike's information criterion for each logit GLMM of pregnancy status.	58
5.8	Estimates of model building for a logit GLMM of pregnancy status by sub- population.	62
6.1	The number of unique identifiers and units within the levels.	65
6.2	The number of observations for each household membership category by year. . .	66
6.3	Estimates for a two level logit GLMM building of pregnancy status with house- hold cluster.	68
6.4	Variance components of female and household random effect, Model (6.3.2). . .	70
6.5	Estimates of model building for a three level logit GLMM of pregnancy status. .	71
7.1	Likelihood estimates of the better fit GLM and GLMM.	75

Chapter 1

Introduction

Section 1.1 discusses the problem statement which briefly describes possible reasons and solutions for the problem. Section 1.2 motivates the possible solutions that were suggested in section 1.1 and further includes the aim and objectives. Section 1.2 also introduces briefly the review of literature. The rest of this chapter discusses the dataset and the analytical procedures (section 1.3); the scientific contributions (in section 1.4); and lastly the ethical consideration (section 1.5).

1.1 Problem Statement

Data structures in social and health sciences are naturally multilevel or nested (Zumbo and Chan, 2014) and often involve nonlinear outcome data. *Generalised linear models* (GLMs) are frequently used to analyse outcome data that are linear or nonlinear (Molenberghs and Verbeke, 2005). However, GLMs do not account for the hierarchy of the data. *Multilevel models* (MLMs), which are sometimes referred to as *Generalised linear mixed models* (GLMMs), have presented a significant accountability to explore information that comes from populations with nested data structure (Goldstein, 2011). However, researchers are continuously using methods for identically and independently distributed data, neglecting the nested data structures (Luke, 2004). For example, scholars such as Malema (2000) and Woodward et al. (2001) have identified factors that are possible causes of teenage pregnancy, but have not taken into account the variability between groupings of pregnancy status and/or females. In most of teenage pregnancy dataset, females are nested within the households and households are nested within villages, hence, the variability between households and/or villages should be taken into consideration (Hox and Roberts, 2011). In this study we applied MLMs or GLMMs on teenage pregnancy datasets to demonstrate the importance of assessing *variance components* (VCs) to account for within cluster differences and similarities.

1.2 Rationale

In social or health science research, it is of substantial interest to investigate the effect the households have on teenage pregnancy. Assessing this impact sheds some light on how household practices generate differences between and within households. This means that, teenage pregnancy will not be compared by the average number of teenage pregnancies per household, but by also studying the effect of households' practices. Furthermore, we would expect households to be nested within villages. As a result, the effect of villages on the teenage pregnancy can explain the geographic variation across villages.

1.2.1 Aim

The aim of this study is to assess, using teenage pregnancy data, the variance components of social science data in South Africa.

1.2.2 Objectives

The objectives of this study are to:

- (a) explore the multilevel structure of teenage pregnancy and specify an appropriate model;
- (b) examine the effects of some covariates on teenage pregnancy using a model that ignores the multilevel structure;
- (c) specify an appropriate MLM that examines the effects of some covariates on teenage pregnancy;
- (d) examine the effects of clustering on teenage pregnancy using *variance partition coefficient* (VPC) and *intraclass correlation coefficient* (ICC); and
- (e) determine the magnitude of the effects of the anticipated effects, in (c), taking into account the clustering effect.

1.2.3 Literature Review

MLMs were introduced in the mid-1980s and since then, they have gained popularity because of their ability to model simultaneously the effect of individuals and contextual information (Goldstein, 2011). MLMs are advantageous over ordinary least square methods because they overcome the assumption of observations independency as well as the correction of overestimation of type-I error (Wang et al., 2011).

Social science data are usually collected over time; hence, are mostly unbalanced (Steele, 2008). MLMs are again advantageous because of their ability to handle unbalanced or incomplete data (Wang et al., 2011). Additionally, social and health sciences data are habitually longitudinal. Nevertheless, scholars such as Molenberghs and Verbeke (2005) have presented both GLM and GLMM for longitudinal binary data.

Mchunu et al. (2012) indicated that 19.2% of the females in the youth group have experienced teenage pregnancy. Even more as a health issue, scholars such as Christofides et al. (2014) reported a human immunodeficiency virus incidence rate ratio of 3.02 for pregnancies of females less than 16 years.

Statistics South Africa (STATSSA) recorded 142452, 143812, 150984, 147120 and 117139 births by teenage mothers in the years 2010, 2011, 2012, 2013 and 2014, respectively (STATSSA, 2014). In South Africa, as in other African countries, teenage pregnancy is an alarming social concern (Nguyen et al., 2016). Furthermore, the areas that are hit hard by this concern are previously disadvantaged villages with poor communities (STATSSA, 2014).

1.3 Methodology and Analytical Procedure

1.3.1 Methodology

International network for the demographic evaluation of populations and their health (INDEPTH) is a network of *Health and demographic surveillance systems* (HDSS). INDEPTH's mission is to provide an understanding of social and health issues in areas of low and middle-income (INDEPTH, 2014). For two decades, INDEPTH collaborated with some centers that collect information on pregnancy status from rural villages around South Africa. Reduction of teenage pregnancy is one of the foremost objectives of South Africa's goals. The choice of teenage pregnancy data is motivated by the fact that teenage pregnancy is an important social and health problem affecting the public (Langille, 2007).

- *Data*

Africa center HDSS in South Africa have collected census data from 1995 to 2015 in the Mpukunyoni rural area, which is part of the Umkhanyakude district of KwaZulu-Natal. Mpukunyoni is approximately 430 square kilometers and has 11000 households. Teenage pregnancy was defined by Kumar et al. (2007) as pregnancy occurring between the maternal ages of 13-19 completed years of delivery. Hence, this study considers 1707 females aged 13-19 from Mpukunyoni area during the cohort years 2011 to 2015.

- *Variables and levels*

The outcome variable is the pregnancy status of female teenagers. Our levels are measurements of teenage pregnancies at level 1, female teenagers at level 2 and households at level 3. Some characteristics of female teenagers will also be included in the analysis as covariates. Both females and households will be included in the analysis for level effect on teenage pregnancy status.

1.3.2 Analytical Procedures

Researchers who opt to use ordinary least square methods to analyse nested data often encounter misleading inferences such as, biased predictor coefficients (Chung and Beretvas, 2012) and underestimation of standard error (Aarts et al., 2014). Due to these distorted estimates, researchers can substantively interpret the statistical significance results erroneously (Pornprasertmanit et al., 2014). Given these actualities and the data, the phases that follow compose the analytical procedure of this work.

- *Phase i*

This work explored the multilevel structure of pregnancy data, in order to observe the nested pattern that can either be purely hierarchical or multiple membership. This exploration will further give more understanding of the covariates to be used. Thereafter, an appropriate MLM together with their underlying assumptions will be specified.

- *Phase ii*

This work further used the likelihood ratio tests to examine the effects of levels (female teenagers and/or households). That is, testing the alternative against the null joint hypothesis that there are zero variances at all levels. Thereafter, we test the alternative against the null hypothesis for each level that has zero variation.

- *Phase iii*

The VPCs for each level were calculated in order to evaluate the ratio of variation in teenage pregnancy statuses that is unexplained by the covariates lying at each level. Thereafter, the ICC will also be calculated. The ICC measures the degree of similarity that is expected to be between teenage pregnancy statuses within households and/or female teenagers.

- *Phase iv*

In this last phase, this work tested whether the relationship between some significant covariate/s and teenage pregnancy varies across households and/or female teenagers. This was achieved by adding random slope effects of the corresponding covariant at a particular level.

1.4 Scientific Contribution

The results of this study can be used in dealing with the complexity of multilevel social science data structures and analysis to social science data scientists and researchers. Data scientists would realise the relevance of proper planning when gathering information at different levels of human society and/or social relationships. Moreover, this would also accommodate other future complex scientific methods to analyse social science data. Social science researchers will realise the ways of analysing multilevel social science data, which is using data to its maximum potential. Moreover, this study will add value to scientific approaches of understanding and dealing with analyses of social and health issues.

1.5 Ethical Consideration

This study uses secondary data that involve people residing at Mpukunyoni, KwaZulu-Natal. In this regard, the required data include information on human subjects. This means that we have assured that

- the data are kept in a secured environment and that only sanctioned users have access to the information,
- we will not issue or allow others to release the files or data therein to any individual,
- we will not use or allow others to use the information, except only for the listed objectives of this study,
- we will not issue, or allow others to issue, any data that identifies individuals, households or village directly or indirectly, and
- the dataset remains the property of Africa center.

Chapter 2

Literature Review

In this chapter, sections 2.1, 2.2 and 2.3 discuss the history, reasons for multilevel models and possible multilevel data structure, respectively. This chapter further elaborates on some possible outcome data for social studies such as pregnancy in section 2.4, while section 2.5 highlights application of multilevel modelling to various studies over the years. The last section of this chapter (section 2.7) highlights some of the statistical software packages that are able to fit multilevel models.

2.1 History

Over the past 30 years, scholars such as De Leeuw and Kreft (1986), Luke (2004), De Leeuw et al. (2008), Goldstein (2011), Hox and Roberts (2011), and Wang et al. (2011) have presented solid histories and extensive theories on *multilevel models* (MLMs). MLMs have gained popularity since their introduction in the mid 1980's (Wang et al., 2011). Although MLMs were first applied in educational and sociology studies, Wang et al. (2011) point out that these models can be applied in many study areas. Other application areas of these models include, but are not confined to psychology, public health, and economics (Bini et al., 2009).

In preference to MLM, researchers use various names such as, random coefficient model, hierarchical model, mixed linear model, variance component model, and random effect model (De Leeuw et al., 2008; Goldstein, 2011). Scholars refer to MLM as a statistical technique that is used by researchers as a methodology to analyse data with nested sources of variability (Goldstein, 2011; Singer, 1998). Example of such data can be people in families, students in schools, employees in firms, and animals in litters (Snijders, 2011).

2.2 Reasons for multilevel models

2.2.1 Theoretical reasons

According to Luke (2004), data structures in many studies are naturally multilevel; hence, appropriate modelling techniques should be considered for such data. Researchers have long acknowledged, before 1980's, the problems of ignoring hierarchical structures (Luke, 2004). De Leeuw et al. (2008) also indicated that social science, economics, and bio-statistics researchers were concerned, in the mid-1980's, about failure to model *hierarchical data structure* (HDS). As this failure resulted into incorrect inference, researchers in economics and bio-statistics resorted to using *analysis of variance* (ANOVA) models (De Leeuw et al., 2008). However, others opted to perform ordinary least square analysis for each group-level, which also ignored similar characteristics that are shared at a particular group-level.

Luke (2004) further stated that the existence of mathematical concepts of MLMs did not stop researchers to use simplistic single-level methods. This was because the statistical software to execute the MLM analysis were not yet developed. For example, researchers in social science collect data from multiple population and erroneously ignore the nested nature when doing the analysis (Luke, 2004). In addition, Luke (2004) emphasised that relationships discovered at

one particular group-level are not always similar to other group-level. For this understanding, scholars have a responsibility to employ appropriate methodologies such as MLM to assess group-level effects for nested data structure.

2.2.2 Statistical reasons

Most of the mathematical statistical reasons as to why researchers prefer MLMs when analysing HDSs are mainly because of misleading inferences that other techniques give. Goldstein (1986) argued that the reality of hierarchical data presented problems of model specification because of lack of independence between measurements, however, MLMs are able to deal this problem.

Multilevel models are sometimes called mixed models (Snijders, 2011) and are different from contextual analysis. Contextual analysis is an analysis that was developed in the social science to analyse mainly the effects of social context based on individual behaviour (Snijders, 2011). Furthermore, Snijders (2011) defined mixed models as statistical models that are the combination of regression analysis and ANOVA where some of the coefficients are assumed to be fixed and others are random.

Many social researchers were previously using single-level statistical techniques on the data that are naturally multilevel by pooling the group level information into individual level; hence ignoring the groups (Luke, 2004). These allowed the use of multiple regression, which however created two mathematical problems (Luke, 2004). Firstly, the individual error term of the model carries the un-modelled contextual information which violates the assumption of correlated errors (Luke, 2004; Hox and Roberts, 2011). Secondly, the method assumes that all regression coefficients are equal for all groups. Moreover, some researchers resorted to using ANOVA or *analysis of covariance* (ANCOVA) to encounter for grouping of individuals' information. However, the statistical problem arises as the number of groups increases, that is, all groups are treated as fixed effects. Even more, the ANCOVA deals only with balanced data/design. The issues discussed reduce the power of the model.

When dealing with data that follows a hierarchical structure, observations that belong to the same group are generally dependent. Statistically, this means that intra-class correlation is non-zero; hence, single level models are inappropriate to analyse such data (Hox and Roberts, 2011). Also, using ordinary least squares regression to analyse multilevel data underestimates the standard errors of conventional statistical tests (Hox and Roberts, 2011).

2.3 Hierarchical and non-hierarchical data structures

2.3.1 Hierarchical data structures

HDSs are noticeable in human and biological sciences (De Deleeuw et al., 2007; Goldstein, 2011; Hox and Roberts, 2011). The most commonly cited example of a hierarchically structured data is in education, in which students are grouped in classes, classes grouped in schools, schools are grouped in districts, and so on (De Deleeuw et al., 2007; Bini et al., 2009; Creemers et al., 2010; Goldstein, 2011). Some of the experimental designs lead to hierarchically structured data, such as, clinical trials selected from a number of randomly chosen groups of individuals. Goldstein (2011) refers to hierarchical data as consisting of units nested at different levels or units grouped at different levels.

Other examples of HDS are seen in the health and social sciences, where population (e.g. people or animals) are grouped within social structures (e.g. families); political structures (e.g. political parties); cultural structures (e.g. ethnic groups); or physical environment (e.g. ecological or biological environments). Shepelev (2011) is of an opinion that the identification of HDS is similar to identifying personalising reference characteristics. Some researchers refer to MLMs as hierarchical linear models because of their data structure name “hierarchical” (O’Connell and Reed, 2012).

The application of MLMs in many studies is restricted to situations involving a purely hierarchical data structure (Johnson, 2012).

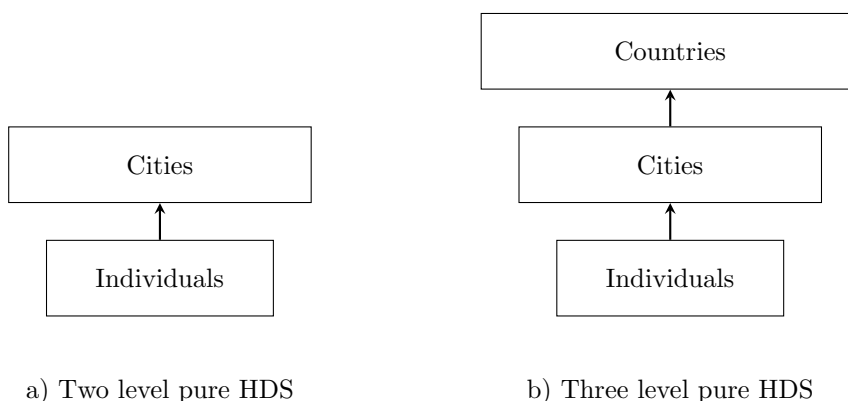


Figure 2.1: Classification diagram for two and three level pure hierarchical data structures.

Figure 2.1 shows the two pure HDSs where from a group of individuals, an individual is nested within one and only one city. Moreover, a city is also nested within one and only one country.

2.3.2 Non-hierarchical data structures

In most studies that involve multilevel modelling, researchers assume that the structures of the populations from which the data have been drawn are hierarchical, whereas in some cases this cannot be justified (De Deleeuw et al., 2007; Johnson, 2012). De Deleeuw et al. (2007), and Goldstein (2011) suggested and discussed other multilevel data structures that are sometimes confused with purely HDS. For example, the work by Aunsmo et al. (2009) used a complex data structure that is both hierarchical and cross-classified which clearly indicates the difference between the two. This section will discuss non-hierarchical cross-classified and multiple membership data structures.

- *Two-level cross-classified data structure*

The cross-classified models result from a *cross-classified data structure* (CCDS). A CCDS occurs when a unit is classified within more than one group at the same hierarchy level (Zaccarin and Rivellini, 2002; Bini et al., 2009). Figure 2.2 shows a social science CCDS example where individuals are classified by both their respective families and the organisations they work for.

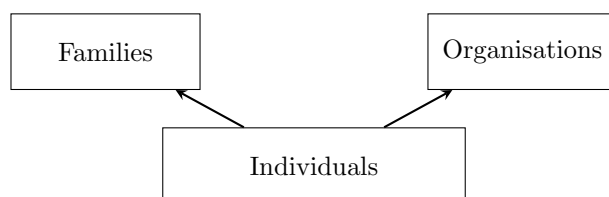


Figure 2.2: A classification diagram for cross-classified individuals within families and organisations.

In Figure 2.2, it is clear that both the organisations and families are level 2 clusters, hence individuals are cross-classified by these level 2 clusters. This means that there is pure hierarchy from individuals to families they come from and individuals to organisations they work for. One other example is in health where patients are cross-classified by general practitioner and hospital. There are more complex two-level cross-classification data structures such as longitudinal data.

- *Multiple membership*

Multiple membership data structures (MMDSs) are like the hierarchical structure, the difference is that the units are members of more than one higher-level unit, thus students may enrol in more than one institution or school (Goldstein, 2011). In order to illustrate a basic structure for two-level multiple memberships, consider examples in: a) education, where students change institution over the course of their education and each institution has an effect on their edu-

cation; and b) health, where children at the hospital are seen by several nurses and doctors at the time of their treatment. In these examples, more than one higher-level unit from the same classification influences the lower level units being the students or children at the hospital. For example, a classification diagram that can represent the aforementioned education example, where st denotes student and $inst$ denotes institution, is demonstrated in Figure 2.3:

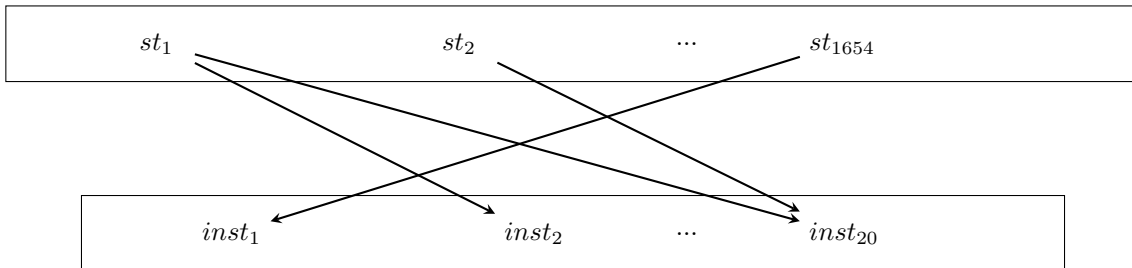


Figure 2.3: A two-level classification diagram for MMDS of students within institutions

The arrows pointing to institutions (at level 2) clearly show that students are nested within multiple institutions (e.g. student 1 is nested within institutions 2 and 20). This example can be extended to a longitudinal study where average marks of students are recorded for a given number of years, say three years. This means that for each student there will be a measurement for each of the three year at level 1, thereby extending Figure 2.3 to a three level HDS, thus

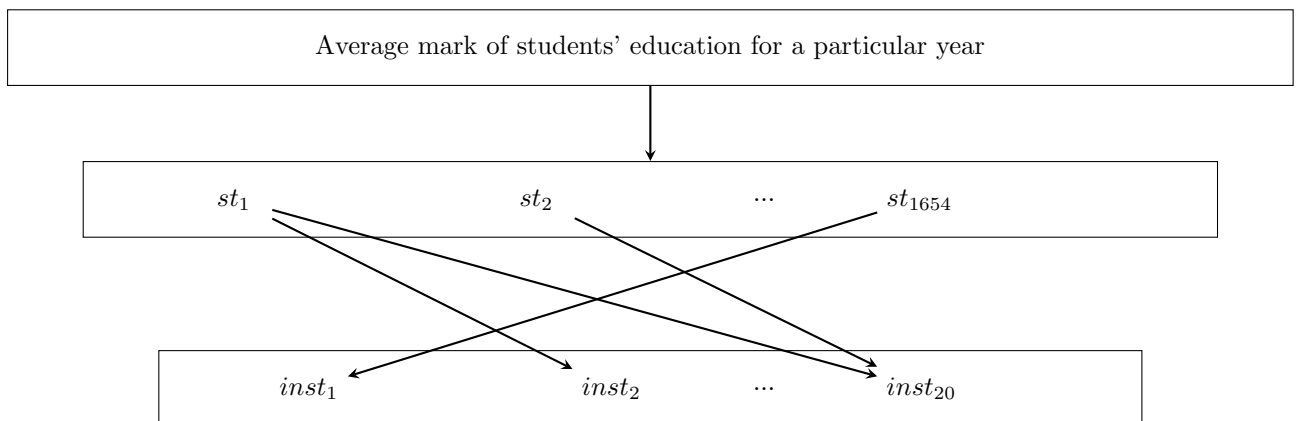


Figure 2.4: A three-level classification diagram for multiple membership dataset of yearly average mark for students within institutions

In Figure 2.4, the measurements are purely nested within students at level 2, while students are not purely nested within institutions. Other complex data structures arises when there is a mixture of two different non-hierarchical data structures.

2.4 Application of multilevel models to the study of teenage pregnancy

2.4.1 Nonlinear outcome data

Goldstein (2011) is of an opinion that other kinds of data structures are represented well in terms of nonlinear rather than linear models. Nonlinear models are generally considered when discrete response data are modelled. MLMs do not exist only where response or dependent variable is continuously distributed, but also when it is categorical. Hall and Clutter (2004), Goldstein (2011), and Tolvanen et al. (2011) also indicated that nonlinear mixed effects models are resourceful in growth data.

