Modelling Poverty in Zimbabwe Based on the Demographic Health Survey Dataset Using GLMs and GAMMs.

By

Precious Mtshali

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: Prof Shaun Ramroop

Thesis Co-supervisor: Prof Henry Mwambi

School of Mathematics, Statistics and Computer Science

Pietermaritzburg Campus

2020

Declaration

I, Precious Mtshali declare that this thesis title *Modelling Poverty in Zimbabwe Using Demographic Healthy Survey Dataset* is my own. I confirm that:

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

[redacted] 2020-11-30

Precious Mtshali (Student) Date

[redacted] 2020-11-30

Prof Shaun Ramroop (Supervisor) Date

Prof Henry Mwambi (Co-supervisor) Date

Disclaimer

This document describes work undertaken as part of a master's degree of study at the University of KwaZulu-Natal (UKZN), Pietermaritzburg Campus under the supervision of Prof Shaun Ramroop and Prof Henry Mwambi.

I, Precious Mtshali, declare that this thesis is my own. It has not been submitted to any other university.

Acknowledgement

I would first like to thank the one above all, God, who makes everything possible for me. If it was not or his Love and Mercy, I would not be here today.

I would like to express my deeply held gratitude to Prof Shaun Ramroop, my supervisor, who gave up all his time to help me academically, and more importantly, for his supervisory style, which taught me to be independent and instilled self-belief. I also thank my co-supervisor, Prof Henry Mwambi, for his support throughout my research. I am very honoured to be under their guidance.

I would also like to thank Dr Habyarimana Faustin and all the UKZN statistics lecturers for all the support they have given me over the years. Lastly, I would like to thank my family and friends for their love, support and belief in me. The research was financially supported by Deltas Sub-Saharan Africa Consortium for Advanced Biostatistics Training (SSACAB) and the South African Statistical Association (SASA) scholarships.

Abstract

Zimbabwe has been in a state of political, economic, and social crisis for the past 15 years. In 2004, 80% of Zimbabweans were living below the national poverty line. By January 2009, only 6% of the population held jobs in the formal sector. Living in poverty may lead to stressful conditions that are linked to poor mental health problems in adults and developmental issues in children. This study investigates the risk factors that affect poverty status in Zimbabwe and makes recommendations for current policy on poverty, using statistical models such as generalized linear models (GLMs) and generalized additive mixed models (GAMMs). This study makes use of the Zimbabwe 2015 Demographic and Healthy Survey Dataset (DHS). The index was created using 29 variables questions from a principal component analysis. The first component was taken and the factor score was used. There was a cutoff below the median and above the median. Hence, the dichotomous response variable was socioeconomic status (SES) (1=Poor, 2=Not poor).The DHS data has explanatory variables such as the level of education, sex of the household head and age of the household head, size of the household head, and place of residence and sex of the household head. The results in both models (GLMs and GLMMs) reveal that these demographic factors are key determinants of poverty of households in Zimbabwe. This study demonstrates that the government of Zimbabwe needs to pay attention and intervene by looking into the demographic factors that affect poverty status.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The definition of poverty is very complex. A definition is difficult to formulate because poverty means different things to different people. Some people may define poverty as a lack of income resulting in the absence of a car or refrigerator, while others may define it as a lack of basic service, formal housing, and employment. According to the Oxford English Dictionary (1989), the adjective “poor “means “lacking adequate money or means to live comfortably”. The noun “poverty” is defined as “the state of being poor” and as a “want of the necessities of life”. The World Bank defines poverty as “the inability to attain a minimum standard of living” (World Bank 1990:26). This definition comprises low level of earnings, inaccessible healthcare facility, poor hygienic condition, lack of portable drinking water, high level of illiteracy rate, poor security and protection from preventable crime among others.

According to Teal (2011) there are four main causes of poverty:

1. Poor governance: Governance does not solely reflect a government, but also the civil societies, networks or markets that exercise power over the management of a country’s social and economic resources for development. The high rates of poverty are usually found within countries with corrupt leaders, weak state institutions and no rule of law.

2. Environmental degradation: This can have huge impact on poverty rates and well-being of people. If the resources are depleted due to climate change, natural disasters and deforestation, then citizens are more likely to be living in poverty. The environmental problems can lead to shortage of food, water and materials for housing as well as other essential resources. Without these above items, the poor people will not only remain poor, but it also increases their chances of premature death (Teal, 2011).

3. Discrimination, racism and prejudice: These all fit together to constitute a prime cause of poverty. This problem seen all over the world, even in the United States, where people are

treated unequal because of a person's skin color and religion. Those who discriminated against often do not get similar opportunities and benefits as the rest of the country.

4. Lack of employment opportunities: The high rate of unemployment is the main cause of poverty in the world (Teal, 2011).

Several theories have been put forward to explain the causes of poverty. Theoretical literature categorizes poverty as caused by individual deficiencies, cultural belief systems, economic, social and political distortions, geographical disparities, and cumulative and cyclical interdependencies (Bradshaw, 2007). Individual deficiency theorists blame poverty on the poor, claiming that poor people are poor because they are lazy (Bradshaw, 2007).

Poverty may lead to stressful conditions that are linked to poor mental health problems in adults and developmental issues in children. Most of the disease burden in low-income countries finds its roots in the consequences of poverty, such as poor nutrition, indoor air pollution and lack of access to proper sanitation and health education. The WHO estimates that diseases associated with poverty account for 45 per cent of the disease burden in the poorest countries (WHO 2002). Diseases such as TB, Malaria, Diarrhoea and HIV/AIDS.

Other studies shows that poverty increased the HIV spread (Whiteside 2002). There is a distinct relationship between poverty and communicable disease epidemics. According to Stillwaggon (2000) Study shows that HIV prevalence is highly correlated with falling calorie consumption, falling protein consumption, unequal distribution of income and other variables conventionally associated with susceptibility to infectious disease. The causal chain runs from macro-factors, which result in poverty through the community, household and individual, into the capacity of the individual's immune system.

Poverty is a world phenomenon even though is endemic in developing countries than the developed world. Records have it that seventy-five per cent of the world's poorest countries are located in Africa. In the last 30 years, extreme poverty incidence globally has decreased (from 40% to below 20%) but has little effect in African countries. Currently, in sub-Sahara Africa, more than 40% of people live in extreme poverty (World Bank, 2004).

1.1 Country Background

The Republic of Zimbabwe is located in Southern Africa between the Zambezi and Limpopo rivers; it is in an entirely landlocked country. The country is bounded by South Africa to the south, Zambia to the northwest, Botswana to the southwest and Mozambique to the east. The country covers 150 871 square miles (Ayad *et al.*, 1997). It is located between the Tropic of Cancer and Tropic of Capricorn. Zimbabwe is a country known for its dramatic landscape and diverse wildlife, much of it within parks, reserves and safari areas. On the Zambezi River, the Victoria Falls makes a thundering 108 meters drop into the narrow Batoka Gorge, where there are white-water rafting and bungee jumping activities. According to statistics in 2017, the population of Zimbabwe was estimated to be 16 53 million people. The map below (Figure 1.1) shows the location of Zimbabwe in Africa.

Figure 1.1 The location of Zimbabwe in Africa

Source: https://www.worldatlas.com/webimage/countrys/eu.htm

Zimbabwe has been in a state of political, economic, and social crisis for the past 15 years. In 2004, 80% of Zimbabweans were living below the national poverty line. By January 2009, only 6% of the population held jobs in the formal sector World Bank (2009). According to World Bank

(2014) Zimbabwe is one of the countries in Africa with a high level of poverty, and the number of poor people has increased in recent years by 1.3 million, with the rate of unemployment rising to approximately 11.7 million unemployed people. A high incidence of unemployment is among the key distinguishing features of poverty. The economy cannot generate enough opportunities to meet the needs of the labour force leading to increased poverty.

According to Zimbabwe poverty report 2017, poor people are identified as those who are facing consumption shortfalls. However, poverty is not a single economic condition and goes beyond consumption deficit. Poverty in Zimbabwe remains a persistent problem. The main causes of poverty in Zimbabwe have been identified as a lack of sufficient credit, infrastructure and social service. The last assessment by the World Bank (2007b) for Zimbabwe and more recent World Bank report (World Bank 2011, 2012a, 2013d) flagged the sluggish response towards poverty as a concern. The assessment for Zimbabwe (World Bank, 2007) revealed a stagnant level of poverty at around 33–36% between 2001 and 2007. The basic need and extreme poverty headcount rates for the Zimbabwe were 28.2 per cent and 9.7 per cent in 2011 and 2012, respectively. The headcount rates are based on the official National Bureau of Statistics (NBS) based on basic needs and food poverty line.

High mortality rates and corresponding low life expectancy are important dimensions of poverty. Poverty-related diseases cause a higher level of mortality in low-income countries than high-income countries (Stevens 2014). In 1999, infant mortality was estimated at 99 per 1000 live births and under-five mortality was 158 per 100 live births in Zimbabwe. The diseases that kill infants and under-fives are malaria, anemia, and pneumonia (WHO 1999). In 2009, 50% of the population between the ages of 15 and 49 were infected with HIV/AIDS and there are over one million AIDS orphans. This is consisted to what was reported by the WHO in relation to the diseases, that related to poverty.

Zimbabwe is also facing the challenge of poor social services such as the lack of clean drinking water (Estache, 2017). Statistics have estimated that 68% of the urban population have at least some kind of access to piped water, but less than half obtain 24 hours' service. However, about

45% of rural areas have access to safe water and about 30% do not. In 1991/1992, about 53% of the population was using unprotected water sources and the percentage of people infected with chorale and other waterborne diseases was extremely higher.

Zimbabwe is also currently experiencing out-migration of young people from low productivity agriculture to urban informal service sectors, but where productivity is just as low. Unemployment is high and growing rapidly, especially in the urban areas and among the youth. According to the Zimbabwe Poverty Atlas (2015), the overall unemployment rate for young people, aged 15–34 was 15.3 per cent, while that for young women was much higher, at 20.3 per cent, than that for young men, at 9.8 per cent. The causes or triggers of migration in Zimbabwe appear to be associated strongly with poverty (Dzingirai et al, 2014). The literature suggests that migrants tend to be those who are no longer employed as a result of closure of industries (Raftopolous, 2011) and are people living on less than a dollar a day, (Bracking and Sachikonye 2006, Raftoplous 2011). Furthermore, migrants are drawn from households whose consumption expenditure per capita is below the food poverty line. Poverty is indeed severe in Zimbabwe: the United Nations Country Office report (UN, 2014) suggests that almost 80% of the rural population is poor, compared to just under 40% in urban areas.

According to Chinake (1997), reports that Africa can escape poverty if agricultural development is accorded a priority in policy. Making Effective agriculture programs to improve storage schemes and increase local purchase need to pay more attention to rural industrialization. The Zimbabwean government is in the process of carrying out an assessment of poverty in the country, building on various studies already underway by UNICEF and the World Bank. The World Bank has set a goal to end extreme poverty by 2030 and to promote shared prosperity in every society and it is examining the feasibility of these objectives for sub-Saharan Africa, however each country Sub-Saharan need to supplement such international efforts with its own government policies to mitigate the afflictions of poverty in society.

1.2 Aims and Objectives

The aims and objectives of the research are two fold:

- The first objective is to model the poverty in Zimbabwe by creating a dichotomous poverty index and to investigate which socio-demographic factors are related to poverty using statistical models such as generalized linear models and generalized additive mixed models.
- The second objective is to make recommendations to current policy on poverty based on the results.

1.3 Literature Review

The skill of modelling poverty seem to be preoccupied in getting the best criteria for the judgment of the poverty status of individuals. Rouband and Razafindrakoto (2003) assert that there is link of the objective and subjective poverty measures and more argue that the various forms of poverty are not reducible one against the other. Apart from being, obsessed with monetary approach for measuring poverty there has been a growing literature which tries to come up with an index of multidimensional poverty facet (Bourguignon and Chakravarty, 2003). However there is little conclusion so far and as Kanbur and Squire (1999) argue there is no material difference in the number of poor identified as poor by employing different approaches. This seems to be convincing for at least the hard core poor where they are poor are in every dimension. Moreover after comparing different definitions of poverty and their implication to poverty modelling (Rouband and Razafindrakoto, 2003) argue that the traditional approach of monetary approach to measurement of poverty seems justified as it is the one most correlated with the other subjective measures . The devil is not on the usage of money metric unit for the determination of absolute poverty line rather on the mechanism employed for the derivation of such a line (Ravallion, 1996).The debate on the definition and measurement of poverty is really far from settled (Ravallion, 1996 and Laderchi, R.C. et al, 2003).

