Statistical analysis of the school attendance rate among under 20 South African learners

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Thabang Goodman Chabalala

December, 2020

Statistical analysis of the school attendance rate among under 20 South African learners

by

Thabang Goodman Chabalala

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
MASTER OF SCIENCE
in
STATISTICS

Thesis Supervisor: Ms Danielle Roberts

Thesis Co-supervisor: Prof Temesgen Zewotir

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Thabang Goodman Chabalala, declare that

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

Thabang Goodman Chabalala (Student)	Date
Ms Danielle Roberts (Supervisor)	Date
Prof Temesgen Zewotir (Co-supervisor)	Date

Disclaimer

This document describes work undertaken as a Masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Abstract

School attendance is very crucial for the growth and development of the mindset of a child. The development of the mindset and provision of training to learners is an investment of a better future for the country. The government even made school attendance compulsory because of the fruits it bears in the future. But in the past, many studies have reflected a problem with school attendance and mostly the financial constrains appearing as the hindrance towards school attendance. Which is why the government has taken the initiative to make school attendance free for those who doesn't afford to pay for it. This has reduced a greater number of individuals who had a wish to attend school but with no funds to pay for it and allowed an opportunity for those who need it. But still the country is experiencing individuals who are in school going age but not attending school. Some of these individuals are enrolled for school but choose not to attend. This brings many questions now about the factors affecting school attendance of learners. Which brings us to the aim of this study which is to identify factors affecting school attendance of learners at the basic education level.

In identification of these factors, the study make use of different statistical models which accommodate the binary response. The models used in the study include Correspondence Analysis(CA), Survey Logistic Regression(SLR), Generalized Linear Mixed Model(GLMM) and Generalized Additive Mixed Model(GAMM). The results suggest that the likelihood of school non-attendance is associated with Northern Cape and Western Cape which are mostly dominated by Coloured/Indian/Asian

race groups sharing "Other" relationship to household head and have no parents presence. Moreover, the female learners with mothers not alive and coming from families with salaries and pension/grant as source of income are less likely to attend school. While learners coming from all other provinces except the two specified above, African/Black by race, sharing child/grandchild relationship to household head, have both parents alive, deviating from household with high wealth index z-score and have total income above R25000 are more likely to attend school. This is a clear indication that the initiatives which were applied by the government and results of the past studies have assisted in improving school attendance, but still more initiatives are needed to cover the areas which are still reflecting poor school attendance in order to meet the aims of the Millennium Development Goals.

Acknowledgements

I would firstly like to thank my main supervisor, Ms Danielle Roberts for her endless push towards helping me and giving guidance at all times of need in fulfilling this thesis project. Further on, all thanks to my co-supervisor Prof Zewotir for giving me the opportunity and also his wise guidance towards fulfilment of my work. More gratitude to my both supervisors for taking the role as parents in pushing me to work hard and with the presentations which assisted me in understanding my work even further. I would not forget to thank SASA for the funds they have given me to fulfil my studies, this has played a very crucial role in my life and in helping me moving forward smoothly with my research.

I wish to thank also the University of KwaZulu Natal, my School (Maths, Stats and Comp Sci), the Head of my department Prof Delia North and everyone in my School (Statistics) for their continuing support and words of encouragement towards me. Lastly wish to thank my family and friends especially my mother and fiancé for giving me time to work freely with my school work and understanding whenever im busy with my school work.

Contents

List of Figures

List of Tables

Abbreviations

AOR	Adjusted Odds Ratio
GLM	Generalized Linear Model
GHS	General Household Survey
Stat-SA	Statistics South Africa
PML	Pseudo-Likelihood estimation
PSU	Primary Sampling Unit
ROC	Receiver Operating Characteristic
AIC	Akaike Information Criteria
SC	Schwartz Criteria
FPC	Finite Population Correction
MLE	Maximum Likelihood Estimation
GLMM	Generalized Linear Mixed Model
SA	South Africa
Metro	Metropolitan
CA	Correspondence Analysis
MCA	Multiple Correspondence Analysis
SLR	Survey Logistic Regression
QL	Quasi Likelihood
ROC	Receiving Operating Characteristic
CI	Confidence interval
OR	Odds Ratio
SAGHS	South African General Household Survey
REML	Restricted Maximum Likelihood Estimation
PQL	Practical Quantitation Limit

MQL	Closed Loop Report
BIC	Bayesian Information Criteria
ML	Maximum Likelihood
GAMM	Generalized Additive Mixed Models
GMRF	Gaussian Markov Random Field

Chapter 1

Introduction

School attendance is a baseline factor in determining student success and is crucial for the growth and development of the mindset of a learner (Department of Education, South Africa, 2014). The development of the mindset and provision of training to learners is an investment of a better future for the individual, and in turn, their country. Education is a tool to empower people, improve an individual's earning potential, promote a healthy population, reduce poverty and crime, and build a competitive economy (Koledade, 2008). Education investment decisions are primarily made by parents in the hopes that their children will provide a better life someday for their families (Koledade, 2008). It has been well documented that strong academic attendance correlates with good academic success (Meador, 2017). Campbell (2006) shows that education is the most valuable and efficient weapon to combat diseases such as HIV and AIDS, Cancer, diabetes, and other chronic illnesses. Thus, poor education attendance has possible chances of poverty and malnutrition within communities, resulting in communities with a high number of health related issues (Campbell, 2006). This is attributed to less educated individuals often having a poor understanding of health related issues, where they are not sufficiently equipped with the knowledge and ability to understand how to face such issues.

Education motivates self-assurance and improves one's capabilities and potential.

Every individual requires a set of skills to survive in this competitive world and progress to a successful future. These skills develop through education and training, and knowledge helps individuals understand the environment that live in, together with its needs (Balfanz & Byrnes, 2012). Furthermore, education aids in equipping individuals with skills of development and understanding of modern technology (UKessays, 2018). In addition to school attendance being vital for a learner's academic success and development, being in school aids in keeping the learner out of trouble as well as danger. When a learner is at school, they are generally in a safe environment where they are prevented from the risks of trafficking and abuse. Moreover, attending school is more likely to prevent the learner from engaging in criminal activities. Anderson (2014) discussed the correlation between youth dropouts and juvenile criminal behaviour.

There are numerous reasons shown in studies that contribute to poor school attendance in educational facilities. These reasons include a lack of information about the importance of education and a lack of employment for qualified individuals, as the lack of jobs for individuals who are qualified increases the negative mindset towards school attendance. The mindset develops from the knowledge that lives among communities that education is essential for an individual to get employment, so when older peers do not get employment after being qualified, some learners see no need to attend school (Department of Basic Education, 2016). The study by Phineas Reuckert (2019) outlines the factors that affect education enrolment across the globe, where such factors contribute to poor or no attendance in schools. Financial issues appear to be the top-ranked barrier that decreases school attendance. A lack of trained and experienced teachers, which is a problem that many schools in South Africa (SA) face, also affects a learner's tendency to attend school as these teachers are not effective in student learning (Phineas Reuckert, 2019). The lack of classrooms, learning material, exclusion of learners with disabilities, distance to school, and poor nutrition also harms the school attendance of learners (Phineas

Reuckert, 2019). Although, financial constraints remain the central issue towards non-attendance of learners in South Africa (Stats-SA, 2017). Other issues are poor academic success, peer pressure, lack of parental control, and family commitments, such as pregnancy and marriage.

All South Africans have the right to basic education and the Bill of Rights obliges the government to continuously make education available and accessible through reasonable measures.The education system in South Africa has phases that are primary, secondary, and higher education, where the primary and secondary phases form part of a learner's basic education. The percentage of individuals that attend basic educational institutions has increased yearly since 1996 (National Department of Health (NDoH), Statistics South Africa (Stats SA), South African Medical Research Council (SAMRC), and ICF, 2019). The percentages of learners attending government schools increased from 0.4% in 2002 to 65.9% in 2014, before stalling and largely moving to 66% in 2017 (Stats-SA, 2017). This is a tremendous increase, even though there are numerous problems faced by public or government schools nationally. However, the country has to continue to improve and keep up with the education of its youth. Improved access to educational facilities and services has led to a continuous increase in educational attainment in South Africa compared with the past.

Meador (2017) suggested that schools should be challenged to develop attendance policies and programs together with parents, which will be used to improve education attendance in different areas around the country. This will allow partnership in overcoming the issue of absenteeism and school non-attendance, enabling parents to take part in the issues facing their children (Meador, 2017). In addition to the role of the parent and school in addressing school attendance, the government also has a role to play. The South African government uses short-term, medium-term, and long-term strategic goals to improve education attendance. These strategic goals

have a vision that by 2030, SA will have universal early childhood education, high-quality schooling, further education and training (quality basic education) under goal 4 of Sustainable Development Goals (Department of Education, South Africa, 2014). These government strategies are outlined to improve school attendance, and provide multiple outcomes. These outcomes include funding plans to lower financial constraints towards school attendance, improving the quality of teaching by providing infrastructure, learning material, and supporting teachers' training before the appointment. The improved planning for the extension of early childhood development provides credible outcomes focused on planning and accounting for system-building. The intervention of artificial intelligence and technology is also part of the strategies that the South African government plans to use to draw more learners to school and to improve understanding of educational concepts to all learners (Department of Education, South Africa, 2014).

However, despite these efforts, there are still learners in the country that do not attend school for various reasons. Thus, it is important to investigate the factors that contribute to school attendance and non-attendance. Such studies can assist the government in understanding the issues related to poor attendance and develop appropriate strategies to face these issues.

1.1 Objectives of the Thesis

The aim of this thesis is to make use of appropriate statistical models to investigate school attendance rates among South Africa learners under the age of 20 years old. The specific objectives are as follows:

- To explore the school attendance rates of learners in South Africa and across the provinces of South Africa as well as according to various socio-economic and demographic factors.
- To determine the factors that are significantly associated with school atten-

dance of a learner, as well as determine which factors contribute to a lower likelihood of school attendance.

- To examine the spatial variation in the likelihood of school attendance of a learner across the provinces of South Africa.
- To determine which provinces have a higher or lower likelihood associated with school attendance of a learner.

1.2 Structure of the Thesis

This thesis is structured in the following manner:
Chapter 1 covers the introduction to the topic and outlines the significance of the study, as well as the aims and objectives. Chapter 2 outlines the source, description and exploration of data used in this thesis. It also presents the results of multiple correspondence analysis, which provides a graphical representation of cross-tabulations between categories of the qualitative variables. Chapter 3 gives an overview of the survey logistic regression model and its application to the data used in this thesis. Further on, chapter 4 gives an overview of generalized linear mixed models, which extends on a generalized linear model. Chapter 5 gives an overview of generalized additive mixed model, which accounts for possible spatial variation and autocorrelation existing in the data. Lastly, Chapter 6 concludes and discusses the results of the various statistical approaches, and presents the limitations of the study.

Chapter 2

Data Description and Exploration

This chapter outlines the source, description, and exploration of the data that was used in this thesis.

2.1 Data description

The data used in this study is drawn from the South African General Household Survey (SAGHS) that was conducted by Statistics South Africa (STATS-SA) from January to December in the year 2017. The survey was aimed at determining the progress of development within the country by measuring the performance of programs and the quality of service delivery within the country. The SAGHS was nationally represented where the data was collected based on a stratified two-stage cluster sampling design. South Africa was stratified into its 9 provinces which were then further subdivided into the different geo-type within metro/non-metro areas. These geo-types included urban, traditional, and farm areas. The first stage of sampling involved selecting the primary sampling units/cluster with a probability proportional to size, and the second stage involved sampling of the dwelling units based on systematic sampling. The final data set used in this thesis consisted of 21 033 observations from individuals of the ages 5 to 19 years old who had not yet completed their matric (Grade 12 year).

2.2 Study Variables

The response variable considered in this thesis was based on whether or not the learner was currently attending a basic educational institution at the time of the survey. This response, which will be referred to as 'school attendance', is binary. The explanatory variables considered were based on a range of individual-level, household-level and geographical factors, as shown in Figure 2.1 on the next page. The household wealth index Z-score was a composite measure based on the ownership of durable goods in the household.

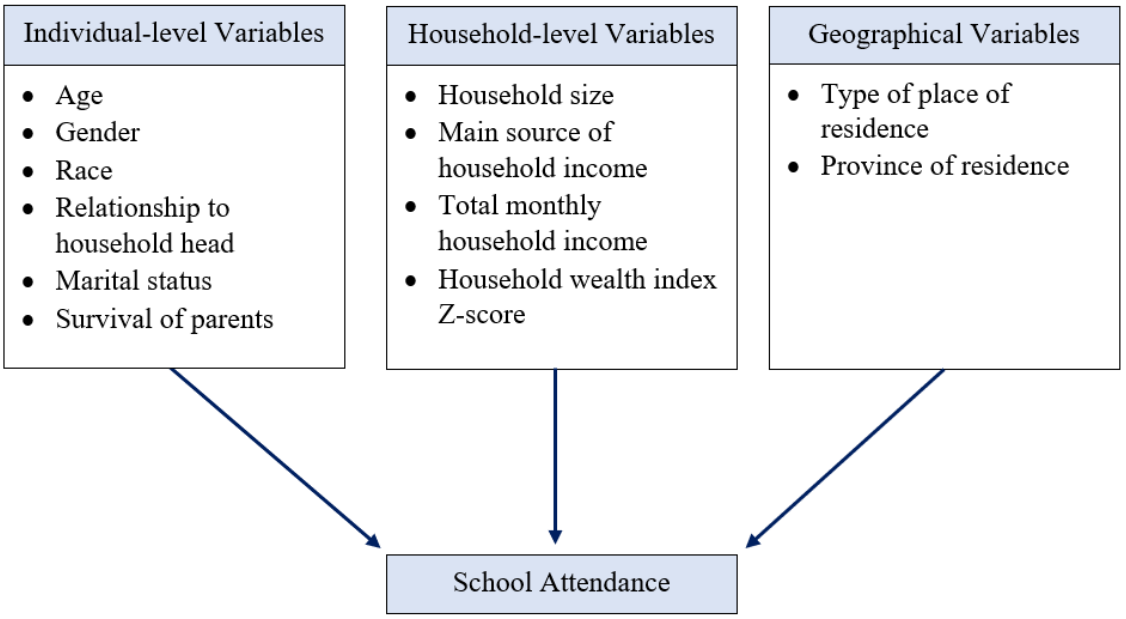

Figure 2.1: Potential factors associated with school attendance

2.3 Data exploration

Before any advanced statistical modelling is done, it is important to get an understanding of the data that is being used. In this section, an exploratory data analysis is performed. Table 2.1 presents the distribution of the sample according to school attendance. At the time of the survey, 93.4% of learners between the ages of 5 and 19 years old were attending school.

Table 2.1: The summary of learner's current school attendance within South Africa

Current school attendance	Frequency	Percentage (%)
Yes	19 646	93.4
No	1387	6.6

Figure 2.2 below breaks down school attendance according to the province of residence. Limpopo province had the highest percentage of learners attending school at 97.6%, followed by the Free State at 94.9%. The Northern Cape and the Western Cape had the lowest rates of school attendance among the learners, with only 88.4% and 89.5% of the learners that reside in these provinces attending school, respectively. In addition, Figure 2.2 demonstrates possible spatial variation in school attendance rates across the difference provinces of South Africa.

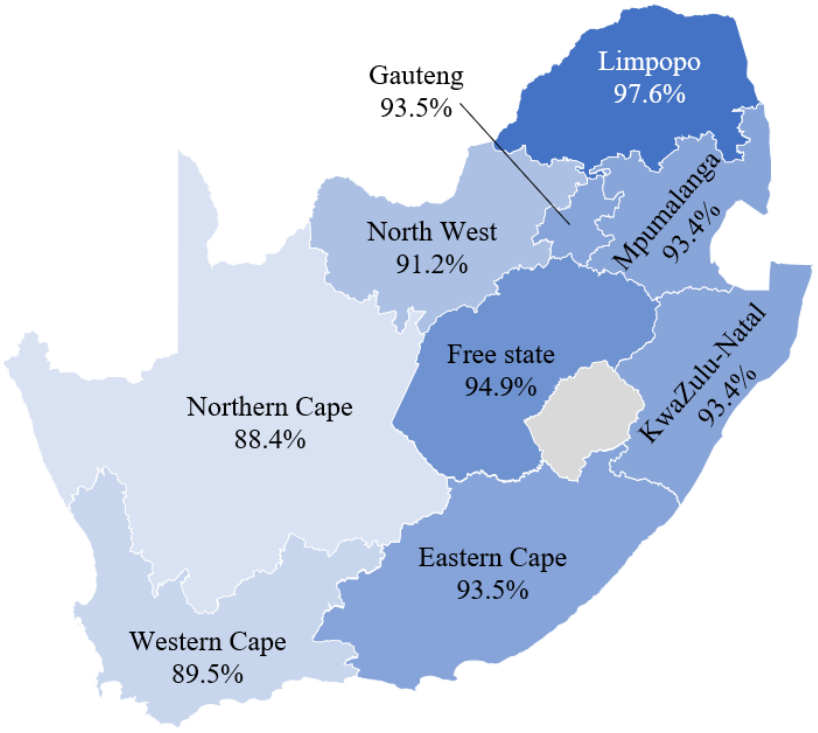

Figure 2.2: School attendance rate according to province of residence

During the survey, the type of area of residence was categorised as a metro area or non-metro area. Figure 2.3 presents the rate of school attendance and non-attendance according to these areas of residence. This figure reflects that the school attendance rate was slightly higher among learners residing in non-metro regions (93.7%).