Hall and Clutter (2004) conducted a study using multivariate multilevel nonlinear mixed effects models in order to describe simultaneously several plot-level time quantity characteristics. Tolvanen et al. (2011) conducted a study that discusses and presents a new approach for the estimation of a nonlinear growth curve component with fixed and random effects in multilevel modelling. This approach can be used to estimate change in longitudinal data. A study conducted by Natesan et al. (2010) discussed the formulation of graded response models in nonlinear MLMs framework. This formulation estimates item parameters while investigating the group-level effects for specific covariates using Bayesian estimation.

The study that was conducted by Bhat (2000) demonstrated that many spatial analyses estimated a discrete response variable. In the work of Bhat (2000), multilevel cross-classified data of a discrete response variable were fitted using a mixed logit model that includes spatial heterogeneity.

2.4.2 Longitudinal outcome data

In section 2.3, we have already illustrated some examples of data structures that are collected from longitudinal studies. MLMs for longitudinal data allow researchers to model the between individual and within individual research questions (Singer and Willett, 2003). A longitudinal study happens in an occasion where a particular response is recorded for some subject over time or a given fixed time (sometimes called fixed occasion). Since MLMs are able to deal with data that are unbalanced, they come handy for the analysis of longitudinal data or repeated measurements (Snijders, 2011). Generally, continuous repeated measures can be modelled using multilevel random coefficient model, while the discrete ones are modelled using the multilevel logistic regression model (Rabe-Hesketh and Skrondal, 2008). In the context of MLMs, time variate measurements are usually at level 1, hence, longitudinal outcome data.

Kwok et al. (2008) conducted a study which validates the use of multilevel modelling and its advantages in analysing longitudinal data. The study by Kwok et al. (2008) used data from a sample of individuals with intra-articular fractures of the lower extremity from the University of Alabama at Birmingham's Injury Control Research Centre. Kwok et al. (2008) illustrated an overview of models and procedures for analysing longitudinal data under the multilevel modelling framework, which includes 1) a simple linear growth model; a model with a time-invariant covariate; and 2) more complicated growth models with different between- and within-individual covariance structures; and nonlinear models.

The work done by Dagum et al. (2009) focused primarily on analysing the university student achievements in order to support planning, control and decision processes in a Higher Education area. Dagum et al. (2009) used data collected from 2001 to 2006 that revealed an important finding on how to deal with longitudinal data. The results used different approaches of analysing longitudinal data, which indicated that a nonlinear growth model provides good estimates.

Lodico et al. (2006), Bini et al. (2009), Aitkin and Aitkin (2011), and Attanasio and Capursi (2011) have improved the quality of longitudinal research using MLMs and latent growth models. Authors such as Kwok et al. (2008), and Gustafsson (2013) demonstrated practises and advantages of analysing longitudinal data using MLM and LGM.

2.5 Evidence for the existence and application of multilevel models over the years

2.5.1 School effectiveness

Goldstein (2011) highlighted that it has been an interesting study by educational researchers to examine students' performance or achievements at schools and universities. Bini et al. (2009) have compiled a book that incorporates researches that elaborate on techniques that deal with educational effectiveness. When using MLMs in a study of school effectiveness, where students are grouped within schools, there are a number of advantages that motivate the use of MLMs (Bini et al., 2009). These are to enable researchers to

- investigate the relationship between explanatory and outcome factors;
- estimate statistically and efficiently the regression coefficients;
- almost accurately estimate standard errors, confidence intervals, and significance tests by using information about the groups; and

- explore the extent to which differences in average examination results between schools are accountable in terms of the characteristics of the students.

The fundamental to methodological progress in studies of school effectiveness depends on the development of suitable models for the analysis of multilevel data (Hall and Clutter, 2004). Most of the researches use multilevel modelling to investigate school and teacher effectiveness (Bini et al., 2009). Hill and Rowe (1996) indicated that identifying key issues in the design of studies for investigating the relative importance of class and school effects should generate different findings for different studies.

Methodological requirements for valid inferences in school effectiveness studies are long term longitudinal data and proper statistical modelling of multilevel data (Goldstein, 1997). The work of Goldstein (1997) highlighted that it is important to use appropriate MLMs to model the complexities of school, class, and student level data.

Fox (2004) conducted a school effectiveness study that is concerned with exploring differences within and between schools. The study suggested that variance component models are generally appropriate for the analysis of school effectiveness in educational research (Fox, 2004). Fox (2004) extends the random effects model in order to handle measurement error using a response model which lead to a random effects item response theory model. The extended random effects models are also usable and suitable for a longitudinal study, where measurements of subjects on the same outcome are observed numerous points in time (Fox, 2004).

A multilevel modelling study conducted by Fielding and Yang (2005) discussed the complexities of educational processes, the structure, and the need for disentangling effects beneath the level of the school. This study also introduced and debated the ordinal response multilevel crossed random-effects models for educational grades, the advanced level grades cross-classified by student and teaching group within a number of institutions (Fielding and Yang, 2005). The work of Fielding and Yang (2005) revealed that multilevel modelling techniques can handle the teachers' effects on several teachers' contributions to provision and on each teacher dealing with several groups. This means that the analysis brought conclusions on the sources of variation in educational progress, and predominantly the effect of teachers.

Proceedings of the integration between multilevel models and agent-based models by Salgado and Marchione (2011) indicated that MLMs are pioneers for dealing with the analyses of two or more levels hierarchical data. Salgado and Marchione (2011) focused on differential school

effectiveness analysis in order to provide a basis for comparison. Although this is the case, it was further deduced that MLMs cannot provide a causal mechanism that can explain the differences in school performances (Salgado and Marchione, 2011).

Costantini and Vitale (2011) performed a longitudinal analysis using multilevel model because of a common issue of undergraduate students obtaining their degree after the expected time at most Italian universities. The results of the study suggested a reform, which intended to reduce the gap between the average number of years in which a student complete the education programme and the official deadline established by the university regulations (Costantini and Vitale, 2011).

Green et al. (2011) applied a multilevel model technique to estimate the level of variation across schools in students' reports of non-physical bully victimisation and identify school-level predictors of bullying. The study further investigated, using multilevel models, the indicators of academic performance, emotional well-being, and school safety at school-level. The results of their study indicated that a specific group of students was significantly associated with bullying, after controlling for individual-level covariates and demographic controls (Green et al., 2011).

2.5.2 Causality

For the reason that units under study are manipulated experimentally using random allocation, Goldstein (2011) argues that causal inferences are more popular in the natural sciences. Goldstein (2011) further indicated that there is an extensive acceptance for results of experiments to be used over space and time. In the studies of causation, numerous concerns addressed by MLMs are somehow straightforward predictions (Goldstein, 2011). For example, in teenage pregnancy study, researchers might be interested in knowing the causes of female differences, and in predicting which teenage group is at risk. Scholars such as Gitelman (2005), VanderWeele (2010), and Gustafsson (2013) have studied causal effects using multilevel modelling technique.

In the study conducted by Gitelman (2005), the overall outcomes depending on group characteristics, group membership, and treatment were developed to provide a structure for stipulating causal effects of treatment in the multilevel setting. The work done by VanderWeele (2010) revealed findings that are classified within the context of multilevel modelling, causal inference, direct and indirect effects interference, longitudinal data, neighbourhood effects, mediation, and potential outcomes. Gustafsson (2013) indicated that, in educational effectiveness research, it is often difficult to make reliable inferences about cause and effect relations. For this reason,

Gustafsson (2013) conducted a study on causal inference in educational effectiveness research, in which three methods were used to investigate the effects of homework on students' achievement.

The study by Gustafsson (2013) identified main categories of threats to valid causal inference from observational data, and discussed designs and analytic approaches that protect against them. These three different methods were applied to dataset of 22 countries that participated in both the trends in international mathematics and science study. The data contain information about a sample of Grade 8 students between 2003 and 2008. The study applied a 2-level regression that separates student-level relations from class-level relations in order to, firstly, investigate the effects of the time spent on homework on mathematics achievement. Secondly, to investigate instrumental variables regression, using teacher-reported homework time to instrument student-reported homework time. Lastly, to examine the differences in analysis that are investigating country-level changes between 2003 and 2008.

2.5.3 Multivariate outcome data

Multivariate models are models that look simultaneously at a number of dependent variables as functions of independent variables (Goldstein, 2011). These models enable the researcher to deal with a wide range of problems, such as missed responses, matrix design for survey, and other methods for dealing with missing data structures. In some cases, measurements are missing by design rather than at random, hence special applications are needed. Like in a rotation designs or matrix sample designs, an individual unit has on it, one subset of measurements made.

Goldstein (2011) noted that researchers use multivariate multilevel model as the basis for handling missing data in multilevel models. These models are able to analyse data even if some of the responses are missing, hence the researcher does not have to go through special procedures of dealing with missing data.

A study conducted by Thum (2003), developed a procedure for measuring how much is gained by students in a pre-test and post-test situation against a target score on the post-test. Thum (2003) further employed a Bayesian implementation of a multivariate mixed model for repeated test scores from individual students. The approach has shown its strength in a straightforward estimation of the productivity index where its uncertainty in the form of a productivity profile was represented. The approach had further simplified a Bayesian effect size analysis that does not appeal to non-central t or F distributions.

2.5.4 Survey data

Green et al. (2011) remarked that in some literature sample survey data, researchers highlight that it is important to consider the clustering in complex sample designs. For example, geographical unit is most of the times the first stage-sampling unit when conducting a survey.

Based on Green et al. (2011), multilevel modelling is more advantageous than the traditional sampling design as it views the population structure of potential interest in itself, and it is mostly utilised for collection and analysis of data based on the higher level units of the population. Even though these models are able to model directly the clustered data, the use of weight when analysing is important for reflecting the sampling design such as patterns of non-response, and in order for the robust population to be obtained, thus allowing for protection against serious model mis-specification (Green et al., 2011). Multilevel models in sample surveys are occasionally used on data that is from a complex survey involving unequal sampling probabilities, multistage sampling, and stratification (Rabe-Hesketh and Skrondal, 2006). It is noteworthy to highlight that it is advantageous to fit a multilevel model to explore the population data structures even if a survey does not involve clustering or stratification.

Green et al. (2011) highlighted that in most of the sample surveys, clustering and stratification sampling are involved because of their importance in increasing precision for a given cost for a given total sample size. When comparing clustering and stratification, the former is able to reduce survey costs and increase standard errors of estimates, while the latter tends to reduce standard errors. Nonetheless, the main key of survey is to find balance between survey cost and standard error. Rabe-Hesketh and Skrondal (2006) is of the opinion that when dealing with MLMs with an arbitrary number of levels, a pseudo-likelihood approach for allowing inverse probability weights can be implemented by using adaptive quadrature.

2.6 Evidence for application of multilevel models in african data

Several studies have applied MLMs on datasets collected across Africa (Ukwuani et al., 2003; Wiysonge et al., 2012; Tomita and Burns, 2013). The studies cover the areas of, but not limited to, epidemiology, psychology, and health. Work by Tomita and Burns (2013) used South African data to measure the magnitude of the variation in depression as an outcome generated by neighborhood-level social capital. On the other hand, the study by Wiysonge et al. (2012) used data from sub-Saharan Africa, in which countries were clustering factor at level 2 of an MLM.

2.7 Software packages for multilevel modelling

It has been noted by researchers such as Kwok et al. (2008), and Van Buuren (2011) that *statistical analysis system* (SAS) and *statistical package for the social sciences* (SPSS) software packages can be used to perform multilevel models and latent growth modelling technique to analyse longitudinal data. In addition, the book of Rabe-Hesketh and Skrondal (2008) have also shown that *Statistics and Data* (STATA) software package and R software (Vaughn, 2008) are also suitable for performing multilevel modelling technique for linear and nonlinear models.

Chapter 3

Data Description and Exploration

In this 3rd chapter we discuss the population, structure and variables that are considered in the study. We further explore the variables of the dataset and their relationships. Finally, this chapter provides a brief summary and conclusion that respond to the first objective of this work.

3.1 Population and Data Structure

3.1.1 Population

In this work, census data for health and demographic surveillance system are used. The data are collected and provided by the Africa Center, *KwaZulu-Natal* (KZN) in South Africa. The data consist of population of female teenagers whose pregnancy status was observed during the census years 2011 to 2015 in Mpukunyoni rural area, KZN. This population of female teenagers include 11544 females aged 13 to 19. These are females who were born between 1992 and 2002, inclusive. For the reason that some groups of females from the population of female teenagers, will be observed for at least one of the census year, there are five sub-populations of female teenagers that are formed by each census year. The variable *spft* is used to denote all five sub-populations of female teenagers. In this work, 2011, 2012, 2013, 2014 and 2015 are respectively representing, the first, second, third, fourth and fifth sub-populations. Table 3.1 shows the observed number of female teenagers for each *spft*.

Table 3.1: Population of female teenagers for each *spft*, 2011 - 2015.

<i>spft</i>	Birth year interval	N (%)
2011	1992-1998	7732 (22.1)
2012	1993-1999	7391 (21.1)
2013	1994-2000	6959 (19.9)
2014	1995-2001	6623 (18.9)
2015	1996-2002	6317 (18.0)
	mean (μ)	7004 (20.0)
	std. (σ)	510 (1.5)

N is the number of female teenagers

Table 3.1 shows that 2011 with 7732 female teenagers has the highest sub-population, while the lowest with 6317 is recorded in the year 2015. This table also shows that the number of female teenagers in Mpukunyoni has been decreasing over years. However, the observed sub-populations are fairly distributed across the years since no sub-population is below or above 2σ of each *spft*.

Moreover, we expect to find some groups of females who are observed only in the 2011 or 2015 sub-populations. For example, a female who was 19 years old in 2011 would not be included in the remainder of the sub-populations, while a 13-year old female who was observed in 2015 would not be observed in the previous sub-populations. This means that female teenagers are

not observed equally; hence, the data are unbalanced (Steele, 2008). Furthermore, some female teenagers might not have been observed for some sub-populations because of death, migration or other reasons. Table 3.2 summarises the possible patterns or combinations of observed number of female teenagers in all sub-populations.

Table 3.2: Number of observed females per pattern of subpopulations.

Pattern of <i>spft</i>					N (%)
2011	2012	2013	2014	2015	
1	1	1	1	1	2775 (24.0)
1	1402 (12.1)
1	1	.	.	.	1368 (11.9)
1	1	1	.	.	1134 (9.8)
1	1	1	1	.	1053 (9.1)
.	1	1	1	1	919 (8.0)
.	.	1	1	1	906 (7.9)
.	.	.	1	1	861 (7.5)
.	.	.	.	1	856 (7.4)
(other patterns)					270 (2.3)
Total					11544 (100.0)

N is the number of female teenagers

From Table 3.2, 2775 (24.0%) female teenagers have been observed in all five sub-populations, followed by 1402 (12.1%) female teenagers who were observed only in 2011. From the third highest to the fifth highest numbers of female teenagers, there are those who were observed in the first two (11.9%), first three (9.8%) and first four (9.1%) sub-populations.. We also note that the number of female teenagers who were observed between the last four sub-populations and the last sub-population has been decreasing. This explains the reduction in the number of female teenagers from 2011 to 2015, which was observed in Table 3.1.

3.1.2 Data structure

Based on the multilevel data structures (discussed in section 2.3), the data for this work form a two level data structure. As we have seen in Table 3.2, there are cases where more than one measurement of teenage pregnancy status is observed for one female, meaning that this measurements are at level 1 while female teenagers (denoted by *id*) are at level 2. We can simply say that measurements of teenage pregnancy are clustered by female teenagers. Table 3.3 presents the number of unique identifiers and the number of units at each level.

Table 3.3: The number of unique identifiers and units within the levels.

Level	Level name	Range	Unique identifiers	Units
1	Measurements of pregnancy status	[2011, 2015]	5	35022
2	Female teenagers	[0, 11544]	11544	11544

In Table 3.3, there are 11544 unique identifications of female teenagers at level 2, thereby confirming the population of female teenagers in this study. Hypothetically, if the same group of female teenagers were observed for five census years, level 1 will have $5 \times 11544 = 57720$ units. However, since the sizes of sub-populations are not equal, only 35022 units were recorded. The number of units, 11544, which is equal to the number of unique identifiers at level 2, means that no female teenager was repeated at level 2. This implies that there is no possibility of multiple membership data structure. Furthermore, cross-classified structure is not possible because there is only one level 2 cluster. A classification diagram, Figure 3.1, summarises a two-level data structure of the dataset for this work.

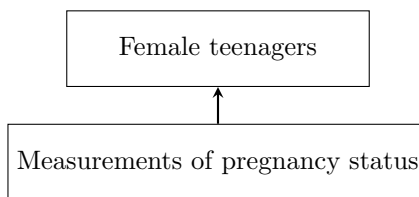


Figure 3.1: Classification diagram for a two-level pregnancy dataset (a).

In Figure 3.1, the arrow pointing to the female teenagers indicates that the measurements of pregnancy status at level 1 are nested within female teenagers at level 2. However, this diagram does not indicate the different sub-populations that are presented in Tables 3.1 and 3.2. For example, in our dataset id_1 's pregnancy status was recorded only in 2011 while the pregnancy status for id_2 and id_{11544} are recorded for the last four and first four sub-populations, respectively. The data structure for the selected ids can be represented in a more elaborate classification diagram (Figure 3.2).

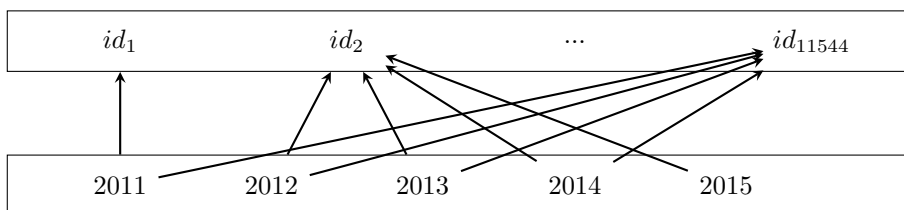


Figure 3.2: Classification diagram for a two-level pregnancy dataset (b).

In Figure 3.2, it is clear that the arrows pointing to id_1 , id_2 and id_{11544} indicate that pregnancy

statuses were recorded for the first, last four and first four sub-populations, respectively.

3.2 Variables of the dataset

3.2.1 Response variable

The response variable of interest, in this work, is the pregnancy status and it is denoted by *ps*. This is a response variable that takes the values 0 “no” or 1 “yes”; responding to a question of females being pregnant or not pregnant. Table 3.4 shows the frequencies of pregnancies of females by census year.

Table 3.4: The number of observations for each pregnancy status category by year.

	2011	2012	2013	2014	2015	Mean
<i>ps</i>	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
no	7464 (96.5)	7194 (97.3)	6791 (97.6)	6483 (97.9)	6232 (98.7)	6833 (97.6)
yes	268 (3.4)	197 (2.7)	168 (2.4)	140 (2.1)	85 (1.3)	172 (2.5)

N is the number of observations for each *ps* category

The observed average proportion of teenage pregnancies from 2011 to 2015 is 2.5% across *spft* (see Table 3.4). Table 3.4 shows that the proportion within *spft* has been decreasing from 2011 to 2015.

3.2.2 Covariates

Our dataset includes three categorical covariates, namely age of the female, the number of household membership and the number of pregnancies the female had before observation. These three covariates are also recorded at each census year.

- *Census year*

In the previous sections census year was presented as a factor variable *spft*, which represents different sub-populations. However, census year can also be treated as a measurement of occasions, denote by *year* that records values 0, 1, 2, 3 and 4, representing census years 2011, 2012, 2013, 2014 and 2015, respectively. This is because some female teenagers are studied at least once across *year*.

However, we can further show the frequency of number of measurements of occasion (denoted by *nyear*). Thus, we compute the number of non-missing response observations per *id* and examine the distribution of these numbers across female teenagers.



Figure 3.3: Histogram for the frequency of number of occasion measurements across females.

Figure 3.3 shows the same number, 2775, as in Table 3.2 for female teenagers who were observed in all five subpopulations. We also note that number of females who were observed only once are also higher than those who were observed two, three or four times.

- *Age of the female teenage*

Female age is a categorical covariate which is denoted by *age* and it varies within *spft* and across *year*. It is coded as 0 (used as a reference category), 1, 2, 3, 4, 5 or 6 for females who are 13, 14, 15, 16, 17, 18 or 19 years old, respectively.

Table 3.5: The number of observations for each female age category within *spft* and across *year*.

<i>age</i>	2011	2012	2013	2014	2015	across <i>year</i>
	N (%)	N (%)	N (%)	N (%)	N (%)	
13	1048 (13.6)	1061 (14.4)	996 (14.3)	899 (13.6)	856 (13.5)	4860 (13.9)
14	1110 (14.4)	1009 (13.6)	1001 (14.4)	942 (14.2)	861 (13.6)	4923 (14.1)
15	1112 (14.4)	1067 (14.4)	957 (13.8)	954 (14.4)	906 (14.3)	4996 (14.3)
16	1085 (14.0)	1072 (14.5)	1016 (14.6)	926 (14.0)	919 (14.6)	5018 (14.3)
17	1072 (13.8)	1041 (14.1)	1026 (14.7)	981 (14.8)	903 (14.3)	5023 (14.3)
18	1155 (14.9)	1035 (14.0)	986 (14.2)	976 (14.7)	940 (14.9)	5092 (14.5)
19	1150 (14.9)	1106 (15.0)	977 (14)	945 (14.3)	932 (14.8)	5110 (14.6)
Total	7732 (22.1)	7391 (21.1)	6959 (19.9)	6623 (18.9)	6317 (18.0)	35022 (100.0)

N is the number of observations for each *age* category

In all the categories of *age*, but 13, the number of observations reduces from 2011 to 2015. We also note fluctuations in the number of observations between ages within *spft* and across

year.

- *The number of household membership for a female*

The variable name *idhhms* denotes the number of households a female is a member to. It takes the category values 0, 1, 2 and 3 which represent 1, 2, 3 and 4 household membership for a female, respectively. This is a categorical variable that also varies within and across *year*.

Table 3.6: The number of females for each household membership category within *spft* and across *year*.