Several studies have been conducted on poverty in nearly every country in the world. The study that was conducted by Eyob. F and Mark H (2004) uses micro level data from Eritrean Household Income and Expenditure survey 1996-97 to examine the determinants of poverty in Eritrea. The DOGEV model was used estimated using poverty index as a dependent variable and explanatory variables (demographic variables, community variables, labour force variables, remittance, schooling, access to social services by controlling for regional differences). It was shown that the DOGEV is an attractive model from class of discrete choice models for modelling determinants of poverty across poverty categories. The study presents evidence of captivity of households in poverty in Eritrea. These captivities may be explained by demand factors such as occupation and number of hours worked or some social and behavioral problems. The result shown that the education impacts welfare differently across poverty categories and there are pockets of poverty in the educated population sub group. Effect of household size is not the same across poverty categories. Contrary to the evidence in the literature, the relationship between age and probability of being poor was found to be convex to the origin. Regional unemployment was found to be positively associated with poverty. Remittances, house ownership and access to sewage and sanitation facilities were found to be highly negatively related to poverty. The study also finds out that there is captivity in poverty category and a significant correlation between poverty orderings which renders usage of standard multinomial/ordered logit in poverty analysis less defensible.

The literature review revealed that a number of previous studies used these variables, Filmer and Prichett's (2001), method was followed. A study was conducted on construct indices of living standards in rural Bangladesh that could be useful to study health outcomes or identify target populations for poverty-alleviation programmes (Snaebjorn, G. Alain, BL. *et al*. 2010). The indices were constructed using principal component analysis of data on household assets and house construction materials (Filmer and Pritchett, 2001). Their robustness and use was tested and found to be internally consistent and correlated with maternal and infant health, nutritional and demographic indicators, and infant mortality. Indices derived from 9 or 10 household asset variables performed well; little was gained by adding more variables but problems emerged if fewer variables were used. A ranking of the most informative assets from this rural, South Asian

context is provided. Living standards consistently and significantly improved over the six-year study period. It was concluded that simple household socioeconomic data, collected under field conditions, could be used for constructing reliable and useful indices of living standards in rural South Asian communities that can assist in the assessment of health, quality of life, and capabilities of households and their members.

The main outcome of poverty identified by both these studies include demographic factors such as age, household size, geographical location, education, employment, gender of household head, environmental factors such as drought, remittances, and asset ownership, among others (Hoddinott, 2006). The study by Pindiriri, (2015) identified drought as one of the major factors affecting livelihoods in Zimbabwe. The two most recent studies on the outcomes of poverty in Zimbabwe are by Sakuhunni (2011) and Manjengwa et al., (2012). They both found primary education to be a significant determinant of poverty, but higher education was insignificant. Furthermore, the established primary education was a significant determinant of poverty in Zimbabwe (Manjengwa et al., 2012). The study applied binary dependent variable models (logit, probit and tobit) (Asogwa et al., 2012).

According to Pindiriri (2015), investigating of the determinants of poverty in Zimbabwe. A two-stage least squares method was employed using the myeloid zinc finger (MZF) dataset. In the study it was found that poverty in Zimbabwe is primarily caused by low household income, low educational achievement of the household head, bigger household size, and household location. The study recommended increasing family planning campaigns, supporting education for the poor, creating employment through implementing investment-friendly policies, and establishing land redistribution policies targeting the poor.

There was a study, that was conducted on agricultural growth, poverty and nutrition in Tanzania (Pauw and Thurlow, 2011). Using a regionalized, recursive dynamic computable general equilibrium and micro simulation model, they found that economic growth does not appear to have significantly improved poverty and nutrition outcomes in Tanzania. The results indicate some inconsistency between recent growth and poverty measurements in Tanzania. The study

also found that the structure of economic growth might have constrained the rate of poverty reduction. Agricultural growth in the country has been driven by large-scale farmers who are less likely to be poor and has been concentrated on crops grown in specific regions of the country. Slow expansion of food crops and livestock also explain the weak relationship between agricultural growth and nutrition outcomes. The findings show that accelerating agricultural growth, particularly in maize, strengthens the growth-poverty relationship and enhances households' caloric availability, while also contributing significantly to growth itself.

According to Jewkes (2002), in a study on intimate partner violence, it was found that poverty and associated stress are key contributors. Although violence occurs in all social groups, it is more frequent and severe in lower income groups across diverse settings such as in the USA, Nicaragua and India(Ellsberg *et al.*, 1999). To the extent that poverty is inherently stressful, it has been argued that intimate partner violence may result from stress (Cosner ,1967), and also that poorer men have fewer resources to reduce stress (Steinmetz, 1987). In Jewkes' study it was found that in South Africa, physical violence was not associated in the expected way with Indicators of socioeconomic status, counting ownership of household goods, gender, occupation and unemployment (Jewkes *et al.*, 2002).

According to Kohn *et al*., (1991), it is their diversity that makes the poor as category even harder to study. Carter and May (1999) conducted a study on poverty, livelihood and class in rural South Africa, a country that ranks as an upper-middle income country with a per-capital GDP of some $3000, but where the majority live also in poverty. Non-parametric regression methods were used to estimate and graphically explore the nature of the likelihood mapping between endowments and real incomes. This study found that stratification of the rural population into livelihood classes based on shared livelihood strategies reveals that economic well-being differs systematically across livelihood classes. It suggests that the poor and not poor gain their livelihood from rather distinctive portfolios of activities and enjoy rather different sets of economic endowment and social claims.

Chapter 2

Data Distribution and Exploratory Analysis

2.1 Data Distribution and Methods

2.1.1 Data Distribution

This study uses part of the data from the Zimbabwe Demographic Healthy Survey (DHS) for the year 2015. The 2015 Zimbabwe Demographic and Health Survey (2015 ZDHS) presents the major findings of a nationally representative survey with a sample of more than 11,000 household. 2015 ZDHS used three questionnaires such as a household questionnaire, a questionnaire for individual women aged from 15 to 49 and a questionnaire for individual men aged from 15 to 54. Data collection was carried out from July 2015 to December 2015. The questionnaires were administered throughout the country among the selected household and selected men and women.

The 2015 ZDHS sample was designed to yield representative information for most indicators for the country as a whole, for urban and rural areas, and for each of Zimbabwe's ten provinces: Manicaland, Mashonaland Central, Mashonaland East, Mashonaland West, Matabeleland North, Matabeleland South, Midlands, Masvingo, Harare, and Bulawayo. The 2012 Zimbabwe Population Census was used as the sampling frame for the 2015 ZDHS.

The 2015 ZDHS sample was selected with a stratified, two-stage cluster design, with census enumeration areas (EAs) as the sampling units for the first stage. The 2015 ZDHS sample included 400 EAs—166 in urban areas and 234 in rural areas. The principle reason of non- response among both eligible men and women was the failure to find them at home after repeated visits to the household.

2.1.2 Methods

The formalization of the relationship between the outcome and independent variables was done using two modelling techniques namely:

- Generalized Linear Models (GLMs)
- Generalized Additive Mixed Models (GAMMs)

The Statistical Package for Social Science (SPSS), Statistical Analysis System (SAS) and R were used to fit the statistical models. The results were then interpreted and are discussed. The explanatory data analysis (EDA) approach was used to summarize the main characteristics of the cross-tabulation without using statistical models. The surveyed variables analyzed are sex of head of household, age of the household head, region, residence, highest education levels of household head and the number in the household. In this study, Filmer and Prichett's (2001) method was followed. The index was created using 29 variables questions from a principal component analysis. The first component was taken and the factor score was used. There was a cutoff below the median and above the median. Hence, the response variable was socioeconomic status (1=Poor, 2=Not poor).

2.2 Cross-tabulation Analysis

Cross-tabulation is one of the most useful analysis tools and it is a mainstay of the marketing industry. This simple frequency analysis and cross-tabulation analysis account for more than 90% of all research analysis (A. Aprameya, 2016). The cross-tabulation analysis is the most often used to analyze categorical data. The cross-tabulation tables provide a wealth of information about the relationship between dependent and independent variables. This type of analysis has its own unique language, using terms such as banner, stabs, chi-square statistics and expected values.

2.3 Chi-square Test Statistics

The chi-square statistic is a primary statistic used for testing the statistical significance of the cross-tabulation table. Chi-square examination whether or not the two variables are independent. Therefore, the null hypothesis was not rejected if the variables are independent

meaning that there is no relationship between the response and explanatory variables. If the variables are related, then the results of the statistical test will be significant. Therefore, we reject the null hypothesis meaning so that we can state there is some relationship between the variables. In our case, the chi-square test of independent was used to examine the relationship between explanatory variables and socioeconomic status with the use of cross-tabulation.

This represents the cross-tabulation of socioeconomic status (SES) versus demographic variables such as type of place of the household head, region, sex of the household head, household members and age of household head in Zimbabwe. Tables and bar graphs are presented in this sub section for interpretation of socioeconomic status versus explanatory variables. Both tables and bar graphs were obtained from SPSS. Also obtained were chi-square tables to check the possibility of including certain variables in the statistical models and also to check the association between socioeconomic status and demographic variables. Under the null hypothesis, H_0: the rows and columns are independent against H_1: where the rows and columns are dependent.

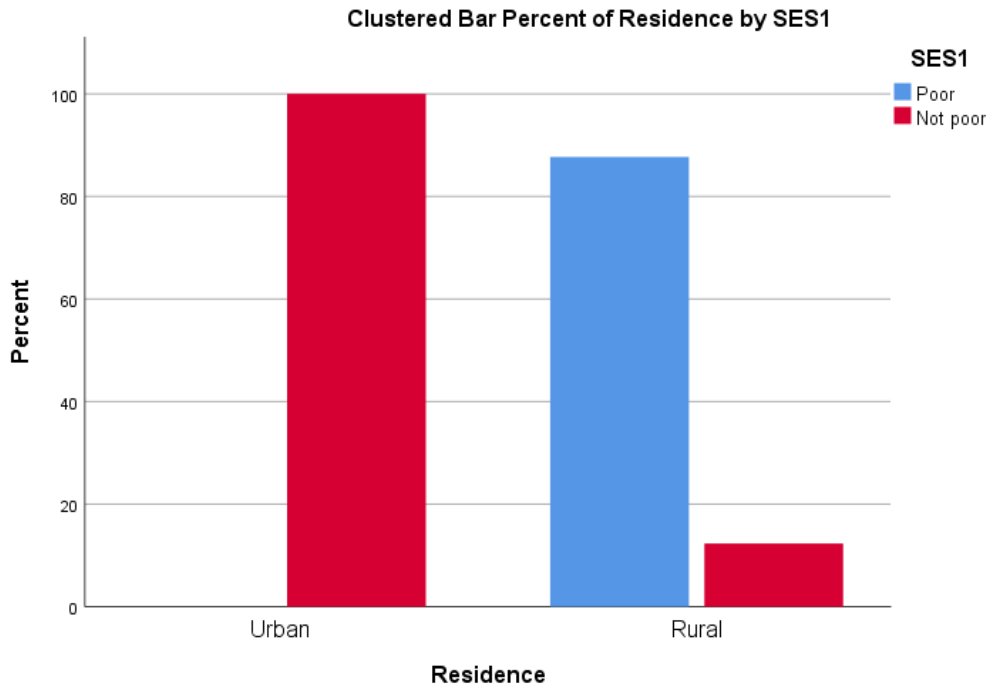

Figure 2.1 Clustered bar graph showing percentage of household of SES in rural and urban areas

The distribution of type of place and socio economic status in rural areas and poor household is Approximate to 99% and not poor is $\sim 0.1\%$. Figure 2.1 shows that there is higher percentage of poor households in rural areas. Also shown that all people who live in urban areas can afford their needs, as the percentage of not poor is approximate to 90%. Figure 2.1 also confirms that lot of people who are staying in rural areas are poor, as the percentage of poor is approximately to 88%.

Table 2.1 Cross-tabulation of the type of place of residence and socioeconomic status

Variables	Poor	Not poor
Urban	0.1%	99%
Rural	87.7%	12.3%
Total	61.4%	38.6%

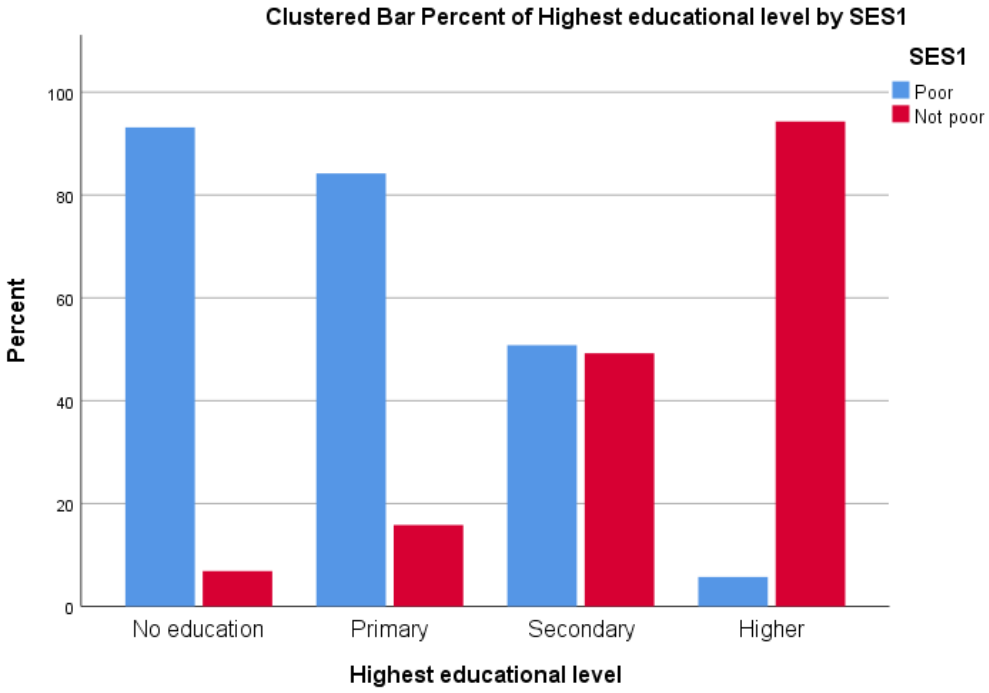

Figure 2.2 The distribution of poor in educational levels

From Figure 2.2 above it can be seen that there is a higher percentage of poor people who are uneducated (≈90%) and a lower percentage of uneducated people who are not poor (≈10%).In primary education level there is also a higher percentage of poor, approximately ≈82%, compared to those who are not poor. The secondary education level can be seen to have almost the same percentage of poor and not poor with not much difference between poor and not poor people. Figure 2.2 also shows that there is a higher percentage (≈95%) of not poor people who are educated or who are at the higher education level in Zimbabwe and there is very small percentage of poor at this education level. Figure 2.2 confirms that as level of education increases the poverty status decreases.