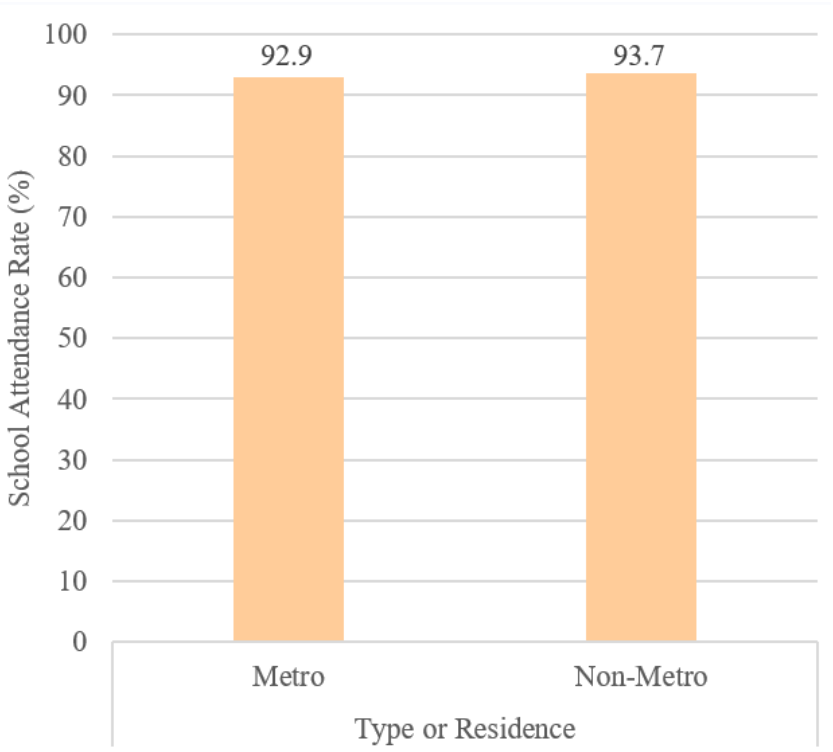

Figure 2.3: School attendance rate according to type of regional area

Figure 2.4 displays the rates of school attendance according to the age of the learner. School attendance was above 90% for learners aged 6 to 16 years, after which the attendance rate dropped down to below 60% among the older learners. This is possibly due to the stigma that is often associated with a learner being older than their peers, resulting in the learners in the older age groups simply not returning to school after dropping out for various reasons.

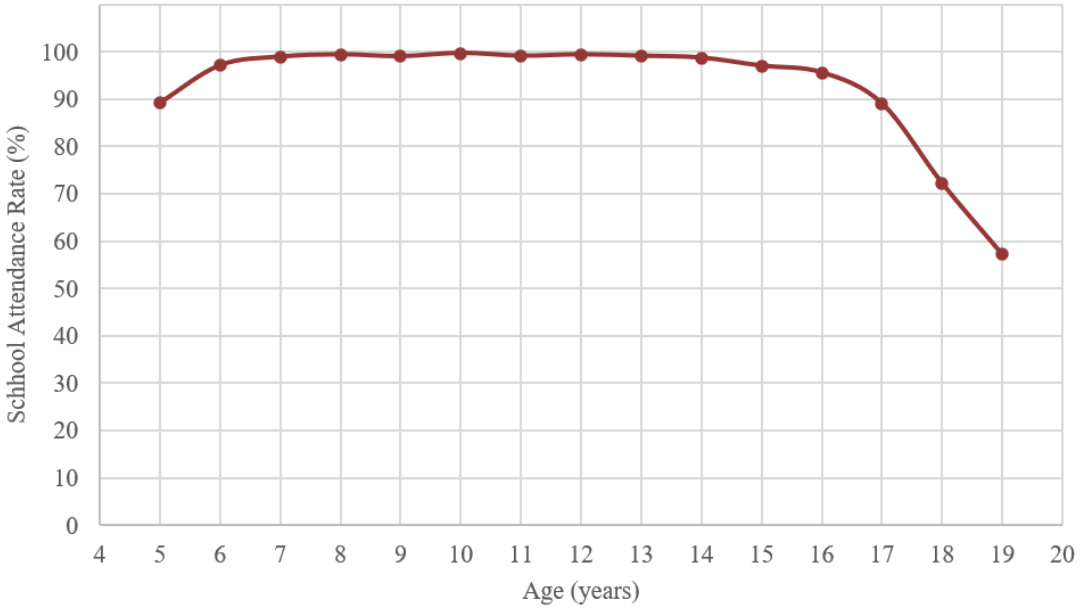

Figure 2.4: School attendance according to different age groups

Figure 2.5 below shows the attendance rate according to the learner's race group. The rates of attendance for learners within the different population groups were similar, except for those from the Coloured population group which had a substantially lower attendance rate of 86.9%.

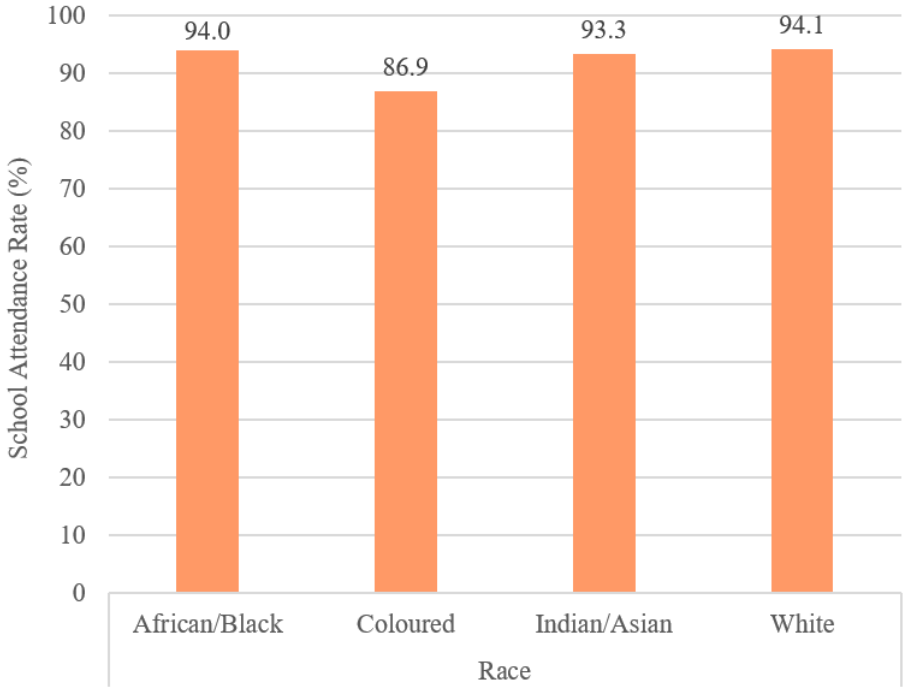

Figure 2.5: School attendance rate according to population group

The rates of school attendance according to a learner's marital status and gender are presented in Figure 2.6. While attendance was substantially lower for learners who were not single, which included those who were married, living with a partner or divorced, these learners only comprised of less than 2% of the sample. Not much difference is observed in the rate of attendance between male and female learners.

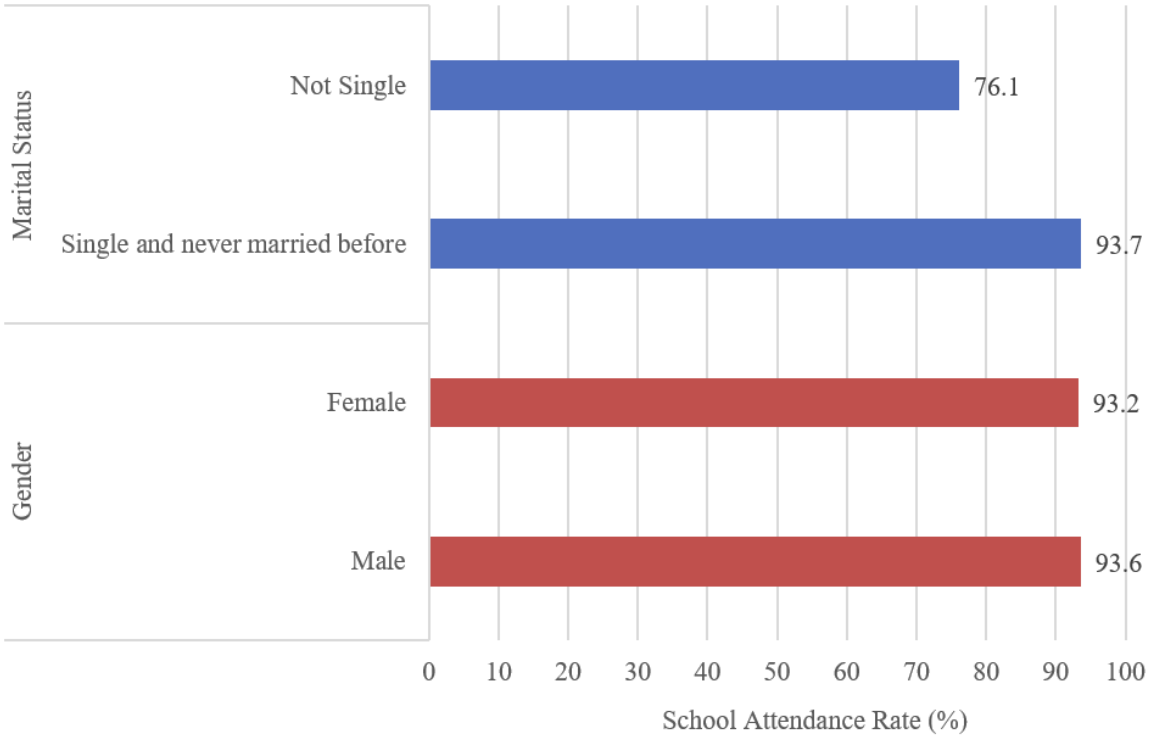

Figure 2.6: School attendance rate according marital status and gender

Figure 2.7 displays the school attendance rate according to the relationship that a learner had with the household head. Attendance was higher among learner's whose grandparent was the head of the household (95.5%). The rate of attendance was lowest among learner's whose parents or grandparents were not the head of the household. This 'other' category may also include learners who were the household head themselves. The added responsibility of heading a household may restrict a learner's ability to attend school.

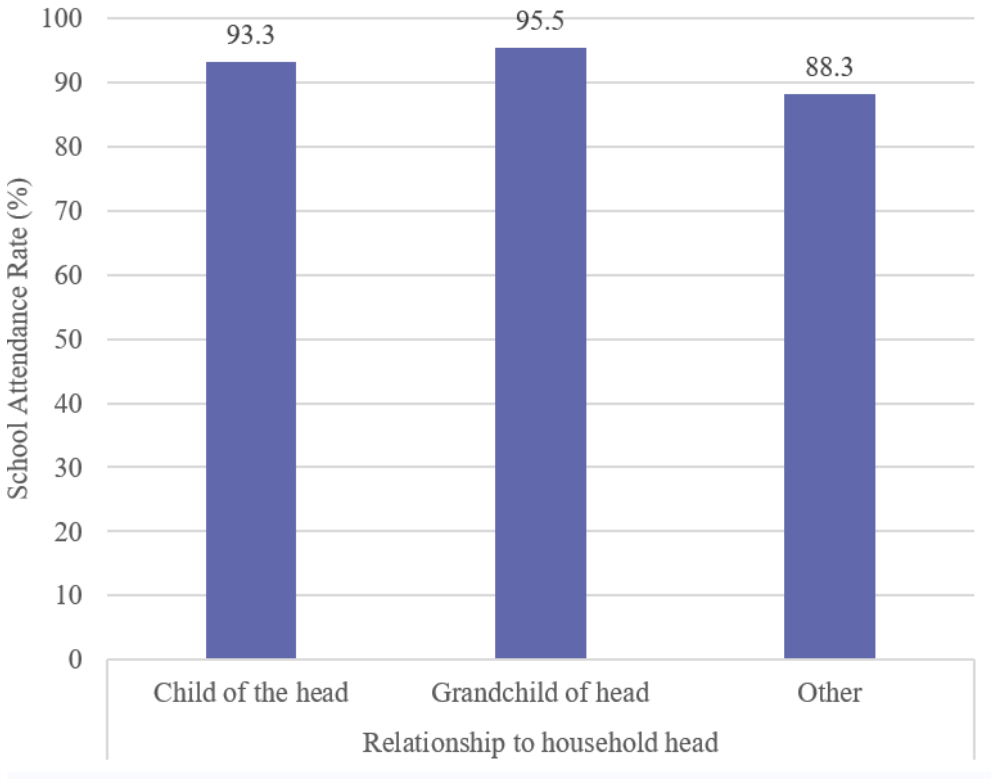

Figure 2.7: School attendance rate according to the relationship a learner shares with the household head

Figure 2.8 gives the rates of school attendance according to the survival status of each parent of the learner. The figure clearly depicts higher rates of attendance among learners whose mother or father was still alive. While the data does not give an indication of whether the surviving parent was actually present in the learner's life, we know the opposite to be true where a deceased parent was unable to be present. This has may contribute to the decline in the rate of school attendance, where the learner may be unable to attend school due to emotional reasons associated with losing a parent or due to the added responsibility required of them in the household.

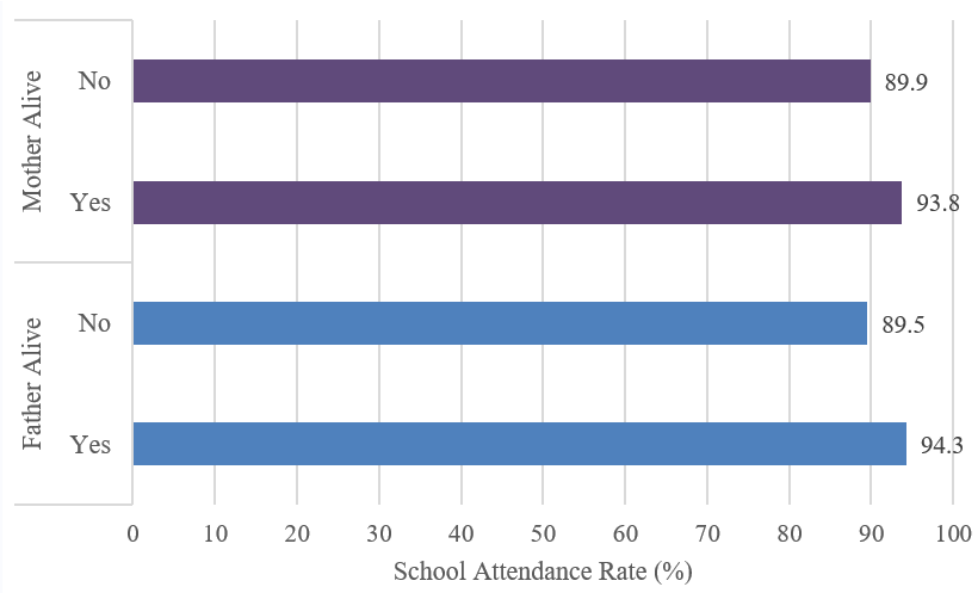

Figure 2.8: School attendance rate according to the survival status of each parent

Figure 2.9 presents the school attendance rates according to the main source of income in the learner's household as well as the household's total monthly income, which was categorized as over or under R25,000. The rates of attendance did not differ by much across any of the categories of these two factors.

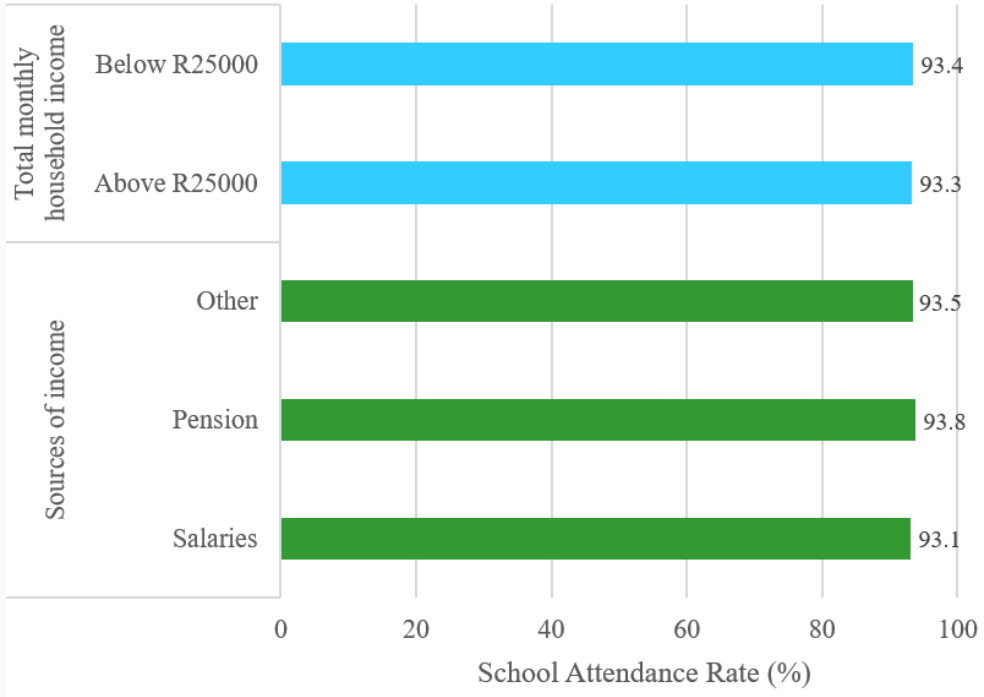

Figure 2.9: School attendance rate according to source of household income and total monthly income

Box plots for the size of the household and the household wealth index Z-score are presented in Figure 2.10 according to learners who were attending school as well as not attending school. The distribution of the household size was fairly similar for both groups of learners, however there were a few outliers in the upper end of the scale for those learners who were not attending school. Similarly, the distribution of the household wealth index Z-score was almost the same for both groups of learners, with a few outliers on the lower end of the scale for learners who were not attending school. This figure suggests that learners who do not attend school come from much poorer socio-economic backgrounds compared to those who do attend school.

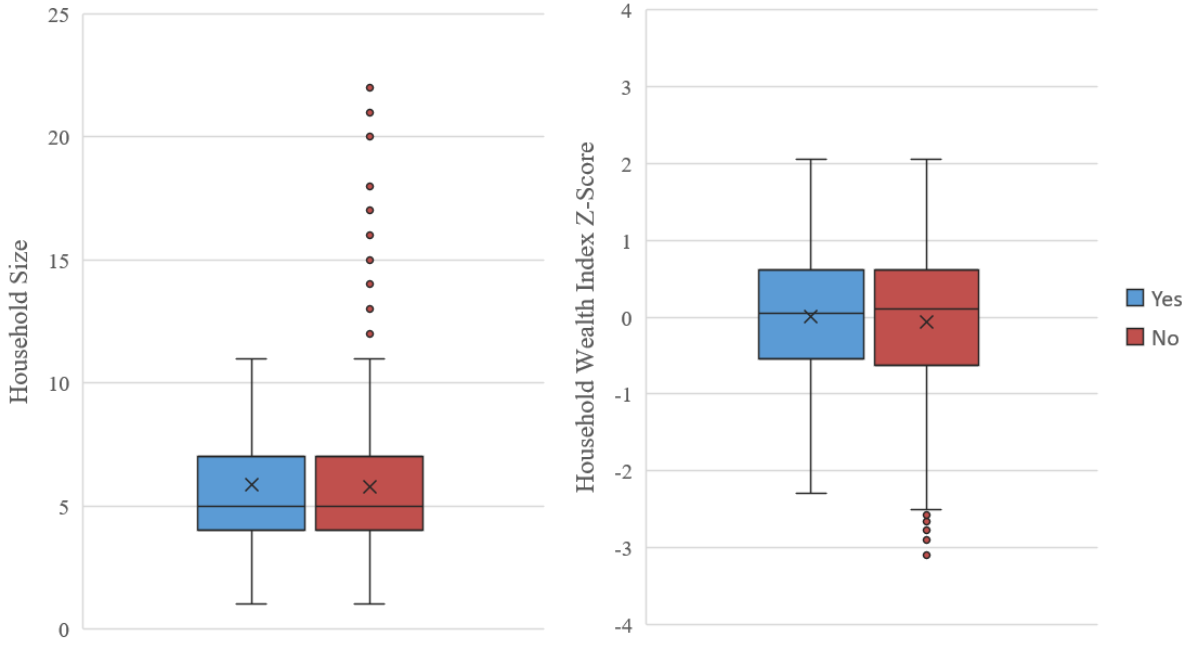

Figure 2.10: Box plots of household size and household wealth index Z-score according to school attendance

2.4 Correspondence Analysis

Correspondence analysis (CA) is a statistical technique to provide a graphical representation of cross-tabular data in the form of numerical frequencies. CA is used to gain insight into the relative relationships between and within two groups of variables, based on data given in a contingency table. The distance between category points in a plot reflects the relationships between categories, with similar categories plotted closer to each other for each variable (Greenacre, 2017).