<i>idhhms</i>	2011	2012	2013	2014	2015	across <i>year</i>
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
1	6815 (88.1)	6633 (89.7)	6366 (91.5)	6149(92.9)	5913(93.6)	31876(91.0)
2	820 (10.6)	704 (9.5)	560 (8.1)	452(6.8)	390(6.2)	2926(8.4)
3	91 (1.2)	50 (0.7)	33 (0.5)	22(0.3)	14(0.2)	210(0.6)
4	6 (0.1)	4 (0.1)	0 (0.0)	0(0.0)	0(0.0)	10(0.0)
Total	7732(100.0)	7391(100.0)	6959(100.0)	6623(100.0)	6317(100.0)	35022(100.0)

N is the number of observations for each *idhhms* category

For both within *spft* and across *year*, the observations are decrease as household membership increases (see Table 3.6). We also see, in this table, that from 2013 to 2015 there were no female teenagers belonging to 4 households.

- *The number of pregnancies the female teenager had before census year of observation*

The number of pregnancies the female teenager had before census year of observation is denoted by *nch* and it is also categorised by 4 values (0, 1, 2 and 3).

Table 3.7: The number of pregnancies the female teenager had before year of observation within *spft* and across *year*.

<i>nch</i>	2011	2012	2013	2014	2015	across <i>year</i>
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
0	7381 (95.5)	7078 (95.8)	6737 (96.8)	6405 (96.7)	6137 (97.2)	33738 (96.3)
1	330 (4.3)	294 (4)	207 (3)	208 (3.1)	177 (2.8)	1216 (3.5)
2	20 (0.3)	19 (0.3)	14 (0.2)	9 (0.1)	3 (0.1)	65 (0.2)
3	1 (0.0)	0 (0.0)	1 (0.0)	1 (0.0)	0 (0.0)	3 (0.0)
Total	7732 (100.0)	7391 (100.0)	6959 (100.0)	6623 (100.0)	6317 (100.0)	35022 (100.0)

N is the number of observations for each category of *nch*

Table 3.7 shows that 330, 20 and 1 female teenager(s in *spft*₁ had already been pregnant once,

twice and thrice before they were observed in 2011, respectively.

3.3 Data exploration

This section starts by examining the descriptive statistics of the response variable, which include assessing its mean and standard deviation. Thereafter, this section will examine the relationship between *ps* and all covariates.

3.3.1 Examining the mean and standard deviation of variables

Table 3.8 presents the mean, standard deviation, minimum and maximum of *ps* by within *spft* and across *year*.

Table 3.8: Descriptive summary of teenage pregnancy status and all four covariates.

	N	Mean	Std. Dev.	Min	Max
2011	7732	0.034	0.183	0	1
2012	7391	0.027	0.161	0	1
2013	6959	0.024	0.153	0	1
2014	6623	0.021	0.144	0	1
2015	6317	0.013	0.115	0	1
across <i>year</i>	35022	0.025	0.155	0	1

N is the number of observed pregnancy status

The estimated mean across census years is 0.025, which deduces that there is approximately 2.5% chance of a teenage pregnancy across *year*. Using the estimated value ($\hat{\pi}$), the standard deviation in Table 3.8 is expected to be equal to $\hat{\pi}(1 - \hat{\pi}) = 0.025(0.975)$, which is indeed the same. In order to confirm the value of the mean as well as the possible alternatives of the response variable, we use the results in Table 3.4. Clearly the results in Table 3.8 confirm that indeed the 2.5% is distributed to the females whose *ps* is yes. Table 3.8 also shows that there is a variation of pregnancy status within *spft*. This variation of *ps* can be examined for each covariate across census years and within each sub-population.

3.3.2 Exploring the relationship of pregnancy status and the covariates

Unlike in linear regression where we could plot a line graph to examine a bivariate relationship between the response and covariates, such a line graph will not be very informative. For this reason, a line graph of proportions of teenage pregnancy across and within census years by each covariate could be calculated and plotted against the across categories of the covariates.

The relationships between the proportions of teenage pregnancy *spft* and across *year* by *age*, *idhhms* and *nch* are, respectively, presented using line graphs in Figures 3.4, 3.5 and 3.6.

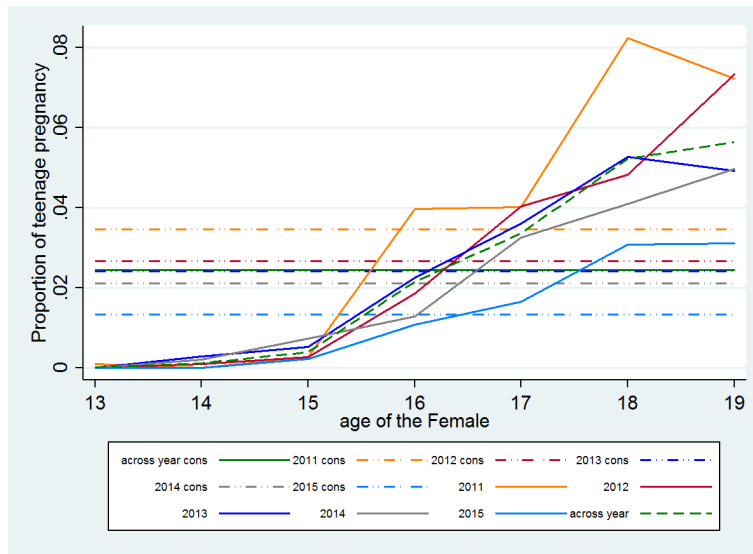


Figure 3.4: The relationship proportion of teenage pregnancy and *age*.

In Figures 3.4 - 3.6, the constant lines (dash-dot) are proportions of pregnancy status which are the mean values for each sub-population in Table 3.8. The green dash lines, in Figures 3.4, 3.5 and 3.6, represent the proportions that are averaged over each covariate *age*, *idhhms* and *nch*, respectively. On the other hand, the rest of the non-constant line plots are averaged over each sub-population and each covariate.

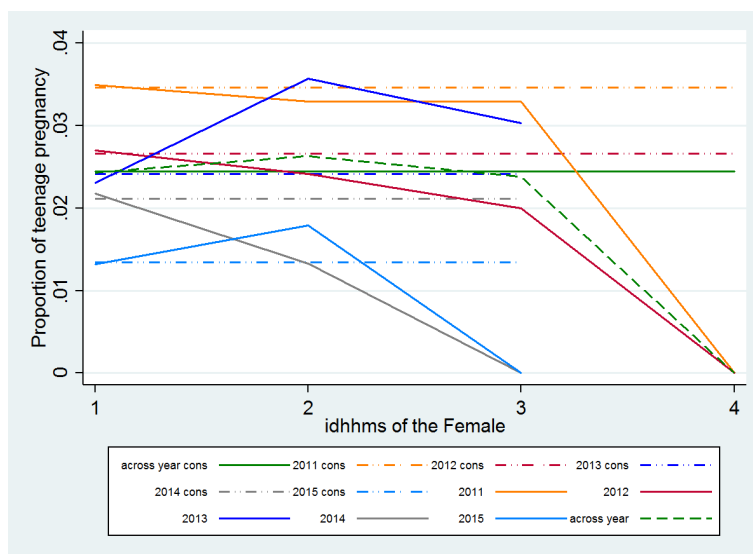


Figure 3.5: The relationship proportion of teenage pregnancy and *idhhms*.

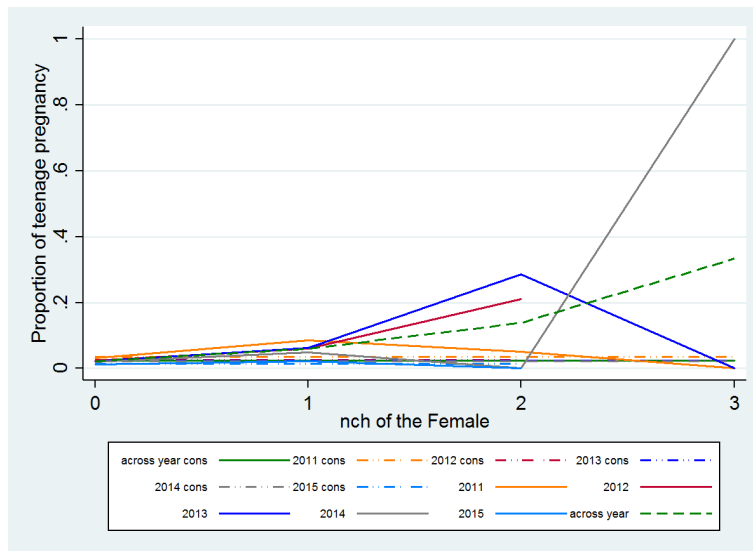


Figure 3.6: The relationship proportion of teenage pregnancy and *nch*.

Figure 3.6 shows a small variation of proportion of pregnancy status between sub-populations when the number of pregnancies the female had before she was observed is between 0 and 1. This same case is also seen for ages 13-15 in Figure 3.4.

The proportions for female teenagers age between 13 and 15 are all lower than the average proportion for each sub-population. This means that, on average, the female teenagers age between 13 and 15 are less likely to get pregnant. In Figures 3.4, 3.5 and 3.6, we also note that proportions of teenage pregnancy increases as age increase within each *spft* and across *year*. This suggests a positive linear relationship between teenage pregnancy and age of female teenagers. The proportions of 2012 (red solid line) are almost similar to the across *year* by age (green dash line) in all figures (see Figures 3.4, 3.5 and 3.6).

Nonetheless, Figure 3.5 shows a high variation in proportions between sub-populations for each category of *idhhms*. The same case of high variation is also observed in Figure 3.6 for *nch* between 2 and 3. Although the linear relationship of proportions and *idhhms* is not clear, it becomes clear from the second category *idhhms* that there is a negative relationship (see Figure 3.5). The highest proportions of approximately 10% is observed when the number of pregnancies the female had before she was observed is 3 in *spft*₄ (see Figure 3.6). This is followed by a proportion which is just above 8% for females who are 18 years old in *spft*₁ (see Figure 3.4).

3.4 Summary

In summary, we have observed teenage pregnancy status (denoted by *ps*) of 11544 female teenagers across five census years. However, for each census year a female from a population of female teenagers of size 11544 was observed at most once within each census year but at least once across (see table 3.2). The data in this work are purely hierarchical, in which multiple teenage pregnancy measurements at level 1 are clustered by females (denoted by *id*) at level 2. There are 35022 (data lines of the dataset) teenage pregnancy measurements that were observed across census years.

In addition to the variables *ps* and *id*, the dataset for this work include 5 other variables that are treated as covariates (*spft*, *year*, *age*, *idhhms* and *nch*). The variable *ps* is the response variable that records whether a female teenager is pregnant or not. All five covariates are categorical but three (*age*, *idhhms* and *nch*) are female characteristics that changes over time covariate *year*. Even more, *spft* is treated as a factor covariate in order to account for different sub-populations.

Based on the observed relationship between main covariates and teenage pregnancy, teenage pregnancy increases as *age* increases but reduces as *year* and *idhhms* increases. However, there is no clear relationship between *nch* and teenage pregnancy. We have also seen that these relationships vary across the sub-populations, hence it important to investigate interaction of each covariate with *year*.

Going forward, this work will use generalised linear models and their extensions in order to model teenage pregnancy status given the nature of the variables *id*, *spft*, *year*, *age*, *idhhms* and *nch*.

Chapter 4

Prediction of the Probability of Teenage Pregnancy using a Logit GLM

The sections of this chapter include introduction on generalised linear models (GLMs), methodology of GLM for binary data, specification and fitting a multiple logistic regression model, and conclusion.

4.1 Introduction

In the context of *generalised linear models* (GLMs) the response variables Y_i are assumed to be generated from some specific distribution in the exponential family (Molenberghs and Verbeke, 2005). In our study the response variable is binary, hence the section on methodology of GLM in this chapter will cover linear predictor, link function, distribution and maximum likelihood estimation for binary data.

The logit and probit link functions can be utilised to perform analysis for discrete outcome data (Rabe-Hesketh and Skrondal, 2008). Nonetheless, we exclusively consider the logit function as it is sufficient to map the *logistic regression model* (LRM) to a linear predictor. The Bernoulli probability density function will be considered for the discussion of the distribution of LRM because it deals with response data that are having two possible outcomes (Czepiel, 2002; Molenberghs and Verbeke, 2005). More often, in GLMs for either continuous or discrete response, the expected value of the response given some covariates is usually of interest (Molenberghs and Verbeke, 2005). For example, for a simple linear regression model written as $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; researchers are more interested in a linear predictor $\beta_0 + \beta_1 x_i$, which is the expectation of y_i given the covariates x_i .

4.2 Methodology of generalised linear model for binary data

4.2.1 Definition of generalised linear models

To define GLM in general, we first let an independent set of response variables, Y_1, Y_2, \dots, Y_N , correspond to p -dimensional vectors of covariate, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$. We further assume that the Y_i 's are having a probability density function $f(y_i|\theta_i, \rho)$ belonging to an exponential family. The GLM is then defined as

$$E(y_i|x_i) = \mu = g(\mathbf{x}'_i\boldsymbol{\beta}) \quad (4.2.1)$$

wherein $E(y_i|x_i)$ is the expected value of y_i given the covariates x_i and μ is the mean. This expectation can be written as some known function, $g(\cdot)$, of a covariate vector, \mathbf{x}_i and a p -dimension vector with fixed unknown coefficients, $\boldsymbol{\beta}$. Alternatively, we can write the linear function by use of the inverse function of $h(\cdot)$, sometimes written as $g^{-1}(\cdot)$, thus

$$h(E(y_i|x_i)) = g^{-1}(E(y_i|x_i)) = \mathbf{x}'_i\boldsymbol{\beta}.$$

4.2.2 Formulation of generalised linear models for binary data

Binary data are obtained in a setup where their data take two possible options, in which a specific option is true and another is false. These options can include, for instance, a trial succeeding or failing; or a person being dead or alive. Habitually, data scientists and analysts would represent these options as 0 and 1, where 0 would mean “no” and 1 “yes”. In a binary response data setup, the expectation of a particular response given some covariates is a probability value, thereby predicting a value between 0 and 1. This probability is commonly referred to as probability of success and it is expressed mathematically as

$$E(y_i|x_i) = \mu = P(y_i = 1|x_i) \quad (4.2.2)$$

where μ is the mean. Conversely, $P(y_i = 0|x_i)$ (equal to $1 - P(y_i = 1|x_i)$) is the probability of failure. For the reason that $P(y_i = 1|x_i)$ is restricted between the values 0 and 1, the $E(y_i|x_i)$ results in a nonlinear predictor function that can be written as,

$$P(y_i = 1|x_i) = g(\beta_0 + \beta_1 x_i), \quad (4.2.3)$$

same as in Model (4.2.1). By introducing a link function $h(\cdot)$ such that $g(\cdot)$ is an inverse link function, it follows that function (4.2.3) is equivalent to

$$h(P(y_i = 1|x_i)) = \beta_0 + \beta_1 x_i. \quad (4.2.4)$$

As in linear regression, Model (4.2.4) allows us to examine the relationship that some covariates (e.g. dummy and/or polynomial) can have on the binary responses. We will make use of the logit function. The logit function is defined as

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

Therefore, the expression (4.2.4) can be derived to

$$\begin{aligned} \text{logit}(P(y_i = 1|x_i)) &\equiv \log\left(\frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)}\right) \\ &= \beta_0 + \beta_1 x_i, \end{aligned} \quad (4.2.5)$$

which is called a logistic regression model with one covariate. In a case where p covariates $x_{i1}, x_{i2}, \dots, x_{ip}$ are considered, expression (4.2.5) can be extended to

$$\begin{aligned} \log\left[\frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)}\right] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \\ &= \sum_{k=0}^p \beta_k x_{ik}, \end{aligned} \quad (4.2.6)$$

where $x_{i0} = 1$. A logistic regression model with $p > 1$ covariates is called a *multiple logistic regression model* (MLRM). The left hand side of equation (4.2.6) can be interpreted as the log of the odds of a success against a failure. The positive value of the coefficient β indicates an increase in the odds of success while a negative value indicates a decrease. The odds of success given some covariates can thus be derived as

$$\text{Odds}(y_i = 1|x_i) = e^{\sum_{k=0}^p \beta_k x_{ik}}. \quad (4.2.7)$$

Furthermore, the probability of success given some covariates is given by

$$P(y_i = 1|x_i) = \frac{\text{Odds}(y_i = 1|x_i)}{1 + \text{Odds}(y_i = 1|x_i)}. \quad (4.2.8)$$

As a result, the probability of success can be predicted using this function

$$P(y_i = 1|x_i) = \frac{e^{\sum_{k=0}^p \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^p \beta_k x_{ik}}}, \quad (4.2.9)$$

which is referred to as the inverse logit function of the linear predictor ($\text{logit}^{-1}[P(y_i = 1|x_i)]$).

4.2.3 Distribution of binary data

Following the concept and notation illustrated by Molenberghs and Verbeke (2005), a probability density function that is of exponential family can be expressed as,

$$f(y) \equiv f(y|\theta, \phi) = e^{\phi^{-1}[y\theta - \psi(\theta)] + c(y, \phi)}, \quad (4.2.10)$$

A Bernoulli probability density for an independent set of response variables, Y_1, Y_2, \dots, Y_N , where Y_i takes on values 0 or 1 and follows a Bernoulli distribution with $P(y_i = 1|x_i) = \pi_i$,

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

is written as

$$f(y_i|\theta, \phi) = e^{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)}, \quad (4.2.11)$$

and can be rewritten as

$$f(y_i) = \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}. \quad (4.2.12)$$

In expression (4.2.12), the term ϕ is equal to 1 and θ is a function of π . The mean, $\mu = \pi$, and the variance of the mean, $v(\mu) = \pi(1 - \pi)$, are derived using the first and second moments. The variance, $v(\mu)$, is generally written as $\phi v(\mu)$ and it is called the variance function. Estimation of the regression coefficients is done using the principle of maximum likelihood, which will be discussed in subsection 4.2.4.

4.2.4 Maximum likelihood estimation of binary data

The likelihood function is the probability density function. The only difference is that the likelihood function, $L(\cdot)$, is a function of the parameter, $\boldsymbol{\beta}$, given the fixed responses \mathbf{y} . A GLM likelihood function can be derived using the exponential density function in expression (4.2.10) as,

$$\begin{aligned} f(\mathbf{y}_i|\boldsymbol{\theta}_i, \phi) &= e^{\phi^{-1}[y_1\theta_1 - \psi(\theta_1)] + c(y_1, \phi)} \times \dots \times e^{\phi^{-1}[y_N\theta_N - \psi(\theta_N)] + c(y_N, \phi)} \\ &= e^{\phi^{-1} \sum_{i=1}^N [y_i\theta_i - \psi(\theta_i)] + \sum_{i=1}^N c(y_i, \phi)} \\ &= L(\boldsymbol{\beta}|\mathbf{y}_i, \phi). \end{aligned} \quad (4.2.13)$$

The log-likelihood function $l(\boldsymbol{\beta}|\mathbf{y}_i, \phi)$, which is the natural logarithm of the likelihood function ($\log[L(\boldsymbol{\beta}|\mathbf{y}_i, \phi)]$), is commonly used in practice to derive the likelihood estimate, as this simplifies calculations. The log-likelihood function for general GLM is

$$l(\boldsymbol{\beta}|\mathbf{y}_i, \phi) = \frac{1}{\phi} \sum_{i=1}^N [y_i\theta_i - \psi(\theta_i)] + \sum_{i=1}^N c(y_i, \phi). \quad (4.2.14)$$

We can derive the log-likelihood function for a probability density function of a Bernoulli distribution by substituting the following terms

$$y_i\theta_i - \psi(\theta_i) = y \log \left(\frac{\pi}{1 - \pi} \right) \text{ and } \psi(\theta_i) = \log(\pi - 1)$$

as done in functions (4.2.12) and (4.2.14). Thus,

$$\begin{aligned} l(\boldsymbol{\beta}|\mathbf{y}_i, \phi) &= \sum_{i=1}^N y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^N \log(1 - \pi_i) \\ &= \sum_{i=1}^N y_i \log \left(e^{\sum_{k=0}^p \beta_k x_{ik}} \right) + \sum_{i=1}^N \log \left(1 - \frac{e^{\sum_{k=0}^p \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^p \beta_k x_{ik}}} \right) \\ &= \sum_{i=1}^N y_i \sum_{k=0}^p \beta_k x_{ik} + \sum_{i=1}^N \log \left(\frac{1}{1 + e^{\sum_{k=0}^p \beta_k x_{ik}}} \right) \\ &= \sum_{i=1}^N \left(y_i \left[\sum_{k=0}^p \beta_k x_{ik} \right] - \log \left(1 + e^{\sum_{k=0}^p \beta_k x_{ik}} \right) \right) \end{aligned} \quad (4.2.15)$$

is the log-likelihood function for a logistic regression with a Bernoulli distribution.

4.3 Specifying and fitting a logit generalised linear model

Our analysis will use the data with an intention to predict the log odds of teenage pregnancy for females during the census years of 2011 to 2015 at the area of Mpukunyoni, KNZ. This dataset has already described and explored in chapter 3. The purpose of this analysis is to estimate,

using maximum likelihood, the coefficients of several covariates that are suspected to have an effect on teenage pregnancy. We will in this manner use these coefficients to predict and picture the relationship of the predicted teenage pregnancy probabilities and values of the covariates. Furthermore, we will likewise take note of the *Akaike information criterion* (AIC) value that we will use to select the best model.

The analysis includes selected variables, namely *year*, *age*, *idhhms* and *nch* from our dataset. The other covariates that will be added during the model building are interactions between year and other covariates.

In the subsections of this section, we will build a logit GLM or LRM. Thus, we shall fit a logit GLM to predict the probability of teenage pregnancy ($P(p\hat{s}_i)$) given the covariates, where $i = 1, 2, \dots, 35022$ and $ps_i = 1$ indicates a female teenager who is pregnant. We use the notation β_{year} , β_{age} , β_{ihm} and β_{nch} to denote the fixed effects of *year*, *age*, *idhhms* and *nch*. Furthermore, the effects of the interaction of *year* and other covariates (*year* × *age*, *year* × *idhhms* and *year* × *nch*) is denoted by β_{yage} , β_{yihm} and β_{ynch} , respectively. The inclusion of covariates in the final model depends on whether the coefficient is significant at 5% level of significance in the univariate model.