Table 2.2 Cross-tabulation of the highest education level and socioeconomic status

Variables	Poor	Not poor
No education; Preschool	93.1%	6.9%
Primary	84.2%	15.8%
Secondary	50.8%	49.2%
Higher	5.7%	94.3%
Total	61.4%	38.7%

Table 2.2 confirms the exact percentages that Figure 2.2 above approximates. There is high percentage of poor uneducated people (93.1%), followed by a high percentage (84.2%) at primary level. Table 2.2 also confirms that at secondary level the percentage of poor and not poor is not much different (1.6%); it's almost equal. At the higher education, level there is a small percentage of poor (5.7%) compared to not poor (94.3%). This suggests that almost everyone who has a higher education level has a higher chance of not being poor.

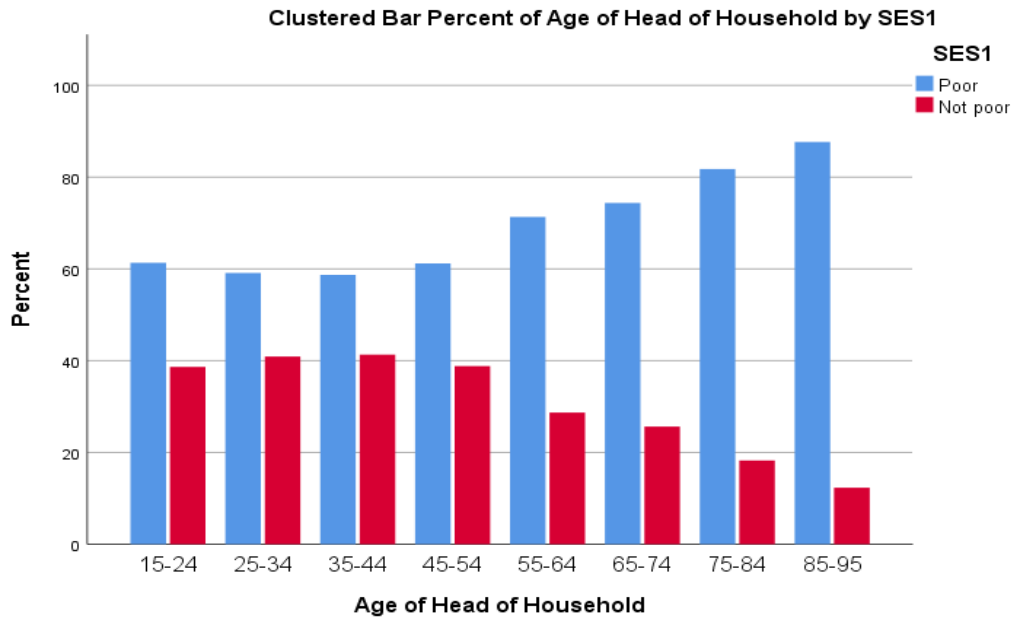

Figure 2.3 The distribution of poor for the age of household head

The results in Figure 2.3 above indicate that from age 15 to 95 the percentage of poor increases, that is, older people have a greater chance of being poor. However, as age increases the percentage of not poor decreases.

Table 2.3 Cross-tabulation of age of the household head and SES

Variables	Poor	Not poor
15–24	61.3%	38.7%
25–34	59.1%	40.9%
35–44	58.7%	41.3%
45–54	61.2%	28.7%
55–64	71.3%	25.6%
65–74	74.4%	17.2%
75–84	81.8%	12.3%
85–95	87.7%	12.3%
Total	61.4%	38.6%

Table 2.3 confirms that there is higher percentage of poor old people, at the age of 85–95 year (87.7%), followed by the age of 75–84 years (81.8%). The percentage of poverty increases as age increases.

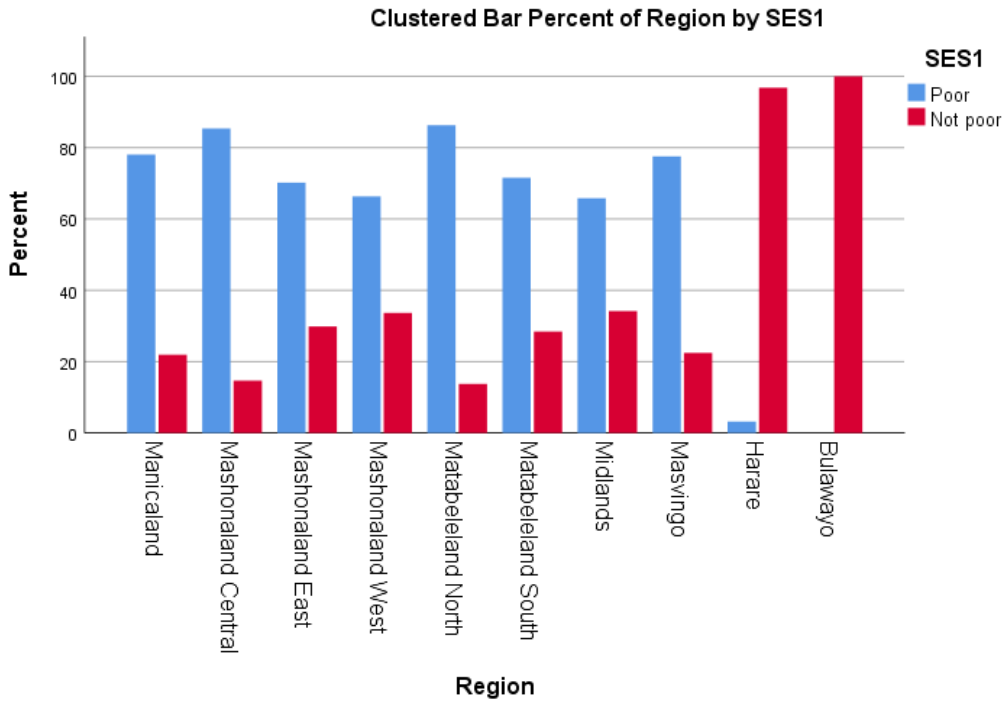

Figure 2.4 The distribution of poor among regions

Results in Figure 2.5 show that Mashonaland Central and Matabeleland North have a higher almost equal percentage of poor (≈85%) than other regions, followed by Manicaland and Masvingo (≈78%). Figure 2.5 also shows that Bulawayo does not seem to have any poor households, followed by Harare (≈5%). Thus, the percentage of not poor households is very high in the Bulawayo and Harare regions.

Table 2.3 Cross-tabulation of the regions and socioeconomic status

Variables	Poor	Not poor
Manicaland	78.1%	21.9%
Mashonaland Central	85.3%	14.7%
Mashonaland East	70.2%	29.8%
Mashonaland West	66.3%	33.7%
Matabeleland North	86.3%	13.7%
Matabeleland South	71.5%	28.5%
Midlands	65.8%	34.2%
Masvingo	77.5%	22.5 %
Harare	3.2%	96.8%
Bulawayo	0.1%	99.9%
Total	61.4%	38.6%

The results in Table 2.4 confirm that Matabeleland North and Mashonaland Central have the highest percentage of poor (86.3% and 85.3% respectively) compared to other regions in Zimbabwe. Manicaland and Masvingo follow these regions with a high percentage of poor (78.1% and 77. 5% respectively). Table 2.4 also confirms that there are no poor people or households in the Bulawayo region (0.1%), followed by Harare where there are very few poor people (3.2% poor), which tells us that a very high number of people in Bulawayo and Harare are wealthy or that they can afford their needs. The overall data for the regions indicate that there is higher percentage of poor compared to not poor.

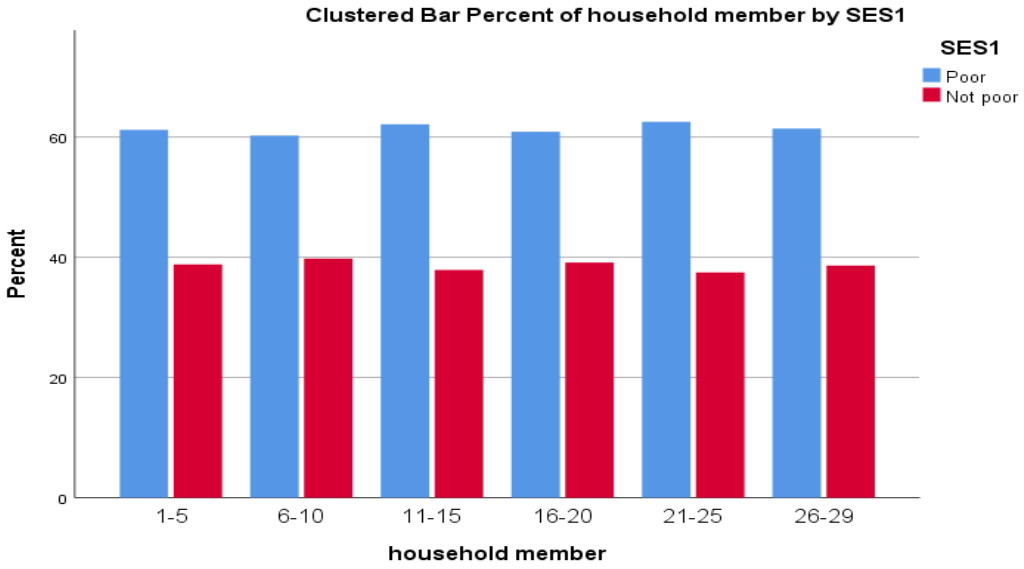

Figure 2.5 The distribution of poor for the size of household

The distribution of poverty according to the size of household shown in Figure 2.5 approximates a higher percentage of poor in all household membership sizes, with small differences between them. However, it can be seen that memberships of 11–15 and 21–25 have the highest percentage of poor (approximately≈63 %), followed by memberships of 1–5 and 26–29 (≈61%).

Table 2.4 Cross-tabulation of number of household members and socioeconomic status

Variables	Poor	Not poor
1–5	61.2%	38.8%
6–10	62.2%	39.8%
11–15	62.1%	37.9%
16–20	60.9%	39.1%
21–25	62.5%	37.8%
26–29	61.4%	38.6%
Total	61.4%	38.6%

Table 2.5 confirms the results of the bar graph in Figure 2.5 above, that the percentage of poor in all households is not much different.

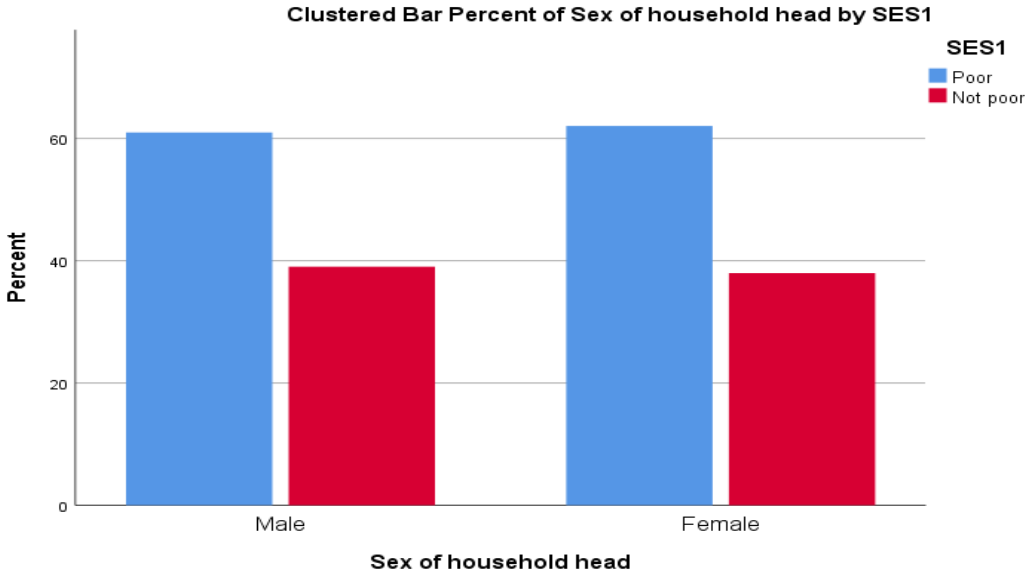

Figure 2.6 The distribution of poor and not poor among gender

Figure 2.6 shows the distribution of poor according to gender, indicating that the percentage of poor is almost equal for both genders, at approximately ≈61%, and the percentage of not poor in both genders is also equal (≈39%). This suggests that there is an equal chance of being either poor or not poor in both genders and also shows that there is a higher percentage of poor all households.