There are existing forms of CA which includes Dual Scaling and Multiple Correspondence Analysis. Multiple correspondence analysis (MCA) allows the exploration of the patterns of relationships between categories of several categorical variables. MCA consists of joint graphical displays that produce two dual displays with row and column geometries that have similar interpretations (Greenacre, 2017). It also incorporates the diagrammatic view of the association between categories through bi-plots. The bi-plots help to visualize associations present between categories (Khangar, 2017). MCA can be considered as a type of principal component analysis for categorical variables. As the data used in this thesis consists of multiple categorical variables, MCA will be used to explore the associations between such variables.

2.4.1 Multiple Correspondence Analysis Theory

MCA is obtained by using a standard correspondence analysis on an indicator matrix (Valentin, 2007). Assume we have n respondents with k categorical variables and l_j distinct values for variable j. Then, we define an $n \times l_j$ indicator matrix $\boldsymbol{P}_j$. Concatenating the $\boldsymbol{P}_j$'s forms the $n \times l$ matrix $\boldsymbol{P}$, which is a respondents-by-categories table that has as many rows as respondents, and l is the sum of l_j (Valentin, 2007). The elements of $\boldsymbol{P}$ are ones in the positions to indicate the categories of response of each respondent and zero elsewhere. $\boldsymbol{P}$ can be divided up by its grand total nk to obtain the correspondence matrix $\boldsymbol{F} = \frac{1}{nk}\boldsymbol{P}$, which contains the relative frequencies. This

gives $\mathbf{1}_n' \boldsymbol{F} \mathbf{1}_l = 1$, where $\mathbf{1}_i$ is an $i \times 1$ vector of ones. Performing CA on the indicator matrix provides two sets of factor scores: one for the rows and one for the columns (Valentin, 2007). The vectors/factor scores obtained in the rows and columns are given by $\boldsymbol{r} = \boldsymbol{F}\mathbf{1}_l$ and $\boldsymbol{c} = \boldsymbol{F}'\mathbf{1}_n$, respectively, which are called marginals. These marginals are collectively called masses. Furthermore, these vectors are scaled so that their variance is equal to their corresponding eigenvalues (Valentin, 2007). We assume for the diagonal matrices of the masses are defined by $\boldsymbol{D}_r = diag(\boldsymbol{r})$ for the rows and $\boldsymbol{D}_c = diag(\boldsymbol{c})$ for the columns.

In the application of MCA, we first compute the probability matrix using the total $N = nk$, given by $\boldsymbol{Q} = N^{-1}\boldsymbol{P}$. The factor scores are obtained from the following singular value decomposition

$$\boldsymbol{D}_r^{-\frac{1}{2}}(\boldsymbol{Q} - \boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-\frac{1}{2}} = \boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{W}', \tag{2.1}$$

where $\boldsymbol{\Delta}$ represents the diagonal matrix of singular values, V and W are the diagonal matrices of masses for rows and columns with an exponent of negative half respectively. Moreover, the $\boldsymbol{\Upsilon} = \boldsymbol{\Delta}^2$ is the matrix of eigenvalues and the row and column factor scores are obtained respectively as follows:

$$\boldsymbol{R} = \boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{V}\boldsymbol{\Delta},$$

and

$$\boldsymbol{S} = \boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{\Delta}. \tag{2.2}$$

The above equations will then lead to the calculation of the squared distance (χ^2) of the rows and columns to their respective barycentre, which is given in the following form

$$\boldsymbol{d}_r = diag\{\boldsymbol{R}\boldsymbol{R}'\},$$

and

$$\boldsymbol{d}_c = diag\{\boldsymbol{S}\boldsymbol{S}'\}, \tag{2.3}$$

respectively. In CA, the total variance, referred to as inertia, of the factor scores is proportional to the independence Chi-square statistic of a cross-tabulation.

Squared cosines can be used to locate the factors that are important for a given observation or variable. The squared cosine between the i^{th} row and factor m and the j^{th} column and factor m are respectively given by

$$C_{i,m} = \frac{f_{i,m}^2}{d_{r,i}^2},$$

and

$$C_{j,m} = \frac{g_{j,m}^2}{d_{c,j}^2}, \tag{2.4}$$

where f and g are functions and $d_{r,i}^2$ and $d_{c,j}^2$ represent element i of $\boldsymbol{d}_r$ and element j of $\boldsymbol{d}_c$, respectively. The contributions of the i^{th} row and the j^{th} column to factor m are obtained in the following manner, respectively

$$b_{i,m} = \frac{f_{i,m}^2}{\lambda_m},$$

and

$$b_{j,m} = \frac{g_{j.m}^2}{\lambda_m}. \tag{2.5}$$

These contributions of the rows and columns help locate the observations or variables that are of importance to a given factor. Supplementary elements can be projected onto the factors using the transition formula (Valentin, 2007). Suppose we let $\boldsymbol{i}'_{sup}$ be the supplementary row and $\boldsymbol{j}_{sup}$ be the supplementary column to be projected, then the coordinates of supplementary functions $\boldsymbol{f}_{sup}$ and $\boldsymbol{g}_{sup}$ are given as

$$\boldsymbol{f}_{sup} = (\boldsymbol{i}'_{sup}\mathbf{1})\boldsymbol{i}'_{sub}\boldsymbol{S}\boldsymbol{\Delta}^{-1},$$

and

$$\boldsymbol{g}_{sup} = (\boldsymbol{j}'_{sup}\mathbf{1})\boldsymbol{j}'_{sub}\boldsymbol{R}\boldsymbol{\Delta}^{-1}, \tag{2.6}$$

respectively. CA produces the element scores for rows and columns, however MCA requires these scores to be re-scaled. The Burt matrix is the $l \times l$ table, which is obtained by $\boldsymbol{B} = \boldsymbol{P}'\boldsymbol{P}$. This matrix serves to give CA the same factors as the analysis of $\boldsymbol{P}$, however in a computationally easier way. The Burt matrix also plays an important role in providing eigenvalues which provide a better approximation of the of inertia described by the factors than the ones of $\boldsymbol{P}$.

2.4.2 Eigenvalue correction for MCA

MCA codes data by creating several binary columns for each variable with the constraint that one and only one of the columns gets the value of 1 (Valentin, 2007). This method of coding creates additional dimensions as one nominal variable is coded with multiple columns. The additional dimensions cause the solution space variance to be expanded, which causes the inertia described by the first dimension to be under-scaled. That leads to the variation produced by the first dimension to be under-estimated. However, this under-scaling can be corrected using correction formulas Valentin (2007). These formulas take into account that all the eigenvalues that are smaller than $\frac{1}{k}$ are accounting for the extra dimensions and that MCA is equivalent to the analysis of the Burt matrix, which has eigenvalues that are equal to the square of the eigenvalues from the analysis of $\boldsymbol{P}$ (Valentin, 2007). That is, if we denote λ_m as the eigenvalues found from the analysis $\boldsymbol{P}$, then the rectified eigenvalues are given by

$${}_c\lambda_m = \begin{cases} \left[\left(\frac{k}{k-1}\right)\left(\lambda_m - \frac{1}{k}\right)\right]^2 & \text{if } \lambda_m > \frac{1}{k}, \\ 0 & \text{if } \lambda_m \leqslant \frac{1}{k}, \end{cases} \tag{2.7}$$

where ${}_c\lambda_m$ represents the rectified eigenvalues for factor m. The above rectification allows for better estimations of the inertia. The percentages of inertia can be computed by dividing an eigenvalue by the sum of all of the eigenvalues. However, a

suggestion for a better estimate of inertia is provided in Greenacre (1993), which allows for evaluation of the percentage of inertia respective to the average inertia of the off-diagonal blocks of Burt matrix (Valentin, 2007). This average inertia (γ) can be calculated as follows

$$\gamma = \frac{k}{k-1} \times \left(\sum_m \lambda_m^2 - \frac{l-k}{k} \right)^2 \tag{2.8}$$

Based on this approach, the percentage of inertia would be obtained by

$$\tau_c = \frac{{}_c\lambda}{\gamma} \text{ instead of } \frac{{}_c\lambda}{\sum\limits_m {}_c\lambda_m} \tag{2.9}$$

2.4.3 Application of MCA to school attendance data

In this section, MCA is applied to the SAGHS data. All the categorical variables were incorporated into the analysis, namely school attendance, province, type of place of residence, race, gender, marital status, relationship to the household head, the survival status of each parent, the main source of household income and the total monthly household income. MCA locates all of the categories of the variables in a Euclidean space.

Figure 2.11 displays the results of the inertia and chi-squared decomposition for MCA applied to the data. A total of 22 dimensions explain the full variation of the data. The result suggests that dimension 1 accounts for 10.6% of the variation, while dimension 2 accounts for 7.05% of the variation, which is 17.66% of the entire variation. These inertiae are relatively low and thus indicate possible instability in the individual axes. The first column of results in Figure 2.11 presents the singular value for each dimension, which indicates the relative importance of each dimension in explaining the inertia, or proportion of variation. The singular values can be considered as the correlation between the rows and columns of the contingency table.

Figure 2.12 displays the scree plot of singular values for each of the dimensions. The variation decreases faster for dimensions 1 to 2, then decreases less rapidly for dimensions beyond dimension 2. The first two dimensions of the space are plotted to examine the associations among the categories. This plot is presented in Figure 2.13.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.46049	0.21205	57372	10.60	10.60
0.37563	0.14110	38175	7.05	17.66
0.36701	0.13470	36444	6.73	24.39
0.34471	0.11883	32149	5.94	30.33
0.31699	0.10048	27186	5.02	35.36
0.31416	0.09870	26703	4.93	40.29
0.31226	0.09751	26382	4.88	45.17
0.30522	0.09316	25206	4.66	49.83
0.30360	0.09217	24939	4.61	54.43
0.30157	0.09094	24606	4.55	58.98
0.29984	0.08990	24324	4.50	63.48
0.29858	0.08915	24121	4.46	67.93
0.29758	0.08856	23960	4.43	72.36
0.28714	0.08245	22308	4.12	76.48
0.28562	0.08158	22072	4.08	80.56
0.28040	0.07862	21272	3.93	84.49
0.26481	0.07012	18972	3.51	88.00
0.24861	0.06181	16722	3.09	91.09
0.23833	0.05680	15368	2.84	93.93
0.23178	0.05372	14535	2.69	96.62
0.19699	0.03880	10499	1.94	98.56
0.16985	0.02885	7806	1.44	100.00
	2.00000	541121	100.00	

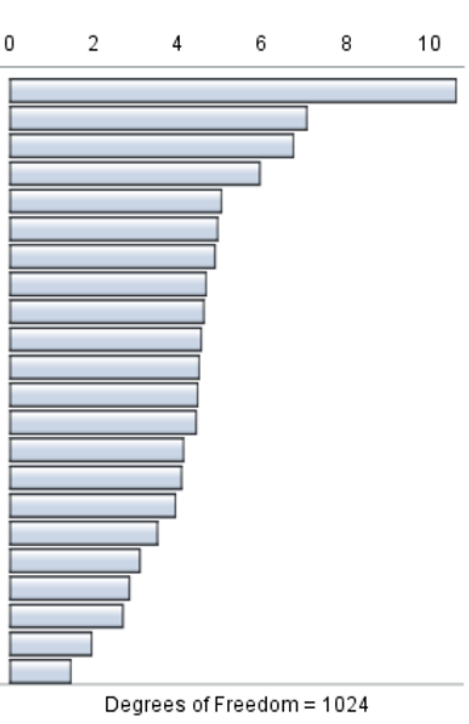

Figure 2.11: The table of inertia and singular value decomposition

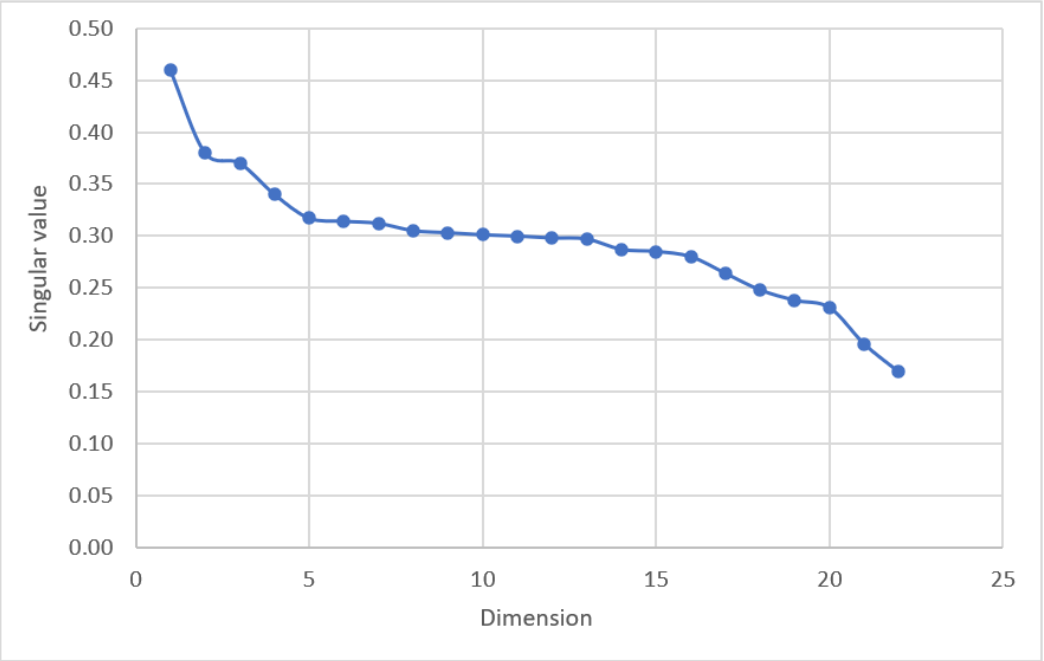

Figure 2.12: Scree plot of singular values

From Figure 2.13 of the first two dimensions, it is observed that the categories in the lower hemisphere, given by the Northern Cape, Western Cape and the Coloured race group, are grouped close together and thus are associated. In addition, these categories are closer to non-attendance in school compared to attendance. Similarly, the categories in the right side of the upper hemisphere, namely the Indian and White race groups, Gauteng, and above R2500 for total household income, can be considered as associated, and closer to school attendance compared to non-attendance. The categories of father not alive and mother not alive appear to be close to and thus associated with other relationship to the head of the household (other than being the child or grandchild of the head of the household).

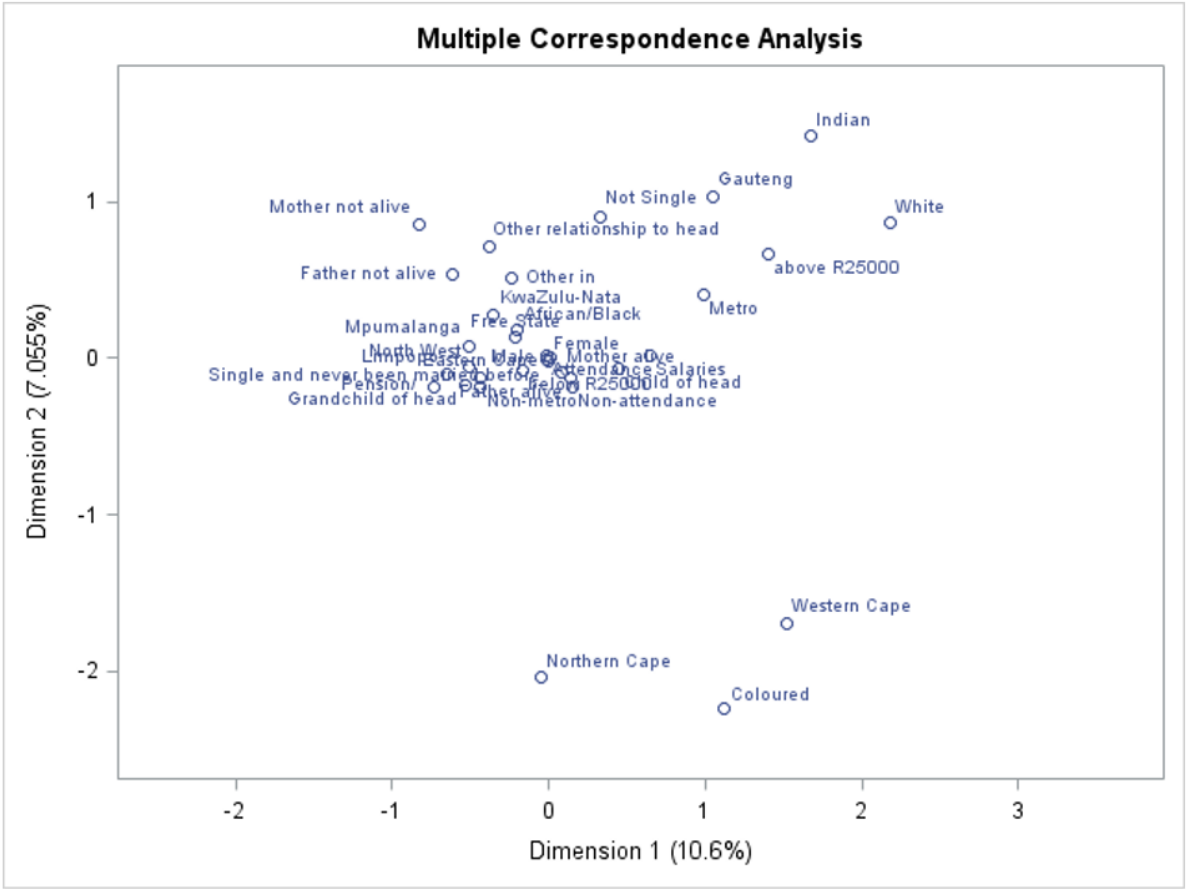

Figure 2.13: Multiple correspondence analysis plot for the first two dimensions

2.4.4 Summary

This chapter introduced the data used in this thesis as well as the variables of interest. These variables were explored in relation to the response, namely school attendance. The school attendance rate among the surveyed learners was fairly high at 93.4%. However, the fact that there are still a number of learners without a matric who are not attending school is alarming considering all the efforts that the South African government has made in making basic education accessible to everyone. The rates of school attendance also differed substantially across the different provinces, where the western part of South Africa experienced the lowest rates of school attendance. The rate of school attendance declined dramatically among learners of the older age groups (16 to 19 years of age), and was lowest among the

Coloured race group. Very little difference in the rate of school attendance was seen between male and female learners, and the rate of school attendance was lowest among learners who had a deceased parent. Learners of a low socio-economic background tended not to attend school.