4.3.1 Model building

The model assumes a single and unique population across census years (*year*). Our model building start by fitting a null empty logit GLM. This is a model with no covariates and it is written as

$$\text{logit}(\pi_i) = \beta_0 \quad (4.3.1)$$

where \mathbf{x}_i is a vector of covariates and β_0 is the intercept. The estimates of Model (4.3.1) shows that on average the log odds of teenage pregnancy is -3.684, which is the magnitude of β_0 (see Table 4.1). It is expected that when β_0 is exponentiated and divided by one plus the exponentiated value ($e^{-3.684}/(1 + e^{-3.684}) = 0.0245$), it is equal to the across *year* proportion of teenage pregnancy as presented in Table 3.8.

The next step of the logit GLM building we takes into consideration that *year* is an important covariate or grouping factor; hence, it is added first before adding other covariates, thus

$$\text{logit}(\pi_i) = \beta_0 + \beta_{year}year_i. \quad (4.3.2)$$

The likelihood estimates of Model (4.3.2) are also shown in Table 4.1. At 5% level of significance, Model (4.3.2) estimates show that **year** is a significant negative effect with a magnitude of approximately 0.207 on the log odds of teenage pregnancy (see Table 4.1). Moreover, the inclusion of **year** increases the magnitude of β_0 to -3.332. This means that the probability of teenage pregnancy for a female in 2011 is estimated as $e^{-3.332}/(1 + e^{-3.332}) = 0.0345$ and it is approximately equal to the proportion in Table 3.8. Since the effect of **year** reduces the log odd of teenage pregnancy by 3.332, the probability of teenage pregnancy in 2012, 2013, 2014 and 2015 will be approximately 0.028, 0.023, 0.019 and 0.015.

The logit GLM building is continued by adding the effect of **idhhms** and **nch** to Model (4.3.2), thus

$$\text{logit}(\pi_i) = \beta_0 + \beta_{\text{year}}\text{year}_i + \beta_{\text{age}}\text{age}_i + \beta_{\text{idhhms}}\text{idhhms}_i + \beta_{\text{nch}}\text{nch}_i \quad (4.3.3)$$

Table 4.1 also shows that by including other main covariates (Model (4.3.2)), only **age** effect is significant at 5% significant level. However, **idhhms** and **nch** are not significant at 5% level of significance. The effect of these insignificant covariates will be removed one-by-one based on the z statistics closer to zero (see Table 4.1). The estimates were obtained by fitting models

$$\text{logit}(\pi_i) = \beta_0 + \beta_{\text{year}}\text{year}_i + \beta_{\text{age}}\text{age}_i + \beta_{\text{nch}}\text{nch}_i \quad (4.3.4)$$

and

$$\text{logit}(\pi_i) = \beta_0 + \beta_{\text{year}}\text{year}_i + \beta_{\text{age}}\text{age}_i. \quad (4.3.5)$$

The estimates of Model (4.3.5) in Table 4.1 show that the intercept and main covariates **year** and **age** are significant with all p -values of less than 0.001.

Moreover, one other interest that was explored in Chapter 3 was whether the effect of age on teenage pregnancy depends on the census year. We therefore add the interaction effect of the two significant covariates **year** and **age** to Model (4.3.5). This interaction covariate is denoted by **year** \times **age**.

$$\text{logit}(\pi_i) = \beta_0 + \beta_{\text{year}}\text{year}_i + \beta_{\text{age}}\text{age}_i + \beta_{\text{yage}}\text{year}_i \times \text{age}_i \quad (4.3.6)$$

Model (4.3.6) estimates indicate that the effect of **year** \times **age** is non-significant. This means that Model (4.3.5) is the only model with significant fixed effects.

Table 4.1: Estimates of model building for a logit GLM of pregnancy status.

Covariate	Model					
	(4.3.1)	(4.3.2)	(4.3.3)	(4.3.4)	(4.3.5)	(4.3.6)
Intercept	-3.684*** (-106.59)	-3.332*** (-64.27)	-5.525*** (-44.20)	-5.528*** (-44.47)	-5.542*** (-44.61)	-5.588*** (-31.39)
<i>year</i>		-0.207*** (-8.12)	-0.208*** (-8.07)	-0.207*** (-8.07)	-0.210*** (-8.17)	-0.180* (-2.07)
<i>age</i>			0.549*** (22.73)	0.549*** (22.73)	0.557*** (23.42)	0.567*** (15.84)
<i>idhhms</i>			-0.0189 (-0.18)			
<i>nch</i>			0.191 (1.78)	0.191 (1.78)		
<i>year</i> × <i>age</i>						-0.01 (-0.36)

z statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3.2 Model selection and interpretation of a better fit model

Model selection is done using AIC value which provides a statistic that determines power of the fitted model (Akaike, 2011). The AIC value is calculated using the formula $AIC = -2\ln[l(\beta)] + 2k$, where k is the number of estimated coefficients (Akaike, 2011). A lower AIC value indicates a better fit. The AIC results for all models in Table 4.1 are summarised in Table 4.2.

Table 4.2: Akaike’s information criterion for each logit GLM of pregnancy status.

Covariate	Model					
	(4.3.1)	(4.3.2)	(4.3.3)	(4.3.4)	(4.3.5)	(4.3.6)
Observations	35022	35022	35022	35022	35022	35022
$l(\beta)$	-4029.84	-3995.66	-3619.52	-3619.54	-3621.04	-3620.973
df	1	2	5	4	3	4
AIC	8061.67	7995.31	7249.05	7247.08	7248.08	7249.95

Based on the AIC results in Table 4.2, Model (4.3.4) has the lowest AIC value of 7247.08. However, this model include an insignificant effect of *nch*; hence, the model cannot be selected as

a better fit. On the other hand Model (4.3.5) in Table 4.1 has the lowest AIC value, after Model (4.3.4), of 7248.08; hence Model (4.3.5) is better fit logit GLM to predict teenage pregnancy. Table 4.3 presents the likelihood estimates of a better fit logit GLM of pregnancy status.

Table 4.3: Likelihood estimates of a better fit logit GLM of pregnancy status.

Covariate	Effect	Std. Err.	z -value	p -value	[95% Conf. Interval]
Intercept	-5.542	0.124	-44.610	0.000	[-5.785, -5.298]
<i>year</i>	-0.210	0.026	-8.170	0.000	[-0.260, -0.159]
<i>age</i>	0.557	0.024	23.420	0.000	[0.510, 0.604]

Based on the results in Table 4.3, the final model that is of better fit is written as,

$$\text{logit}(\pi_i) = -5.542 - 0.21\text{year}_i + 0.557\text{age}_i. \quad (4.3.7)$$

Model (4.3.7) will be interpreted and used to predict the probabilities of teenage pregnancy given *year* and *age*.

Based on Model (4.3.7), the magnitude of $\beta_0 = -5.542$ suggests that the probability of teenage pregnancy for a 13-year old female teenager in 2011 is estimated as $e^{-5.542}/(1 + e^{-5.542}) = 0.0039$. For every 1 additional year to the age of a female teenager, the log odds of teenage pregnancy will increase by 0.557, in 2011. However, this log odds will be reduced every year by a value 0.210, which means that the corresponding log odds of teenage pregnancy for the year 2012, 2013, 2014 and 2015 are -5.752, -5.962, -6.172 and -6.382, respectively.

4.3.3 Prediction of the probability of teenage pregnancy

The predictions of the probabilities using Model (4.3.7) are calculated individually by substituting the data point of the subjected main covariates into Equation (4.2.9) while holding other predictors at their base (meaning category zero). Nonetheless, Model (4.3.7) shows that the probabilities of teenage pregnancy are predicted by *year* and *age*.

The interaction covariate between *year* and *age* was not significant; therefore the predictions of teenage pregnancy based on *age* by *year* would not be valid. However, probability predictions that are averaged over *age* will be done for when *year* is held at zero (base) and when it is not.

Figure 4.1 shows predicted probabilities of teenage pregnancy averaged over *age* across *year* and within 2011.

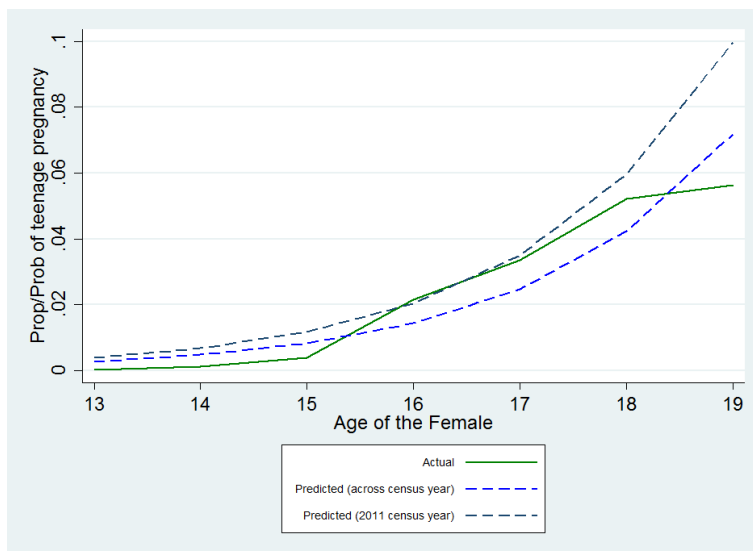


Figure 4.1: The relationship of probabilities and proportion of teenage pregnancy with female’s ages.

In Figure 4.1, the dotted lines are the predicted probabilities whereas the solid one is the actual proportion that is averaged over age. As expected, the risk of teenage pregnancy is higher for older females. The probability predictions across census year seem to be below the one for 2011. This means that, in 2011, the risk of teenage pregnancy was higher for older females than when census years are averaged. Nonetheless, these probabilities are slightly different from the proportions; hence, they are slightly acceptable.

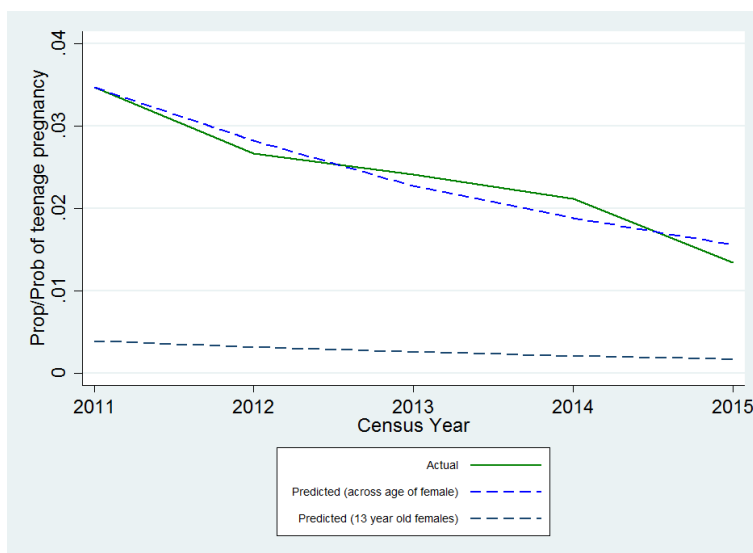


Figure 4.2: The relationship of probabilities and proportion of teenage pregnancy with census year.

Figure 4.2 further shows the predicted probabilities of teenage pregnancy averaged over *year*

across *age* and within 13-year old females. Although the plotted line in Figure 4.2 shows that the risk of teenage pregnancy decreases as census year increases, these probabilities are much lower for 13-year old females. Given this and the predicted probabilities across female age, this confirms that older female teenagers are at higher risk of teenage pregnancy than those younger ones. Over and above, by excluding the effect of age 14 to 19 years, the risk of teenage pregnancy is definitely underestimated.

4.4 Logit generalised linear model building with census year as a factor covariate

In the previous logit GLM population differences effect across census years have been pushed aside. However, this section treats census year (*year*) as a factor covariate, *spft*.

4.4.1 Adding to the effect of different sub-populations to a null logit GLM

With the reference category of *spft* set as 2011, a logit GLM with covariate *year* only will be written as

$$\text{logit}(\pi_i) = \beta_0 + (\beta_{p2}2012_i + \beta_{p3}2013_i + \beta_{p4}2014_i + \beta_{p5}2015_i) \quad (4.4.1)$$

where β_0 is the fixed intercept. On the other hand,

- β_{p2} is the slope of some female in 2012 sub-population (here there is a zero effects of female teenagers in 2011, 2013, 2014 and 2015),
- β_{p3} is the slope of some female in 2013 sub-population (here there is a zero effects of female teenagers in 2011, 2012, 2014 and 2015),
- β_{p4} is the slope of some female in 2014 sub-population (here there is a zero effects of female teenagers in 2011, 2012, 2013 and 2015) and
- β_{p5} is the slope of some female in 2015 sub-population (here there is a zero effects of female teenagers in 2011, 2012, 2013 and 2014).

The results of Model (4.4.1) are presented in Table 4.4. The shows the maximum likelihood estimates of the categories of *spft* on log odds of teenage pregnancy. Based on these estimates, the probabilities of teenage pregnancy for female teenagers in 2011, 2012, 2013, 2014 and 2015 sub-populations are 0.034, 0.027, 0.024, 0.021 and 0.013, respectively. This values tie with the actual proportions of teenage pregnancy in Table 3.8. However, the value of $AIC = 7996.857$

Table 4.4: Likelihood estimates of Model (4.4.1).

Covariate	Effect	Std. Err.	z-value	p-value	[95% Conf. Interval]
Intercept	-3.327	0.062	-53.51	0.000	[-3.449, -3.205]
<i>spft</i>					
2012	-0.271	0.095	-2.840	0.004	[-0.458, -0.084]
2013	-0.373	0.100	-3.730	0.000	[-0.568, -0.177]
2014	-0.508	0.106	-4.810	0.000	[-0.716, -0.301]
2015	-0.968	0.126	-7.700	0.000	[-1.214, -0.722]

$l(\beta) = -3993.429$, Observations = 35022, AIC = 7996.857

for Model (4.4.1) is larger than that value of Model (4.3.2) by a difference of 1.545. This means that Model (4.4.1) predicts teenage pregnancy better than Model (4.3.2).

4.4.2 Adding to the effect age of the female to a logit GLM with sub-population effect
 Since the models in section 4.3 suggested that age of the female (*age*) is a significant factor of teenage pregnancy risk, the effect of age can be added to Model (4.4.1), thus

$$\text{logit}(\pi_i) = \beta_0 + (\beta_{p2}2012_i + \beta_{p3}2013_i + \beta_{p4}2014_i + \beta_{p5}2015_i) + \beta_{age}age_i. \quad (4.4.2)$$

The effect of the female's age is still represented by the magnitude of β_{age} . Table 4.5 presents the results of the Model (4.4.2).

Table 4.5: Likelihood estimates of Model (4.4.2).

Covariate	Effect	Std. Err.	z-value	p-value	[95% Conf. Interval]
Intercept	-5.545	0.129	-42.900	0.000	[-5.798, -5.291]
<i>spft</i>					
2012	-0.270	0.097	-2.790	0.005	[-0.459, -0.08]
2013	-0.355	0.101	-3.500	0.000	[-0.553, -0.156]
2014	-0.509	0.107	-4.760	0.000	[-0.719, -0.299]
2015	-0.990	0.127	-7.810	0.000	[-1.238, -0.741]
<i>age</i>	0.557	0.024	23.430	0.000	[0.511, 0.604]

$l(\beta) = -3618.443$, Observations = 35022, AIC = 7248.885

The results in Table 4.5 show that age of a female together with all the sub-population categories are significant at 5% level of significance. Although this is the case, Model (4.4.2) is not better

than Model (4.3.5) because of the AIC value that is large by 0.805. This means that Model (4.3.5) is still a better fit logit GLM for teenage pregnancy.

4.4.3 Adding to the interaction effect of age of the female and each sub-population

Since *age* and *spft* are risk factors of teenage pregnancy, Model (4.4.2) can be extended such that it investigates the effect of their interaction. That is,

$$\begin{aligned} \text{logit}(\pi_i) = & \beta_0 + (\beta_{p2}2012_i + \beta_{p3}2013_i + \beta_{p4}2014_i + \beta_{p5}2015_i) + \beta_{age}age_i \\ & + (\beta_{page2}2012 \times age_i + \beta_{page3}2013 \times age_i \\ & + \beta_{page4}2014 \times age_i + \beta_{page5}2015 \times age_i) \end{aligned} \quad (4.4.3)$$

where β_{page1} , β_{page2} , β_{page3} and β_{page4} are, respectively, the effect of $2012 \times age_i$, $2013 \times age_i$, $2014 \times age_i$ and $2015 \times age_i$. Even in this case $2011 \times age_i$ is the reference category of the variable *spft* \times *age*.

The result in Table 4.6 shows the likelihood estimates for fitting Model (4.4.3). The interaction

Table 4.6: Likelihood estimates of Model (4.4.3).

Covariate	Effect	Std. Err.	z-value	p-value	[95% Conf. Interval]
Intercept	-5.518	0.211	-26.180	0.000	[-5.931, -5.105]
<i>spft</i>					
2012	-0.636	0.342	-1.860	0.063	[-1.305, 0.034]
2013	-0.113	0.324	-0.350	0.727	[-0.749, 0.522]
2014	-0.425	0.354	-1.200	0.230	[-1.120, 0.270]
2015	-1.136	0.445	-2.550	0.011	[-2.008, -0.263]
<i>age</i>					
	0.552	0.043	12.970	0.000	[0.468, 0.635]
<i>spft</i> \times <i>age</i>					
2012 \times <i>age</i>	0.076	0.068	1.120	0.263	[-0.057, 0.209]
2013 \times <i>age</i>	-0.052	0.066	-0.780	0.433	[-0.182, 0.078]
2014 \times <i>age</i>	-0.018	0.072	-0.250	0.804	[-0.158, 0.123]
2015 \times <i>age</i>	0.030	0.089	0.340	0.732	[-0.143, 0.204]

$l(\beta) = -3616.754$, Observations = 35022, AIC = 7253.508

effect of age of the female and each sub-population are not risk factors of teenage pregnancy since the *p*-value for each category of *spft* \times *age* is more than 0.05. This suggests, as in the

estimate of Model (4.4.3), that *age* and *spft* do not depend on each other to predict the risk of teenage pregnancy.

4.5 Summary

As a conclusion to this chapter, we fitted a logit GLM that predicts teenage pregnancy. The estimates suggest that pregnancy status can be predicted using both census year and age of the female teenager. However, we have also discovered that the interaction of the predictors is not important for predicting teenage pregnancy. This means that these predictors do not depend on each other to predict teenage pregnancy. All logit GLM that fitted the effect of *idhhms* and *nch* indicated that these covariates are not risk factors of teenage pregnancy. Although a logit GLM that takes account of the population differences was fitted, Model (4.3.7) that estimates the effects of *year* and *age* is a better model for predicting teenage pregnancy.

Using Model (4.3.7), the predicted probabilities of teenage pregnancy that are averaged over age or across census year slightly tie with the observed proportions. Nonetheless, given that the models in this chapter are fitted under the assumption that every teenage pregnancy status is for one unique female; thus, all \mathbf{ps}_i are independent of each other. In view of the facts of data structure for this study, some females have been observed more than once across the five census years. This means that there are different \mathbf{ps}_i for the same female, say \mathbf{ps}_{ti} (where $t = 0, 1, \dots, 4$ representing categories of *year/spft* and $i = 1, 2, \dots, 11011$ representing each unique female). The response term, \mathbf{ps}_{ti} , allows us to identify a teenage female at a specific census year.

Despite the fact that other predicted probabilities are acceptable, it is clear that the clustering effect has been pushed aside; hence the predictions might be inequitable. The issue of clustering will be covered in the next chapter where the same data are analysed using generalised linear mixed model to account for clustering effect.

Chapter 5

Prediction of the Probability of Teenage Pregnancy using a Logit GLMM

This chapter introduces the generalised linear mixed models (GLMM) and then discusses the methodology of GLMM for the context of binary data. The work in this chapter will further fit two level GLMMs to model teenage pregnancy, in which a better model will be selected and interpreted

5.1 Introduction

We have seen in chapter 3 that some females have been observed more than once over different years; making our data longitudinal. Our investigation in chapter 4 has just fitted the model that permits estimations of predictor variable to change across females, along these lines accepting that these estimations are independent. In any case, one method for composing a model that can consider the variability of responses within females is by modifying the Model (4.2.6) as

$$\log \left[\frac{P(y_{ti} = 1|x_{ti})}{1 - P(y_{ti} = 1|x_{ti})} \right] = \sum_{k=0}^p \beta_k x_{tik}, \quad (5.1.1)$$

where $x_{ti0} = 1$ and y_{ti} represent the teenage pregnancy status at **year** “census year” of t ($t=0,1,2,3,4$) within **id** “female id” i ($i = 1, 2, \dots, N$). Model (5.1.1) is not different from the model used in chapter 4, however, it can be used to estimate parameters of a GLM with unknown correlation between responses (Rabe-Hesketh and Skrondal, 2008). Even more, Model (5.1.1) is a fixed effect model and this model is not designed to appropriately fit correlated data such as longitudinal and/or clustered.

Chapter 3 uncovered that information required in this review are with three levels of hierarchy; hence, two clusters are to be taken into account when discussing the methodology of *generalised linear mixed model* (GLMM). For straightforwardness and consistency, both the methodology and analysis of this chapter will utilise indexes t , i and j for level 1, level 2 and level 3, respectively.

5.2 Methodology of generalised linear mixed model for binary data

The GLMM is a model that estimates the degree of dependence among responses of the same cluster (Molenberghs and Verbeke, 2005). GLMMs are sometimes referred to as multilevel generalised linear models. GLMMs are an extension of GLM that adds a random cluster effect to account for the correlation of the data. Formulation of GLMM model for binary data is similar to Model (5.1.1) except that a random effect must be included. In this chapter, the general formulation of this extension is also adopted from the concept and formulation used by Molenberghs and Verbeke (2005). This section will first layout the general formulation of GLMM, thereafter use the Bernoulli distribution and logit link to setup a GLMM for binary response data. This section will further discuss random intercept and slope of a logit model together with parameters and variance components estimation.