Table 2.5 Cross-tabulation of gender of the household and socioeconomic status

Variables	Poor	Not poor
Male	61.0%	39.0%
Female	62.0%	38.0%
Total	61.4%	38.6%

The results in Table 2.6 above confirm that the percentage of the poor between male and female is almost equal, with a difference of 1%, which means that females are 1% poorer than males.

Table 2.6 Chi-square test table.

Explanatory variables	x^2-value	df	p-value
Sex of head of household	164.155	1	0.00
Region	2073.384	9	0.00
Age of head of household	355.948	7	0.00
Number of household members	114.196	5	0.00
Type of place of residence	3342.585	1	0.00
Highest education level	418.468	3	0.00

Results in Table 2.7 indicate that all explanatory variables above have p-values which are less than 0.05. This implies that there is a significant relationship between all explanatory variables and poverty status.

2.4 Summary of Exploratory Data Analysis

The results show that the percentage of poor is higher than not poor in all variables. Using the chi-square test, it was also found that all the explanatory variables have p-values that are less than 0.05. Therefore, H_0 was rejected at 5% level of significance. All explanatory variables were found to be significant. This demonstrates that all explanatory variables have a significant relationship with response variables (poverty status).

Chapter 3

Generalized Linear Models

3.1 Introduction

The generalized linear models (GLMs) theory was formulated by (McCullagh and Nelder, 1989a) as a way of unifying various statistical models, including linear regression, logistic regression, and others. The GLMs are a family of important models for categorical as well as continuous responses in statistics. Thus, GLMs are defined as a flexible generalization of ordinary least squares regression. The class of GLMs is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions. However, the general linear model assuming the normal distributed errors is a special case of the GLM where the link function is the identity. Generalized linear models attempt to accommodate variance heterogeneity and asymmetric, non-normal behavior by offering a range of distribution types that cover at least the more common mean-variance relationships. They are useful for non-normal data, such as binary data. The GLMs are a large class of statistical models used for relating responses to line a combination of predictor variables, and they can also include interaction terms(Dobson and Barnett, 2018). The regression parameters are estimated by using the maximum likelihood method (McCullagh and Nelder, 1989b) and also using the iterated re-weighted the least squares (IRWLS) as the implementing algorithm, when the link function is the canonical link IRWLS reduces to the classical Newton-Raphson method.

3.2 The Model

The linear regression model was formulated by Gill (2001), as a matrix notation. The model was give as:

$$y_i = \boldsymbol{X_i\beta} + \boldsymbol{e_i} \qquad 3.1$$

where i takes value from $i = 1,2 \ldots n$

y is a dependent variable.

β is a vector of unknown parameter.

e_i is assumed to be the normal distribution with a mean of zero and constant variance σ^2 **.** $\boldsymbol{X_i}$ is a vector of k independent variable.

In a linear model Gill (2001), suggests an assumption need to be made which relaxes it.

The assumption made is related to the Gauss-Markov theorem, which is given as in the following regression:

$$y_i = \alpha + \boldsymbol{X_i\beta} + e_i \qquad 3.2$$

The assumptions are given as follows:

1. The relationship between each exploratory variable and the outcome variable is approximately linear in structure.
2. The residuals are independent with a mean zero and constant variance.
3. There is no correlation between any regression variables and exploratory variables (Gill, 2001).

Under the assumption that the linear is a special case GLM, we assume that a random variable, the Y_i, has a normal distribution with a mean u_i and variance σ^2 (Rodrìgues, 2001).

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

The expected value of Y_i is assumed to be a linear function of p-predictors that take on values $x_i' = (x_{1i}, x_{2i}, \ldots\ldots, x_{pi})$for i^{th}case, so we have

$$u_i = \boldsymbol{x_i\beta_i}$$

Where $\boldsymbol{\beta}$ is a vector of unknown parameters.

In the case of the linear model which is a special case of the GLMs the three components are:

1. A random component:

   The random component refers to the probability distribution of the random variable Y_i. In the case of the linear model variables Y_i are usually assumed to be independent normal distribution with a mean $E(y_i) = \mu_i$ and constant variance σ^2, thus the y_i are iid N(u_i, σ^2).

2. A systematic component:

   This component refers to the explanatory variables $(X_1, X_2, \dots X_K)$ as a combination of linear predictors(Nelder and Wedderburn, 1972). In other words covariates $X_j, j = 1,2,..,k$ combine with the linear coefficient to form the linear predictor η_i.

$$\eta_i = \alpha + \beta_{i1}X_{i1} + \beta_{i2}X_{i2}, \dots, \beta_{ip}X_{ip} \tag{3.3}$$

3. The link functions $g(u)$:

   The link function specifies the link between the random and the systematic components. It equates a function of the mean response μ_i to the linear predictor $\eta_i = \boldsymbol{X'_i\beta}$. The link functions g describes how the mean μ_i is related to linear predictor η_i. As shown in equation (3.4).

$$g(u_i) = \eta_i = \alpha + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \cdots + \beta_{ip}X_{ip} \tag{3.4}$$

3.3 The Canonical Link Functions

The link function can be defined as the process of linking a transformation of the mean response to the explanatory variables (O'Connell, 2006). Table 3.1 shows examples of link functions in GLMs.

Table 3.1 Table of common examples of link functions in generalized linear models and their inverses (Nelder and Beker, 1972).

Link Functions	$g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$log_e(\mu_i)$	$e^{-\eta_i}$
Log-Log	$log_e(-log_e(\mu_i))$	$\exp[\exp(\eta_i)]$
Logit	$log_e\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{1+e^{-\eta_i}}$
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Complementary log-log	$log_e[-log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

3.4 The Exponential Family

The exponential family consists of a set of distributions for both discrete and continuous random variables. The exponential family is also known as a general class of distribution that includes the normal distribution as a special case (Olsson, 2002). If we take a continuous random variable Y_i from a distribution that is a member of the exponential family, and it depends on the parameter θ,ϕ, then the probability density function (pdf) for Y_i can be expressed as:

$$f(y_i, \boldsymbol{\theta_i}, \phi) = \exp\left\{\frac{y_i\boldsymbol{\theta_i} - b(\boldsymbol{\theta_i})}{a(\phi)} + c(y_i, \phi)\right\} \qquad i = 1,2,\dots,n \tag{3.5}$$

where $a(\phi)$ and $b(\theta_i)$ are known functions, then $c(y_i, \phi)$ is some function of y_i and θ_i. The parameter ϕ is a dispersion parameter, and θ_i is canonical parameter. If y_i has a distribution in the exponential family, then the mean and variance are given as:

$$E(y_i) = \mu_i = b'(\boldsymbol{\theta_i}) \tag{3.6}$$

And

$$Var(y_i) = \sigma^2 = b''(\boldsymbol{\theta_i})$$

(3.7)

where $b'(\theta_i)$ is the first derivative and $b''(\theta_i)$ is the second derivative of $b(\theta_i)$ (Rodriguez, 2001). If general expression of the exponential distribution in terms of $a, b\ and\ \phi$, is:

$$f(y, \boldsymbol{\theta}, \phi) = \exp\left\{\frac{\boldsymbol{\theta} y - b(\boldsymbol{\theta})}{a(\phi)} + c(y, \phi)\right\} \tag{3.8}$$

then

$$\int f(y, \boldsymbol{\theta}, \phi) dy = 1$$

Differentiating both sides with respect to θ, then we get:

$$\frac{\partial}{\partial \theta}\left[\int \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\} dy\right] = 0 \tag{3.9}$$

$$\int \frac{\partial}{\partial \theta} \exp\left\{\frac{\boldsymbol{\theta} y - b(\boldsymbol{\theta})}{a(\phi)} + c(y, \phi)\right\} dy = 0$$

$$\int \left[\frac{y - b'(\theta)}{a(\phi)}\right] f(y, \theta, \phi) \mathrm{dy} = 0$$

$$\int \frac{y f(y,\theta,\phi)}{a(\phi)} dy - \int \frac{b\prime(\theta) f(y,\theta,\phi)}{a(\phi)} = 0 \tag{3.10}$$

$$\int \frac{y f(y, \boldsymbol{\theta}, \phi)}{a(\phi)} dy = \int \frac{b'(\theta) f(y, \boldsymbol{\theta}, \phi)}{a(\phi)}$$

$$\int y f(y, \boldsymbol{\theta}, \phi)\, dy = \int b'(\boldsymbol{\theta}) f(y, \boldsymbol{\theta}, \phi)$$

$$\int y f(y, \boldsymbol{\theta}, \phi)\, dy = b'(\theta) \int f(y, \boldsymbol{\theta}, \phi)\, dy$$

$$E(y) = b'(\boldsymbol{\theta}) \quad \text{Or} \quad \boldsymbol{\mu} = E(y) \tag{3.11}$$

Taking the second derivative with respect to $\boldsymbol{\theta}$, we have:

$$\int \left[\frac{y - b'(\boldsymbol{\theta})}{a(\phi)}\right] f(y, \boldsymbol{\theta}, \phi)\mathrm{dy} = 0$$

$$\int \left\{\left[\frac{y-b'(\theta)}{a(\phi)}\right]^2 f(y, \boldsymbol{\theta}, \phi) - \frac{b''(\theta)}{a(\phi)} f(y, \boldsymbol{\theta}, \phi)\right\} dy = 0, \tag{3.12}$$

$$\int \left[\frac{y - b'(\boldsymbol{\theta})}{a(\phi)}\right]^2 f(y, \boldsymbol{\theta}, \phi)dy = \frac{b''(\boldsymbol{\theta})}{a(\phi)} \int f(y, \boldsymbol{\theta}, \phi)\mathrm{dy},$$

$$\frac{1}{a(\phi)^2} \int [y - b'(\boldsymbol{\theta})]^2 f(y, \boldsymbol{\theta}, \phi)\mathrm{dy} = \frac{b''(\boldsymbol{\theta})}{a(\phi)},$$

$$\frac{Var(y)}{a(\phi)^2} = \frac{b''(\boldsymbol{a})}{a(\phi)}$$

$$Var(y) = a(\phi)b''(\boldsymbol{\theta}) \qquad \text{where } V(\boldsymbol{\mu}) = b''(\boldsymbol{\theta}) \tag{3.13}$$

We note that in general mean and variance are dependent since

$$\begin{aligned} Var(y) &= a(\phi)[b^{-1}(\boldsymbol{\mu_i})] \\ &= a(\phi)V(\boldsymbol{\mu}) \end{aligned} \tag{3.14}$$

The function $V(\boldsymbol{\mu})$ called the variance function. The function $b'^{-1}(.)$, which express $\boldsymbol{\theta}$ as a function of μ is the link function and $b'(.)$ is the inverse link function. The several distributions, which belong to this structure and for classification purposes we briefly relate to above

formulation to the Normal, Poisson, Binomial and Bernoulli distributions which all, fall under the exponential family of distributions.

For example, suppose that y is normally distributed with mean μ and variance σ^2(i.e $y \sim (\mu, \sigma)$). Then distribution is given by

$$f(y, \mu, \sigma^2) = \frac{1}{(2\boldsymbol{\pi}\boldsymbol{\sigma}^{\boldsymbol{2}})^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{\mu})^2\right)$$

$$= exp\left\{log\left(\frac{1}{(2\boldsymbol{\pi}\boldsymbol{\sigma}^{\boldsymbol{2}})^{\frac{1}{2}}}\right) exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{\mu})^2\right)\right\} \quad (3.15)$$

$$= exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{2\sigma} - \frac{1}{2}\log(2\boldsymbol{\pi}\sigma^2) - \frac{y^2}{2\sigma^2}\right\}$$

where $\boldsymbol{\theta} = \boldsymbol{\mu}, b(\boldsymbol{\theta}) = \frac{\theta^2}{2}$, $a(\phi) = \sigma^2$, $C(y, \phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$ and the canonical link is the density.

Therefore, the mean and the variance is given by

$$E(y) = b'(\boldsymbol{\theta}) = \boldsymbol{\mu}$$

and

$$Var(y) = a(\phi)Var(\boldsymbol{\mu}) = \sigma^2$$

which is independent on $\boldsymbol{\mu}$**.** The variance function is $V(\mu) = 1$, and the dispersion parameter is $a(\phi) = \sigma^2$.