The SAGHS data was collected based on a complex survey design. This means that statistical models that assume the data was collected based on simple random sampling are not appropriate for the analysis of the SAGHS data. The next chapters consider different statistical approaches to deal with such data.

Chapter 3

Survey Logistic Regression Models

As the response variable of interest in this study is binary, indicating whether or not the learner was attending school at the time of the survey, we need to make use of an appropriate statistical model for this type of response. A general linear model assumes that the response variable is continuous and follows a normal distribution. Thus, it is not suitable in our case. A possible method for modelling a non-normal binary response is via a generalized linear model (GLM). This chapter presents an overview of the GLM. In addition, an extension of the GLM to model data from complex survey designs is presented, namely the survey logistic regression (SLR) model. The results of the SLR model applied to the SAGHS data are also presented in this chapter.

3.1 GLM model

The GLM consists of three components:

- **A Random Component**: This consists of the independent response variable Y_i.
- **A Systematic component**: This component involves the linear predictor, η_i,

which is related to a set of explanatory variables

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \\ &= \boldsymbol{x}_i' \boldsymbol{\beta},\end{aligned}$$

where $\boldsymbol{x}_i = (1, x_{1i}, ..., x_{pi})'$ is a $(p+1)$-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is a vector of the unknown regression coefficients. The distribution of Y_i depends on $\boldsymbol{x}_i$ through this linear predictor.

- **A link function**: This component is a monotonic and differentiable function, g, that links the mean of the response response, $\mu_i = E(y_i)$, to the linear predictor, η_i, as follows

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

If Y_i for $i = 1, \ldots, n$ is a response variable from a distribution that is a member of the exponential family, then the probability density function for Y_i is given by

$$f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{3.1}$$

where θ_i is the canonical parameter and $b(\theta_i)$, $a_i(\phi)$ and $c(y_i, \phi)$ are known functions. If ϕ is known, this is an exponential family model with a canonical link θ. It may or may not be a two-parameter exponential family if ϕ is unknown. It can also be easily shown that if the distribution belongs to an exponential distribution, it has a mean and variance of the form

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

$$Var(Y_i) = a_i(\phi) b''(\theta_i) = a_i(\phi) v(\mu_i),$$

where the $b'(\theta_i)$ and $b''(\theta_i) = v(\mu_i)$ are the 1^{st} derivative and 2^{nd} derivatives of the function $b(\theta_i)$ with respect to θ_i, respectively. The GLM has a property of a variance that can vary across the responses (non-constant variance). This is controlled by the function $a_i(\phi)$. When it is greater than 1, the model is over-dispersed, when it

is less than 1, then model is under-dispersed (Nelder & Wedderburn, 1972). The Poisson, Binomial, Chi-Square and Gamma distributions are examples that belong to the exponential family.

3.1.1 Parameter Estimation

The method of maximum likelihood (ML) is used for the parameter estimation of GLMs, where the log-likelihood function for a single observation is given by

$$\ell_i = \ln\ f(y_i, \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \tag{3.2}$$

As $Y_i, i = 1, ..., n$, are independent, the joint log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i. \tag{3.3}$$

The ML estimation of parameters $\beta_j, j = 0, ..., p$ is the solution to the equation

$$\frac{\partial \ell_i}{\partial \beta_j} = 0. \tag{3.4}$$

In order to obtain a solution for the above equation, the chain rule is applied as follows

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{3.5}$$

Using Equation 3.2, the following is obtained

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}. \tag{3.6}$$

As $\mu_i = b'(\theta_i),\ Var(Y_i) = a_i(\phi)v(\mu_i),$ and $\eta_i = \sum_j \beta_j x_{ij}$, we get

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = v(\mu_i),$$

and

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

This leads to

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{v(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij},\end{aligned}$$

where W_i is the iterative weight given by

$$\begin{aligned}W_i &= \frac{1}{a_i(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 v_i^{-1} \\ &= \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,\end{aligned} \tag{3.7}$$

where $v_i = v(\mu_i)$ is the variance function. As $\eta_i = g(\mu_i)$, $\dfrac{\partial \mu_i}{\partial \eta_i}$ depends on the link function of the model.

Therefore, the ML estimate for $\boldsymbol{\beta}$ are obtained by solving the following equation

$$\sum_{i=1}^{n} (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij} = 0. \tag{3.8}$$

Equation 3.8 is a non-linear function of $\boldsymbol{\beta}$, therefore iterative procedures such as Newton Raphson and Fisher Score are required to solve this equation. A basic overview of these procedures is given below.

Newton Raphson

The Newton Raphson iterative equation is given by

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1} \boldsymbol{U}^{(t)}, \tag{3.9}$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the approximation of $\boldsymbol{\theta}$ at the t^{th} iteration. $\boldsymbol{U}^{(t)} = \dfrac{\partial \ell_p}{\partial \boldsymbol{\theta}}$ evaluated at $\hat{\boldsymbol{\theta}}^{(t)}$, where $\boldsymbol{U}$ is called the score. $\boldsymbol{H}^{(t)}$ is the Hessian matrix, $\boldsymbol{H}$, with the following elements evaluated at $\hat{\boldsymbol{\theta}}^{(t)}$:

$$\boldsymbol{H}_{jk} = \frac{\partial^2 \ell_p}{\partial \theta_j \partial \theta_k}. \tag{3.10}$$

Fisher Score

The Fisher score iterative equation is given by

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.11}$$

where $\boldsymbol{\mathcal{I}} = -E(\boldsymbol{H})$ is known as the information matrix. Both iterative procedures require an appropriate starting value, $\hat{\boldsymbol{\theta}}^{(0)}$, after which the process will continue until the difference between the successive approximations is very small. Thus, when the algorithm converges.

Applying Newton Raphson to Equation 3.8 to obtain the parameter estimates for the GLM, it follows that the iterative equation is given by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.12}$$

and the iterative Fisher Score equation given by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.13}$$

with information matrix

$$\begin{aligned}\boldsymbol{\mathcal{I}} &= -E(\boldsymbol{H}) \\ &= -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) \\ &= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X},\end{aligned} \tag{3.14}$$

where $\boldsymbol{W}$ is a weight matrix with diagonal elements equal to that given in Equation 3.7. Equation 3.13 can be written as

$$\begin{aligned}\boldsymbol{\mathcal{I}}^{(t)}\hat{\boldsymbol{\beta}}^{(t+1)} &= \boldsymbol{\mathcal{I}}^{(t)}\hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{U}^{(t)} \\ &= \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)},\end{aligned} \tag{3.15}$$

where $\boldsymbol{W}^{(t)}$ is the weight matrix evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$, and $\boldsymbol{z}^{(t)}$ has the following elements evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right). \tag{3.16}$$

z_i is known as the adjusted dependent variable or the working variable. Thus, the following is obtained

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}. \tag{3.17}$$

Each iteration step is as a result of a weighted least squares regression of z_i on the independent variables x_i, with weight W_i. It follows that the asymptotic variance of the estimate of $\boldsymbol{\beta}$ is the inverse of the information matrix from Equation 3.14 and can be estimated by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{3.18}$$

where $\widehat{\boldsymbol{W}}$ is $\boldsymbol{W}$ evaluated at $\hat{\boldsymbol{\beta}}$ and depends on the link function of the model. The dispersion parameter, ϕ, in function $a_i(\phi)$ that is used in the calculation of W_i re-

solves to 0. The estimate of $\boldsymbol{\beta}$ is therefore the same under any value of ϕ. However, the value of ϕ is necessary for the calculation of $\widehat{Var}(\hat{\boldsymbol{\beta}})$. Thus, when ϕ is unknown, it can be estimated using a moment estimator (Nelder & Wedderburn, 1972), given by

$$\hat{\phi} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{\omega_i (y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{3.19}$$

where ω_i is the weight defined in Equation 3.1.

3.1.2 Goodness-of-fit

The goodness-of-fit measures how discrepant the estimated values are to the observations. The discrepancy is regarded as the random error occurring during the estimation of parameter values in the model. The estimated values are never equal to those observed, thus, the question arises of how discrepant they are, as the smaller discrepancy is negligible, however, the larger discrepancy is not. The measures of discrepancy may be formed in various ways, although our concern will be primarily based on the ones formed from the logarithm of a ratio of likelihoods, which is called the deviance (McCulloch et al., 2001).

Assume for a fitted model, there are $p+1$ parameters and $\ell(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y})$ is the log-likelihood function maximized based on $\hat{\boldsymbol{\beta}}$ for a fixed value of the dispersion parameter ϕ, and $\ell(\boldsymbol{y}, \phi, \boldsymbol{y})$ is the maximum log-likelihood obtained under the saturated model where the number of parameters equals the number of observations. Then, the scaled deviance is

$$D^s = \frac{-2\left[\ell(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})\right]}{\phi}. \tag{3.20}$$

If $\phi = 1$, the deviance reduces to

$$D^s = -2\left[\ell(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})\right]. \tag{3.21}$$

The scaled deviance converges asymptotically to a χ^2 distribution with $n - p - 1$ degrees of freedom. Thus, the fitted model is rejected if the calculated deviance is greater that or equal to $\chi^2_{n-p-1;\alpha}$, using an α level of significance. The *Generalized Pearson's Chi-Square* is another commonly used measure of goodness-of-fit. This is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{3.22}$$

which asymptotically follows a χ^2 distribution with $n - p - 1$ degrees of freedom, where $v(\hat{\mu}_i)$ is the estimated variance function for the distribution. As with the deviance, the smaller the value of the χ^2 statistic, the better the fitting the model (Nelder & Wedderburn, 1972).

3.1.3 Likelihood Ratio Test

In order to test if a particular predictor variable does not have a significant effect on the response variable (which means their corresponding regression parameter is equal to zero), while controlling for the other predictor variables in the model, the deviances of the full model and the reduced model can be compared. The test statistic is derived by considering the following

$$D_{reduced} - D_{full}. \tag{3.23}$$

As both deviances above involve the log-likelihood for the saturated model, the test statistic is formed as follows

$$\chi^2 = -2\left[\text{log-likelihood}(reduced\,model) - \text{log-likelihood}(full\,model)\right], \tag{3.24}$$

which has an asymptotic χ^2 distribution with a degrees of freedom equal to the difference in the number of parameters between the full model and the reduced model. This test is known as the *Likelihood Ratio Test*.

3.1.4 Wald Test

The Wald test is a hypothesis test for parameters that have been estimated by the maximum likelihood. This test allows a hypothesis test on a single parameter β_j to be performed. The test statistic for this test is

$$z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}, \tag{3.25}$$

where the standard error of β_j is the square root of the diagonal elements in the inverse of the information matrix given in Equation 3.14. The null hypothesis for the Wald test is H_0: $\beta_j = 0$, which infers the variable has no effect on the response. Thus, the null hypothesis is rejected for large values of the test statistic, suggesting that the corresponding variable is significant to the model and thus has a significant effect on the response.

3.2 Quasi-Likelihood function

The main purpose of many analyses is to show how the mean response is affected by several covariates. In estimation of parameters for a model using maximum likelihood, the distribution of the response is required. However, sometimes there is insufficient information about the data for us to specify the model and in that sense, the parameters are estimated using quasi-likelihood estimation (QL), where only the relationship between the mean and the variance of the observations need to be specified (Elder, 1996).

The following relation is defined to determine the quasi-likelihood (specifically quasi-log-likelihood) function $Q(y_i; \mu_i)$ for each observation

$$\frac{\partial Q(y_i; \mu_i)}{\partial \mu_i} = \frac{\omega_i (y_i - \mu_i)}{\phi \, v(\mu_i)}, \tag{3.26}$$

where ω_i is the known weight associated with observation y_i. Thus, from Equation

3.26, the following is obtained

$$Q(y_i;\mu_i) = \int_{y_i}^{\mu_i} \frac{\omega_i(y_i - t)}{\phi\, v(\mu_i)} dt + \text{some function of } y_i. \tag{3.27}$$

Therefore, using the Fisher Scoring iterative procedure, the maximum quasi-likelihood estimates of $\boldsymbol{\beta}$ can be obtained. Equation 3.19 can be used to estimate ϕ.

The QL above requires observations to be independent. However, it can be extended for observations that are correlated. The properties of the quasi-likelihood function are similar to that of the ordinary log-likelihood function. Therefore, the goodness-of-fit and hypothesis tests that were previously discussed are still valid for this method. When the distribution of the response comes from an exponential family, the log-likelihood function and the quasi-log-likelihood function are identical (Wedderburn, 1974).

3.3 Logistic regression

The logistic regression model is a special case of a generalized linear model for a model binary response. In this study, the response variable, given by school attendance is binary, which can be coded as follows:

$$Y_i = \begin{cases} 1 & \text{the learner attends school (success),} \\ 0 & \text{if the learner does not attend school (failure).} \end{cases} \tag{3.28}$$

Therefore, Y_i follows a Bernoulli distribution where $P(Y_i = 1) = \pi_i$ is the probability of success and $P(Y_i = 0) = 1 - \pi_i$ is the probability of failure. This means that

$$E(Y_i) = \pi_i, \tag{3.29}$$

and

$$Var(Y_i) = \pi_i(1 - \pi_i). \tag{3.30}$$

Since π_i is a probability, it is limited to $0 \leq \pi_i \leq 1$. Thus, a model for $E(Y_i)$ that restircts its value between 0 and 1 needs to be used. The logistic regression model is such a model, and is given by

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \qquad i = 1, 2, ..., n, \tag{3.31}$$

where $\boldsymbol{x}_i'$ is the vector of explanatory variables corresponding to the i^{th} observation and $\boldsymbol{\beta}$ is a vector of unknown parameters (Alan Agresti, 2007). The left hand side of Equation 3.31 is referred to as the logit link, denoted by η_i in the GLM. Thus, Equation 3.29 will become

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}. \tag{3.32}$$

This is known as the logistic regression model which is a class of the GLM with a logit link. The value of the link, η_i, is allowed to range freely while restricting that of $E(Y_i) = \pi_i = \mu_i$ to between 0 and 1. The maximum likelihood estimates of $\boldsymbol{\beta}$ can be found using the iterative equations discussed previously.

3.4 Survey logistic regression

Logistic regression is a method to analyse the relationship between a binary response and a set of covariates. However, this method does not account for study design (SAS Institute, 2004). In addition, logistic regression assumes that the observations were obtained based on simple random sampling. Therefore, in the case of survey data that was colected based on a stratified, multi-stage approach, logistic regression is no longer appropriate as failure to account for the survey design can result in overestimation of standard errors which can lead to incorrect results (Heeringa et al., 2010). However, the classical logistic regression model can be extended to

incorporate th survey design in order to make valid inferences. This results in the survey logistic regression model (SLR), which is a design-based statistical approach (Heeringa et al., 2010).

Survey logistic regression uses unique methods in estimating parameters, and corresponding variances (Heeringa et al., 2010). The most common methods of variance estimation are the Tylor series approximation which is based on linearisation, Jackknife repeated replication, and balanced repeated replication. Previous studies have shown that there is not one of these estimation procedures that is better than the other. However, it depends on the design of the model and information provided in the data (Heeringa et al., 2010). Lastly, linearisation and re-sampling techniques have been proven to be asymptotically equivalent, meaning that they produce approximately the same results.

3.4.1 The model

Consider the survey logistic regression model for a binary response variable $Y_{hij}, j = 1, ..., n_{hi}; i = 1, ..., n_h; h = 1, ..., H$, which equals 1 if the j^{th} learner in the i^{th} household, within the h^{th} cluster, attends school and 0 otherwise. Also, let $\pi_{hij} = P(Y_{hij}) = 1$ be the probability that this learner attends school. Then the survey logistic regression model is given by

$$logit(\pi_{hij}) = \boldsymbol{x}'_{hij}\boldsymbol{\beta}, \tag{3.33}$$

with

$$\pi_{hij} = \frac{\exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})}, \tag{3.34}$$

where $\boldsymbol{x}_{hij}$ is the row of the design matrix corresponding to the response of the j^{th} learner in the i^{th} household within the h^{th} cluster, and $\boldsymbol{\beta}$ is the vector of unknown regression coefficients to be estimated.

Maximum likelihood estimation used to estimate parameters of an ordinary logistic regression model is not appropriate in this case as the calculation of standard errors of parameter estimates can be complicated for data obtained from complex survey designs (Vittinghoff et al., 2011). The survey logistic regression model has the same form as the ordinary logistic regression model, where the probability distribution of the response variable is given by

$$f(y_{hij}) = \pi_{hij}^{y_{hij}}(1-\pi_{hij})^{1-y_{hij}}. \tag{3.35}$$

The leads to

$$E(Y_{hij}) = \pi_{hij} = \frac{e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}, \tag{3.36}$$

and

$$Var(Y_{hij}) = \pi_{hij}(1-\pi_{hij}) = \frac{e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}})^2}. \tag{3.37}$$

Therefore, the log-likelihood function is then given by

$$\ell = \ln L(y) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{h_i}} \ln f(y_{hij}). \tag{3.38}$$

This log-likelihood function does not incorporate the sampling weights, which means that the maximum likelihood estimates of the model's parameters based on this function will only be valid for simple random sampling where observations are unweighted. Incorporating the sampling weights into the log-likelihood function leads to a pseudo-likelihood function and the method of estimation that uses this function is called the pseudo-maximum likelihood (PML) method.