5.2.1 Formulation of GLMM

For the reason that the responses vary within a specific cluster, we let $t = 1, \dots, n_i$ represent the level 1 units, where $i = 1, \dots, N$ denote the cluster units in level 2. As in GLM, Y_{ti} are assumed to be independent with cluster specific regression parameter. Following the exponential family defined in Equation (4.2.10), Y_{ti} can be assumed to have the following densities

$$f_i(y_i) \equiv f(y_{ij}|\theta_{ti}, \phi) = e^{\phi^{-1}[y_{ti}\theta_{ti} - \psi(\theta_{ti})] + c(y_{ti}, \phi)}. \quad (5.2.1)$$

In Function (5.2.1), θ_{ti} is a natural parameter that can, through a link function, be represented as a linear predictor η_{ti} while ϕ is a scalar parameter. In exponential probability density, the functions $\psi(\cdot)$ and $c(\cdot, \cdot)$ are all known. Just as in the formulation of GLM, the mean μ_{ti} is of interest and it is estimated using a linear predictor with regression parameters that are fixed (β) and that are cluster specific random effects (\mathbf{v}_i). Thus,

$$h(\mu_{ti}) = \mathbf{x}'_{ti}\beta + \mathbf{z}'_{ti}\mathbf{v}_i = \eta_{ti} \quad (5.2.2)$$

where $h(\cdot)$ is some known link function for two covariate vectors \mathbf{x}_{ti} and \mathbf{z}_{ti} . The term \mathbf{v}_i is a vector of random effects that are following multivariate normal distribution with mean zero and variance-covariance matrix Σ_v . Likewise, η_{ti} is a linear predictor. The expected value of the response variable given the random effect and the covariates is equal to $\mu_{ti} = E(Y_{ti}|\mathbf{v}_i, \mathbf{x}_{ti})$. However, Model (5.2.2) is an all-purpose GLMM where $h(\cdot)$ can be replaced with a known specific link function and Function (5.2.1) can correspond to a specific distribution for the response variable at hand.

5.2.2 GLMM for binary response data

Suppose Y_{ti} is a set of independent responses within a specific cluster and it takes values 0 for no or 1 for yes. The indexes t ($t = 1, 2, \dots, n_i$) and i ($i = 1, 2, \dots, N$) represent level 1 units and level 2 units, respectively. The response measurement Y_{ti} can be assumed to follow a Bernoulli distribution with parameter π_{ti} . A logit link function, $\text{logit}(\pi_{ti}) = \log((\pi_{ti}/(1 - \pi_{ti}))$ can be used to map a binary response data to a linear predictor function $h(\cdot)$ in Model (5.2.2). That is,

$$\text{logit}(\pi_{ti}) \equiv \log[\text{odds}(y_{ti} = 1)] = \mathbf{x}'_{ti}\beta + \mathbf{z}'_{ti}\mathbf{v}_i. \quad (5.2.3)$$

where π_{ti} in this case, is the probability of $Y_{ti} = 1$ given the random effect and covariates values that is also represented as $E(Y_{ti}|\mathbf{v}_i, \mathbf{x}_{ti})$. Model (5.2.3) is known as a two level GLMM because the response measurements are clustered within one cluster. The predicted values of π_{ti} are

given by

$$\pi_{ti} = \frac{e^{x'_{ti}\beta + z'_{ti}v_i}}{1 + e^{x'_{ti}\beta + z'_{ti}v_i}} = \mu_{ti}.$$

5.2.3 Random part of a logit GLMM

Considering a data structure where n_i observations are within N clusters, y_{ti} are responses for each observation $t = 1, \dots, n_i$ in cluster $i = 1, \dots, N$. Consider also a covariate x_{ti} as an observation effect. By letting $z_{ti} = 1$ in Model (5.2.3), a two level random intercept model for a binary response y_{ti} can be written as

$$\log[\text{odds}(y_{ti} = 1)] = \beta_0 + \beta_1 x_{ti} + v_{0i}^{(2)} \quad (5.2.4)$$

where $v_{0i} \sim N(0, \sigma_{v_{0(2)}}^2)$ and $\sigma_{v_{0(2)}}^2$ is the parameter that represents the degree of heterogeneity of the N clusters and it is referred to as the level 2 variance of the random intercept.

The value of β_0 , when x_{ti} and v_{0i} are kept at zero, indicates the log odds of the response being equal to one. The value of β_1 is the magnitude that is added on β_0 when there is a 1 unit increase on the covariate x_{ti} while v_{0i} are kept at zero. This means that for the effect of x_{ti} on the log odds of the response being equal to one while v_i are kept at zero, β_1 is a cluster specific effect because it is the effect measured for the observations within a specific cluster (namely the cluster for which $v_{0i} = 0$). The magnitude of v_{0i} represent the effect of the i^{th} cluster and the intercept for such cluster effect is given by $\beta_0 + v_{0i}$. Testing for cluster effect is the null hypothesis that there is no within cluster variation against that there is a significant within cluster variation, thus

$$\begin{aligned} H_0: \sigma_{v_{0(2)}}^2 &= 0 \text{ (no within cluster variation)} \\ H_1: \sigma_{v_{0(2)}}^2 &> 0 \text{ (significant cluster variation).} \end{aligned}$$

A random slope effect of z_{ti} can be added to the model (5.2.4), thus

$$\log[\text{odds}(y_{ti} = 1)] = \beta_0 + \beta_1 x_{ti} + v_{0i}^{(2)} + v_{1i}^{(2)} z_{ti}, \quad (5.2.5)$$

where both the random intercept $v_{0i}^{(2)}$ and random slope $v_{1i}^{(2)}$ are normally distributed with mean zero and variance $\sigma_{v_{0(2)}}^2$ and $\sigma_{v_{1(2)}}^2$. Since the covariance between the cluster intercept and slope $\sigma_{v_{01(2)}}^2$ is assumed not to be zero, then $v_{0i}^{(2)}$ and $v_{1i}^{(2)}$ follow a bivariate normal distribution with mean zero and variance-covariance matrix Σ_v . Just as the random intercept model, $\beta_0 + v_{0i}^{(2)}$ is the slope of the relationship of log odd that $y_{ti} = 1$ when the covariate x_{ti} and the random intercept v_{0i} are kept at zero.

5.2.4 Maximum likelihood parameter estimation of GLMM

Although there are a number of ways such as Bayesian approach to estimate the parameter in GLMMs, this work will formulate the maximum likelihood estimation method. By maximising the marginal likelihood that is obtained by taking the integration of the random effects, one can fit the GLMM. The models involved so far in this chapter are with cluster specific effect say $\mathbf{v}_l^{(L)}$, which are inferences that follow a random effect approach, where L represents the cluster level and l is the cluster index. The $\mathbf{v}_l^{(L)}$ is said to be a random vector where its attributes are independently drawn from a known distribution.

All models in this work assume Bernoulli distribution, which is a special case of binomial distribution. For simplicity, we consider a two level binomial model where y_{ti} are number of successes from a Bernoulli trial, thus only one replication. For N clusters with n_i response measurements, the density function of Y_{tj} given $\mathbf{v}_{[i]}^{(2)}$ in the form of (5.2.1) is given by

$$f(Y_{tj}|\mathbf{v}_{[i]}^{(2)}) = e^{\sum_{t=1}^{n_i} y_{ti}(\mathbf{x}'_{ti}\boldsymbol{\beta} + \mathbf{z}'_{ti}\mathbf{v}_i) - \log(1 + e^{\mathbf{x}'_{ti}\boldsymbol{\beta} + \mathbf{z}'_{ti}\mathbf{v}_i})}, \quad (5.2.6)$$

where

$$y_{ti}\theta_{ti} = y_{ti}(\mathbf{x}'_{ti}\boldsymbol{\beta} + \mathbf{z}'_{ti}\mathbf{v}_i), \quad \psi(\theta_{ti}) = \log(1 + e^{\mathbf{x}'_{ti}\boldsymbol{\beta} + \mathbf{z}'_{ti}\mathbf{v}_i}) \text{ and } c(y_{ti}, \phi) = 1.$$

The index t takes on values $t = 1, \dots, n_i$ which are measurements for units at level 1 for each cluster $i = 1, 2, \dots, N$ units at level 2.

This means for one level 2 random effect $v_i^{(2)} \sim N(0, \sigma_{v(2)}^2)$ and one fixed effect $\boldsymbol{\beta}$ of x_{ti} , the likelihood function that can be derived from the density $\log(\pi_{ti}/(1 - \pi_{ti})) = x_{ti}\boldsymbol{\beta} + v_i^{(2)}$ is

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma_{v(2)}^2 | Y_{tj}) &= \int_{-\infty}^{\infty} \prod_{t,i}^{n_i, N} P(Y_{tj} | \boldsymbol{\beta}, \sigma_{v(2)}^2) d\mathbf{v}_{[i]}^{(2)} \\ &= \int_{-\infty}^{\infty} \prod_{t,i}^{n_i, N} P(Y_{tj} | \boldsymbol{\beta}, \mathbf{v}_{[i]}^{(2)}) f(\mathbf{v}_{[i]}^{(2)} | \sigma_{v(2)}^2) d\mathbf{v}_{[i]}^{(2)} \\ &= \prod_{i=1}^N \int_{-\infty}^{\infty} \prod_{t=1}^{n_i} \frac{e^{y_{ti}(\mathbf{x}'_{ti}\boldsymbol{\beta} + \mathbf{z}'_{ti}\mathbf{v}_i)}}{1 + e^{y_{ti}(\mathbf{x}'_{ti}\boldsymbol{\beta} + \mathbf{z}'_{ti}\mathbf{v}_i)}} \times \frac{e^{-v_i^{(2)}/2\sigma_{v(2)}^2}}{(2\sigma_{v(2)}^2)^{1/2}} dv_i^{(2)}. \end{aligned} \quad (5.2.7)$$

The resulting likelihood function can be computed or evaluated by using the adaptive quadrature or Laplace approximation approaches.

5.2.5 Variance components estimation for binary response data GLMM

Although there are several ways to examine the variance components of mixed models, our study focused on the *variance partition coefficients* (VPC) and *intraclass correlation coefficients*

(ICC). The VPC reports the proportion of the response variance that lies at each level of the model hierarchy while ICC reports the expected degree of similarity between responses within a given cluster. Consider a two level model without any covariates, thus

$$\log[\text{odds}(y_{ti} = 1)] = \beta_0 + v_{0i}^{(2)}, \quad (5.2.8)$$

where β_0 is a fixed intercept that is unknown. The terms $v_{0i}^{(2)}$ are the level 2 random intercepts which are assumed to be normally distributed with mean zero and variance $\sigma_{v0(2)}^2$. The level 1 errors for a logistic model are assumed to be normally distributed with mean zero and variance $e = \pi^2/3 = 3.29$; hence, the total variance is calculated as

$$\text{var}(\log[\text{odds}(y_{ti} = 1)]) = \text{var}(\beta_0 + v_{0i}^{(2)}) = \pi^2/3 + \sigma_{v0(2)}^2 \quad (5.2.9)$$

Variance partition coefficients

The VPC for a two level model reports two types of coefficients; i.e. the level two and level one VPC, calculated as

$$VPC_v^{(2)} = \frac{\sigma_{v0(2)}^2}{\pi^2/3 + \sigma_{v0(2)}^2} \quad \text{and} \quad (5.2.10a)$$

$$VPC_e^{(1)} = \frac{\pi^2/3}{\pi^2/3 + \sigma_{v0(2)}^2}, \quad (5.2.10b)$$

respectively. When the coefficients $VPC_v^{(2)}$ and $VPC_e^{(1)}$ are multiplied by 100, they are interpreted, respectively, as the percentage of the variation that lies between level 2 clusters; and within level two clusters between level 1 units.

Intraclass correlation coefficients

A two level logit GLMM estimates two ICC values, but one is important, thus $ICC_v^{(2)}$. This value measures the correlation between level one measurements that belong to the same cluster at level two. It is then calculated as

$$ICC_v^{(2)} = \rho^{(2)} = \frac{\sigma_{v0(2)}^2}{\pi^2/3 + \sigma_{v0(2)}^2}. \quad (5.2.11)$$

For a two level GLMM this value is similar to the value of $VPC_v^{(2)}$. However, for a GLMM with more than 2 levels $VPC_v^{(2)}$ is not always equal to $ICC_v^{(2)} = \rho^{(2)}$. For example, a three level logit GLMM reports three ICC values but only two ($ICC_v^{(2)}$ and $ICC_v^{(3)}$) are important. The magnitudes of $ICC_v^{(2)}$ and $ICC_v^{(3)}$ for a three level logit GLMM are computed as

$$ICC_v^{(2)} = \rho^{(2)} = \frac{\sigma_{v0(2)}^2 + \sigma_{v0(3)}^2}{\pi^2/3 + \sigma_{v0(2)}^2 + \sigma_{v0(3)}^2} \quad \text{and} \quad (5.2.12a)$$

$$ICC_v^{(3)} = \rho^{(3)} = \frac{\sigma_{v0(3)}^2}{\pi^2/3 + \sigma_{v0(2)}^2 + \sigma_{v0(3)}^2}, \quad (5.2.12b)$$

respectively. Surely given that the numerator of the calculation of $VPC_v^{(2)}$ will include only the $\sigma_{v0(2)}^2$, the results will be different from $ICC_v^{(2)}$. Generally, an m levels ICC can be written as

$$ICC_v^{(L)} = \rho^{(L)} = \frac{\sum_{l=1}^L \sigma_{v0(l)}^2}{\pi^2/3 + \sum_{l=2}^m \sigma_{v0(l)}^2}, \quad (5.2.13)$$

where $L = 1, 2, \dots, m$. L is the level of hierarchy and l is the cluster index. The magnitude of $ICC_v^{(2)}$ closer to 1 indicates that the level 1 measurements are highly correlated for ones that belonging to the same cluster attributes. Due to the fact that the variance components are non-negative, the point estimate say $\hat{\rho}^{(L)}$, will always be between 0 and 1. This means, in order to calculate the confidence interval, the calculation will maintain to use the logit transformation. The $(1 - \alpha)100\%$ confidence interval of the logit of $\rho^{(L)}$ with a standard error of $\hat{SE}\rho^{(L)}$, is given by

$$\left(\text{logit}(\hat{\rho}^{(L)}) - z_{\alpha/2} \frac{SE\hat{\rho}^{(L)}}{\hat{\rho}^{(L)}(1 - \hat{\rho}^{(L)})}, \text{logit}(\hat{\rho}^{(L)}) + z_{\alpha/2} \frac{SE\hat{\rho}^{(L)}}{\hat{\rho}^{(L)}(1 - \hat{\rho}^{(L)})} \right)$$

where $z_{\alpha/2}$ is one minus half of alpha of the standard normal distribution.

5.3 Specifying and fitting a Logit GLMM for teenage pregnancy

5.3.1 Null model

Just like in chapter 4, this section starts by fitting a null logit model, but with female *id* as cluster effect. Thus

$$\text{logit}(\pi_{ti}) = \beta_0 + v_{0i}^{(2)}, \quad (5.3.1)$$

where β_0 is the fixed intercept for log odd of measurements of teenage pregnancies and $v_{0i}^{(2)}$ is the random intercept for female cluster. The results of Model (5.3.1) maximum likelihood estimates are shown in Table 5.1.

Unlike the results of a logit GLM, the GLMM estimates show the fixed effects and the random effects. Table 5.1 shows that the magnitude of the fixed intercept (β_0) is -3.957 with p -value less than 0.001. This means the log odds of teenage pregnancy for an average female ($v_{0i}^{(2)} = 0$) is -3.957. The random effects intercept shows that the variance $\sigma_{v0(2)}^2 = 0.598$ is significant with a p -value less than 0.01. This means that there is strong evidence that teenage pregnancy varies within female cluster. Based on the aforementioned estimates of Model (5.3.1), the intercept

Table 5.1: Maximum likelihood estimates for null mixed logit model with female random effect.

Covariate	Effect	Std. Err.	z -value [$\chi^2_{(1)}$]	p -value	[95% Conf. Interval]
<i>Fixed effects</i>					
Intercept	-3.956	0.099	-40.16	0.000	[-4.150, -3.764]
<i>Random effects</i>					
$\sigma^2_{v_{0(2)}}$	0.598	0.207	[8.69]	0.0016	[0.304, 1.176]

$l(\boldsymbol{\beta}) = -4025.491$, Observations = 35022, AIC = 8054.981

of female i is $-3.957 + v_{0i}^{(2)}$. Thus, the probability of teenage pregnancy for female i can be calculated as

$$\pi_{ti} = \frac{e^{-3.957 + v_{0i}^{(2)}}}{1 + e^{-3.957 + v_{0i}^{(2)}}}$$

The value of β_0 for Model (5.3.1) is less than the one for Model (4.3.1) by 0.273, meaning that an average of 0.273 log odds is accounted for by the random intercept, $v_{0i}^{(2)}$. The random intercepts estimate for each female can be estimated using Model (5.3.1) or by simulating from a normal distribution with mean $\mu_{v_{0(2)}} = 0$ and variance $\sigma^2_{v_{0(2)}} = 0.598$.

We can further examine the female random effect using Model (5.3.1) estimated female level intercepts and their respective standard errors. It is expected that some of the females have the same estimates of the random intercepts. Table 5.2 shows the number of females with the same random intercepts estimates from the lowest to highest.

One can see from Table 5.2 that almost 7.1% of the female teenagers have a positive log odds of teenage pregnancy. Given that most of the females are with negative, these suggest that teenage pregnancy are below average. We could also make a plot of estimates and include the 95% of confidence interval in order to observe which random intercept rank include zero; hence, Figure 5.1. It is clear that the 95% confidence interval for all the estimated random intercepts except one do not overlap the zero red line (see Figure 5.1). In this figure we also notice that the width of the confidence intervals are approximately equal, indicating that the standard errors are also approximately the same for all ranks.

5.3.2 Adding census year fixed effect

As in chapter 4, we have seen that census year (*year*) is an important covariate which is used as measure of occasion and to distinguish between different sub-populations. This model building

Table 5.2: Female random intercept estimates for group of females.

$v_0^{(2)}$	Rank	$v_0^{(2)}$	Std. Err.	N (%)
-0.053	1	-0.053	0.754	2541 (22.011)
-0.043	2	-0.043	0.757	1795 (15.549)
-0.033	3	-0.033	0.761	1938 (16.788)
-0.022	4	-0.022	0.765	2128 (18.434)
-0.011	5	-0.011	0.769	2318 (20.08)
0.506	6	0.506	0.741	227 (1.966)
0.523	7	0.523	0.747	163 (1.412)
0.541	8	0.541	0.753	128 (1.109)
0.559	9	0.559	0.759	181 (1.568)
0.578	10	0.578	0.766	92 (0.797)
1.042	11	1.042	0.722	7 (0.061)
1.069	12	1.069	0.731	13 (0.113)
1.098	13	1.098	0.740	9 (0.078)
1.129	14	1.129	0.750	3 (0.026)
1.588	15	1.588	0.710	1 (0.009)

N denotes the number of females

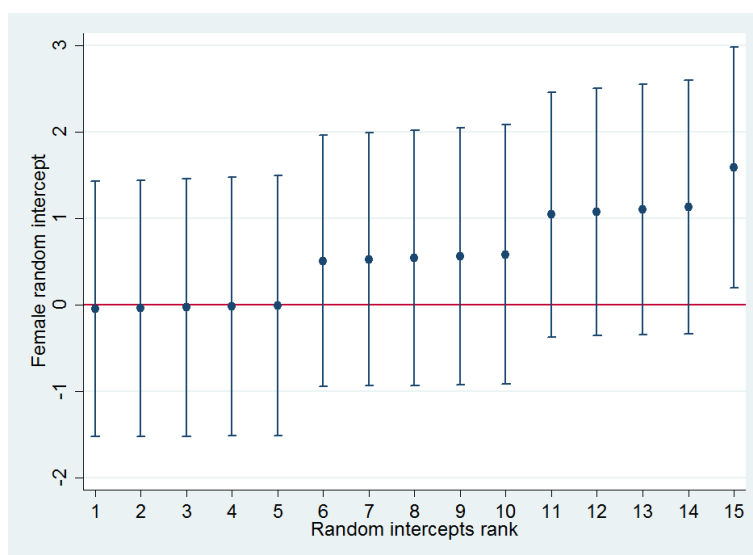


Figure 5.1: Caterpillar plot of random intercept by rank with 95% confidence interval for a null GLMM.

will further extend Model (5.3.1) by including census year (*year*) as covariates. Fitting the next model we use subscript ti on *year* to model census year effect for some female ($i = 1, 2, \dots, 11544$) in a specific census year ($t = 1, 2, 3, 4, 5$). This means that such a model takes into account the

different five sub-populations. The logit GLMM with census year effect is represented as

$$\text{logit}(\pi_{ti}) = \beta_0 + \beta_{\text{year}} \text{year}_{ti} + v_{0i}^{(2)}, \quad (5.3.2)$$

where β_{year} is the magnitude of the fixed effect of census year on the log odds of teenage pregnancy. The likelihood estimates for this model are shown in Table 5.3.

Table 5.3: Maximum likelihood estimates for Model (5.3.2).