For a Poisson distribution with mean μ (i.e $y \sim Poi(\boldsymbol{\mu})$. then distribution is given by

$$f(y, \boldsymbol{\mu}) = \frac{\exp(-\boldsymbol{\mu})\boldsymbol{\mu}^{\boldsymbol{y}}}{y!} \quad (3.16)$$

$$= exp\{ylog\boldsymbol{\mu} - \boldsymbol{\mu} - logy!\}$$

where $\boldsymbol{\theta} = \log \boldsymbol{\mu}$, $b(\theta) = \exp(\boldsymbol{\theta})$, $a(\phi) = 1$ and $C(y, \phi) = logy!$. There canonical link function is log link.

The mean and variance is given by

$$E(y) = b'(\boldsymbol{\theta}) = \boldsymbol{\mu}$$

and

$$Var(y) = a(\phi)Var(\boldsymbol{\mu}) = \boldsymbol{\mu}$$

Which depends on $\boldsymbol{\mu}$. The variance function is $Var(\boldsymbol{\mu}) = \boldsymbol{\mu}$ and the dispersion parameter is $a(\phi) = 1$.

For Binomial distribution with parameters $\boldsymbol{n}$ and $\boldsymbol{\pi}$ (i.e $y \sim Bin(n, \pi)$. The distribution is given as

$$f(y, \boldsymbol{n}, \boldsymbol{\sigma^2}) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

$$= exp\left\{log\left(\binom{n}{y} n^y (1-\pi)^{n-y}\right)\right\} \quad (3.17)$$

$$= exp\left\{ylog\ \left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) + nlog(1-\boldsymbol{\pi}) + log\binom{n}{y}\right\}$$

$\boldsymbol{\theta} = \left(\frac{\pi}{1-\pi}\right)$, $a(\phi) = 1$ and $b'(\theta) = nlog(1 + \exp(\boldsymbol{\theta}))$. The canonical link function is logit.

Thus

$$E(\boldsymbol{Y}) = b'(\theta) = n\left(\frac{\exp(\boldsymbol{\theta})}{1 + \exp(\boldsymbol{\theta})}\right) = n\pi$$

and

$$Var(\boldsymbol{Y}) = a(\phi)b''(\boldsymbol{\theta}) = n\left(\frac{\exp(\boldsymbol{\theta})}{(1+\exp(\theta))^2}\right)$$

where the variance function is $V(\mu) = \boldsymbol{n\pi}(1-\boldsymbol{\pi})$ and the dispersion parameter is $a(\phi) = 1$.

For Bernoulli distribution with mean π. The distribution is given by

$$\begin{aligned} f(y, \boldsymbol{\mu}) &= \pi^{y}(1-\boldsymbol{\pi})^{1-y} \\ &= \exp\{y log\boldsymbol{\pi} + (1-y)\log(1-\boldsymbol{\pi})\} \\ &= exp\left\{ylog\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} \end{aligned} \tag{3.18}$$

where $= log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right)$, $a(\phi) = 1$ and $b(\theta) = -\log(1-\boldsymbol{\pi}) = \log(1+\exp(\boldsymbol{\theta}))$,

since $\pi = \frac{\exp(\boldsymbol{\theta})}{(1+\exp(\boldsymbol{\theta}))}$ and the canonical link is logit link.

Then the mean and variance is given by

$$E(y) = b'(\boldsymbol{\theta}) = \boldsymbol{\pi}$$

and

$$Var(y) = \ a(\phi)Var(\boldsymbol{\mu}) = \boldsymbol{\pi}(1-\boldsymbol{\pi})$$

Which is dependent in $\boldsymbol{\mu}$. In this case, the variance function is $V(\mu) = \pi(1-\pi)$ and the dispersion parameter is $a(\phi) = 1$.

3.4.1 Components of a Generalized Linear Model

The following features consider the generalized linear regression model:

- **A random component:** this is component identifies the response, y_i and assumes a distribution that follows the exponential family. It is given as equation (3.8)
- **Systematic component:** this specifies the explanatory or predictor variables. The covariate is x_i combined linearly with the coefficients to form the linear predictor.

$$\boldsymbol{\eta_i} = \boldsymbol{X\beta} \tag{3.19}$$

where the i^{th} row of X is given by $x_i = (1, x_{i1}, \ldots . x_{ip})'$ with $x_{ij}, i = 1 \ldots n$ Equal to the value of j^{th} explanatory or predictor $j = 1 \ldots p$ and $\boldsymbol{\beta} = (\beta\_0, \ldots ., \beta\_1, \beta\_p)'$
Is a regression coefficient.

- **Link component**: this specifies the relationship between the mean of the random and systematic components: the linear predictor $X_i\beta$ =$\_i$ is a function of the mean parameters μ_i via a link function, $g(\mu_i)$.

$$\eta_i = g(\boldsymbol{\mu_i}) \qquad (3.20)$$
$$= x_i'\boldsymbol{\beta}$$

According to Davis (2002) and McCulloch and Neuhaus (2001), the $g(\mu_i)$ must be monotonic and a different function such as that

$$\eta_{\boldsymbol{i}} = g(\boldsymbol{\mu_i})$$

Thus

$$g(\mu_i) = \sum_j \beta_j x_{ij},\ i = 1 \ldots, N$$

Relating the linear predictor to the mean response follows

$$\mu = g^{-1}(\eta_i) = E(y)$$

We model a function of the mean as a combination of linear predictors. This function $g(.)$ is monotone, which means as the systematic part gets larger, μ gets larger too and again when the systematic part gets smaller, μ gets smaller too. The relationship between $E(y)$ and the Systematic part can be nonlinear.

Table 3.2 The summary for different model components of generalized linear models.

$y \backsim (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$	$Normal(\boldsymbol{\mu}, \boldsymbol{\sigma})$	$Poisson(\boldsymbol{\mu})$	$Binomial(\boldsymbol{n_i}, \boldsymbol{\pi_i})$	Bernoulli $(\pi, \pi(1-\pi)$
$E(y) = b'(\theta)$	$\theta = \mu$	$e^{\theta} = \mu$	$n\pi_i$	π
$Var(y) = b''(\theta)a(\phi)$	σ^2	$e^{\theta} = \mu$	$n\pi_i(1-\pi_1)$	$\pi(1-\pi)$
$b''(\theta)$	1	$e^{\theta} = \mu$	$n\pi_i(1-\pi_1)$	$\pi(1-\pi)$
$a(\phi)$	σ^2	1	1	1
$c(y, \phi)$	$-\frac{1}{2}\left[\frac{y}{\sigma^2} + \ln(2\pi\sigma^2)\right]$	$-\ln(y!)$	$\log(nCy)$	0
$Link\ g(.)$	$Identity$	Log	$Logit$	$Logit$

Table 3.2 also contains the useful distribution for GLMs along with their link function.

3.5 The Parameter Estimate

The concept of GLMs merges different approaches to explaining variation in data in terms of a linear combination of covariates (Agresti, 2020). The GLMs is consists of a single response variable and the predictor variable is a member of exponential family distribution. The GLMs converts the relationship between the linear predictor and the mean response, such that a nonlinear relationship can be modeled as a linear. This admits a model specification allowing for continuous or discrete responses and allows a description of the variance as a function of the mean response. The parameters in GLMs can be estimated by the maximum likelihood method. Maximum likelihood can be used as a theoretical basis for the parameter estimation in GLMs. The likelihood function can be written as:

$$L(y_i, \theta_i) = \prod_{i=1}^{n} exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + C(y_i, \phi)\right) \tag{3.21}$$

$$= \exp\left[\sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + C(y_i, \phi)\right)\right]$$

The log-likelihood function is given as:

$$l(y,\theta) = Log\ L(y_i,\theta) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right) \quad (3.22)$$

Taking the derivative of the log-likelihood function with respect to β_j, where is $0,1,2\ldots,p$, and equating the equation to zero then solve equations simultaneously. p Is the number parameters, then we obtain the score function, which is given as:

$$(U_{\beta_0}, U_{\beta_1}, U_{\beta_2}, \ldots, U_{\beta_j}) \qquad \text{where } U_{\beta_j} = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\frac{\partial l_i}{\partial \beta_j}$$

By the use of a chain rule, we obtain:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial u_i}\frac{\partial u_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \quad (3.23)$$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b\prime(\boldsymbol{\theta}_i)}{a_i(\phi)} = \frac{y_i - u_i}{a_i(\phi)}, \text{ since } u_i = E(y_i) = b'_i(\boldsymbol{\theta_i}) \quad (3.24)$$

and

$$\frac{\partial \theta_i}{\partial u_i} = \frac{1}{b\prime\prime(\boldsymbol{\theta_i})} = \frac{a_i(\phi)}{var(y_i)} \quad (3.25)$$

The factor $\frac{\partial u_i}{\partial \eta_i}$ depends on the link function, where $\eta = X'\boldsymbol{\beta}$

And

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \quad (3.26)$$

where x_{ij} is the j^{th} element of the covariates vector X_i for the i^{th} observation.

Then we substitute into equation (3.24), (3.25) and (3.26) into equation (3.23), to get:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - u_i}{var(y_i) b''(\boldsymbol{\theta_i})\, x_{ij}} = \frac{y_i - u_i}{a_i(\phi)} x_{ij}$$

The system of equations to be solved for $\beta'_j s$ is given as:

$$\frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i - u_i}{a_i(\phi)} x_{ij}\right] = 0 \qquad (3.27)$$

Then the MLE of the parameter vector β is obtained based on solving the estimating equation (3.34). The estimate of β depends on the density function only through the mean and variance function $Var(\mu_i)$. Equation (3.27) can be solved using Newton-Raphson, Fisher's scoring and re-re-weighted lease square (RWLS) algorithm for maximum likelihood estimation (MCcullagh and Nelder, 1989). These algorithms are available in SAS and Stata software. The ML estimation for β is carried out via Newton-Raphson,

$$\beta^{(m+1)} = \beta^{(m)} + \left(l''(\beta^{(m)})\right)^{-1} l'\left(\beta^{(m)}\right), \qquad (3.28)$$

where l is a log-likehood function for entire sample $y_{i,\dots}y_N$. Then we let l, l' and l'' denote the contribution of the observation y_i to the log-likelihood and its derivatives. The Fisher scoring iterative equation is given by

$$\beta^{(m+1)} = \beta^{(m)} + \left[-E\left(l''(\beta^{(m)})\right)\right]^{-1} l'\left(\beta^{(m)}\right) \qquad (3.29)$$

where the expected Hessian matrix becomes

$-E\left(l''\left(\beta^{(m)}\right)\right) = \boldsymbol{H}^{(m)} = \boldsymbol{X'WX}$ and $W = Diag\left\{Var(y_i)\left(\frac{l\text{⍰}_i}{l\mu_i}\right)^2\right\}^{-1}$

where W is a diagonal matrix with main diagonal element and $A = W\left(\frac{l\square}{l\mu}\right)$, then note that A and W are related. Note $\left(\frac{l\square}{l\mu}\right) = Diag\left(\frac{l\square_i}{l\mu_i}\right)$. This takes us to RWLS equation, which is given as:

$$\boldsymbol{\beta}^{(m+1)} = [X'WX]^{-1}X'W_z \tag{3.30}$$

where

$$\rho = \square + \left(\frac{\boldsymbol{l\square}}{\boldsymbol{l\mu}}\right)(y - \mu) \tag{3.31}$$

$$\left(\rho_{i\ldots,}\rho_N\right)'$$

where $\rho_i = \square_i + \left(\frac{l\square_i}{l\mu_i}\right)(y - \mu_i)$ and it's called a linearized form of link function g evaluated at y. Fisher scoring method is similar to the Newton-Raphson method but the difference is that Fisher scoring uses the expected value of matrix called expected information while the Newton Raphson method uses the matrix itself.

3.6 Review of the Logistic Regression models

Logistic regression is also known as logistic modelling or the logit model. It is a statistical method for analyzing a dataset in which the dependent variable is dichotomous or binary that determine outcome. It was introduced by (Cox, 1972) to describe the dependence of a binary variable on a set of discrete and continuous variables. It is a type of predictive model that can be used when the target variable is a categorical variable with two categories.

The logistic regression model, as a non-linear model, is a special case of a GLM (McCullagh and Nelder, 1989), where the assumptions of normality and constant variance of residuals are not satisfied. In logistic regression, it is assumed that the explanatory variables are independent. However, if they are not independent one has to take dependencies into account. Logistic regression has been given a comprehensive history and methodology by Hosmer and Lemeshow (1989). It is applicable to multilevel responses and these may be ordinal or nominal. For ordinal response outcomes, the model function is called cumulative logits by performing ordered logistic

regression using the proportional odds model and for nominal response outcome one forms generalized logit and performs a logistic analysis (McCullagh, 1980). The logistic regression model is a statistical technique for predicting the probability of an event, given a set of predictor variables.