3.4.2 Pseudo-likelihood estimation

The Pseudo-likelihood estimation method works similarly to the maximum likelihood estimation approach where it requires information about the distribution of the response variable, however, it takes into consideration the sampling weights in

the following manner

$$P\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{h_i}} \omega_{hij} \ln f(y_{hij}), \tag{3.39}$$

where $P\ell$ represents the pseudo-log likelihood function and w_{hij} the weight of observation y_{hij}. Thus, based on the probability distribution of the response for the survey logistic logistic regression model, it follows that the Pseudo-log likelihood function is given by

$$P\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{h_i}} \omega_{hij}[y_{hij}\ln(\pi_{hij}) + (1-y_{hij})\ln(1-\pi_{hij})].$$

The above equation is maximized with respect to $\boldsymbol{\beta}$ in order to obtain the model's parameter estimates. It can be shown that this process leads to the following set of estimating equations

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{h_i}} w_{hij}(y_{hij}-\pi_{hij})\boldsymbol{x}'_{hij} = \boldsymbol{0}. \tag{3.40}$$

These equations are non-linear functions of $\boldsymbol{\beta}$ and thus iterative procedures are required. We will once again consider Newton-Raphson and Fisher Scoring iterative procedures. Incorporating the above equation into the Newton iterative Equation 3.12, the following is obtained

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{S}(\hat{\boldsymbol{\beta}})^{(t)}, \tag{3.41}$$

where $\boldsymbol{S}(\hat{\boldsymbol{\beta}})$ is Equation 3.40 evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$ and $\boldsymbol{H}^{(t)}$ is the Hessian matrix, $\boldsymbol{H}$, evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$, such that

$$\boldsymbol{H} = \frac{\partial^2 P\ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}. \tag{3.42}$$

Furthermore, incorporating Equation 3.40 into the Fisher Scoring iterative Equation 3.13, we obtain the following

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{I}^{(t)})^{-1}\boldsymbol{S}(\hat{\boldsymbol{\beta}})^{(t)}. \tag{3.43}$$

It can be shown that parameter estimates obtained using this PML estimation method are consistent (Heeringa et al., 2010). While there are various methods of variance estimation for the survey logistic regression model, only the Taylor series approximation method will be considered in this thesis.

3.4.3 Tylor series approximation

The estimated variances of the Pseudo Maximum Likelihood parameter estimates are no longer equal to the inverse of the information matrix, as is the case of the unweighted generalized linear model discussed in Section 3.1. Thus, Equation 3.18 does not hold for $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ when weighting is involved. A Taylor series approximation to obtain the variance estimates was proposed by Binder (1983), as follows.

As the parameter estimates, $\widehat{\boldsymbol{\beta}}$, are defined by $\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = 0$, the first order Taylor series expansion of $\boldsymbol{S}(\widehat{\boldsymbol{\beta}})$ at $\widehat{\boldsymbol{\beta}}$ equal to its population parameter value, $\boldsymbol{\beta}$, is given by

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) \simeq \boldsymbol{S}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta}). \tag{3.44}$$

This leads to

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{S}(\boldsymbol{\beta}) \simeq \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta}). \tag{3.45}$$

By taking the variance of both sides of the above equation, the variance approximation of $\boldsymbol{S}(\widehat{\boldsymbol{\beta}})$ can be expressed as

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] Var(\widehat{\boldsymbol{\beta}}) \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]'. \tag{3.46}$$

This can be written as

$$Var(\widehat{\boldsymbol{\beta}}) = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right]\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1}. \tag{3.47}$$

This then leads to the following sandwich type variance estimator

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left[\mathcal{I}(\widehat{\boldsymbol{\beta}})\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right]\left[\mathcal{I}(\widehat{\boldsymbol{\beta}})\right]^{-1}, \tag{3.48}$$

where $\mathcal{I}(\widehat{\boldsymbol{\beta}}) = \dfrac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \dfrac{\partial^2 P\ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}$ is the information matrix evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and $Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right]$ denotes the variance-covariance matrix for the $p+1$ estimating equations. As each estimating equation is simply a sample total of the individual scores for the n observations, standard formulae can be used to estimate the variances and covariances (Heeringa et al., 2010).

3.4.4 Assessing the model

Goodness-of-Fit

The goodness-of-fit measures discussed for the GLM in Section 3.1.2 are based on data obtained by simple random sampling. Thus, different methods are required to assess the goodness-of-fit of a survey logistic regression model, such as an extension to the Hosmer-Lemeshow goodness-of-fit-test (Hosmer & Lemeshow, 1980). The Hosmer-Lemeshow goodness-of-fit-test involves partitioning observations into g (where g is preferably 10) equal sized groups based on their ordered estimated probabilities, $\hat{\pi}_i$. The Hosmer-Lemeshow test statistic, which has a Chi-square distribution with 8 degrees of freedom, is given by

$$\chi^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - \frac{E_j}{n_j})}, \tag{3.49}$$

where

n_j = number of observations in the j^{th} group,

$O_j = \sum_i y_i$ = observed number of cases in the j^{th} group,

$E_j = \sum_i = \hat{\pi}_i$ = expected number of cases in the j^{th} group.

An extension of this Hosmer-Lemeshow goodness-of-fit test for the SLR model was proposed Archer & Lemeshow (2006). It is called the F-adjusted mean residual test, also referred to as the Archer and Lemeshow goodness-of-fit test, and is estimated as follows.

$$\hat{r}_{ij} = y_{ij} - \hat{\pi}(x_{ij}). \tag{3.50}$$

Then using a grouping strategy based on that proposed by Graubard et al. (1997), the observations are grouped into deciles of risk based on their estimated probabilities and sampling weights (Archer & Lemeshow, 2006). The mean residuals by decile of risk is $\widehat{\boldsymbol{M}}' = (\widehat{M}_1, \widehat{M}_2, ..., \widehat{M}_{10})$ where

$$\widehat{M}_g = \frac{\sum_i \sum_j w_{ij} \hat{r}_{ij}}{\sum_i \sum_j w_{ij}} \qquad g = 1, 2, ..., 10, \tag{3.51}$$

where w_{ij} denotes the sampling weight associated with observation y_{ij}.

The Wald test statistic for testing g categories is given by

$$\widehat{W} = \widehat{\boldsymbol{M}}' \left[\widehat{Var}(\widehat{\boldsymbol{M}}) \right]^{-1} \widehat{\boldsymbol{M}}, \tag{3.52}$$

where $\widehat{Var}(\widehat{\boldsymbol{M}})$ denotes the variance-covariance matrix of $\widehat{\boldsymbol{M}}$, which can be obtained using the Taylor series approximation discussed in Sub-section 3.4.3 (Archer et al., 2007). This test statistic has a Chi-square distribution with $g - 1$ degrees of freedom. However, this Chi-square distribution is known to be an inappropriate appropriate reference distribution. Rather, an F-corrected Wald statistic is used instead, which is given by

$$F = \frac{(f - g + 2)}{fg} W. \tag{3.53}$$

This test statistic follows an approximate F-distribution with $g-1$ and $f-g+2$ numerator and denominator degrees of freedom, respectively. f is the difference between the number of sampled clusters and the number of strata, and g denotes the number of categories (Archer & Lemeshow, 2006). This leads to the F-adjusted mean residual test statistic as follows

$$\widehat{Q}_m = \frac{(f-8)}{10f}\widehat{\boldsymbol{M}'}\left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1}\widehat{\boldsymbol{M}}, \tag{3.54}$$

with $g = 10$ deciles of risk.

In addition to the above, information criteria such as Akaike's Information Criteria (AIC) and Schwarz Criterion (SC) can be used to compare the goodness-of-fit of two nested models.

Testing Model Parameters

The Survey Logistic regression model parameters are estimated using Pseudo-likelihood method, which is an estimate of the true likelihood, thus, inferences about parameters cannot be based on the likelihood ratio test (Hosmer Jr et al., 2013). The Wald test is a more appropriate test, which consist of a null hypothesis in the form H_0 : $\boldsymbol{C\beta} = \boldsymbol{0}$, where $\boldsymbol{C}$ is the matrix of constants that defines the hypothesis being tested. The statistic is given by

$$W = (\boldsymbol{C}\widehat{\boldsymbol{\beta}})'\left[\boldsymbol{C}\,\widehat{Var}(\widehat{\boldsymbol{\beta}})\,\boldsymbol{C}'\right]^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\beta}}), \tag{3.55}$$

where $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$. This test statistic follows a Chi-square distribution with q degrees of freedom, where q is the rank of matrix $\boldsymbol{C}$. Again, as seen in the previous sub-section, it is common to approximate this Wald test statistic to an F-distribution.

3.5 Application of the SLR model

The survey logistic regression model was fitted to the data using PROC SURVEYLOGISTIC SAS 9.4, which incorporates the complex survey design by accounting for stratification, clustering (PSU) and unequal weighting. For this analysis, each learner's observation was weighted according to their household's sampling weight, which was equal to the inverse of the probability of the household being selected to take part in the survey. These sampling weights were adjusted for non-response and missing observations. A Taylor series approximation method was used for variance estimation of the fitted SLR model.

Before the final SLR model was obtained, a univariate SLR model was fitted for each independent variable to assess its association with the response, being school attendance. Variables that were significant at a relaxed p-value of 20% were incorporated into the final SLR model. In addition, in order to avoid possible confounding effects between the independent variables, all two-way interactions between the variables were explored. Three of the significant interaction effects that substantially reduced the model's AIC value were selected for the final SLR model.

The predictive accuracy of the SLR model is assessed using the concordance index (c), which is the area under the receiver operating characteristic (ROC) curve ranging from 0 to 1, where the different ranges inform us about the predictive accuracy of the model, as follows:

- 0 - No association
- (0.5; 0.6) - Poor accuracy
- (0.6; 0.7) - Moderate accuracy
- (0.7; 0.8) - Acceptable accuracy
- >0.8 - Excellent accuracy

The concordance statistic is calculated as $c = [n_t - 0.5(t - n_c - n_d)]t^{-1}$, where n_c represents the number of concordant pairs (a pair is said to be concordant if the observation with low order has the lower mean score than the observation with higher order response), while n_d represents number of discordant pairs and t reflects the number of all possible pairs (SAS Institute Inc., 2013).

Table 3.1 presents the final SLR model. Type of place of residence and marital status were not included in the final model based on the results of the univariate analyses of these two variables with school attendance. The two-way interaction effects in the final model included the interaction between the learner's age and their relationship to household head, the learner's age and the survival status of the father, as lastly the household size and the province of residence. The concordance index of the final SLR model for the SAGHS data was 80.1%, thus the model has an excellent predictive accuracy.

Table 3.1: Analysis of effects for the final SLR model

Effect	F-value	P-value
Main Effects		
Age	115.86	<.001*
Province	4.38	<.001*
Gender	1.09	0.2497
Race	19.86	<.001*
Relationship to the household head	8.80	0.0002*
Survival status of the father	17.30	<.001*
Survival status of the mother	0.35	0.5534
Wealth index Z-score for the household	47.61	<.001*
Household source of income	4.38	0.0126*
Household size	0.58	0.4458
Total household income (p/m)	2.22	0.1360
Interaction Effects		
Age * relationship to household head	10.51	<.001*
Age * Survival status of the father	19.96	<.001*
Household size * province	2.56	0.0080*

*significant at 5% level of significance

Table 3.2 gives the estimated odds ratios with their 95% confidence intervals for the variables that were not included in any of the interaction effects. Significance of the factors were assessed based on the inclusion of 1 in the 95% confidence interval for the odds ratio. No significant differences in the odds of attending school was seen for male and female learners, for an increase in the total household income, the survival status of the mother, and for learners of the White or Indian race group compared to the African race group. Learners from the coloured race group had a significantly lower likelihood of attending school compared to learners in the African race group (OR = 0.423, 95% CI: 0.339-0.527). There was an increased likelihood of attending school as the learner's household wealth index Z-score increased (OR = 1.311, 95% CI: 1.214-1.416). Learners whose household income was based on sources other than

salaries or pensions had a significantly higher odds of attending school compared to those whose household income was primarily through salaries (OR = 1.379, 95% CI: 1.113-1.707).

Table 3.2: Estimated odds ratios (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the SLR model

Variables	Odds Ratio (95% CI)
Gender (ref = Male)	
Female	1.066 (0.945; 1.203)
Race (ref = African/Black)	
Coloured	0.423 (0.339; 0.527)*
Indian	0.635 (0.347; 1.164)
White	0.739 (0.496; 1.101)
Survival status of the mother (ref = No)	
Yes	0.941 (0.771; 1.150)
Wealth index Z-score for the household	1.311 (1.214; 1.416)*
Household source of income (ref = Salaries)	
Pension	1.049 (0.884; 1.245)
Other	1.379 (1.113; 1.707)*
Total Household income (ref = Below R25000(p/m))	
Above R25000	1.188 (0.947; 1.490)

*significant at 5% level of significance

The effects of the interaction between age of the learner and relationship to the head of household, the interaction between the age of the learner and survival status of their father, as well as the interaction between the size of the learner's household and province of residence on the log-odds of school attendance are presented in Figures 3.1, 3.2 and 3.3, respectively. A positive log-odds is associated with a higher likelihood of school attendance, and a negative log-odds is associated with a decreased likelihood of school attendance. Based on Figure 3.1, there was a decline in the log-odds of school attendance with an increase in age for all categories of relationship to the head of household. After the age of 9 years of age, learners with other relation-

ships to the household head had the lowest log-odds of attending school. Similarly, from Figure 3.2, it is observed that the log-odds of school attendance declined with an increase in age for learners whose fathers were alive as well as not alive. Learners under the age of 16 years old whose father was still alive surprisingly had a lower likelihood of attending school. However, learners over the age of 16 years old whose father was not alive had the lowest likelihood of attending school. Based on Figure 3.3, there was an increase in the log-odds of school attendance as the household size increased for learners residing in Limpopo, Gauteng and Mpumalanga, while the log-odds declined with an increase in household size for learners residing in the other provinces. Learners residing in Limpopo province had the highest likelihood of attending school.

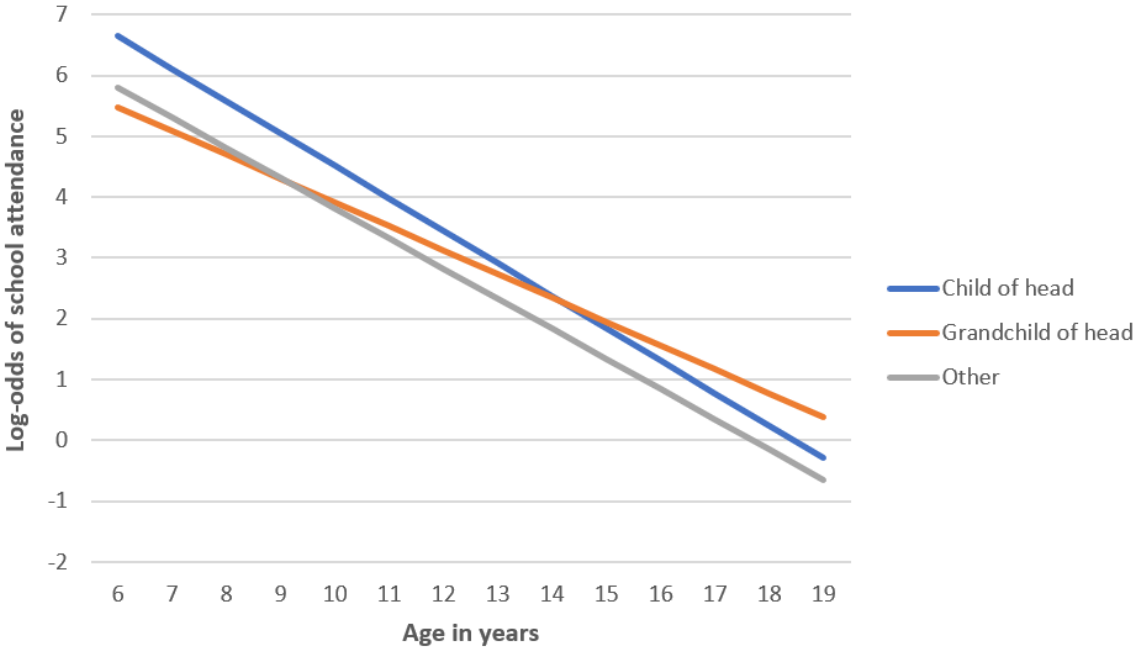

Figure 3.1: The estimated log-odds of school attendance for the interaction between the age of the learner and relationship to the head of household

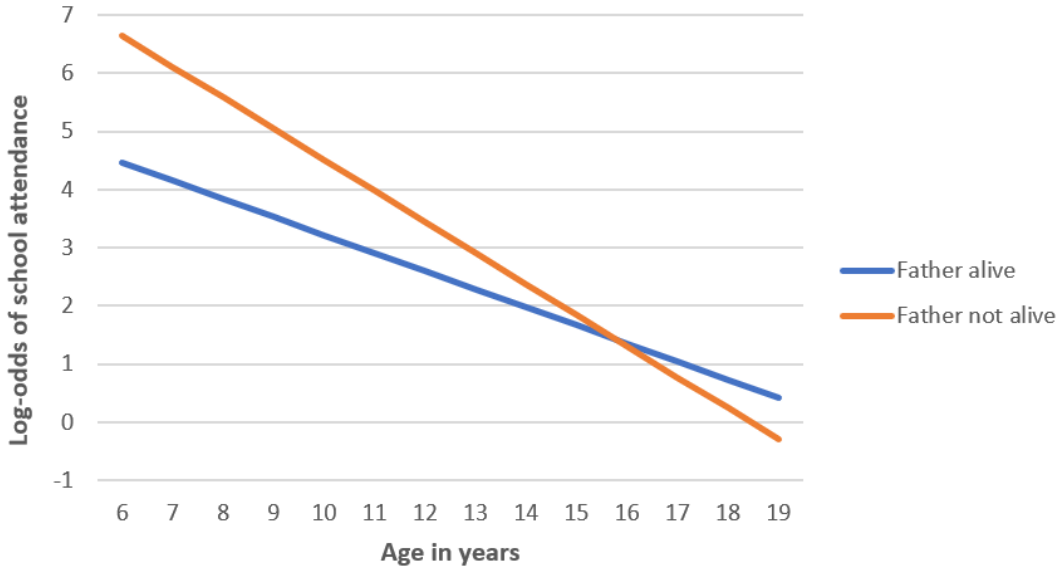

Figure 3.2: The estimated log-odds of school attendance for the interaction between the age of the learner and survival status of their father