Covariate	Effect	Std. Err.	z -value [$\chi^2_{(1)}$]	p -value	[95% Conf. Interval]
<i>Fixed effects</i>					
Intercept	-3.501	0.106	-33.08	0.000	[-3.709, -3.294]
year	-0.203	0.026	-7.86	0.000	[-0.253, -0.152]
<i>Random effects</i>					
$\sigma^2_{v0(2)}$	0.350	0.191	[3.36]	0.0334	[0.120, 1.021]

$$l(\beta) = -3993.976, \text{ Observations} = 35022, \text{ AIC} = 7993.952$$

The estimates of Model (5.3.2) indicate that the fixed effect of **year** equal -0.203 is significant with a p -value less than 0.001 (see Table 5.4). Although this is the case, we see that the variation of teenage pregnancy estimated by Model (5.3.2) has reduced by approximately 41.5% of the variance $\sigma^2_{v0(2)}$ in Model (5.3.1). This means that the distribution of census years (**year**) strongly differ across females. Even in these estimates for Model (5.3.2), there is evidence that teenage pregnancy varies within females because $\sigma^2_{v0(2)}$ is significant with a p -value of less 0.05. We could again use the caterpillar plot to examine the random intercepts estimates and change in the number of ranks (see Figure 5.2). We see in this plot that the number of ranks of random intercepts has increased from the estimates of using the null GLMM; justifying the difference in the distribution of census years across females. We also notice that all the 95% confidence interval overlap the zero line.

We could also compare the estimates of Models (4.3.2) and (5.3.2). The magnitude of the β_0 and effect of **year** on the log odds of teenage pregnancy for an average female, in Model (4.3.2), has respectively reduced by 0.169 and increased by 0.004. This means about 0.165 log odds of teenage pregnancy that was observed in Model (4.3.2) estimates is accounted for by the random intercept for an average female in 2012.

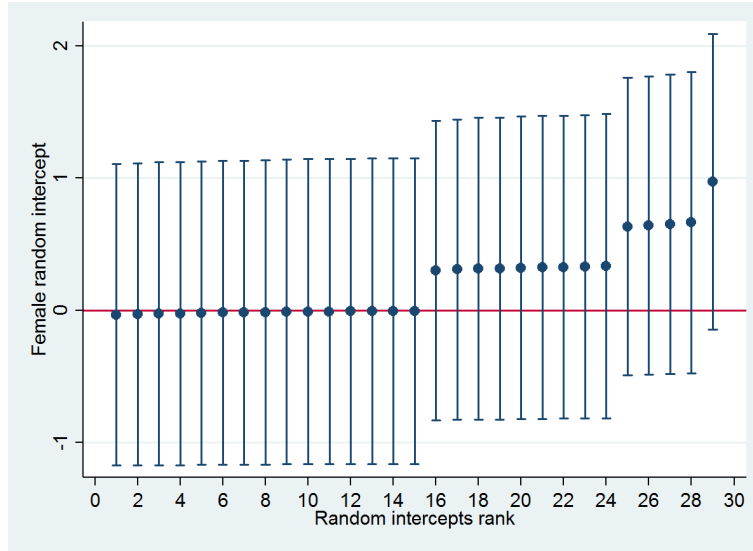


Figure 5.2: Caterpillar plot of random intercept by rank with 95% confidence interval for Model 5.3.2.

5.3.3 Adding other main covariates effect

Going forward, Model (5.3.2) will be extended by adding simultaneously the main covariates *age*, *idhhms* and *nch*, thus

$$\text{logit}(\pi_{ti}) = \beta_0 + \beta_{\text{year}}\text{year}_{ti} + \beta_{\text{age}}\text{age}_{ti} + \beta_{\text{ihm}}\text{idhhms}_{ti} + \beta_{\text{nch}}\text{nch}_{ti} + v_{0i}^{(2)}, \quad (5.3.3)$$

where β_{age} , β_{ihm} and β_{nch} denote the fixed effects of *age*, *idhhms* and *nch*, respectively for a specific female in a particular census year. The covariates effects will then be removed one-by-one from Model (5.3.3) based on significant status at 5% level of significance.

The likelihood estimates of Model (5.3.3) that are presented in Table 5.4 indicate that both fixed effects of *idhhms* and *nch* are insignificant at 5% level of significance. However, *idhhms* has a z -value closer to zero; therefore at 5% level of significance, there is a strong evidence against its effect on the log odds of teenage pregnancy.

Based on the aforementioned findings, the next model to be fitted will remove the effect of *idhhms*, thus

$$\text{logit}(\pi_{ti}) = \beta_0 + \beta_{\text{year}}\text{year}_{ti} + \beta_{\text{age}}\text{age}_{ti} + \beta_{\text{nch}}\text{nch}_{ti} + v_{0i}^{(2)}. \quad (5.3.4)$$

The magnitude of the fixed effect of *nch* still suggests that there is no effect of *nch* on the log odds of teenage pregnancy at 5% level of significance (see Table 5.4). This means that another model can be fitted without the effect of *nch*, thus

$$\text{logit}(\pi_{ti}) = \beta_0 + \beta_{\text{year}}\text{year}_{ti} + \beta_{\text{age}}\text{age}_{ti} + v_{0i}^{(2)}. \quad (5.3.5)$$

For Model (5.3.5) estimates in Table 5.4, we see that all the fixed effect including the fixed intercept are significant with a p -value less than 0.001.

Table 5.4: Estimates of model building for a logit GLMM of pregnancy status.

Covariate	Model		
	(5.3.3)	(5.3.4)	(5.3.5)
<i>Fixed effects</i>			
Intercept	-5.528*** (-44.19)	-5.599*** (-44.47)	-5.654*** (-35.87)
<i>year</i>	-0.208*** (-8.07)	-0.207*** (-8.07)	-0.211*** (-8.15)
<i>age</i>	0.549*** (22.73)	0.549*** (22.73)	0.559*** (23.16)
<i>idhhms</i>	-0.0189 (-0.18)		
<i>nch</i>	0.191 (1.78)	0.191(1.78)	
<i>Random-effects</i>			
$\sigma_{v0(2)}^2$	5.23×10^{-11} [0.00]	6.38×10^{-10} [0.00]	0.111 [0.37]

z statistics in round parentheses () $\chi^2_{(1)}$ statistics in square parentheses []

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

With regard to Models (5.3.3) and (5.3.4), the variance of the $\sigma_{v0(2)}^2$ is very small (closer to zero) such that the fixed effect estimates are approximately equal to Model (3) and (4) estimates in Table 4.1. This suggests that female clustering has no significant impact on the teenage pregnancy. Model (5.3.5) has estimated a variance of teenage pregnancy within females as $\sigma_{v0(2)}^2 = 0.111$. Nonetheless, all the variance estimates in Table 5.4 are not significant since their χ^2 values at 1 degree of freedom have p -values greater than 0.05. This means by including the fixed effect of *age*, *idhhms* and/or *nch*, teenage pregnancy does not vary within females at 5% level of significance.

5.3.4 Variance components

Unlike GLM, GLMM provides an additional statistic called ICC. For the reason that $\sigma_{v0(2)}^2$ of Models (5.3.3) and (5.3.4) are very much closer to zero, the ICC values for these models will not be estimated. Table 5.5, shows the ICC estimates for Models (5.3.1), 5.3.2 and (5.3.5) in Table 5.7.

The two level null model, Model (5.3.1), allows us to evaluate correlation between teenage pregnancy status of the same female together with the extent of between female variation in teenage pregnancy status. Table 5.5 reports a female cluster ICC value for Models (5.3.1),

Table 5.5: ICC estimates for each model.

Model	ICC	Std. Err.	[95% Conf. Interval]
(5.3.1)	0.154	0.045	[0.084, 0.263]
(5.3.2)	0.096	0.047	[0.035, 0.237]
(5.3.5)	0.032	0.051	[0.001, 0.453]

(5.3.2) and (5.3.5) of 0.154, 0.096, and 0.32 respectively. These values suggest that there is 15.4%, 9.6% and 3.2% of the variation of teenage pregnancy which is accounted for by females given, respectively,

- that there is no fixed effect,
- the effect of census year, and
- the effect of census year and age of a female.

5.3.5 Adding the random effect of census year

However before these predictions, we can fit another model that allows census year effect to vary across females. Thus

$$\text{logit}(\pi_{ti}) = \beta_0 + \beta_{\text{year}} \text{year}_{ti} + v_{0i}^{(2)} + v_{1i}^{(2)} \text{year}_{tj}, \quad (5.3.6)$$

where $v_{1i}^{(2)}$ is said to be the random effect of **year** and it is normally distributed with mean zero and variance $\sigma_{v1(2)}^2$. The estimates of these model are shown in Table 5.6.

Table 5.6: Maximum likelihood estimates for null mixed logit model with female random effect.

Covariate	Effect	Std. Err.	z -value $[\chi_{(1)}^2]$	p -value	[95% Conf. Interval]
<i>Fixed effects</i>					
Intercept	-3.501	0.106	-33.08	0.000	[-3.709, -3.294]
year	-0.203	0.026	-7.86	0.000	[-0.253, -0.152]
<i>Random effects</i>					
$\sigma_{v1(2)}^2$	9.27×10^{-17}	2.24×10^{-9}	[3.36]	[0.186]	[0.000, 0.000]
$\sigma_{v0(2)}^2$	0.350	0.191			[0.120, 1.021]

$l(\beta) = -3993.976$, Observations = 35022, AIC = 7993.952

In these results it is clear that census year does not vary across females. This is because the likelihood ratio test for the random slope effect ($v_{1i}^{(2)}$) shows a χ^2 with 2 degrees of freedom of 3.36 having a p -value = 0.186 which is greater than 0.05.

5.3.6 Model Selection

The AIC results for Models (5.3.1) to (5.3.5) are shown in Table 5.7.

Table 5.7: Akaike's information criterion for each logit GLMM of pregnancy status.

Covariate	Model				
	(5.3.1)	(5.3.2)	(5.3.3)	(5.3.4)	(5.3.5)
Observations	35022	35022	35022	35022	35022
$l(\beta)$	-4025.49	-3993.98	-3619.52	-3619.54	-3620.85
Parameters	2	3	6	5	4
AIC	8054.98	7993.95	7251.05	7247.08	7249.71

Based on these results, Model (5.3.4) followed by Model (5.3.5) and then Model (5.3.3) seem to have a better fit based on their AIC values. Furthermore, the Models (5.3.3) and (5.3.4) have the same AIC values with Models (4.3.3) and (4.3.4) estimates in Table 4.1, respectively. In both models, the fixed effects *nch*, *idhhms* and/or the random intercept $v_{0i}^{(2)}$ are not significant. These models are not of better fit when compared to Model (5.3.2).

Even more, Model (5.3.6) had the same AIC value of 7993.95 but $v_{1i}^{(2)}$ was not significant. This means that Model (5.3.2) is a better logit GLMM fit for teenage pregnancy data for this work; hence it will be used for probability predictions.

5.3.7 Interpretation of a better fit logit GLMM for teenage pregnancy

Based on the results of the logit GLMM for this work, Model (5.3.2) will be used to predict probability teenage pregnancy. The equation resulting from the likelihood estimates Model (5.3.2) is written as

$$\eta_{ti} = -3.501 - 0.203year_{ti} + v_{0i}^{(2)},$$

where $v_{0i}^{(2)}$ is normally distributed with $\mu_{v0(2)} = 0$ and $\sigma_{v0(2)}^2 = 0.35$. The estimated fixed intercept of -3.501, which is the log odds that female is pregnant during census year 2011, means that the probability of a teenage female being pregnant is equal to $e^{-3.501}/(1 + e^{-3.501}) = 0.029$.

The magnitude of the effect of census year equal -0.203 means that for each 1 year increase

to a census year, the log odds of teenage pregnancy is expected to decrease by 0.203, when controlling for female difference. The estimate of the random intercept line for each i th female can be estimated as $-3.501 + \hat{v}_{0i}^{(2)}$. This means that the log odds of teenage pregnancy are expected to differ for female i by a value $\hat{v}_{0i}^{(2)}$.

5.3.8 Probability predictions of teenage pregnancy

To predict the probabilities, we take the inverse logit of η_{ti} , thus

$$\pi_{ti} = \frac{e^{\eta_{ti}}}{1 + e^{\eta_{ti}}},$$

where π_{ti} is the probability of teenage pregnancy for some female i in census year t . The probability predictions can be calculated for cluster specific and probability averaged.

- *Cluster specific probability predictions*

For these predictions we use the estimated random intercepts of the female cluster to substitute the term $v_{0i}^{(2)}$ and therefore compute probabilities. From Table 5.2, Model (5.3.2) has predicted 29 different random intercepts. Meaning we expect to have at least $5 \times 29 = 145$ predicted probability values.

- *Population averaged probability predictions*

There are two ways to compute population averaged probabilities. The first one is to average the cluster specific probabilities for each census year, thus,

$$\pi_t = \frac{1}{t \text{ population size}} \sum_{i=1}^{t \text{ population size}} \pi_{ti},$$

where $t = 0, 1, 2, 3, 4$.

The second one is to simulate the M values of $v_{0i}^{(2)}$ from a normal distribution $N(0, 0.35)$ for each of the $G = 29$ groups of females per census year. For $M = 1000$, it means for each group of females per census year, we can compute

$$\pi_{tg}^{(m)} = \frac{e^{-3.501 - 0.203 \text{year}_{tg} + v_{0g}^{(2)m}}}{1 + e^{-3.501 - 0.203 \text{year}_{tg} + v_{0g}^{(2)m}}},$$

where $m = 1, 2, \dots, 1000$, $t = 0, 2, 3, 4$ and $g = 1, 2, \dots, 29$. Thereafter, average within each census year the probabilities, $\pi_{tg}^{(m)}$, thus

$$\pi_t = \frac{1}{M \times G} \sum_{g=1}^{29} \sum_{m=1}^{1000} \pi_{tg}^{(m)}$$

where $t = 0, 1, 2, 3, 4$. These simulated values will be substituted in the equation to predict π_{ti} and then averaged over census year.

Figures 5.3 and 5.4 are plotted using all the aforementioned calculations. Figure 5.3 compares the cluster specific and population averaged predictions while Figure 5.4 compares the population averaged predictions using the estimated and simulated random intercepts.

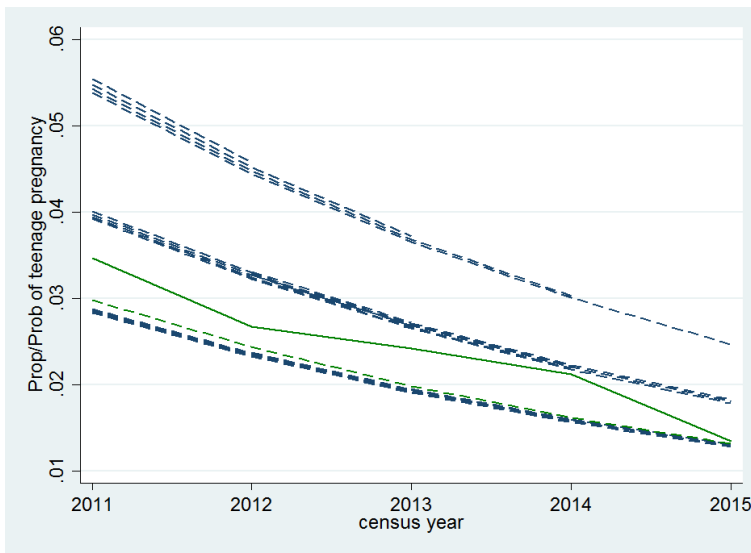


Figure 5.3: Cluster specific and population averaged probability of teenage pregnancy by census year using estimated random intercepts.

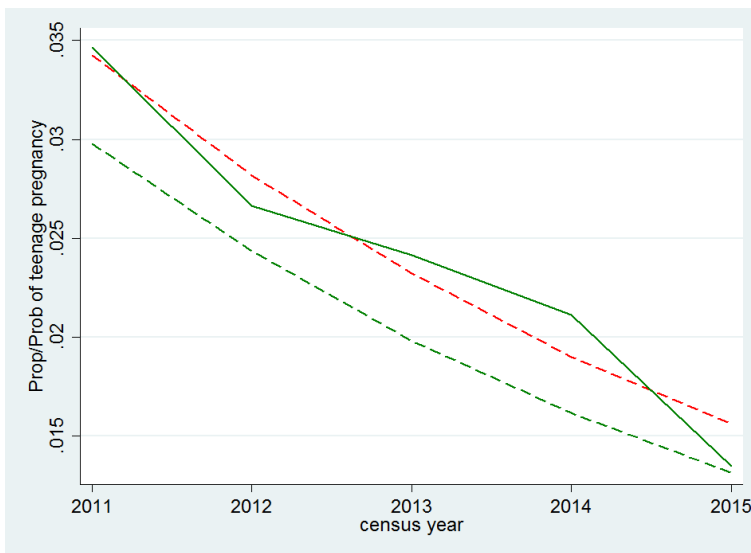


Figure 5.4: Comparison of the population averaged probability of teenage pregnancy by census year using the estimated and simulated random intercepts.

In Figure 5.3, each blue dash line represents the cluster specific probabilities for each group of

females by census year. The green dash line (in Figures 5.3 and 5.4) is the population averaged predicted probabilities over years that used the estimated random intercepts while the red dash line (in Figure 5.4) used the simulated random intercepts. The green solid line is the observed proportion of teenage pregnancy over census year.

As expected, all the predicted probabilities are reducing over year. The cluster specific predictions form some three groups with probability of teenage pregnancy in 2011 below 0.03 (first group), approximately 0.04 (second group) and above 0.05 but below 0.06 (third group). Nonetheless, it is also clear that within this group the probabilities in 2011 still differ. This means that indeed teenage pregnancy varies within females when census year is taken into account. We also see that a single line of probability predictions for cluster specific is way above the rest of the population averaged predictions. However, in all the predicted blue dash lines, we also observe, approximately, a constant change of probabilities of teenage pregnancy across female groups.

Judging from the green population averaged dash line, the number of females with probabilities in 2011 below 0.03 is extremely higher than that of females with probabilities approximately 0.04 and above 0.05 combined. This is because the green population averaged, the dash line, slightly ties with the first group.

Moreover, the red line seems to tie with the observed proportions but just about 0.04 above the green dash line. Over and above, the probabilities of teenage pregnancy are on average below 0.03 when using teenage pregnancy of our dataset, but below 0.035 when teenage pregnancy status are sampled from a normally distributed population.

5.4 A Logit GLMM building for teenage pregnancy by sub-population

As in chapter 4, this chapter also fits a logit GLMM that takes into account of the sub-population differences. This means in Model (5.3.2), *year* effect will be replaced with *spft*, thus

$$\text{logit}(\pi_{ti}) = \beta_0 + (\beta_{p2}2012_i + \beta_{p3}2013_i + \beta_{p4}2014_i + \beta_{p5}2015_i) + v_{0i}^{(2)}, \quad (5.4.1)$$

where the effect of a female in 2012, 2013, 2014 and 2014 are β_{p2} , β_{p3} , β_{p4} and β_{p5} , respectively. The fixed and the random intercepts are, respectively, represented by β_0 and $v_{0i}^{(2)}$. Model (5.4.1) was further extended by including the effect of *age*.

$$\text{logit}(\pi_{ti}) = \beta_0 + (\beta_{p2}2012_i + \beta_{p3}2013_i + \beta_{p4}2014_i + \beta_{p5}2015_i) + \beta_{age}age_{ti} + v_{0i}^{(2)}, \quad (5.4.2)$$

Table 5.8 shows Models (5.4.1 and 5.4.2) likelihood estimates.

Table 5.8: Estimates of model building for a logit GLMM of pregnancy status by sub-population.

Covariate	Model	
	(5.4.1)	(5.4.2)
<i>Fixed effects</i>		
Intercept	-3.499*** (-31.44)	-5.601*** (-35.11)
<i>spft</i>		
2012	-0.269** (-2.80)	-0.273** (-2.81)
2013	-0.363*** (-3.60)	-0.357*** (-3.51)
2014	-0.495*** (-4.64)	-0.514*** (-4.77)
2015	-0.953*** (-7.51)	-0.995*** (-7.80)
<i>age</i>		0.559*** (23.17)
<i>Random effects</i>		
$\sigma_{v0(2)}^2$	0.356* [3.45]	0.109 [0.36]
<i>Akaike's information criterion estimates</i>		
Observations	35022	35022
$l(\beta)$	-3991.704	-3618.261
df	6	7
AIC	7995.409	7250.523
<i>t</i> statistics in round parentheses ()		
$\chi_{(1)}^2$ statistics in square parentheses []		
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$		

In both estimates of Models (5.4.1 and 5.4.2) the fixed effects and the intercept are all highly significant because their p -values are less than 0.01 (see Table 5.8). However, the random intercept for Model (5.4.1) estimates is significant while the one for Model (5.4.2) is not. Even though Model (5.4.1) seems to have significant effects and intercepts, its AIC value of 7995.409 suggests that Model (5.3.2) with AIC of 7993.952 predicts teenage pregnancy better.

5.5 Summary

This chapter has addressed the last three objectives of the work and specified a multilevel model or generalised linear mixed model which was used to model teenage pregnancy data. The model revealed that only census year (*year*) is a significant effect on teenage pregnancy at

5% level of significance. This means that age of the female teenager, the number of household a female teenager belongs to and the number of children the female teenager had before year of observation are not significant predictors of teenage pregnancy when female cluster is taken into account.

Further more, this chapter has also suggested that the effect of census year does not vary across females. This means for every female the risk of teenage pregnancy changes the same for all the females as census year changes.

The variance components for the better fit model suggested that 9.6% variation of teenage pregnancy which is accounted for by female cluster given 0.203 negative effect census year. Which also means that there is a 0.096 correlation of teenage pregnancy status for same female teenagers. This correlation is considered as moderate or reasonable given that 29 groups of random intercepts estimate were formed out of 11544 females.

Chapter 6

Model Extensions and More Complex Data Structure

In this chapter, section 6.1 extends a two level teenage pregnancy dataset that was used in chapters 4 and 5 to a three level dataset. The chapter will further, in section 6.2, specify and build logit generalised linear mixed model with household cluster effect. Summary and conclusion will also be provided towards the end of this chapter

6.1 Population and Data structure

In chapters 4 and 5 our models used a two level dataset with measurements of teenage pregnancy at level 1 and female teenagers at level 2. However, as we have seen in chapter 3, the variable *idhhms* shows that each female belongs to one or more household. In this chapter, we expand our data to a three level data structure where households are at level three. Table 6.1 is the extension of Table 3.3 by including household cluster.

Table 6.1: The number of unique identifiers and units within the levels.