The likelihood equation for a logistic regression model does not always have a finite solution. Sometimes there is a no unique maximum on the boundary of the parameter space at infinity. The existence, uniqueness, and finiteness of maximum likelihood estimates for the logistic regression model depend on the patterns of data points in the observation space (Santner and Duffy, 1986).

3.6.1 The Logistic Regression Models

We consider the explanatory variable of interest denoted by the vector $x = (x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi})$ for the i^{th} individual. Then let the probability that the event of poverty poor be denoted by $p(y_i = 1) = \pi_i$ for the i^{th} individual and let the event of that an individual is not poor be denoted by $(y_i = 0) = 1 - \pi_i$.

Logistic regression does not involve assumptions such as the linearity and normality of the dependent variables and residuals. This method is based on the log transformation of the odds and is given by the

$$logit(\pi_i) = \log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \tag{3.32}$$

where β_0 is the intercept parameter, $1 - \pi_i$ is the probability of the even and β is the vector of slope parameter. The purpose of logistic regression in this study is to find the parameters $\beta_0, \beta_1, \dots, \beta_p$ that best fit the data relating poverty in Zimbabwe using 2016 Demography heath Survey data set.

3.7 The Assumption of Logistic Regression

In order to have effective analysis, the model should satisfy the following four assumptions.

a. It does not need a linear relationship between the independent and dependent variables.
b. The error terms must be independent, since logistic regression requires each observation to be independent.
c. Logistic regression assumes linearity of independent variables and log odds; it requires that the independent variables are linearly related to the log odds, or else the logistic regression underestimates the strength of the relationship and rejects the relationship easily, that is, being not significant where it should be significant. A solution to this problem is the categorization of the independent variables, that is, transforming metric variables to ordinal level and then including them in the logistic regression model. Another approach would be to use discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and no multicollinearity are met.
d. Logistic regression requires quite a large sample, because in the case where sample size is small the hypothesis tests about maximum likelihood estimates will be less powerful than ordinary least squares.

3.8 Parameter Estimation for Logistic Regression

In this study, poverty status, which is the dependent variable (response variable) $y_i(i = 1,2,..,n)$, is dichotomous and probability distribution is Bernoulli or binary. This probability distribution for the dependent variables can be expressed as, $y_i = Bernoulli(\pi_i)$ and p predictor variables (Czepiel, 2002; Lomeshow and Hosmer, 2000; Wood, 2006). The maximum likelihood estimate can be obtained when we let:

$$Y_i = y_i \mid x_{i1}, x_{2i}, \dots, x_{pi}\,; \sim Bernoulli(\pi_i) \quad where\ i = 1,2,..n, \tag{3.33}$$

The probability mass function (PMF) for the Bernoulli distribution can be given by:

$$P\big(Y_i = y_i \mid x_{11}, x_{2i}, \dots, x_{pi}\big) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad i = 1,2, \dots n, \tag{3.34}$$

Where π_i is the probability of success and $(1-\pi_i)$ is the probability of failure and π_i which is the mean reponse for function of $x_i = x_{1i}, x_{2i}, .. x_{pi}$, under the GLM formulation with a logit link. According to this, π_i can be taken as the probability of being poor and $(1-\pi_i)$ as the probability of being not poor. The likelihood of the response variables for all observations is given as:

$$L = P_r(y_1, y_2, y_2, \dots, y_n)$$

Assuming that the observations are independent, the likelihood function can be expressed as the product of the individual probabilities.

$$L = P_r(y_i)P_r(y_2) \dots P_r(y_n) = \prod_{i=1}^{n} P_r(y_i) \tag{3.35}$$

Since the response variables $y_i's$ are Bernoulli distribution, the likelihood function can be given as:

$$L(\beta \mid Y) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Therefore, substitute $\pi_i's$ in terms of the covariate:

$$L(\beta \mid Y) = \prod_{i=1}^{n} \left[\frac{Exp(X_i\beta')}{1+Exp(X_i\beta')}\right]^{y_i} \left[\frac{1}{1+Exp(X_i\beta')}\right]^{1-y_i} \tag{3.36}$$

where $\boldsymbol{\beta} = (\beta X_1, \beta X_2, \dots., \beta X_p)'$ and X_i is the matrix of covariates. It is complicated to differentiate the likelihood function; therefore, we need to simplify the likelihood by taking the log. The logarithm is a monotonic function; any maximum of the likelihood function will be the maximum of the log likelihood function (Czepiel, 2002). The log likelihood is given by:

$$l(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n}(1-y_i)\log(1-\pi_i) \tag{3.37}$$

To find the parameter estimates, we must find the first derivative of log-likelihood with respect to β and equate it to zero. The maximum likelihood estimates for β can be found by equating each of the $K+1$ equations to zero and solve for β_k (Czepiel, 2002). Assuming there are K

independent variables. The critical point will be the maximum if the matrix of second derivative is negative definite. For a matrix to be negative definite it means that every element on the diagonal of the matrix is less than zero, while positive means that every element on the diagonal of the matrix is greater than zero (Czepiel, 2002; Moet, 2010).

3.9 Testing Hypothesis for $\boldsymbol{\beta}$

To test for the significance of the regression parameters the Wald test for individual parameters can be used.

3.9.1 The Wald Test

A Wald test is useful if we want to test the statistical significance of a coefficient (β_j) in the model. The Wald statistic is given by:

$$Z = \frac{\widehat{\beta_J}}{SE(\widehat{\beta_J})}$$

where $\widehat{\beta_J}$ is the parameter estimate on fitting the model and $SE(\widehat{\beta_J})$ is the standard error of the estimate, when the Z-value is squared, the square approximately follows the chi-square distribution with degrees of freedom equal to one. Once the model is fitted, we compare the test value, with a pre-determined critical value at a given level of significance. In order for exploratory variables to be significant in the model, the statistical test must be greater than the critical value. According to Agresti (2018) the likelihood-ratio test is more reliable for small sample sizes than the Wald test.

3.9.2 Likelihood-Ratio Test

The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the simple model (L_0), over the maximized value of the likelihood function for the full model(L_1). The likelihood-ratio test statistics can be given as:

$$-2\ log\left(\frac{L_0}{L_1}\right) = -2\log(L_0) - 2\log(L_1) \quad (3.38)$$
$$= -2(l_0 - l_1) \sim x_k^2$$

where k is the degree of freedom calculated as the difference in the number of parameters between the simple and the full model. The test is used of nested models. The likelihood-ratio test is the most recommended test to use in backward stepwise elimination.

3.9.3 Chi-Square Test

A chi-squared test is used to examination whether the two variables are independent.

When calculating the chi-square test statistic we need to calculate the expected count two or more groups, population. Once the expected values has been computed, the chi-square test statistic is computed as:

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

The distribution is the chi-square distribution with degrees of freedom $(r-1)(c-1)$, where r is the number of rows in the two-way table and c is the number of columns.

3.9.4 Odds Ratio Test

An odds ratio (OR) is used to measure the association between two predictors and the response of interest. The odds ratio compares the odds of an event between two groups. When the outcome variable is binary in a logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables. A logit is defined as logarithm of the odds. If p is the probability of an event and $(1-p)$ is the probability of not observing the event, then the odds of the event happening as opposed to not happens is $\left(\frac{p}{1-p}\right)$. This linaer relationship can be written as :

$$\log(p) = \log\left(\frac{p}{1-p}\right) \tag{3.39}$$

Logit transform is mostly used in logistic regression and for fitting linear models to categorical data.

3.10 Model Selection

The Hosmer-Lemeshow statistic is not produced in PROC SURVEYLOGISTIC. Since this statistics is not yet available, however can also, use Akaike's Information Criterion (AIC) and Schwarz Criterion (SC) to compare the goodness of fit of two nested models (Moeti, 2010).

Akaike's Information Criterion:

One way to evaluate a model is to use the Information Criterion (IC). This criterion at tempts to quantify how well the model has predicted the data. The Akaike's Information Criterion (AIC) is a useful statistic for comparing the relative fit of different models. This statistic was proposed by Akaike (1974) and is given by

$$AIC = -2logL + 2(k + s)$$

where k is the total number of responses, minus one and s is the number of explanatory variables. This method penalizes the log likelihood for the number of parameters estimated (Akaike, 1974). A model that minimizes the AIC is preferred. The method is particularly useful when comparing non-nested models.

Schwarz Criterion:

This is a different to AIC for comparing non-nested models. Schwarz Criterion (SC) also known as Bayesian Information Criterion (BIC) and was proposed by Schwarz et al. (1978). SC is given by

$$SC = -2logL + (k + S)\log(\sum_j f_j)$$

where s and k has the same function in the AIC, f_j is a frequency value of the j^{th} observations.

SC produces more severe penalization on the likelihood for estimating more parameters (Allison, 2012). The model chosen is the one which leads to the minimum SC. While doing a model selection, we can narrow down the options before comparing models. This can be done by building the regression model step by step using selection procedure of variables that enters the model. The AIC and SC statistics give two different ways of adjusting the –2 log likelihood statistic for the number of terms in the model and the number of observations used. These statistics should be used when comparing different models for the same data (for example, when you use the METHOD=STEPWISE option in the MODEL statement in SAS); lower values of the statistic show a more appropriate model.

3.11 Fitting a logistic Regression Model in SAS

Fitting a logistic regression model to binary data has a similar procedure with ordinal data, with the only difference being the link function (binary link=logit, order link=cumulative logit). The results below were obtained using PROC SURVEYLOGISTIC in SAS 9.4. We used survey logistic regression in our survey dataset to measure the relationship between a categorical dependent variable (poverty status) and independent variables (sex of head of household, region, residence, (Size, age and education) and to check whether the model is a good fit.

3.11.1 Model Fitting

PROC SURVEYLOGISTIC automatically generates model fit statistics and Testing Global Null Hypothesis: BETA=0 tables, which give the various criteria (AIC and SC) based on the likelihood for fitting a model with intercept only and for fitting a model with intercept and explanatory variables. PROC SURVEYLOGISTIC also generates Type 3 Analysis of Effects tables if the model contains an effect involving a CLASS variable. This table gives the degree of freedom, the Wald chi-square statistic, and the p- value for each effect in the model

3.11.2 Model Checking

The PROC SURVEYLOGISTIC in SAS does not produce plots and Hosmer-Lemeshow statistics, so one may use the AIC and SC to check if the model is a good fit or not. The AIC of the full model (contains intercept and other variables) is smaller compare to the AIC of the reduced model (only the intercept); this indicates that the fitted model better explains the data. Referred to table 3.3.

3.12 Interpretation of the Coefficient of the Model and the Odds Ratio

Table 3.3 Model fit statistics

Model Fit Statistics		
Criterion	**Intercept only**	**Intercept and Covariates**
AIC	28606.465	8513.898
SC	28614.407	8728.340
–2 Log L	28604.465	8459.898

Table 3.4 Type 3 analysis of effects

Type 3 Analysis of Effects			
Effect	**F Value**	**Num DF**	**Pr > F**
Region	88.38	9	<.0001
Education	221.34	3	<.0001
Sex	10.61	1	0.0011
Age	8.19	7	<.0001
Size	2.14	5	0.0572
RES	510.97	1	<.0001

The results in Table 3.4 above confirm that all explanatory variables are significant at 5% level of significance, except size of household, which is not significant at $p = 0.0572$, which is greater

than 0.05. All the explanatory variables that are significant have an influence on the poverty status in Zimbabwe.

Table 3.5 Analysis of maximum likelihood estimation and odds ratio table

Parameter		Estimates	OR	S.E	95%CL	P-value
Intercept	Poor	-7.6422		0.1408		<.0001**
Region (ref: Manicaland)	Bulawayo	0.0267	0.802	0.1226	(0.585;1.098)	0.0827
	Harare	-2.0405	0.101	0.0783	(0.081;0.127)	<.0001**
	Mashonaland Central	0.0866	0.851	0.0796	(0.678;1.070)	0.2766
	Mashonaland East	-0.3535	0.548	0.0743	(0.442;0.681)	<.0001**
	Mashonaland West	0.3448	1.102	0.0824	(0.875;1.388)	<.0001**
	Masvingo	0.4184	1.186	0.0828	(0.938;1.501)	<.0001**
	Matabeleland North	1.1445	2.452	0.1195	(1.812;3.318)	<.0001**
	Matabeleland South	-0.1506	0.672	0.0892	(0.524;0.860)	0.0912
	Midlands	0.2761	1.029	0.0880	(0.808;1.311)	0.0017**
Education (ref: higher)	No Edu/preschool	3.7693	1.137	0.7527	(0.582,1.125)	<.0001**
	Primary	0.5028	1.653	0.2575	(0.561;0.751)	0.0509
	Secondary	-0.5568	0.573	0.2559	0.234,0.758	0.0296*
Sex (ref: Female)	Male	-0.0994	0.820	0.0305	0.727,0.924	0.0011*
Age (ref: 15–24)	25–34	-0.4184	0.598	0.0976	(0.389;0.920)	0.9542
	35–44	-0.4299	0.396	0.0898	(0.259;0.605)	<.0001**
	45–54	-0.4298	0.391	0.0965	(0.254;0.603)	<.0001**
	55–64	-0.2544	0.391	0.1165	(0.248;0.618)	0.0002**
	65–74	0.1526	0.466	0.1545	(0.280;0.776)	0.0997
	75–84	0.8768	0.700	0.5438	(0.360;1.363)	0.05715
	85–95	-0.0079	1.445	0.4478	(0.490;4.258)	0.0502
Residence (ref: urban)	Rural	0.0947	1.110	0.1413	(0.216;0.756)	<.0001**
Size (ref: 16–20)	1–5	-0.0079	0.970	0.0631	(0.820;1.196)	0.8998
	6–10	-0.1550	0.855	0.0637	(0.707;1.034)	0.0149*
	11–15	-0.0190	0.979	0.0617	(0.813;1.180)	0.7579
	21–25	-0.0206	0.978	0.0628	(0.811;1.190)	0.7423
	26–29	-0.2009	1.220	0.0745	(0.988;1.507)	0.0079**

Table3. 5 contains the estimated coefficients, standard errors and p-values for the survey logistic regression model. The 95% confidence interval and the logit link function was used in all calculated odds ratios. The intercept for the model was found to be significant at 5% level of significance, since the p-value is less than 0.05.