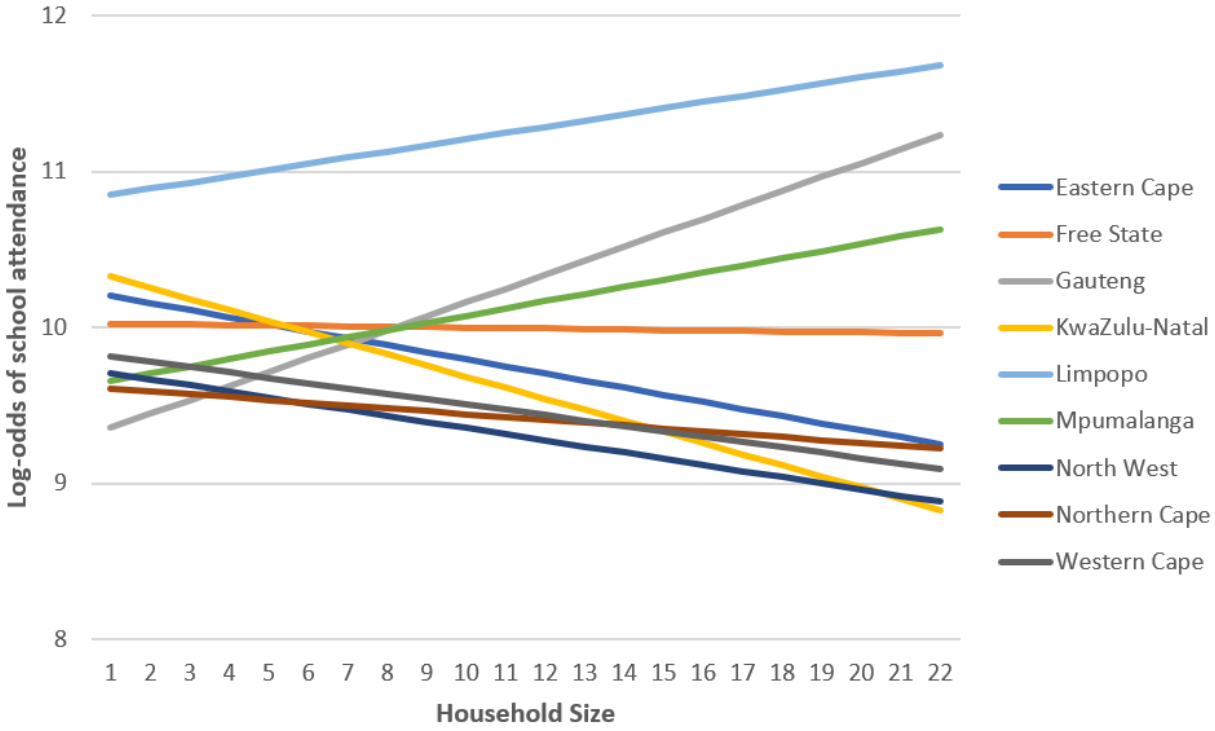

Figure 3.3: The estimated log-odds of school attendance for the interaction between the size of the learner's household and province of residence

3.6 Summary

This chapter presented an overview of generalized linear models, which is used in the analysis of a non-normal response. An extension of this class of models, the survey logistic regression model, was presented to accommodate the analysis of a binary response based on data obtained from complex survey designs. This is a designed-based approach which incorporates the sampling weights in the estimation of the parameters and standard errors of the parameter estimates.

In this chapter, the survey logistic regression model was applied to the SAGHS data to investigate the relationship between the likelihood of a learner attending school and several individual-level, household-level and geographical factors. This model assumes that the observations are independent. However, learners residing in the same communities may be more similar than those form different communities, which may result in possible correlations in the observations. Thus, the next chapter considers a statistical approach to deal with such correlations.

Chapter 4

Generalized Linear Mixed Models

In the case of modelling data from complex survey designs where stratified cluster sampling methods were used, the design of the study is such that the clusters included in the sample represent only a random sample from a population of clusters. In addition, it is possible that observations within the same cluster are more alike compared to those from different clusters. Therefore, we cannot assume that the observations are independent, which is an important assumption of the previously applied SLR model. In this case, the effect of clustering can be included in the model via a random effect, which also allows a possible correlation in the observations to be accounted for. An extension of the generalized linear model to include a random effect is the generalized linear mixed model (GLMM). The GLMM thus includes both fixed and random effects, hence it is a mixed model. This chapter gives an overview of the GLMM and then presents the results of the GLMM applied to the SAGHS data.

4.1 The Model

Suppose Y_{ij} is the j^{th} response, $j = 1, 2..., n_i$, for the i^{th} cluster, $i = 1, 2, ..., m$, and $\boldsymbol{y}_i$ indicates the $n_i \times 1$ vector of responses from the i^{th} cluster. In the GLMM, responses Y_{ij} of $\boldsymbol{y}_i$ are assumed to be conditionally independent given a vector of normally distributed random effects, γ_i. Similar to the GLM, the probability density of the

response comes from the exponential family, and is given by

$$f(y_{ij}|\theta_{ij},\phi) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij},\phi)\right\}. \tag{4.1}$$

The above has the same form as Equation 3.1 in Chapter 3, which means the parameters are similarly defined as those in Equation 3.1.

The conditional mean, μ_{ij}, of Y_{ij} is modelled through a linear predictor η_{ij}, which contains the fixed regression parameters, $\boldsymbol{\beta}$, as well as the subject-specific random effect parameters, γ_i. Therefore,

$$\begin{aligned}\eta_{ij} &= g(\mu_{ij}) = g\big[E(y_{ij}|\gamma_i)\big] \\ &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i,\end{aligned} \tag{4.2}$$

or in matrix form

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}. \tag{4.3}$$

$g(.)$ is the known link function that links the conditional mean of $\boldsymbol{y}$ to the linear predictors. $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix for the fixed effects and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of fixed effects regression coefficients. $\boldsymbol{Z}$ is the $n \times q$ design matrix for the random effects and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effect coefficients. It is assumed $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ depends on unknown variance components.

Two approaches can be used to estimate the parameters in a GLMM, a Bayesian approach and a classical approach that uses maximum likelihood methods (McCulloch & Neuhaus, 2005). In this thesis, the classical approach using a maximum likelihood method will be discussed.

4.2 Estimation

4.2.1 Maximum Likelihood Estimation

The maximum likelihood estimates of a GLMM are obtained by integrating over the distribution of the p-dimensional random effects. The likelihood, based on the contribution of the i^{th} cluster, is given by

$$f_i(y_{ij}|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i|\boldsymbol{G}) d\boldsymbol{\gamma}_i, \tag{4.4}$$

where $f(\boldsymbol{\gamma}_i|\boldsymbol{G})$ represents the distribution of the random effects. This leads to the full likelihood function for $\boldsymbol{\beta}$, $\boldsymbol{G}$ and ϕ, given by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \prod_{i=1}^{m} f_i(y_{ij}|\boldsymbol{\beta}, \boldsymbol{G}, \phi) \\ &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma_i}|\boldsymbol{G}) d\boldsymbol{\gamma}_i. \end{aligned} \tag{4.5}$$

In the case of a normally distributed response, the method of maximum likelihood for the estimation of fixed effects in the GLMM becomes the same as that for a linear mixed model (SAS Institute, 2004). However, for many other cases of the GLMM involving non-normal responses, the likelihood function does not have a closed-form expression (Jiang, 2001). This is as a result of the likelihood involving high-dimensional integrals which cannot be evaluated analytically. Thus, approximations are required to evaluate the likelihood function (Jiang, 2001). The are multiple methods of approximation, three of which are

- Approximation of integrand
- Approximate the integral
- Approximate the data

For the purpose of this thesis, only a method that approximates the integrand will be considered.

4.2.2 Laplace Approximation

A common method used for an approximation of the likelihood function is the Laplace approximation, which is based on an approximation of the integrand (Jiang, 2007). Suppose an integral has the form

$$\int e^{Q(\boldsymbol{x})} dx, \tag{4.6}$$

where $Q(\boldsymbol{x})$ is a known, unimodal function, and $\boldsymbol{x}$ is a $q \times 1$ vector of variables. If $Q(\boldsymbol{x})$ is minimized when $\boldsymbol{x} = \widehat{\boldsymbol{x}}$, then the second-order Taylor series expansion of $Q(\boldsymbol{x})$ around $\widehat{\boldsymbol{x}}$ is given by

$$Q(\boldsymbol{x}) \approx Q(\widehat{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{x}})' Q''(\widehat{\boldsymbol{x}})(\boldsymbol{x} - \widehat{\boldsymbol{x}}), \tag{4.7}$$

where $Q''(\widehat{\boldsymbol{x}})$ is the Hessian of Q evaluated at $\widehat{\boldsymbol{x}}$. This leads to an approximation of Equation 4.6:

$$\int e^{Q(\boldsymbol{x})} dx \approx (2\pi)^{\frac{q}{2}} |Q''(\widehat{\boldsymbol{x}})|^{-\frac{1}{2}} e^{-Q'(\widehat{\boldsymbol{x}})} \tag{4.8}$$

The approximation to this integral uses as many different estimates of $\widehat{\boldsymbol{x}}$ as necessary according to the different modes of function Q. As $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$, it can be shown that the integral in the likelihood Equation 4.5 is proportional to the integral in Equation 4.6, where function Q is given by

$$Q(\gamma) = \phi^{-1} \sum_{j=1}^{n_i} \left[y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) - b(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) \right] - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}. \tag{4.9}$$

This implies that the Laplace approximation method can be applied to the likelihood function for a GLMM. This approximation method is better for large cluster sizes and is improved by adding higher-order terms to the Taylor series expansion.

4.3 Model Selection

The likelihood ratio test and Wald test (F-test) can be used for inference concerning the fixed effect parameter estimates obtained through numerical approximations. The likelihood ratio test can be applied to the data in comparing two nested models with different mean structures but consisting of the same variance-covariance structure. In a similar manner, the likelihood ratio test can be used in comparing nested models with unique covariance structures but consisting of the same mean, and the inferences of variance-covariance components remain valid for Wald test approximation. However, if the variance parameter that is being tested takes values on the boundary of the parameter space, the normal approximation fails. This means that the test statistics for these tests will not have the traditional Chi-square distribution under the null hypothesis (Zhang & Lin, 2008). However, when testing the null hypothesis of no random effects, Zhang & Lin (2008) have show that the test statistic will be a mixture of Chi-square distributions, rather that the classical single Chi-square distribution.

4.4 Application of the GLMM

The GLIMMIX procedure in SAS, which is used to estimate and make inferences for generalized linear mixed models, was used to fit a GLMM to the SAGHS data. PROC GLIMMIX extends the GLM by incorporating normally distributed random effects. To account for the effect of clustering, where learners residing in the same cluster may be more alike with regards to school attendance, a random intercept term which varies at cluster level was incorporated into the GLMM.

The Laplace approximation procedure was used to fit the final GLMM as it is likelihood based, which makes it possible to compare the models using selection criteria, such as the AIC and BIC. Another advantage of using the Laplace approximation is that it is computationally less demanding. The same main effects and two-way inter-

action effects that were incorporated into the SLR model were incorporated into the final GLMM. The importance of random intercept was assessed by testing whether the covariance parameter was equal to zero using the COVTEST in SAS. Table 4.1 below shows the results of testing the effectiveness of the random effect in the model. Clearly, the null hypothesis of the covariance parameter equal to zero was rejected as the p-value is small, thus the cluster effect was significant in the model. This result suggests that learners residing within the same communities may share common factors that affect their school attendance. Such a common factor may be the school itself which may contribute to the likelihood of a learner attending school, as learners within the same community are likely to attend the same school.

Table 4.1: The test for covariance parameters based on likelihood

Label	DF	-2 Log Like	Chi-Sq	Pr>Chi-Sq
No G-side effects	1	8348.37	12.50	0.0002

After considering the significance of the random effect in the model, we have to select/specify the covariance structure which best suits the random effect. There are a variety of considerations when selecting the covariance structure of random effect (Kincaid, 2019), this includes the number of parameters, the best covariance structure (which will allow for easy interpretation), diagnostic results, and effects on the fixed effects. The covariance structure chosen must limit the number of parameters as much as possible to make the model less complex, as the complexity of the model sacrifices power and efficiency. This does not mean we are trying to find the simplest model as that can increase the possibility of making type I error in the fixed effects (Kincaid, 2019). Rather, the best covariance structure is selected based on the one that minimizes the AIC value. There are numerous covariance structures to consider, but it is not required to check all of them. The most used structures include Unstructured (UN), Autoregressive order 1 (AR(1)), Variance Component (VC), Compound Symmetry (CS), Toeplitz (Toep), Heterogeneous AR(1), Heterogeneous CS, and Heterogeneous Toep. The GLMM was fitted based on each of these variance structures. The Variance Component structure produced the minimum AIC value and was cho-

sen as the best covariance structure for the model. The variance component for this cluster effect was estimated at 0.2383 with a standard error of 0.07198 as seen in Table 4.2. This estimate is relatively far from zero, thus confirming the importance for this random effect in the model.

Table 4.2: Covariance parameter estimate for the final GLMM

Cov Param	Subject	Estimate	Standard error
Intercept	Cluster (PSU)	0.2383	0.07198

Table 4.3 presents the final GLMM. This model found the same significant variables as the SLR model, except for the main source of household income which was only significant at a 10% significance level. In addition, the total household income per month was significant at a 10% significance level.

Table 4.3: Analysis of effects for the final GLMM

Effect	F-value	P-value
Main Effects		
Age	568.55	<0.001*
Province	3.49	0.0005*
Gender	0.67	0.4117
Race	18.06	<.001*
Relationship to the household head	25.44	<.001*
Survival status of the father	42.09	<.001*
Survival status of the mother	0.05	0.8186
Wealth index Z-score for the household	54.71	<.001*
Household source of income	2.39	0.0919
Household size	0.81	0.3671
Total household income (p/m)	3.31	0.0688
Interaction Effects		
Age * relationship to household head	29.59	<.001*
Age * Survival status of the father	48.42	<.001*
Household size * province	2.18	0.0258*

*significant at 5% level of significance

Table 4.4 presents the estimated odds ratios with their 95% confidence intervals for the variables that were not included in any of the interaction effects in the fitted GLMM. These results were similar to that of the SLR model, where learners from the coloured race group had a significantly lower likelihood of attending school compared to learners in the African race group (OR = 0.423, 95% CI: 0.335-0.534). In addition, as the household wealth index Z-score increased, there was an significant increase in the likelihood of a learner attending school (OR = 1.320, 95% CI: 1.226-1.420). However, unlike the results of the SLR model, there were no significant differences in the likelihood of a learner attending school for the various sources of household income.

Table 4.4: Estimated odds ratios (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the GLMM

Variables	Odds Ratio (95% CI)
Gender (ref = Male)	
Female	1.052 (0.933; 1.186)
Race (ref = African/Black)	
Coloured	0.423 (0.335; 0.534)*
Indian	0.573 (0.309; 1.060)
White	0.725 (0.486; 1.081)
Survival status of the mother (ref = No)	
Yes	0.977 (0.804; 1.188)
Wealth index Z-score for the household	1.320 (1.226; 1.420)*
Household source of income (ref = Salaries)	
Pension	1.056 (0.908; 1.229)
Other	1.049 (0.884; 1.245)
Total Household income (ref = Below R25000(p/m))	
Above R25000	1.224 (0.985; 1.522)

*significant at 5% level of significance

The effects of the two-way interactions are presented in Figures 4.1, 4.2 and 4.3. The interaction effect between the age of the learner and relationship to the head

of household revealed that learners who were the child or grandchild of the household head had the highest log-odds of attending school from the age of 9 years old (Figure 4.1). However, regardless of the relationship of the learner with the head of the household, the likelihood of attending school decreased with an increase in age. This is also observed in Figure 4.2 for the interaction effect between the age of the learner and the survival status of their father, although, the likelihood of attending school was higher for learners under the age of 16 years old whose father was deceased. Figure 4.3 displays the effect of the interaction between the household size and province of residence. As revealed by this interaction effect in the SLR model, there was an increase in the log-odds of school attendance as the household size increased for learners residing in Limpopo, Gauteng and Mpumalanga, while the log-odds declined with an increase in household size for learners residing in the other provinces. Learners residing in Limpopo province had the highest likelihood of attending school.