Level	Level name	Range	Unique identifiers	Units
1	Measurements of pregnancy status	[2011, 2015]	5	38398
2	Female teenagers	[1, 11544]	11544	11544
3	Households	[1, 7863]	7863	7863

From Table 6.1 we observe that the number of level one units have increased from 35022 to 38398. This is because some female teenagers belong to more than one household. Figure 6.1 show a classification diagram for three level data structure. We also see that there are 7863 households while female population is still 11544.

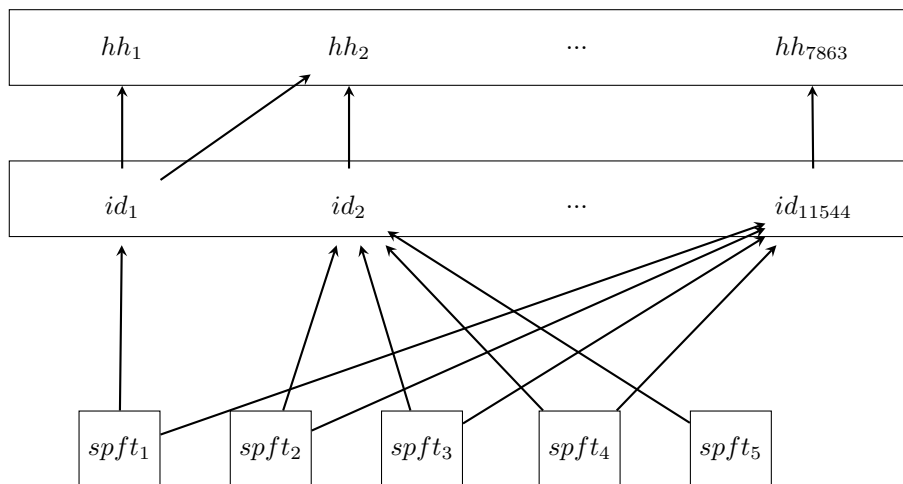


Figure 6.1: Classification diagram for the three level dataset.

It is clear from Figure 6.1 that females are nested within households, for example, id_2 is nested within hh_2 . It is also clear that some females are nested within multiple households, for example, id_1 is nested within hh_1 and hh_2 . This means there exists a multiple membership classification of females at level three; hence a multiple membership data structure. Moreover, in some cases different females are clustered by the same household, for example, id_1 and id_2 are nested within hh_2 . This means, it is of interest to check whether there is any variation in teenage pregnancy

for females coming from the same household.

Table 6.2 indicate that several female belongs to multiple households.

Table 6.2: The number of observations for each household membership category by year.

<i>idhhms</i>	2011	2012	2013	2014	2015	across <i>year</i>
	N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
1	6815 (77.87)	6633 (80.82)	6366 (83.93)	6149 (86.37)	5913 (87.8)	31876 (83.01)
2	1640 (18.74)	1408 (17.16)	1120 (14.77)	904 (12.70)	780 (11.58)	5852 (15.24)
3	273 (3.12)	150 (1.83)	99 (1.31)	66 (0.93)	42 (0.62)	630 (1.64)
4	24 (0.27)	16 (0.19)	0 (0.00)	0 (0.00)	0 (0.00)	40 (0.10)
Total	8752 (100.0)	8207 (100.0)	7585 (100.0)	7119 (100.0)	6735 (100.0)	38398 (100.0)

N is the number of observations for each *idhhms* category

Contrary to the results from Table 3.6, our data here shows that there are 8752, 8207, 7585, 7119 and 6735 data lines for sub-populations of 2011, 2012, 2013, 2014 and 2015, receptively (see Table 6.2). By including the household level, the number of observations have increased by 3376 from 35022. The female teenagers with one household membership in Table 3.6 are the same as in Table 6.2 for each sub-population and across census year. However, for household membership equalling to 2, 3 and 4, the number of females in Table 3.6 are multiplied by 2, 3 and 4 for each sub-population and across census year, respectively. This means that for 2011 sub-population, data across main covariates was repeated for 917 female teenagers, which is calculated as $(1640/2) + (273/3) + (24/4)$ from Table 6.2. By performing this calculation for the total column, we can see that information for 3146 females were duplicated, of which 5852, 630 and 40 observations were respectively duplicated once, twice and thrice.

Given that there was some changes in the number of observations, it would be interesting to still check the relationship between the proportions of teenage pregnancy and all main covariates. The Figures 6.2 (a) to 6.2 (d) show the relationship plot of proportions of teenage pregnancy, respectively, with *year*, *age*, *idhhms* and *nch*. The calculation of the proportions in these plots will be for both two level data and the three level multilevel datasets in which they are averaged over each covariate.

Figure 6.2 shows some differences of the two datasets for the relationship between proportions and census year from 2011 to 2014. The other differences, though slightly, are observed for 18 year old females and for females with 1 and 2 children before year of observation. On the other hand, the relationship between the number of household membership and proportion of teenage

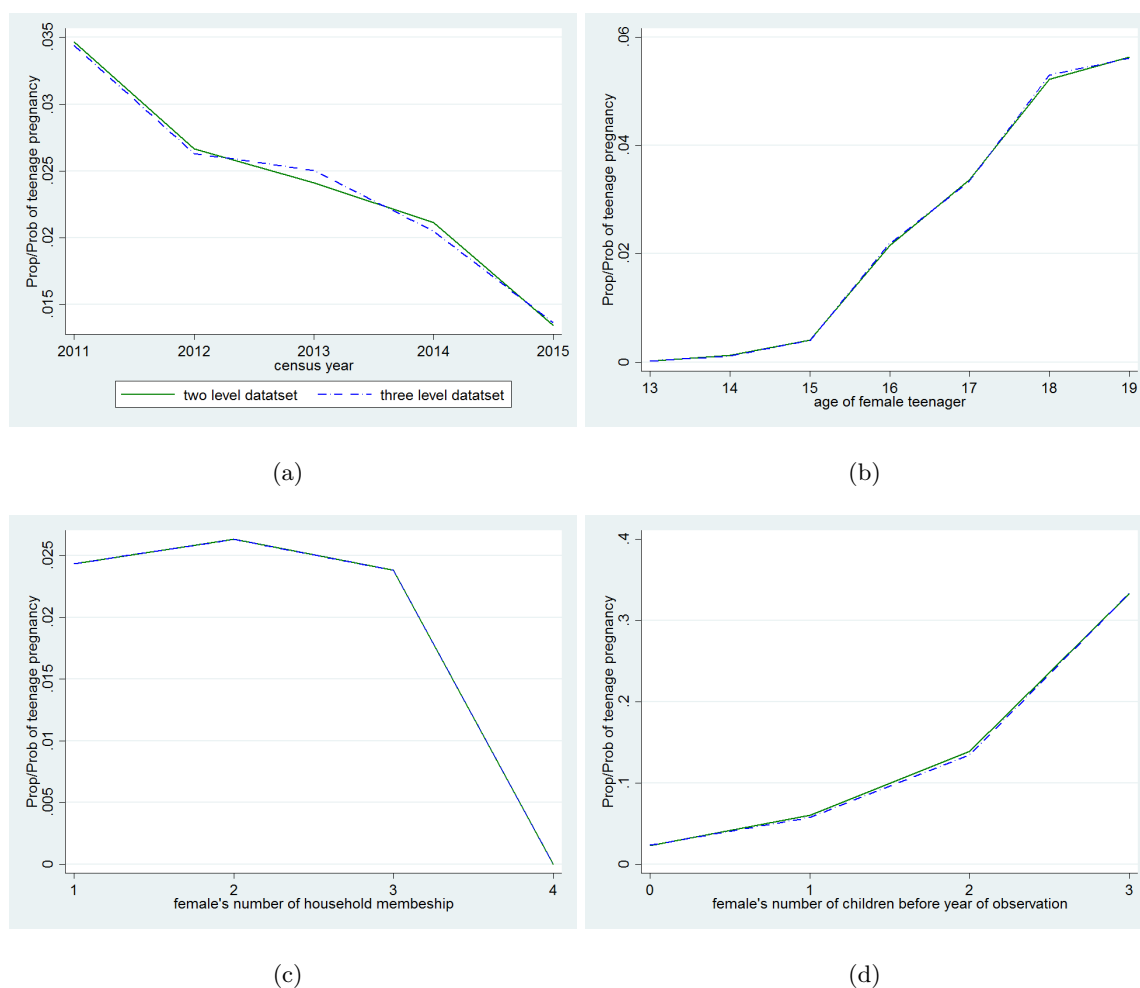


Figure 6.2: The relationship of probabilities of teenage pregnancy with *year*, *age*, *idhhms* and *nch*.

pregnancy for both datasets are the same. These slight differences might mean that the effect of these covariates would not differ much from the ones that were estimated in chapters 4 and/or 5.

6.2 Specifying and fitting a Logit GLMM for teenage pregnancy with household cluster effect

The models that will be fitted in this chapter follow the same methodology that was illustrated in chapter 5 for GLMM. This means that the effect of multiple membership at household level will be ignored; hence, the assumption made would be that a female belongs to one household at each census year. This section starts by building a two level logit model with a household cluster because some groups of teenage pregnancy measurements are also clustered by some specific household. Thereafter, this section will fit a three level logit model with both female

and household cluster.

6.3 Model building of a two level logit GLMM for teenage pregnancy

The same concept that was used in chapters 4 and 5 of fitting first the null model followed by adding *year* then adding other main covariates is also used. Table 6.3 shows the results of a two level logit GLMM building with only household cluster.

Table 6.3: Estimates for a two level logit GLMM building of pregnancy status with household cluster.

Covariate	Model					
	(6.3.1a)	(6.3.1b)	(6.3.1c)	(6.3.1d)	(6.3.1e)	(6.3.1f)
<i>Fixed effects</i>						
Intercept	-3.753*** (-55.06)	-3.372*** (-43.38)	-5.523*** (-46.21)	-5.534*** (-46.80)	-5.544*** (-46.94)	-5.577*** (-33.24)
<i>year</i>		-0.203*** (-8.32)	-0.208*** (-8.42)	-0.206*** (-8.40)	-0.208*** (-8.47)	-0.186* (-2.24)
<i>age</i>			0.551*** (23.94)	0.551*** (23.94)	0.556*** (24.56)	0.563*** (16.70)
<i>idhhms</i>			-0.0451 (-0.61)			
<i>nch</i>			0.140 (1.32)	0.141 (1.33)		
<i>year</i> × <i>age</i>						-0.00454 (-0.27)
<i>Random effects</i>						
$\sigma_{v0(3)}^2$	0.158 [1.22]	0.0691 [0.56]	2.50e-32 [0.00]	9.46e-33 [0.00]	1.57e-32 [0.00]	4.69e-29 [0.00]

z statistics in round parentheses (), $\chi^2_{(1)}$ statistics in square parentheses []

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results of Models (6.3.1a)-(6.3.1f) in Table 6.3 were obtained by fitting the following two level logit GLMM,

$$\text{logit}(\pi_{tj}) = \beta_0 + v_{0j}^{(3)}, \tag{6.3.1a}$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + v_{0j}^{(3)}, \tag{6.3.1b}$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + \beta_{ihm}idhhms_{tj} + \beta_{nch}nch_{tj} + v_{0j}^{(3)}, \tag{6.3.1c}$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + \beta_{nch}nch_{tj} + v_{0j}^{(3)}, \quad (6.3.1d)$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + v_{0j}^{(3)}, \text{ and} \quad (6.3.1e)$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + \beta_{yage}year_{ij} \times age_{ij} + v_{0j}^{(3)}, \quad (6.3.1f)$$

respectively. The term β_0 is a fixed intercept while $v_{0j}^{(3)}$ is the random intercept of the household cluster, in which $j = 1, 2, \dots, 7863$. The random intercept is normally distributed with mean zero and variance $\sigma_{v0(3)}^2$, where (3) indicates household cluster. The fixed effects of **year**, **age**, **idhhms**, **nch** and **year** \times **age** are the magnitude of β_{year} , β_{age} , β_{ihm} and β_{nch} .

As in chapters 4 and 5, the main fixed effects of **idhhms**, **nch** and the interaction effect **year** \times **age** are not significant at 5% level of significance. Even more, none of the models produce evidence that teenage pregnancy varies within households. This is because all variances are insignificant with p -values of more than 0.05 level of significance. For this reason, none of the logit GLMM with household cluster can be used for probability predictions of teenage pregnancy.

6.3.1 Model building for three level logit GLMM

Just as other model building in this dissertation, this analysis starts by fitting a null model of teenage pregnancy, but this time with both female and household cluster, thus

$$\text{logit}(\pi_{tij}) = \beta_0 + v_{0i}^{(2)} + v_{0ij}^{(3)}, \quad (6.3.2)$$

where β_0 is a fixed intercept. The random intercept $v_{0i}^{(2)}$ of female cluster and $v_{0ij}^{(3)}$ of household cluster are assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix Σ_v . The estimates of Model (6.3.2) that are shown in Table 6.5 indicate that the fixed intercept β_0 is significant at 5% level of significance.

The estimated variance within female cluster between households was observed as 0.59 while within household variance, $\sigma_{v0(3)}^2 = 2.46 \times 10^{-32}$, is very close to zero (see Model (6.3.2) results in Table 6.5). These results also shows that the random intercept $v_{0i}^{(2)}$ is significant at 5% level of significance while $v_{0j}^{(3)}$ is not. This means that teenage pregnancy does not vary within household; hence, there is no effect of the household cluster.

For illustration purposes, we examine the magnitude of the VPC and ICC, in which we will show that a three level mixed model report the $VPC_v^{(2)}$ value that is different from the $ICC_v^{(2)}$ value. Table 6.4 shows the variance components estimates of Model (6.3.2).

The $ICC_v^{(3)} = 6.34 \times 10^{-33}$ suggest that there is an extremely small correlation of teenage

Table 6.4: Variance components of female and household random effect, Model (6.3.2).

Components	Coef.	Std. Err.	[95% Conf. Interval]
<i>Females level</i>			
$ICC_v^{(2)}$	0.152	0.043	[0.085, 0.257]
$VPC_v^{(2)}$	0.152		
<i>Households level</i>			
$ICC_v^{(3)}$	6.34×10^{-33}	8.93×10^{-18}	[.,1]
$VPC_v^{(3)}$	6.34×10^{-33}		

pregnancies in the same household. For the reason that $VPC_v^{(3)} = ICC_v^{(3)} = 6.34 \times 10^{-33}$, the variation of teenage pregnancy that lies between is approximately 0. However, $ICC_v^{(2)} = 0.152$, means that there is a reasonable correlation of teenage pregnancy for the same female. In Table 6.4, the values of $VPC_v^{(2)}$ are 6.34×10^{-33} and 0.152; hence, 15.2% of the variation in teenage pregnancy lies within households between females. Although it was expected that the magnitudes of $VPC_v^{(2)}$ and $ICC_v^{(2)}$ are different, it is not the case. This is because the value of $\sigma_{v0(3)}^2$ is very small such that it is approximately zero.

The model fitted for the estimates of Model (6.3.2) in Table 6.5 added a fixed effect β_{year} of census year to Model (6.3.2), thus

$$\text{logit}(\pi_{tij}) = \beta_0 + \beta_{year}year_{tj} + v_{0i}^{(2)} + v_{0ij}^{(3)}. \quad (6.3.3)$$

The results indicated that both the fixed intercept and the effect of census year were significant at 5% level of significance. However, both the random intercept effect of female and household cluster are insignificant; hence, Model (6.3.3) cannot be used as a logit GLMM that predicts teenage pregnancy.

Model building of the three level logit GLMM, further added the effect of other main covariates (*idhhms* and *nch*) to Model (6.3.3), thus

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + \beta_{ihm}idhhms_{tj} + \beta_{nch}nch_{tj} + v_{0j}^{(3)}, \quad (6.3.4a)$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + \beta_{nch}nch_{tj} + v_{0j}^{(3)}, \text{ and} \quad (6.3.4b)$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + v_{0j}^{(3)}. \quad (6.3.4c)$$

$$\text{logit}(\pi_{tj}) = \beta_0 + \beta_{year}year_{tj} + \beta_{age}age_{tj} + \beta_{yage}year_{ij} \times age_{ij} + v_{0j}^{(3)}. \quad (6.3.4d)$$

The effects of *idhhms* (in the Model (6.3.4a)) and *nch* (in the Models (6.3.4a) and (6.3.4b)) were found not to be significant at 5% level of significance. On the other hand, Model (6.3.4c) indicated that the fixed effects of *year* and *age* were significant with *p*-values of less than 0.001; hence Model (6.3.4d). Nonetheless, the interaction effect of census year and age of the female was also found insignificant at 5% significance level.

Table 6.5: Estimates of model building for a three level logit GLMM of pregnancy status.

Covariate	Model					
	(6.3.2)	(6.3.3)	(6.3.4a)	(6.3.4b)	(6.3.4c)	(6.3.4d)
<i>Fixed effects</i>						
Intercept	-3.949*** (-41.89)	-3.506*** (-34.69)	-5.523*** (-46.21)	-5.534*** (-46.80)	-5.601*** (-37.62)	-5.641*** (-28.86)
<i>year</i>		-0.200*** (-8.10)	-0.208*** (-8.42)	-0.206*** (-8.40)	-0.209*** (-8.45)	-0.184* (-2.20)
<i>age</i>			0.551*** (23.94)	0.551*** (23.94)	0.558*** (24.29)	0.566*** (16.51)
<i>idhhms</i>			-0.0451 (-0.61)			
<i>nch</i>			0.140 (1.32)	0.141 (1.33)		
<i>year</i> × <i>age</i>						-0.00531 (-0.32)
<i>Random effects</i>						
$\sigma_{v0(2)}^2$	0.590** [2.99]	0.348 [1.90]	4.67e-32 [0.00]	1.65e-30 [0.00]	0.109 [0.63]	0.112 [0.65]
$\sigma_{v0(3)}^2$	2.46e-32 [0.00]	1.60e-33 [0.00]	9.55e-33 [0.00]	3.60e-33 [0.00]	4.92e-30 [0.00]	2.52e-29 [0.00]

z statistics in round parentheses (), $\chi^2_{(1)}$ statistics in square parentheses []

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Apart from the reason that the fact of the ignorance of multiple membership data structure, random part of the estimates indicates that the variance at level 3 are neglectable as they are very close to zero. Hence, we will not use the three-level models for the prediction of teenage pregnancy.

6.4 Summary

The intention of this chapter was to build a two and three level logit GLMM with household cluster that can be used to predict teenage pregnancy. However, the results obtained from the analysis indicated that household cluster has no effect on teenage pregnancy. This is because the variances at this cluster were very small (approximately zero). Chung and Beretvas (2012) indicated that ignoring the fact of multiple membership at a particular cluster underestimates the variance at that cluster. This means that the small variance at household level might have been caused by the assumption of a pure hierarchy of females within households for the dataset in this chapter.

Nonetheless, the work in this chapter has illustrated how one can fit a three level logit GLMM on a pure hierarchical data structure. Even more, this chapter leaves us with a question as to whether, by using appropriate MLM to fit a multiple membership data structure, the household cluster will indeed not have an effect on teenage pregnancy.

Chapter 7

Summary, Conclusions and Recommendations

This chapter summarises the research work embraced in all six chapters. The chapter will thereafter draw conclusions and suggest recommendations for further studies.

7.1 Summary

7.1.1 Chapters 1 and 2

The overall goal of this dissertation is to demonstrate the importance of using generalised linear mixed models to fit and assess variance components of the nested nature of social science data. Chapter 1 specified that the intended aim was stretched out into five objectives. The first objective of this study was accomplished in chapter 3 that explored the multilevel structure of teenage pregnancy and chapter 5 that specified appropriate model to fit the presumed data structure. In exploring the data structure of this work, chapter 3 suggested a two level hierarchical structure with at most five measurements of teenage pregnancy that are purely nested within females. This data structure was explored in chapter 2, in which possible multilevel data structures were discussed.

7.1.2 Chapters 3

In chapter 2, a pure two level hierarchical data structure is observed when level one units (measurements of teenage pregnancy) are nested into one and only one level two unit (female teenager) (Goldstein, 2011). This exploration of data also revealed seven variables in which one is the response variable, one is a cluster variable and five are covariates. Pregnancy status, denoted by *ps* is a binary response variable that can be either “yes” or “no”. The variable *id* that denotes female teenagers is the cluster at level two. The covariate *spft* is the only factor variable and it indicates the 2011, 2012, 2013, 2014 and 2015 sub-population of female teenagers. Census year denoted by *year* is also a covariate that captures the years 2011, 2012, 2013, 2014 and 2015 but respectively substituted by categories 0, 1, 2, 3 and 4 in order to measure time where 0 is the reference category. This means that *spft* and *year* are used interchangeably. The other three covariates denoted by *age*, *idhhms* and *nch* are age of the female, number of households a female belong to and the number of children the female had before year of observation, respectively. The covariates *age*, *idhhms* and *nch* are the only characteristics of the female teenager that are available in teenage pregnancy dataset.

7.1.3 Chapter 4 and 5

Table 7.1 shows the summarised estimates for better fit GLMs and GLMMs in chapter 4 and chapter 5. Secondly, chapter 4 addressed the objective (b) that intended to fit the presumed multilevel data structure. This was conducted with a purpose to examine the effect of the aforementioned covariates but intentionally ignoring the clustering of measurements of teenage

Table 7.1: Likelihood estimates of the better fit GLM and GLMM.