For the region of household, it was found that Harare, Mashonaland East, Mashonaland West, Masvingo, Matabeleland North and Midlands were significant at 5% level of significance. The odds ratio for Harare is 0.101, (with 95% CI: 0.081,0.127). The corresponding odds ratio for Mashonaland East is 0.548 (with 95% CI: 0.442, 0.681). Therefore, Mashonaland East and Harare has a significantly lower likelihood of poverty then Manicaland. For Mashonaland West, the corresponding odds ratio is 1.102 (with 95% CI: 0.875, 1.388). Mashonaland West is 0.102 times more likely to be poor compared with Manicaland. The odds ratio for Masvingo is 1.182 (with 95% CI: 0.938, 1.388) is 0.388 times more likely to be poor compared with Manicaland. The corresponding odds ratio for Matabeleland North is 2.452 (with 95% CI: 1.812, 3.318). Matabeleland has a higher odds ratio; it is 1.452 times more likely to be poor compared with Manicaland. Midlands was found to be 0,029 times poorer than Manicaland.

For education level, Table 3.5 confirms that secondary and no education are significant at 5% level of significance. The corresponding odds ratio for pre/no education is 1.137 (with 95% CI: 0.582, 1.125). Therefore, people with no education are 0.137 times more likely to be poorer when compared with tertiary educated people. The odds ratio for secondary education is 0.858 (with 95% CI: 0.234, 0.758). Uneducated people have a higher odds ratio value than other levels, which indicates that there is a higher percentage of poor uneducated people than other education levels.

The results in Table 3.5 shows that in the category sex of head of household, male-headed household are significant at 5% level of significance (p=0.0011). The poverty odds for male headed household is 0.820 (with 95% CI: 0.727, 0.92), therefore male-headed household is 0.180 times lower than the poverty odds for female-headed household.

The age of head of household is significant at age 35–44, 45–54 and 55-64 years, and has a p-value less than 0.05. The corresponding poverty odds ratio for age group 35–44 years is 0.601

times the poverty odds for the age of 15-24 years. The poverty odds for age group 45–54 and 55-64 was both estimated to be 0.601 times the poverty odds for age group 15–24 years.

Table 3.5 confirms that in the residence category, rural areas were found to be significant at 5% level of significance (p-value= 0.001). The corresponding poverty odds ratio for rural areas is 1.110, indicating that the poverty odds for rural area is 0.110 times the poverty odds in Urban areas.

For the number of household members (size), the results in Table 3.5 show that a household size of 6–10 and 26–29 were found to be significant to poverty status at 5% level of significance. The poverty odds ratio for size 6-10 was estimated to be 0.855 times the poverty odds for size 16-20, while the poverty odds for size 26-29 was 1.220 times the poverty odds for size 16-20.

Chapter 4

Generalized Additive Mixed Models

4.1 Introduction

In the previous chapter, the survey data was modelled using statistical model such as generalized survey logistic regression (with binary outcome), which accounts for the survey design. There are different types of parametric models such as generalized linear mixed models, multivariate joint models, spatial multivariate joint models, but in this chapter, the generalized additive mixed model is used. The parametric models offer a strong tool for modelling the relationship between the outcome variables and predictor variables when their assumptions hold. Other parametric models may suffer from inflexibility in modelling complicated relationships between the outcome variables and the predictor variables in some applications. The parametric mean assumption may not always be desirable, as suitable functional forms of the predictor variables may not be known in advance and the response variables may depend on the covariates in a complicated manner (Lin and Zhang, 1999). The generalized additive mixed model (GAMM) relaxes the assumption of normality and linearity inherent in linear regression. The flexibility of non-parametric regression for continuous predictor variables, coupled with the linear models for predictor variables, offers ways to reveal structures within the data that may omit linear assumptions.(Lin and Zhang, 1999).

This flexibility of GAMM motivated the current research to use semiparametric logistic mixed model to assess the determinants of poverty. In literature, there exists many nonparametric regression models and smoothing methods for independent data. The most commonly used are splines smoothers, kernel smoothers, locally weighted running-line smoothers and running-mean smoothers. These methods are well detailed in Hastie and Tibshirani (1990); Hardle (1999) and Green and Silverman (1993).

Generalized additive mixed models (GAMMs) are regression models in which the expected value of a response variable is determined by a sum of smooth functions of predictor variables, along

with any parametric and random effects. Generalized additive mixed models are proposed for correlated and over-dispersed data, which arise frequently in studies involving clustered, hierarchical and spatial designs. This class of models allows flexible functional dependence of an outcome variable on covariates by using non-parametric regression. They are useful in modelling situations where the relationship between the response and predictors is complicated to write down in the simple parametric form of a generalized linear mixed model (GLMM) (Lin and Zhang, 1999).

4.2 Review of the Generalized Linear Mixed Model

The generalized additive mixed models (GAMMs) are an extension of the GAMs, incorporating random effect, or an extension of the generalized linear mixed models (GLMMs), (Breslow and Clayton, 1993) that allow the parametric fixed effects to be modelled non-parametrically, using additive smooth functions in a similar spirit (Hastie and Tibshirani, 1990). Generalized linear mixed models are an extension of linear mixed models to allow response variables from different distributions such as binary responses. The GLMMs represent a class of fixed effect regression models for several types of dependent variables (i.e. continuous, dichotomous, counts). According to (McCullagh and Nelder, 1989a) describe these in great detail and indicate that the term “generalized linear mixed model” is due to Nelder and Wedderburn (1935), who described how a collection of seemingly disparate statistical techniques could be unified. Common GLMMs include linear regression, logistic regression, and Poisson regression. Alternatively, one could think of GLMMs as an extension of generalized linear models (e.g. logistic regression) to include both fixed and random effects.

4.2.1 Fixed Effect

According to Snijders (2005), defined fixed effect as the average effect in the whole population expressed by the regression coefficient. The fixed effect can be thought of as “treatment” levels that have been selected for inclusion in the study (Littell *et al.,* 2002*)*. The fixed effect used when the interest is only in analyzing the impact of variables that vary over time and are further used in analyzing the relationship between predictor and outcome variables within an entity. Each

entity has its own individual characteristics that may or may not affect the predictor variables (Torres-Reyna 2007).

4.2.1 Random Effect

The random effect is the variation across entities and assumed random and uncorrelated with the predictor or independent variables in the model. Furthermore, the random effect is chosen to control specific factors that are expected to cause random variation in the coefficients (Kinney and Dunson, 2006). In the case of balanced data, random factors do not cause inferential problems for a test of fixed effects. However, for unbalanced data, improper treatment can lead to mistaken inferences about treatment effect (Littell, *et al.* 2002*)*. (Gitelman et al., 2003) conducted a study on PET (Positron Emission Tomography) using 12 subjects that were drawn randomly from a large population. Subjects were asked to either repeat a heard letter or respond with a word that began with that letter. The random effect, in this case, was subject variable. Hence, when drawing inferences about population sampling, variability must be taken into account.

4.3 Generalized Additive Mixed Models

Generalized additive mixed models are an extension of the generalized additive model incorporating random effect or an extension of the generalized linear mixed models (GLMM) (Breslow and Clayton, 1993) that allow the parametric fixed effects to be modelled non-parametrically using additive smooth functions in a similar spirit (Hastie and Tibshirani, 1990).

Suppose that observations of the j^{th} of k units consists of an outcome variable y_i and p covariates $x_j = (1, x_{j1}, \dots, x_{jp})^T$ associated with fixed effect and a vector with $q \times 1$ of covariates Z_j associated with random effects. Therefore, the GAMM was formulated by (Lin and Zhang, 1999) as follows:

$$g(\boldsymbol{\mu_j}) = \boldsymbol{\beta_0} + f_i(\boldsymbol{x_{ij}}) + \cdots + f_p(\boldsymbol{x_{jp}}) + Z_i\boldsymbol{b} \tag{4.1}$$

Where g (.) is a monotonic differentiable link function, $\mu_j = E(y_i \mid b)$, $f_j(.)$ is a centered twice-difference smooth function, random effect b is assumed to be distributed as $N\{0, k(\vartheta)\}$ and $\boldsymbol{\vartheta}$ a $c \times 1$ vector of variance components. A fundamental feature of GAMM (4.1) over GAM is that the additive non-parametric functions are used to model observations (Lin and Zhang, 1999). If $f_j(.)$ is a linear function, the GAMM (4.1) is reduced to a generalized linear mixed model (GLMM) (Breslow and Clayton, 1993).

For a given variance component $\boldsymbol{\vartheta}$, the log-quasi-likelihood function of $(\beta_0, f_j, \vartheta, j = 1,2, \dots k)$ is given by (Manjengwa *et al.*, 2012):

$$\text{Exp}[l\{\boldsymbol{\beta_0}, f_1(.), \dots, f_k(.), \vartheta\}]\, \alpha |\boldsymbol{k}|^{-\frac{1}{2}} \int exp\left\{\frac{-1}{2} \sum_{j=1}^{k} d_j\left(y_i, \mu_j\right) - \frac{1}{2} b' \boldsymbol{k^{-1}} \boldsymbol{b}\right\} d\boldsymbol{b} \qquad (4.2)$$

where $\boldsymbol{Y_i} = (y_1, y_2, \dots, y_k)$ and $d_j(y_i, \mu_j)\ \alpha - 2\int_{y_j}^{\mu_j} m_j(y_j - \mu)/(v)\mu du$ define the conditional deviance function of $\{\beta_0, f_j(.), \vartheta\}$ given b. Inference in generalized additive mixed models includes inference on the non-parametric $f_j(.)$; that needs the estimation of a smoothing parameter and inference on the variance components ϑ. There is a close connection between a linear mixed model and the smoothing spline estimators (Green and Silverman, 1993), (Lin and Zhang, 1999).

4.3.1 Natural Cubic Smoothing Spline estimation

As the $f_j(.)$'s are infinite dimensional unknown functions, and are thus estimated by using natural cubic smoothing splines. The derivation is λ and ϑ and the natural cubic smoothing spline estimators of the $f_j(.)$ maximize the penalized log-quasi-likelihood as follows (Greenland, 1994) and (Lin and Zhang, 1999):

$$l\{\beta_0, f_1(.), \dots, f_k(.), \vartheta\} = \frac{1}{2}\sum_{i=1}^{k} \lambda_i \int_{s_i}^{t_i} f''(x^2) dx \qquad (4.3)$$

$$= l\{\beta_0, f_1(.), \dots, f_k(.), \vartheta\} - \frac{1}{2}\sum_{i=1}^{k} \lambda_i f_i^T B_i f_i$$

Where (s_i, t_j) defines the range of the i^{th}covariate and λ_i is a vector of smoothing parameters and controls the trade-off between the goodness of fit and smoothness of estimated function. Hence, f_i is a $r_I \times 1$ unknown vector or values of $f_i(.)$, estimated at the r_i ordered distinct values of the x_{ij}, where $i = 1, \dots, n$ and B_i is the corresponding non-negative definite smoothing matrix (Green and Silverman, 1993). The GAMMs equation given in (4.1) can be calculated in matrix form as:

$$g(\boldsymbol{\mu_i}) = 1\boldsymbol{\beta_0} + N_i f_i + N_i f_i + \cdots + N_k f_k + \boldsymbol{Z}b \tag{4.4}$$

where $g(\boldsymbol{\mu_i}) = \{g(\boldsymbol{\mu_1}), g(\boldsymbol{\mu_2}), \dots g(\boldsymbol{\mu_n})\}$, 1 is an $n \times 1$ vector of 1_s and N_j is an $n \times r_i$ incidence matrix defined in the same way as that given by Green and Silverman (1994), such that the i^{th} components of $N_i f_i$ is $f_i(x_{ij})$ and $\boldsymbol{Z_i} = (z_1, z_2, .., z_n)^T$. Numerical integration is required to estimate equation (4.2), except for the Gaussian case. It is often complicated to calculate the full natural cubic smoothing splice estimators of f_i by directly maximizing expression (4.4). Hence Lin and Zhang formulated a method to solve this problem, called double penalized quasi-likelihood. The method is discussed in 4.2.2.