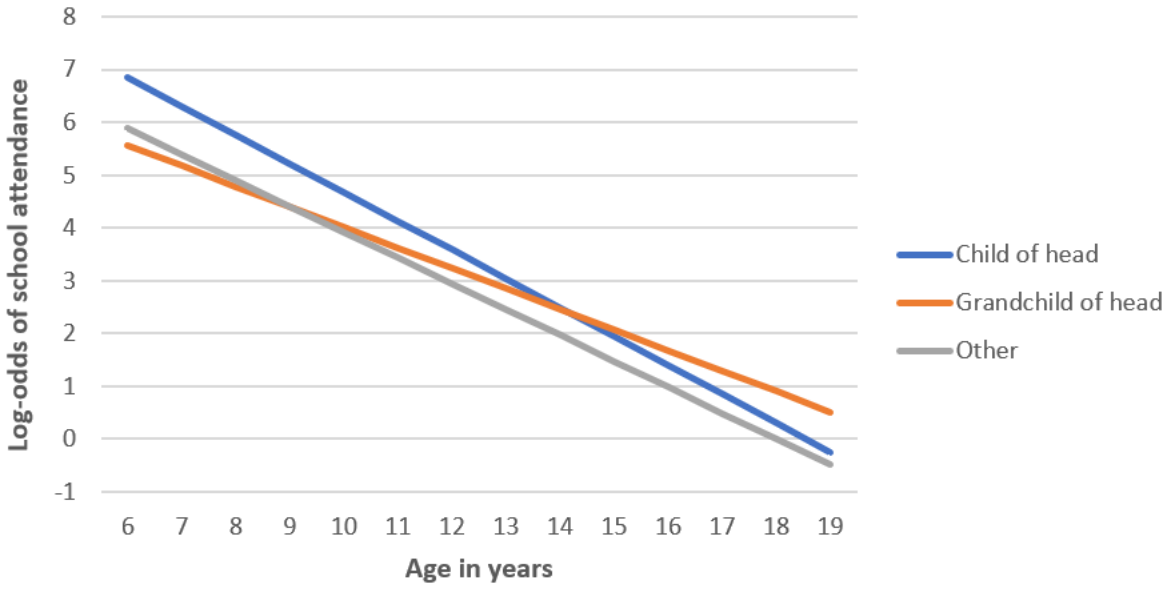

Figure 4.1: The estimated log-odds of school attendance for the interaction between the age of the learner and relationship to the head of household

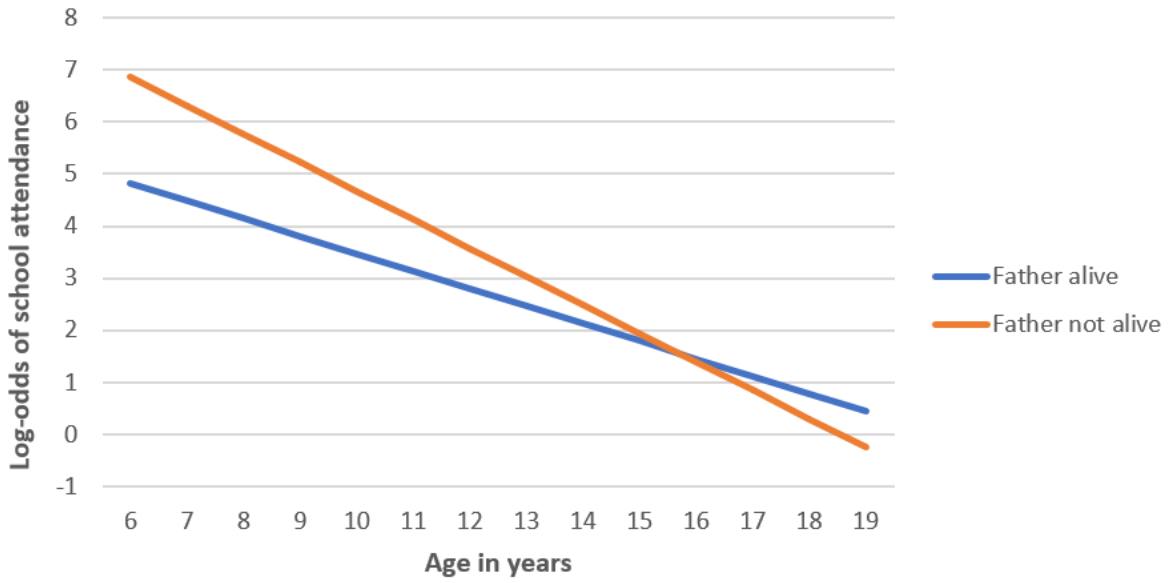

Figure 4.2: The estimated log-odds of school attendance for the interaction between the age of the learner and survival status of their father

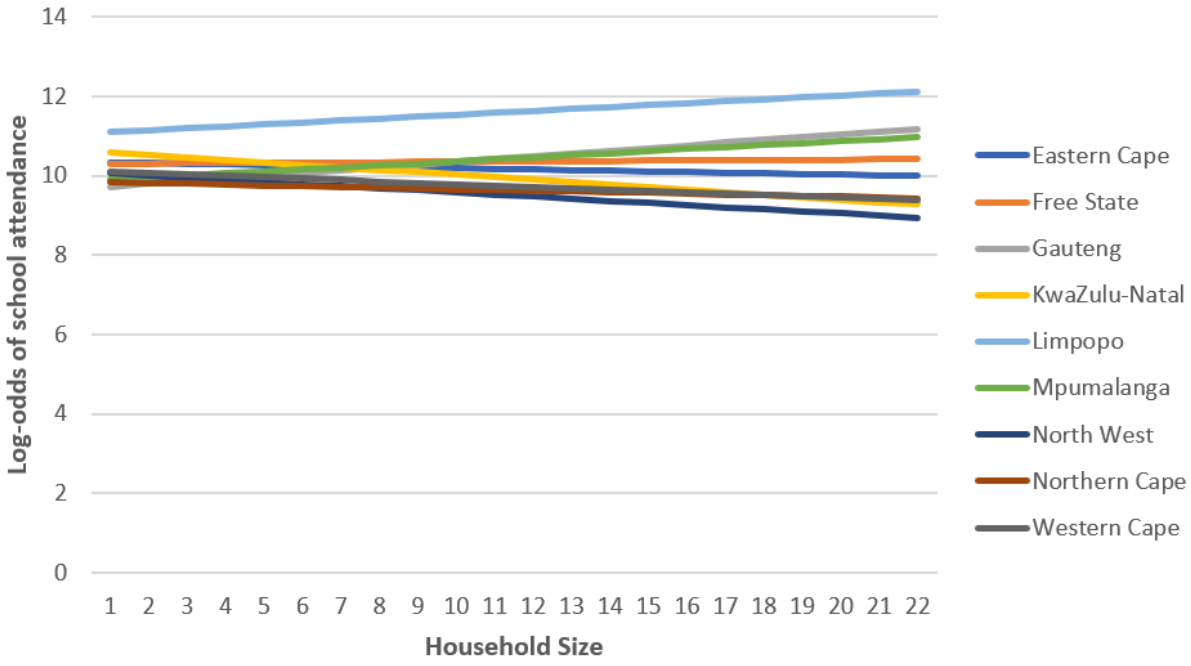

Figure 4.3: The estimated log-odds of school attendance for the interaction between the size of the learner's household and province of residence

4.5 Summary

The generalized linear mixed model is an extension of GLM and includes both fixed and random effects. A random effect assists in accounting for possible correlations in the observations, as well as the fact that the clusters included in the data only represent a random sample selected from a population of clusters. The application of the GLMM is known as a model-based approach where interest is also on estimating the proportion of variation in the response variable that is attributable to each of the multiple levels of sampling (Heeringa et al., 2010). While the SLR model was design-based and did not account for the effect of clustering, similar results were revealed between the two approaches. The older learners were less likely to attend school, and learners from the Coloured race group had the lowest likelihood of attending school. A varying likelihood of school attendance was observed for the different provinces of residence.

Chapter 5

Generalized Additive Mixed Model

The models discussed in the previous chapters, namely the SLR model and the GLMM, incorporated the effect that a province has on the likelihood of school attendance as fixed. However, it is possible that spatial autocorrelation is present, where neighbouring provinces may have similar effects compared to non-neighbouring provinces. Thus, in this chapter, we discuss an extension of the generalized linear mixed model, which is the generalized additive mixed model (GAMM). The GAMM enables one to account for spatial variation and spatial autocorrelation by incorporating a spatial effect as a non-linear random effect in the model. This spatial effect can be decomposed into two: a structured spatial effect and an unstructured spatial effect. The structured spatial effect accounts for spatial autocorrelation through a conditionally autoregressive (CAR) structure which makes use of the neighbourhood structure of the provinces, where two provinces are neighbours if they share a border. However, the unstructured spatial effect accounts for spatial variation between the provinces that is due to the effects of unmeasured factors that are not spatially related. This unstructured spatial effect assumes an independently and identically distributed normal distribution.

In addition to the spatial effects, the GAMM also allows one to explore the non-linear relationship between the continuous covariates and the response. The SLR model and the GLMM assumed a linear relationship between all of the explanatory variables and school attendance, however, it is possible that some of the continuous covariates have a non-linear effect on the likelihood of school attendance. These non-linear effects in the GAMM are represented by unknown smooth functions that are approximated using P-splines with B-splines basis functions (Eilers & Marx, 1994). In chapter, an overview of the GAMM is presented, as well as the results of the GAMM applied to the SAGHS data.

5.1 The Model

The generalized additive mixed model is an extension of GLMM which uses non-parametric functions to model covariates and geospatial effects while accounting for correlation by adding random effects as predictors (Wood, 2017). The GAMM was considered due to the hierarchical and spatially distributed nature of the data. Similar to the SLR model and GLMM, the GAMM can model the probability of a the k^{th} learner residing in household j and province i attending school, given by $P(Y_{ijkl} = 1) = \pi_{ijkl}$, by making use of the logit link function. The model is as follows:

$$logit(\pi_{ijk}) = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \sum_{r=1}^{P} f_r(z_{ijk}) + f_{spat}(s_i), \tag{5.1}$$

where $\boldsymbol{\beta}$ is the vector of linear fixed effects of the covariates that are modelled parametrically, $f_r(.)$, $r = 1, \ldots, P$, are the unknown smooth functions which represents the non-linear effects of p continuous covariates modelled non-parametrically, and lastly, $f_{spat}(s_i)$ is the non-linear spatial effect of the i^{th} province.

5.2 Estimation of the smooth terms

The smooth functions f_r are estimated using the penalised splines (P-splines) with B-splines basis functions (Eilers & Marx, 1996). This approach presumes that the unknown functions can be estimated using the polynomial spline of the 1^{st} degree with knots $z_r^{\{min\}} < \xi_{r0} < \xi_{r1} < \ldots < \xi_{rn_r} < z_r^{\{max\}}$ that are spaced equally. The spline is written in terms of a linear combination of $M_r = n_r + v$ B-spline basis functions, B_{rm}, and regression coefficients of α_{rm}. The function of M_r is made by n_r, which represents the k^{th} order divided difference and v represents the corresponding sequence of integers as follows:

$$f_r(z_r) = \sum_{m=1}^{M_r} \alpha_{rm} B_{rm}(z_r). \tag{5.2}$$

The choice of the number of knots is considered an important aspect in approximation of the smooth function. This is because too many knots may result in estimated curves that over-fit the data, leading to functions which are too rough. While, too few results may not be flexible enough to capture variability in the data (Fahrmeir et al., 2004). Thus, the way to overcome this problem is by estimating a moderately large number of 20-40 equally spaced knots to ensure flexibility. In addition, the roughness penalty is defined based on 1^{st} and 2^{nd} order differences of the adjacent B-spline coefficients that guarantee sufficient smoothness of the fitted curves (Fahrmeir et al., 2004). The above mentioned information leads to penalised likelihood estimation that consists of penalty terms given as:

$$P(\lambda_r) = \frac{1}{2}\lambda_r \sum_{m=v+1}^{M_r} (\Delta^v \alpha_{rm})^2, \; v = 1, 2, \tag{5.3}$$

where λ_r is the smoothing parameter and Δ^v is the differencing operator of order v. The 1^{st} order differences penalise abrupt jumps $\alpha_{rm} - \alpha_{r,m-1}$ between successive parameters, while the 2^{nd} order differences penalise deviations from the linear trend $2\alpha_{r,m-1} - \alpha_{r,m-2}$.

5.3 Estimation of Spatial effects

The spatial effect $f_{spat}(s_i)$ of province s_i in which the learner resides, $s \in (1, \ldots, 9)$, represents the effects of unobserved factors that have not been included in the model and also accounts for spatial autocorrelation. This spatial effect may be decomposed into a spatially correlated (structured) and an uncorrelated (unstructured) effect as follows:

$$f_{spat}(s_i) = f_{str}(s_i) + f_{unstr}(s_i), \tag{5.4}$$

where the structured spatial effect $f_{str}(s_i)$ accounts for the assumption that learners in neighbouring provinces are more likely to be correlated with regards to their outcomes. However, the unstructured spatial effect $f_{unstr}(s_i)$ accounts for the spatial variation due to effects of unmeasured province-specific factors that are not spatially related.

A conditional autoregressive (CAR) approach to model the structured spatial effect is commonly used, where the specification of the structured spatial effect is based on a Markov process approach adopted from time series analysis. The Markov process states that the value of a variable at time $t + 1$ depends only on the previous value. This idea is extended to the spatial domain by assuming $f_{str}(s_i)$ depends only on a set of neighbours. In other words, $f_{str}(s_i)$ depends on $f_{str}(s_j)$ only if province s_j is in the neighbourhood set, $\mathcal{N}_i$, of s_i. For this conditional autoregressive approach, the process $f_{str}(s_i)$ is called a Markov random field (MRF). The models are constructed for the distributions of $f_{str}(s_i)|f_{str}(s_j)$, $s_j \in \mathcal{N}_i$ (Schabenberger & Gotway, 2017). It is common to assume each of these conditional distributions is Gaussian, as follows

$$f_{str}(s_i)|f_{str}(s_j),\ i \neq j,\ \sim N\left(\frac{1}{n_{s_i}} \sum_{s_j \in \mathcal{N}_i} f_{str}(s_j), \frac{1}{n_{s_i}\tau^2_{str}}\right), \tag{5.5}$$

where n_{s_i} is the number of neighbours of province s_i. Thus, the conditional mean of $f_{str}(s_i)$ is an average of the function evaluations $f_{str}(s_j)$ of neighbouring provinces.

The variance component tau^2_{str} controls the smoothness of the spatial effect and accounts for spatial variation between the provinces, it is also used to capture the amount of variation explained by this spatial structure.

The unstructured spatial effect $f_{unstr}(s_i)$ is incorporated into the GAMM as an independently and identically distributed random effect, i.e.

$$f_{unstr}(s_i) \sim N\left(0, \frac{1}{\tau^2_{unstr}}\right), \tag{5.6}$$

where the variance component tau^2_{unstr} once again controls the smoothness of this effect. This function for the random effect can also be approximated by a linear combination of B-spline basis functions given in Equation 5.2. However, the regression coefficients α_{rm} are i.i.d. random effects (Kneib et al., 2008).

5.4 Method of REML

The restricted maximum likelihood (REML) method is a modification of the maximum likelihood method first introduced by Patterson & Thompson (1971). It is used as an alternative for finding the estimates of the variance components in $\boldsymbol{G}$ for a mixed effect model, where it takes into consideration the loss of degrees of freedom from the estimation of $\boldsymbol{\beta}$, therefore producing unbiased estimates of the variance components. Thus, it is a useful procedure for inferences about variance components, particularly in GLMMs. REML optimization is also a common method for fitting GAMMs with smoothing parameters. A thorough overview of the REML method for GAMMs is given by Wood (2017).

5.5 Application of the GAMM

For the application of the GAMM, we made use of the R *mgcv* package where a REML method of estimation was used. The shapefile of the South African provinces

were imported into R so that the neighbourhood structure could be constructed. This neighbourhood structure was then utilised in estimating the structured spatial effect. In order to determine the best fitting model, the non-linear effect of all of the continuous covariates was explored. Those that did not display a significant non-linear effect on the log-odds of school attendance were entered into the final GAMM as linear fixed effects, along with the categorical explanatory variables. The only two continuous covariates that displayed a significant non-linear effect on the log-odds of school attendance were the age of the learner and the number of members that reside in the learner's household (household size). Furthermore, the necessity for the inclusion of the spatial effects were assessed by comparing the AIC and adjusted R-square values of the models with and without the spatial effects. The spatial model significantly improved the AIC as well as improved the adjusted R-square value by 1.2%, thus, the results discussed in this section are based on the GAMM with the structured and unstructured spatial effects according to the province of residence.

5.5.1 Results of the fixed effects

The odds ratios (OR) and corresponding 95% confidence intervals (CI) for the fixed effects are given in Table 5.1. Based on these results, the gender of the learner still did not have a significant effect on the likelihood of attending school, as was seen in the results from the SLR model and GLMM. There was a significantly lower odds of attending school for learners from the White and Coloured race groups compared to learners in the African/Black race group. There was no significant difference in the likelihood of attending school for learners with either parent alive or deceased. Learners from wealthier households had a significantly higher odds of attending school compared to those from poorer households (OR = 1.358, 95% CI: 1.259-1.465). Learners from households with other primary sources of income, other than pension or salaries, had a significantly higher likelihood of attending school compared to those in salaried households (OR = 1.355, 95% CI: 1.091-1.683). Similarly, the odds of attending school was higher among learners in households with a total income above R25 000 per month (OR = 1.506, 95% CI: 1.198-1.893).

5.5.2 Results of the non-linear and spatial effects

Table 5.2 provides the significance of each of the non-linear and spatial effects. The non-linear effect of the learner's age and household size as well as the unstructured spatial effect were significant at a 5% level of significance. However, the structured spatial effect was insignificant. This suggests that the unstructured spatial effect (provincial-level random effect) was more dominant, thus suggesting that there are unmeasured province-specific factors that are affecting the likelihood of a learner attending school, where such factors are not spatially related or common to other provinces.

Table 5.1: Odds ratio estimates (OR) and corresponding 95% confidence intervals (CI) for the fixed effects of the GAMM

Variables	Odds Ratio (95% CI)
Gender (ref = Male)	
Female	1.063 (0.937; 1.206)
Race (ref = African/Black)	
Coloured	0.359 (0.284; 0.454)*
Indian	0.565 (0.300; 1.064)
White	0.604 (0.401; 0.910)*
Survival status of the mother (ref = No)	
Yes	0.915 (0.747; 1.120)
Survival status of the father (ref = No)	
Yes	1.121 (0.962; 1.306)
Wealth index Z-score for the household	1.358 (1.259; 1.465)*
Household source of income (ref = Salaries)	
Pension	1.006 (0.861; 1.175)
Other	1.355 (1.091; 1.683)*
Total Household income (ref = Below R25000(p/m))	
Above R25000	1.506 (1.198; 1.893)*

*significant at 5% level of significance

Table 5.2: The approximate significance of smooth terms for the final GAMM

Variable	Chi-square	P-value
Structured Spatial Effect	200.715	0.3539
Unstructured Spatial Effect	8.066	<0.0001
Age	1904.455	<0.0001
Household size	22.957	0.0002

Figure 5.1 shows the non-linear effect that a learner's age in years and their household size has on the log-odds of attending school, along with the 95% confidence intervals. There was an increased log-odds associated with learners of the ages of just over 6 to 15 years old, after which the log-odds decreased and was negative. Thus, learners older than 15 years were associated with a lower likelihood of attending school. Similarly, learners in the younger age group of 5 to 6 years old had a lower likelihood of attending school. Based on the estimated non-linear effect of the household size, the log-odds was negative for households with approximately 8 or more members. Therefore, learners residing in households with 8 or more members had a lower likelihood of attending school. The widening of the confidence interval for the larger household sizes is as a consequence of very few learners in the sample residing in households of such sizes.

Figure 5.2 reveals the estimated log-odds of school attendance for the structured spatial effect according to the province of residence. While the overall spatial effect was not significant, this figure showed a varying effect across the provinces. In particular, provinces in the Eastern part of the country displayed an increased log-odds and therefore contribute to a higher likelihood of school attendance, with Limpopo province displaying the highest estimated effect. However, provinces in the Western part of the country revealed a much lower, negative log-odds. Thus, these provinces contribute to a much lower likelihood of school attendance for learners residing in them. Furthermore, this figure reveals that neighbouring provinces seem to have similar effects on school attendance compared to provinces further away from each other. Thus, it was necessary to account for spatial autocorrelation, even though the effect was not strong.