Covariate	Model			
	(4.3.2)	(5.3.2)	(4.3.5)	(5.3.5)
<i>Fixed effects</i>				
Intercept	-3.332*** (-64.27)	-3.501*** (-33.09)	-5.542*** (-44.61)	-5.599*** (-35.88)
year	-0.207*** (-8.12)	-0.203*** (-7.86)	-0.210*** (-8.17)	-0.211*** (-8.15)
age			0.557*** (23.42)	0.559*** (23.16)
<i>Random effects</i>				
$\sigma_{v0(2)}^2$		0.350* [3.36]		0.111 [0.181]
<i>Variance Components</i>				
ICC/VPC		0.096		0.032
<i>Akaike's information criterion estimates</i>				
df	2	3	3	4
AIC	7995.312	7993.952	7248.077	7249.705

t statistics in round parentheses (), $\chi^2_{(1)}$ statistics in square parentheses []

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

pregnancy at female cluster. Chapter 4 builds two logit generalised linear models of which the first fits the effect **year** together with the effect of three other covariates. The results revealed that among the four effects, only **year** and **age** are risk factors of teenage pregnancy at 0.05 significance level. This means that, at 5% level of significance, there was no evidence of the effect of **idhhms**, **nch** and the interaction of effect of **year** and **age** on teenage pregnancy. The second logit GLM fits the effect of **spft** together with the effect of the resulting significant covariate **year**. Even in the model that accounts for sub-population differences, only **spft** and **age** were significant. The estimates of the effect of the categories of **spft** tied exactly with the observed proportions of teenage pregnancy. Nonetheless, the final logit GLM with **year** effect stands out to be the better one than that with **spft** effect because of their AIC values. The predictions using the better fit logit GLM, $\text{logit}(\pi_i) = -5.542 - 0.21\text{year}_i + 0.557\text{age}_i$, indicated that the probabilities of teenage pregnancy are high for older female teenagers in 2011 and across census year. These predictions also indicated that teenage pregnancy in 2011 was higher at each female age compared to across census year. On the other hand, the probabilities of teenage pregnancy are very low for 13 year old females for each census year compared to 14-19 year old females. This means that by not considering the 14-19 year old females the predicted probabilities will differ drastically with the observed proportions. Moreover, the predicted

probabilities slightly tie with the observed proportion even though clustering has been ignored.

Thirdly, chapter 5 addressed objective (c) in which a two level multilevel or generalised linear mixed model was used to examine the effects that the aforementioned covariates has on teenage pregnancy. Chapter 5 also built two models, one with *year* and another with *spft* fixed effect. As in chapter 4, the model with *year* effect was better, thus $\text{logit}(\pi_{ti}) = -3.501 - 0.203\text{year}_{ti} + v_{0i}^{(2)}$. However, different from the results of a better logit GLM, the better logit GLMM suggested that there was no effect of *age*, *idhhms* and *nch* at 5% level of significance. This means that *year* was the only significant effect on teenage pregnancy. However, the better logit GLMM also included the female cluster effect that indicated that teenage pregnancy varies within females. Taking into consideration that the better logit GLMM fitted the nested nature of the pregnancy data, this model is considered a better fit despite the comparison of the AIC values with the better fit logit GLM. Supposably if age of the female teenager was not one of the covariates, a logit GLM with only *year* would be considered as a better fit logit GLM, thus $\text{logit}(\pi_i) = -3.332 - 0.207\text{year}_i$. By comparing the supposed estimate with the ones of a better fit logit GLMM, the fixed intercept is over estimated by 0.169 while the fixed effect of census year is under estimated by 0.04 when female cluster is ignored. Although the differences are not that much, it is a fact that by ignoring the nested nature of the data can indeed lead to misleading estimates; leading to incorrect conclusions. Even more, the AIC of the supposed logit GLM is higher than that of a better fit logit GLMM by 1.36; hence, the prediction of probabilities are compromised when clustering effect is pushed aside.

To address the fourth objective of this work, chapter 5 also computed the variance partition and intraclass correlation coefficients in order to examine the effects of female cluster on teenage pregnancy. The better fit logit GLMM indicated that there is a significant effect of the female cluster on teenage pregnancy. This conclusion was made because the p -value of the χ^2 with 1 degrees of freedom of critical value for the variance of teenage pregnancy within female cluster was found to be less than 0.05. Chapter 5 further used the variance to calculate the VPC and ICC of the female cluster that are both 0.096. This value means that 9.6% of the variation in teenage pregnancy lies between the females and also that there is, from a scale of 0 to 1, there is 0.096 correlation of teenage pregnancies for the same females. Although this suggests reasonable differences of teenage pregnancy within female cluster, there are very small similarities of teenage pregnancies for the same females.

Furthermore, after the realisation of a significant effect of census year, this study is also able

to address the objective (e) that intended to assess whether this effect varies across females. However, the variance of census year across female cluster was not significant. Meaning that the effect of census year on teenage pregnancy does not differ across female teenagers.

7.2 Conclusions and Recommendations

Clearly, the analyses conducted and presented in chapters 3, 4 and 5 have addressed all objectives of this dissertation; thereby accomplishing the main aim of this work. That is, given that data in social and health science are nested in nature it is important to use appropriate models to assess variance components. Although generalised linear models with a logit link are able to model binary data, it is evident that logit generalised linear mixed models are able to fit binary data and also assess variance components. Based on the results comparison of chapters 4 and 5, it is clear that the differences and similarities that are generated by clustering are essential. This means that it is important that both data scientists and researchers should see the relevance of having some knowledge on GLMM or MLM statistical methods prior to collection and analysis planning.

Moreover, Chapter 6 expanded the two level data structure to a three level with household at level 3. In this three level data structure, a complex structure than that of pure hierarchical was observed where some females are clustered by more than one household. This means there exists a multiple membership classification of females at level three; hence, a multiple membership data structure that was discussed in chapter 2. Nonetheless, the analysis undertaken in chapter 6 ignored the fact of multiple membership assuming pure hierarchy between female and household level. After fitting a two level (household at level 2) and three level (with female at level 2 and household at level 3) logit GLMM, the results thereof suggested that the household cluster has no effect on teenage pregnancy.

Although chapter 6 revealed that female teenagers are nested within multiple households, the scope of this study did not cover multiple membership multilevel models. The reason for this non-coverage is also due to the large number of household which make the computations of the household random effect complex and time consuming. Given this challenges, the future line of this work will focus on social and health science data that are either multiple membership data structures and/or a mixture of the three data structures which were discussed in chapter 2. This will also require statisticians and statistics software developers to use approximation method which are able to estimate both the fixed and the random effects, faster and effectively.

References

- Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V. and van der Sluis, S. (2014), ‘A solution to dependency: using multilevel analysis to accommodate nested data’, *Nature Neuroscience* **17**(4), 491–496.
- Aitkin, M. and Aitkin, I. (2011), *Statistical modeling of the national assessment of educational progress*, Springer.
- Akaike, H. (2011), Akaike’s information criterion, *in* ‘International Encyclopedia of Statistical Science’, Springer, pp. 25–25.
- Attanasio, M. and Capursi, V. (2011), *Statistical Methods for the Evaluation of University Systems*, Springer.
- Aunsmo, A., Øvretveit, S., Breck, O., Valle, P. S., Larssen, R. B. and Sandberg, M. (2009), ‘Modelling sources of variation and risk factors for spinal deformity in farmed atlantic salmon using hierarchical-and cross-classified multilevel models’, *Preventive Veterinary Medicine* **90**(1), 137–145.
- Bhat, C. R. (2000), ‘A multi-level cross-classified model for discrete response variables’, *Transportation Research Part B: Methodological* **34**(7), 567–582.
- Bini, M., Piccolo, D., Monari, P. and Salmaso, L. (2009), *Statistical methods for the evaluation of educational services and quality of products*, Physica-Verlag.
- Christofides, N. J., Jewkes, R. K., Dunkle, K. L., Nduna, M., Shai, N. J. and Sterk, C. (2014), ‘Early adolescent pregnancy increases risk of incident hiv infection in the eastern cape, south africa: a longitudinal study’, *Journal of the International AIDS Society* **17**(1).
- Chung, H. and Beretvas, S. N. (2012), ‘The impact of ignoring multiple membership data structures in multilevel models’, *British Journal of Mathematical and Statistical Psychology* **65**(2), 185–200.

- Costantini, P. and Vitale, M. P. (2011), Analyzing undergraduate student graduation delay: A longitudinal perspective, in 'Statistical methods for the evaluation of university systems', Springer, pp. 145–159.
- Creemers, B., Kyriakides, L. and Sammons, P. (2010), *Methodological Advances in Educational Effectiveness Research*, Quantitative Methodology Series, Taylor & Francis.
- Czepiel, S. A. (2002), 'Maximum likelihood estimation of logistic regression models: theory and implementation', Available from <https://czep.net/stat/mlelr.pdf>. Accessed on 03.12.2016.
- Dagum, E. B., Bianconcini, S. and Monari, P. (2009), Nonlinearity in the analysis of longitudinal data, in 'Statistical methods for the evaluation of educational services and quality of products', Springer, pp. 47–60.
- de Deleeuw, J., Goldstein, H. and Meijer, E. (2007), *Handbook of Multilevel Analysis*, Springer.
- De Leeuw, J. and Kreft, I. (1986), 'Random coefficient models for multilevel analysis', *Journal of Educational and Behavioral Statistics* **11**(1), 57–85.
- De Leeuw, J., Meijer, E. and Goldstein, H. (2008), *Handbook of multilevel analysis*, Springer.
- Fielding, A. and Yang, M. (2005), 'Generalized linear mixed models for ordered responses in complex multilevel structures: Effects beneath the school or college in education', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**(1), 159–183.
- Fox, J.-P. (2004), 'Modelling response error in school effectiveness research', *Statistica Neerlandica* **58**(2), 138–160.
- Gitelman, A. I. (2005), 'Estimating causal effects from multilevel group-allocation data', *Journal of Educational and Behavioral Statistics* **30**(4), 397–412.
- Goldstein, H. (1986), 'Multilevel mixed linear model analysis using iterative generalized least squares', *Biometrika* **73**(1), 43–56.
- Goldstein, H. (1997), 'Methods in school effectiveness research', *School Effectiveness and School Improvement* **8**(4), 369–395.
- Goldstein, H. (2011), *Multilevel statistical models*, Vol. 922, John Wiley & Sons.
- Green, J. G., Dunn, E. C., Johnson, R. M. and Molnar, B. E. (2011), 'A multilevel investigation of the association between school context and adolescent nonphysical bullying', *Journal of School Violence* **10**(2), 133–149.

- Gustafsson, J.-E. (2013), 'Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement 1', *School Effectiveness and School Improvement* **24**(3), 275–295.
- Hall, D. B. and Clutter, M. (2004), 'Multivariate multilevel nonlinear mixed effects models for timber yield predictions', *Biometrics* **60**(1), 16–24.
- Hill, P. W. and Rowe, K. J. (1996), 'Multilevel modelling in school effectiveness research', *School Effectiveness and School Improvement* **7**(1), 1–34.
- Hox, J. and Roberts, J. K. (2011), *Handbook of advanced multilevel analysis*, Psychology Press.
- INDEPTH, N. (2014), Indepth network - annual report, Technical report, INDEPTH Network, Accra.
- Johnson, B. D. (2012), 'Cross-classified multilevel models: an application to the criminal case processing of indicted terrorists', *Journal of Quantitative Criminology* **28**(1), 163–189.
- Kumar, A., Singh, T., Basu, S., Pandey, S. and Bhargava, V. (2007), 'Outcome of teenage pregnancy', *Indian Journal of Pediatrics* **74**(10), 927–931.
- Kwok, O.-M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R. and Yoon, M. (2008), 'Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures.', *Rehabilitation Psychology* **53**(3), 370.
- Langille, D. B. (2007), 'Teenage pregnancy: trends, contributing factors and the physician's role', *Canadian Medical Association Journal* **176**(11), 1601–1602.
- Lodico, M. G., Spaulding, D. T. and Voegtle, K. H. (2006), *Methods in Educational Research: From Theory to Practice*, Research Methods for the Social Sciences, Wiley.
- Luke, D. A. (2004), *Multilevel modeling*, Vol. 143, Sage.
- Malema, R. N. (2000), Risk factors associated with teenage pregnancy at Ga-Dikgale villages in the Northern Province of South Africa, PhD thesis, University of Pretoria.
- Mchunu, G., Peltzer, K., Tutshana, B. and Seutlwadi, L. (2012), 'Adolescent pregnancy and associated factors in south african youth', *African Health Sciences* **12**(4), 426–434.
- Molenberghs, G. and Verbeke, G. (2005), *Models for discrete longitudinal data*, Springer.

- Natesan, P., Limbers, C. and Varni, J. W. (2010), 'Bayesian estimation of graded response multilevel models using gibbs sampling: formulation and illustration', *Educational and Psychological Measurement* **70**(3), 420–439.
- Nguyen, H., Shiu, C. and Farber, N. (2016), 'Prevalence and factors associated with teen pregnancy in vietnam: Results from two national surveys', *Societies* **6**(2), 17.
- O'Connell, A. A. and Reed, S. J. (2012), 'Hierarchical data structures, institutional research, and multilevel modeling', *New Directions for Institutional Research* **2012**(154), 5–22.
- Pornprasertmanit, S., Lee, J. and Preacher, K. J. (2014), 'Ignoring clustering in confirmatory factor analysis: some consequences for model fit and standardized parameter estimates', *Multivariate Behavioral Research* **49**(6), 518–543.
- Rabe-Hesketh, S. and Skrondal, A. (2006), 'Multilevel modelling of complex survey data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(4), 805–827.
- Rabe-Hesketh, S. and Skrondal, A. (2008), *Multilevel and longitudinal modeling using Stata*, STATA press.
- Salgado, M. and Marchione, E. (2011), *Multilevel and agent-based modelling in the analysis of differential school effectiveness*, pp. 50–55.
- Shepelev, I. (2011), 'Identification of the hierarchical data structure', *Pattern Recognition and Image Analysis* **21**(2), 211–214.
- Singer, J. D. (1998), 'Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models', *Journal of Educational and behavioral statistics* **23**(4), 323–355.
- Singer, J. D. and Willett, J. B. (2003), *Applied longitudinal data analysis: Modeling change and event occurrence*, Oxford university press.
- Snijders, T. A. (2011), *Multilevel analysis*, Springer.
- STATSSA (2014), Recorded live births, Technical report, Statistics South Africa.
- Steele, F. (2008), 'Multilevel models for longitudinal data', *Journal of the Royal Statistical Society: series A (statistics in society)* **171**(1), 5–19.
- Thum, Y. M. (2003), 'Measuring progress toward a goal estimating teacher productivity using a multivariate multilevel model for value-added analysis', *School Effectiveness and School Improvement* **32**(2), 153–207.

- Tolvanen, A., Kiuru, N., Leskinen, E., Hakkarainen, K., Inkinen, M., Lonka, K. and Salmela-Aro, K. (2011), 'A new approach for estimating a nonlinear growth component in multilevel modeling', *International Journal of Behavioral Development* **35**(4), 370–379.
- Tomita, A. and Burns, J. K. (2013), 'A multilevel analysis of association between neighborhood social capital and depression: Evidence from the first south african national income dynamics study', *Journal of affective disorders* **144**(1-2), 101–105.
- Ukwuani, F. A., Tsui, A. O. and Suchindran, C. M. (2003), 'Condom use for preventing hiv infection/aids in sub-saharan africa: a comparative multilevel analysis of uganda and tanzania.', *Journal of acquired immune deficiency syndromes (1999)* **34**(2), 203–213.
- Van Buuren, S. (2011), 'Multiple imputation of multilevel data', *Handbook of Advanced Multilevel Analysis* pp. 173–196.
- VanderWeele, T. J. (2010), 'Direct and indirect effects for neighborhood-based clustered and longitudinal data', *School Effectiveness and School Improvement* **38**(4), 515–544.
- Vaughn, B. K. (2008), 'Data analysis using regression and multilevel/hierarchical models, by gelman, a., & hill, j.', *Journal of Educational Measurement* **45**(1), 94–97.
- Wang, J., Xie, H. and Fisher, J. F. (2011), *Multilevel models: applications using SAS®*, Walter de Gruyter.
- Wiysonge, C. S., Uthman, O. A., Ndumbe, P. M. and Hussey, G. D. (2012), 'Individual and contextual factors associated with low childhood immunisation coverage in sub-saharan africa: a multilevel analysis', *PLoS One* **7**(5), e37905.
- Woodward, L., Fergusson, D. M. and Horwood, L. J. (2001), 'Risk factors and life processes associated with teenage pregnancy: Results of a prospective study from birth to 20 years', *Journal of Marriage and Family* **63**(4), 1170–1184.
- Zaccarin, S. and Rivellini, G. (2002), 'Multilevel analysis in social research: an application of a cross-classified model', *Statistical Methods and Applications* **11**(1), 95–108.
- Zumbo, B. D. and Chan, E. K. (2014), *Validity and validation in social, behavioral, and health sciences*, Springer.

Appendices

Appendix A

Fitting Logit GLMs

A.1 Logit generalised linear models stata codes in chapter 4

A.1.1 Load a two level teenage pregnancy dataset

use "file path", replace

A.1.2 Fitting and Building Logit generalised linear model (GLM)

```
//Model (4.3.1): Fit a null logit GLM and store estimates as est1
    eststo: glm ps, family(binomial 1) link(logit)
//Model (4.3.2): Add the effect of year to Model (4.3.1) and store estimates as est2
    eststo: glm ps (year), family(binomial 1) link(logit)
//Model (4.3.3): Add the effects of age, idhhms and nch to Model (4.3.2) and store estimates as est3
    eststo: glm ps (year) age idhhms nch, family(binomial 1) link(logit)
//Model (4.3.4): Remove the effect of idhhms to Model (4.3.3) and store estimates as est4
    eststo: glm ps (year) age nch, family(binomial 1) link(logit)
//Model (4.3.5): Remove the effect of nch to Model (4.3.4) and store estimates as est5
    eststo: glm ps (year) age, family(binomial 1) link(logit)
//Model (4.3.6): Add the effect of year  $\times$  age to Model (4.3.5) and store estimates as est6
    eststo: glm ps (year) age c.year#c.age, family(binomial 1) link(logit)
//Produce AIC statistics for all six logit GLM in Table 4.2
    estimates stats est1 est2 est3 est4 est5 est6
//Clear all stored estimates
    eststo clear
```

Appendix B

Fitting Logit GLMMs

B.1 Logit generalised linear mixed models stata codes in chapter 5

B.1.1 Load a two level teenage pregnancy dataset

use "file path.dta", replace

B.1.2 Null generalised linear mixed model

```
//Model (5.3.1): Fit a null logit GLMM with female cluster (id) and store estimates as est1
    eststo: meqrlogit ps ||id:, variance
//Produce statistics AIC for Model (5.3.1)
    estat ic
```

B.1.3 Adding the census year fixed effect

```
//Model (5.3.2): Fit Model (5.3.1)+ year with female cluster (id) and store estimates as est2
    meqrlogit ps year ||id:, variance
//Produce AIC for Model (5.3.1)
    estat ic
```

B.1.4 Building a logit GLMM that adds, to Model (5.3.2), other main covariates

```
//Model (5.3.3): Add the effect of age, idhhms and nch to Model (5.3.2) and store estimates as est3
    eststo: meqrlogit ps (year) age idhhms nch ||id:, variance
//Model (5.3.4): Remove the effects of idhhms to Model (5.3.3) and store estimates as est4
    eststo: meqrlogit ps (year) age nch ||id:, variance
//Model (5.3.5): Remove the effect of nch to Model (5.3.4) and store estimates as est5
    eststo: meqrlogit ps (year) age ||id:, variance
//Produce AIC statistics for all five logit GLMM
    estimates stats est1 est2 est3 est4 est5
//Clear all stored estimates
    eststo clear
```

B.2 Logit generalised linear mixed models stata codes in chapter 6

B.2.1 Load a three level teenage pregnancy dataset

use "file path.dta", replace

B.2.2 Building a two level logit GLMM with a household cluster

```
//Model (6.3.1a):Null logit GLMM with the effect of  $v_{0j}^{(3)}$  and store estimates as est1
    eststo: meqrlogit ps ||hh:, variance
//Model (6.3.1b): Add the effect of year to Model (6.3.1a) and store estimates as est2
    eststo: meqrlogit ps year ||hh:, variance
//Model (6.3.1c): Add the effects of age, idhhms and nch to Model (6.3.1b) and store estimates as est3
    eststo: meqrlogit ps (year) age idhhms nch ||hh:, variance
//Model (6.3.1d): Remove the effect of idhhms to Model (6.3.1c) and store estimates as est4
    eststo: meqrlogit ps (year) age nch ||hh:, variance
//Model (6.3.1e): Remove the effect of nch to Model (6.3.1d) and store estimates as est5
    eststo: meqrlogit ps (year) age ||hh:, variance
//Model (6.3.1f): Add the interaction effect of year  $\times$  age to Model (6.3.1e) and store estimates as est5
    eststo: meqrlogit ps (year) age c.year#c.age ||hh:, variance
//Produce AIC statistics for all five logit GLMM
    estimates stats est1 est2 est3 est4 est5
//Clear all stored estimates
    eststo clear
```

B.2.3 Building a three level logit GLMM with female cluster and household cluster

```
//Model (6.3.1a):Null logit GLMM with the effect of  $v_{0i}^{(3)}$  and  $v_{0ij}^{(3)}$  and store estimates as est1
    eststo: meqrlogit ps ||hh: ||id:, variance
//Model (6.3.1b): Add the effect of year to Model (6.3.1a) and store estimates as est2
    eststo: meqrlogit ps year ||hh: ||id, variance
//Model (6.3.1c): Add the effects of age, idhhms and nch to Model (6.3.1b) and store estimates as est3
    eststo: meqrlogit ps (year) age idhhms nch ||hh: ||id, variance
//Model (6.3.1d): Remove the effect of idhhms to Model (6.3.1c) and store estimates as est4
    eststo: meqrlogit ps (year) age nch ||hh: ||id, variance
//Model (6.3.1e): Remove the effect of nch to Model (6.3.1d) and store estimates as est5
    eststo: meqrlogit ps (year) age ||hh: ||id, variance
//Model (6.3.1f): Add the interaction effect of year  $\times$  age to Model (6.3.1e) and store estimates as est5
    eststo: meqrlogit ps (year) age c.year#c.age ||hh: ||id, variance
//Produce AIC statistics for all five logit GLMM
    estimates stats est1 est2 est3 est4 est5
//Clear all stored estimates
    eststo clear
```