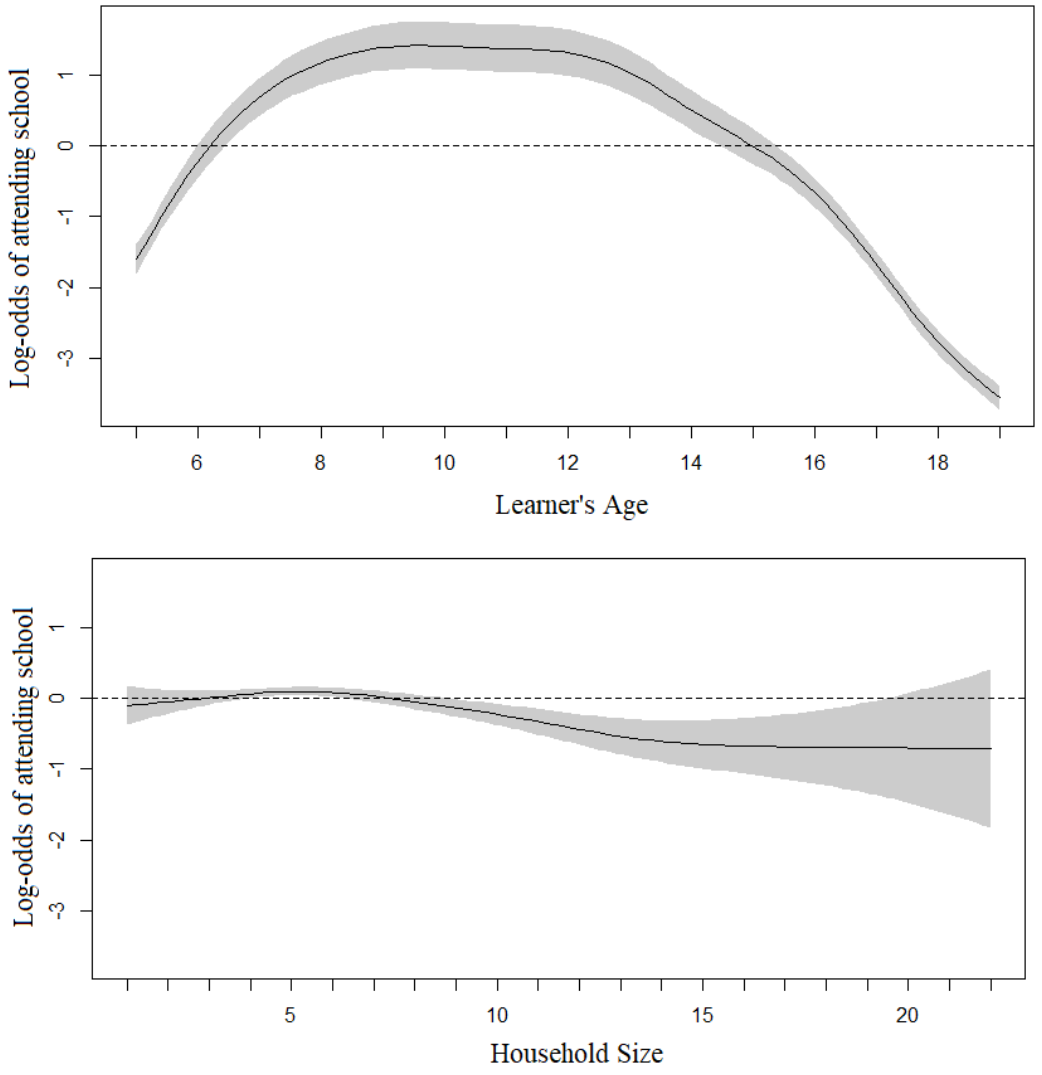

Figure 5.1: Estimated non-linear effect of the learner's age in years (top) and the household size (bottom) on the log-odds of school attendance, along with the 95% confidence intervals.

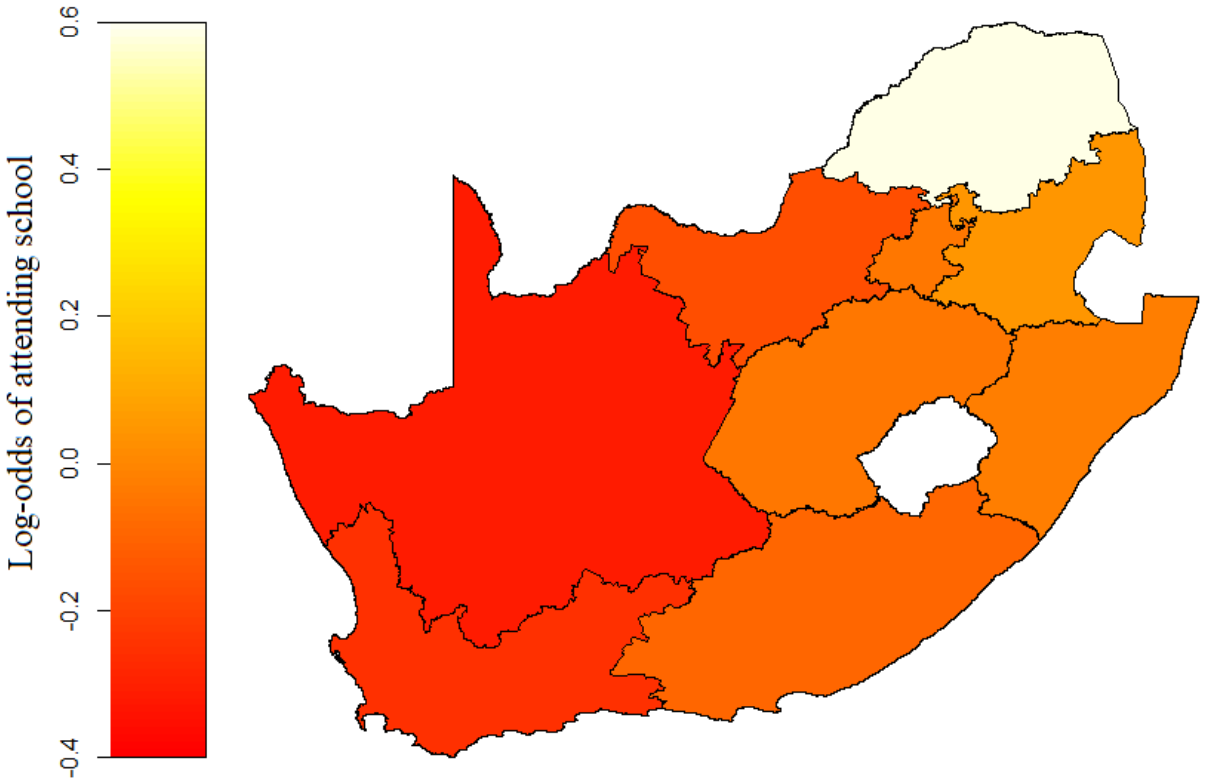

Figure 5.2: The estimated log-odds of school attendance for the structured spatial effect ('Red shading = decreased likelihood' and 'yellow shading = increased likelihood')

5.6 Summary

In this chapter, the effect that the province of residence had on the learner's likelihood of attending school was incorporated as a non-linear spatial effect, rather than as a linear fixed effect, as was done in the SLR model and GLMM. This allowed possible spatial autocorrelation to be accounted for in the observations. This spatial effect was decomposed into a structured and unstructured spatial effect. The structured spatial effect allowed for dependencies in the observations from learners in neighbouring provinces. This structured spatial effect represents the spatial variation in the observations that are due to unmeasured, spatially correlated factors that transcend the boundaries of the provinces. However, the results revealed

that the structured spatial effect was fairly weak in comparison to the unstructured spatial effect, suggesting that the contribution that a particular province has on the likelihood of a learner attending school is due to province-specific factors that are not shared by neighbouring provinces. The further advantage of the GAMM over the SLR model and GLMM was that the non-linear effect of covariates could be explored. This revealed that the learner's age and household size had a significant non-linear effect on the likelihood of attending school.

Chapter 6

Discussion and Conclusion

This thesis aimed to examine the factors affecting school attendance at the basic education level in South Africa for learners under the age of 20 years old who have not yet matriculated. The school attendance rates were investigated across the provinces of South Africa as well as according to various socio-economic and demographic factors. The overall attendance rate was 93.4%, however this rate ranged from 88.4% in the Northern Cape to 97.6% in Limpopo province. Learner's in the Colour race group had the lowest rate of attendance at 86.9%. Three different statistical approaches were used to determine the factors that were significantly associated with the school attendance of a learner, as well as determine which factors contributed to a lower likelihood of school attendance. These approaches included the survey logistic regression model (a design based approach), the generalized linear mixed model (a model based approach) and the generalized additive mixed model (a model based approach which incorporated non-linear and spatial effects).

The results of the survey logistic regression model and generalized linear mixed model largely concurred with each other, where gender had an insignificant effect on the likelihood of attending school. A significant difference in the odds of attending school was seen between learners from the Coloured and African race groups, where Coloured learners had a lower likelihood of attending school. Learners from poorer

households had a substantially lower odds of attending school. This result was in agreement with that of other studies, which showed that financial issues were one of the biggest reasons for non-attendance (Phineas Reuckert, 2019). Significant interactions were found between the age of the learner and their relationship to household head, between age and the survival status of the father, as well as between the household size and province of residence. Both interactions involved with the age of the learner showed a decreases likelihood of school attendance with an increase in age. The household size had a varying effect on the log-odds of school attendance for the difference provinces of residence, however learners residing in Limpopo province had a consistently higher likelihood of attending school compared to those in other provinces.

Both the SLR model and GLMM assumed a linear relationship between the response and covariates. However, upon applying the GAMM, the age of the learner and the household size had a significant non-linear effect on the likelihood of attending school. In addition, the results of the GAMM revealed spatial variation in the likelihood of school attendance across the difference provinces of residence. This spatial variation may be due to unmeasured factors that vary geographically, where such factors contribute to spatial heterogeneity in the observations. The spatial effect was incorporated based on the province of residence. Another way to incorporate a spatial effect is based on the geographical coordinates of the cluster or household of residence. However, due to confidentiality reasons, these coordinates were not available for the data. While the structured provincial-level spatial effect did not have a statistically significant effect on the log-odds of school attendance, the inclusion of this effect reduced the model's AIC. Furthermore, the mapping of this effect revealed a varying log-odds of school attendance, where learners residing in provinces in the Eastern part of the country had an increased log-odds of attending school, and those in provinces in the Western part of the country had a decreased log-odds of attending school. This result is important for policy makers as these provinces should

be targeted for action. The significance of the provincial-level random effect (the unstructured spatial effect) suggests that there are province-specific factors that are primarily responsible for the spatial variation in the likelihood of school attendance of a learner, as compared to factors that are common among learners in neighbouring provinces.

To the best of our knowledge, this is the first nationally representative study assessing the factors that affect school attendance in South Africa. However, the study is not without limitations. Some key information regarding non-attendance of a learner at school was missing from the data. Such as the age at which they stopped attending school, the reason why they stopped attending as well as whether they intend on going back to school. All of this information would have been insightful into further understanding the school attendance rates. In addition, the data was based on a cross sectional design, thus no causal relationship between the variables and school attendance can be established. A future direction of this study includes making use of data from the next SAGHS, where the change in the school attendance rates can be assessed. In particular, making use of spatial-temporal modelling to investigate how the school attendance rates vary over the provinces of residence as well as over time.

References

Alan Agresti (2007). *An Introduction to Categorical Data Analysis*. Department of Statistics University of Florida, Gainesville, Florida, 2nd ed.

Anderson, D. M. (2014). In school and out of trouble? the minimum dropout age and juvenile crime. *The Review of Economics and Statistics*, *96*(2), 318–331.

Archer, K., & Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Strata Journal*, *6*(1), 97–105.

Archer, K., Lemeshow, S., & Hosmer, D. (2007). Goodness-of-fit test for logistic regression models when data are collected using a complex sampling design. *Computational Statistics and Data Analysis*, *51*(9), 4450–4464.

Balfanz, R., & Byrnes, V. (2012). The Importance of Being in School. *A Report on Absenteeism in the Nation's Public Schools*, (2).

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279–292.

Campbell, B. (2006). Why Education is Important. https://pdf4pro.com/view/why-education-is-important-green-bay-packers-59160e.html. [Online; accessed June 2019].

Department of Basic Education (2016). The Introduction of a Three Stream Model into the Basic Education Sector. https://equaleducation.org.za/wp-content/uploads/2018/05/Three-Stream-Report.pdf.

Department of Education, South Africa (2014). Medium Term Strategic Framework 2014-2019. https://www.gcis.gov.za/sites/www.gcis.gov.za/files/docs/resourcecentre/multimedia/mtsf.

Eilers, P. H. C., & Marx, B. D. (1994). Flexible Smoothing with B-splines and Penalties. *11*(2), 89–121.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statist Sci*, *11*(2), 89–121.

Elder, J. A. (1996). *Development of Quasi-Likelihood Techniques for the Analysis of Pseudo-Proportional Data*. Ph.D. thesis, Virginia Commonwealth University.

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat Sin*, *14*, 731–761.

Graubard, B., Kom, E., & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of time-scale. *American Journal of Epidemiology*, *145*(1), 72–80.

Greenacre, M. (1993). *Correspondence Analysis in practice*. Barcelona, Spain.

Greenacre, M. (2017). *Correspondence Analysis in Practice*. Chapman & Hall Interdisciplinary Statistics Series, 3 ed.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Statistics in the Social and Behavioral Sciences Series. Chapman & Hall/CRC. Taylor & Francis Group, LLC.

Hosmer, D., & Lemeshow, S. (1980). Goodness of fit tests for multiple logistic regression model. *Communications in Statistics*, *9*(10), 1043–1069.

Hosmer Jr, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 3rd ed.

Jiang, J. (2001). A nonstandard χ^2-test with application to generalized linear model diagnostics. *Statistics and probability letters*, *53*(1), 101–109.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Science and Business Media, LLC.

Khangar, K. (2017). Multiple Correspondence Analysis and its applications. *Electronic Journal of Applied Statistical Analysis*.

Kincaid, C. (2019). Guidelines for Selecting the Covariance Structure in Mixed Model Analysis. http://www.comsyssas.com.

Kneib, T., Müller, J., & Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environ Ecol Stat*, *15*, 343–364.

Koledade, O. (2008). Differentials in school attendance in South Africa. *A household situational analysis across the provinces*, (1), 9–99.

McCulloch, C. E., Searle, S. R., et al. (2001). *Generalized, Linear and Mixed Models.*. Wiley, New York.

McCulloch, P., & Neuhaus, J. (2005). *Generalized Linear Mixed Models*. John Wiley & Sons, Inc.

Meador, D. (2017). Why School Attendance Matters and Strategies to Improve It. https://www.thoughtco.com/why-school-attendance-matters. [Online; accessed June 2019].

National Department of Health (NDoH), Statistics South Africa (Stats SA), South African Medical Research Council (SAMRC), and ICF (2019). South Africa Demographic and Health Survey 2016. https://www.statssa.gov.za.

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the royal statistical society*, *A135*, 370–384.

Patterson, H., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545–554.

Phineas Reuckert (2019). Barriers to education around the world. https://www.globalcitizen.org/en/content/10-barriers-to-education-around-the-world-2/.

SAS Institute (2004). *SAS/STAT, 9.1, User's Guide.*, vol. 4. SAS publishers, North Carolina.

SAS Institute Inc. (2013). *Exploring SAS Enterprise Miner: Special Collection*. Cary, NC: SAS Institute Inc.

Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis*.

Stats-SA (2017). General Household Survey. http://www.statssa.gov.za/publications/P0318/P03182017.pdf.

UKessays (2018). The Role Of Education In A Country. https://www.ukessays.com/essays/education/the-role-of-education-in-a-country-education-essay.php. [Online; accessed June 2019].

Valentin, H. A. . D. (2007). Multiple Correspondence Analysis . https://www.researchgate.net/publication/239542271.

Vittinghoff, E., Glidden, D., Shiboski, S., & McCulloch, C. (2011). *Regression methods in biostatistics: linear, logistic, survival and repeated measures models*. Springer Science & Business Media.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, *61*, 439–447.

Wood, S. N. (2017). *Generalized Additive Models : An Introduction with R SECOND EDITION*. Tylor and Fransis Group, 2 ed.

Zhang, D., & Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In D. Dunson

(Ed.) *Random Effect and Latent Variable Model Selection*, vol. 192 of *Lecture Notes in Statistics*, (pp. 19–36). Springer New York.

Appendix A

A.1 SAS and R-codes

The SAS and R codes used in the final models for analysis of GHS data:

Final Correspondence Analysis

```
*Perform Multiple Correspondence Analysis*
proc corresp mca observed data=library.filter5-new;
tables race prov q11relsh1 q13aFath1 q14aMoth1;
run;
```

Final Survey Logistic Regression

```
proc surveylogistic data=library.filter5-new;
*Perform Survey Logistic*
class gender prov race q11relsh1 Q12aMARST1 Q13aFATH1(ref='No')
Q14aMOTH1(ref='No') Q89bMain1(ref='Other') totmhinc1 / param=reference;
model q110atte(event='Yes') = age*q11relsh1/*(continuous and Categorical)*/
age*q13afath1/*(continuous and categorical)*/ prov*hholdsz/*(categorical and
Continuous)*/ age prov gender race q11relsh1
q13aFATH1 Q14aMOTH1 WI-Zscore Q89bMain1 hholdsz totmhinc1 / df=infinity ;
strata stratum;
```

```
cluster PSU;
weight house.wgt;
run;
```

Final Generalized Linear Mixed Model

```
*Perform GLMM model*;
proc glimmix data=library.filter5-new method=laplace;
class UqNr PSU gender(ref='Female') prov race q11relsh1 Q13aFATH1(ref='No')
Q14aMOTH1(ref='No') Q89bMain1(ref='Other') totmhinc1;
model q110atte(event='Yes') =age*q11relsh1 age*q13afath1 prov*hholdsz age prov
gender race q11relsh1 q13aFATH1
Q14aMOTH1 WI-Zscore Q89bMain1
hholdsz totmhinc1 / dist=binary link=logit s;
random intercept /subject=PSU type=VC;
covtest "No G-side effects" zeroG;
run;
```

Final Generalized Additive Mixed Model

```
*Perform GAMM model*;
spat3=gam(Q110ATTE WI-Zscore+factor(q11relsh)+factor(q13aFath)+factor(q14aMoth)+
factor(q89bMain)+factor(totmhinc)+factor(Race)+factor(Gender)+s(Prov-ID,bs="mrf",xt=xt)+
s(Prov-Random,bs="re")+s(Age)+s(hholdsz), family = binomial, data = FILTER5.NEW2,
method="REML")
summary(spat3)
AIC(spat3)
plot(spat3,shade=TRUE)
abline(h = 0, lty = 2)
plot(spat3, pages=1 , scale = F, shade = T)
```