4.3.2 Double Penalized Quasi-likelihood (DPQL)

Since f_i is a centred parameter vector, it can be parameterized in the form of β_i and $\alpha_i((r_i - 2)) \times 1$ in a one to one transformation given as:

$$f_i = X_i \beta_i + \beta_i \alpha_i \tag{4.5}$$

where $\boldsymbol{X_i}$ is an $r_i \times 1$ vector containing r_n centered ordered distinct values of the x_{ij}, where $i = 1,2 \dots, n)$, $\beta_i = L_i(L_i L_i^T)^{-1}$and L_i is an $r_i \times (r_i - 2)$ full rank matrix that satisfies $H_i = L_i L_i^T$ and $L_i^T X_i = 0$ by using the identity $f_i^T H_i f_i = \alpha_I^T \alpha_i$. The DPQL with respect to $(\beta_0; f_i)$ and b can be written as:

$$-\frac{1}{2\varphi}\sum_{i=1}^{n} d_i(y_i; \mu_i) - \frac{1}{2}\, b^T k^{-1} b - \frac{1}{2}\, b^T \Gamma^{-1} \alpha \tag{4.6}$$

where $\boldsymbol{\alpha} = (\alpha_1^T, \alpha_2^T, .., \alpha_k^T)^T$ and $\Gamma = diag(\tau_1 I, \tau_2 I, .., \tau_k I)$ with $\tau_i = \frac{1}{\lambda_i}$. A small value of τ_i corresponds to over-smoothing. Putting equation (4.5) into (4.4), expression equation (4.4) suggests that given ϑ and τ,the DPQL estimator $\widehat{f_i}$ can be obtained by fitting the generalized linear mixed model, using the penalized quasi-likelihood approach (Breslow and Clayton, 1993):

$$g(\boldsymbol{\mu}) = X\boldsymbol{\beta} + B\boldsymbol{\alpha} + \boldsymbol{Z}b \tag{4.7}$$

Where $\boldsymbol{X} = (1, N_1X_1, N_2X_2, \dots, N_KX_K)$, $\boldsymbol{B} = (N_1B_1, N_2B_2, \dots, N_KB_K)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_K)^T$ is a $(K+1) \times 1$ vector of regression coefficients; α and b are independent random effects with normal distributions $\alpha \sim N(0, \boldsymbol{\Gamma})$ and $b \sim N(0, K)$. Thus the DPQL estimator $\widehat{f_i}$ is calculated as $\widehat{f_i} = X_i\hat{\beta}_i + \beta_i\widehat{a_i}$ that is a linear combination of the Breslow and Clayton (1993) penalized quasi-likelihood estimators of the fixed effect $\widehat{\beta_i}$ and the random effects $\widehat{a_i}$ in using GLMM from equation (4.7). The maximization of the expression (8.6) with respect to (β, a, b) can be carried on by using the Fisher scoring algorithm to solve:

$$\begin{pmatrix} X^TWX & X^TWB & X^BWZ \\ B^TWX & B^TWB + \Gamma^{-1} & B^TWZ \\ Z^TWX & Z^TWB & Z^TWZ + K^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \\ b \end{pmatrix} = \begin{pmatrix} X^TWY \\ B^TWY \\ Z^TWY \end{pmatrix}, \tag{4.8}$$

Where Y is the working vector defined as $Y = \beta_0 1 + \sum_{j=1}^{P} N_i f_i + Zb\Delta(Y - \mu)$ and $\Delta = diag|g'(\mu_i)|, W = diag[\{(\vartheta_v(\mu_i)g'(\mu_i)^2\}^{-1}]$. An examination of the equation (4.8) shows that it corresponds to the normal equation of the best linear unbiased predictors (BLUPs) of β and (a, b) under linear mixed models.

$$Y = X\beta_0 + \beta a + Zb + \varepsilon, \tag{4.9}$$

wwhere a and b are an independent random effect with normal distribution, $a \sim N(0, \Gamma)$, $b \sim N(0, K)$ and $\varepsilon \sim N(0, W^{-1})$. This proposes that the DPQL estimators $\widehat{f_i}$ and the random effects

estimators $\hat{b}$ can easily be obtained using the BLUPs by iteratively fitting model (4.9) to the working vector Y (Lin and Zhang, 1999).

4.4 Estimating Parameters and Variance Components

Previously, it was assumed that the smoothing parameter λ and the variance component ϑ are known when estimation is made on the non-parametric function f_i. However, they usually need to be estimated from the data.

$$y = f(X) + \varepsilon, \tag{4.10}$$

Under the standard non-parametric regression model, where ε is an independent random error which distributed as $N(0, \sigma^2)$. (Wahba, 1985) and (Kohn *et al*., 1991) suggested estimating the smoothing parameter λ by maximizing a marginal likelihood. The marginal likelihood of $\frac{1}{\lambda}$ is constructed by assuming that $f(X)$ has a previous specification in the form of equation (4.5) with distribution as $a \sim N(0, \tau I)$ and a flat prior for β.

The integration with respect to a and β is given by:

$$exp\{l_N(y; \tau, \sigma^2\} \propto \tau^{\frac{1}{2}} \int exp\left\{l(y; \beta, a, \sigma^2) - \frac{1}{2\tau}\, a^T a\right\} da d\boldsymbol{\beta} \tag{4.11}$$

where $l(y; \boldsymbol{\beta}, a, \sigma^2)$ is the log-likelihood of f under equation (4.10). (Robinson, 1991) and (Silverman, 1985) pointed out that marginal likelihood (4.11) of τ is definitely the restricted maximum likelihood (REML)under the linear mixed model, given as:

$$y = 1\beta_0 + X\beta_1 + Ba + \varepsilon, \tag{4.12}$$

where $a \sim N(0, \tau I)$ and $\varepsilon \sim N(0, \sigma^2 I)$ and B is defined previously, then τ is regarded as the covariance component. Therefore, the marginal estimator of τ is indeed an REML estimator.

Kohn et al. proposed that the maximum marginal likelihood estimator of τ can sometimes perform better than the generalized cross validation (GCV) estimator in estimating non-

parametric function. (Zhang *et al.*, 1998) extended these results to estimate the smoothing parameter λ and variance component ϑ jointly, using REML in case of nonparametric mean function, and their model is given as follows:

$$y = f(X) + Zb + \varepsilon; \tag{4.13}$$

where $f(X)$ represents the value of non-parametric function $f(.)$ evaluated at the design points of $X_{(n\times 1)}, b \sim N\big(0, K(\vartheta)\big)$ and $\varepsilon \sim N\big(0, V(\vartheta)\big)$. When $f(.)$ Is estimated, using a cubic smoothing spline in equation (4.5), (Zhang *et al.*, 1998) rewrite the equation (4.13) as a linear mixed model.

$$y = 1\beta_0 + X\beta_1 + Ba + Zb + \varepsilon; \tag{4.14}$$

where a, b and ε distribution are similar to those in equation (4.13), they proposed τ as an extra variance component in addition to ϑ in the model (4.14) and to estimate ϑ and τ jointly by using REML. In this case, REML corresponds to the marginal likelihood of (τ, ϑ), constructed by assuming that f takes the form of (4.5) with distribution $a \sim N(0, \tau I)$ and a flat prior for β. Therefore, the integration with a and β is given as follows:

$$exp\{l_N(y; \tau, \sigma^2\} \propto K^{\frac{-1}{2}} \Gamma^{\frac{-1}{2}} \int exp\left\{l(y; \beta, a, \sigma^2) - \frac{1}{2} b^T K^{-1} - \frac{1}{2\tau}\, a^T a\right\} da d\boldsymbol{\beta} d\boldsymbol{b} \tag{4.15}$$

where $l(y; \boldsymbol{\beta}, ab) = l(y; f, b)$ is the conditional likelihood (with normal distribution) of f, given the random effect $\boldsymbol{b}$ under the equation (4.13). Note that the marginal log-likelihood $l_N(y; \tau, \vartheta)$ in (4.15) has a closed form. The extension of the marginal likelihood approach to GAMM (4.4) and to estimate τ and ϑ jointly by maximizing a marginal quasi-likelihood was proposed by (Wahba, 1985) and (Zhang *et al.*, 1998). Specifically, the GLMM representation of GAMM in (4.7) suggests that τ may be treated as extra variance components in addition to ϑ. Likewise, for REML (4.14), the marginal quasi-likelihood of (τ, ϑ) can be constructed under the GAMM (4.4) by assuming that f_i takes the form (4.5) with $a_i \sim N(0, \tau_i)$ where $i = 1,2.., q$. Therefore, the integration of a_i and β is given as follows:

$$exp\{l_N(y;\tau,\vartheta)\} \propto |\Lambda|^{-\frac{1}{2}} \int ex[\left\{(y;\boldsymbol{\beta},\alpha,\vartheta) - \frac{1}{2}a^T\Gamma^{-1}a\right\} dad\boldsymbol{\beta} \tag{4.16}$$

$$\propto |K|^{\frac{-1}{2}}|r|^{\frac{-1}{2}} \int \left\{\sum_{i=1}^{n} \frac{-1}{2\phi} d_i(y_i;\mu_i) - \frac{1}{2}b^T K^{-1}b - \frac{1}{2}a^T\Gamma^{-1}a\right\}$$

where $l(y;\boldsymbol{\beta},a,\vartheta) = l(y;\beta_0,f_i,f_2,..,f_k\vartheta)$ is the same as in equation (4.2), based on the Gaussian non-parametric mixed model (4.13), the marginal quasi-likelihood reduces to the Gaussian REML (4.15). Laplace's approximation method is another method used to circumvent this problem. Precisely, taking the quadratic expansion exponent of the integrand of the expression (4.17) about its mode before integration and approximating the deviance statistic $d_i(y;\mu_i)$ by the Pearson $\chi^2 - statistic$ therefore, the approximate marginal log-quasi-likelihood is given as:

$$l_N(y;\tau,\vartheta) \approx -\frac{1}{2}log|V| - \frac{1}{2}log(X^TV^{-1}X) - \frac{1}{2}\big(Y - X\hat{\beta}\big)\big(Y - X\hat{\beta}^TV^{-1}\big), \tag{4.17}$$

Where $V = B\boldsymbol{\Gamma}B^T + ZKZ^T + W^{-1}$. The equation (4.17) above corresponds to the REML log-likelihood of the vector y under the linear mixed model in equation (4.9) with a and b as the random effect, and τ and ϑ as the variance component. Hence, τ and ϑ can be estimated by iteratively fitting equation (4.9), using REML.

4.4.3 Model

We consider a sample of N independent random multivariate response $y_1 = (y_{i1},\ldots..y_{in})$; $i = (1,..,N)$, where y_{ij} is the j^{th} response to the i^{th}cluster or subject. We assume that y_{ij} depends on a $p \times 1$ vector of fixed covariates x_{ij} associated with a vector of fixed effect $\boldsymbol{\beta} = \big(\boldsymbol{\beta_1},\ldots.\boldsymbol{\beta}_p\big)'$ and on a $q \times 1$ vector fixed covariate z_{ij} associated with multivariate $q \times 1$ random effect $\boldsymbol{d}_i$. The normality of $f(y_{ij}|b_i,\beta)$, assumed that $f(y_{ij}|b_i,\beta)$ is a member of the exponential family distribution (McCullagh and Nelder, 1989a). Refer to equation (3.5). The conditional mean and variance is given by:

$$E(\boldsymbol{y} \mid \boldsymbol{\theta}) = \sum \frac{\partial b(\theta_{ij})}{\partial \theta_i}$$

(4.18)

$$var(\boldsymbol{y} \mid \boldsymbol{\theta}) = \sum \frac{\partial^2 b(\theta_{ij})}{\partial \theta_{ij}^2}$$

(4.19)

The GLMM with both random and fixed effect is given by:

$$g(\boldsymbol{\theta_{ij}}) = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \boldsymbol{Z}_{ij}'U_{ij}' \tag{4.20}$$

where $\eta_{ij} = g(\theta_{ij})$, g is the link function and U_{ij} is the vector of random effect.

The GLMM can be easily fitted by the DPQL estimator, f_i, defined in section 3.2 using the SAS software. In the previous chapter, we fit the survey logistic regression (PROC SURVEYLOGISTIC) using SAS software, which produces similar result as the GLMM (PROC GLIMMIX). In this chapter, we fit the GAMM model using R software.

4.6 GAMM in R

The procedures for model estimation discussed for fitting the GAMM can be used when fitting the semi-parametric logistic mixed models. The R library $mgcv$ includes $gamm$, which fits the GAMMs based on linear mixed mode, as implemented in the $nlme$ library, fits the specified GAMM to data, by a call to lme in the normal errors identity link case, or otherwise by a call to $gammPQL$. The estimates are only approximate maximum likelihood estimations (Lin & Zhang 1999). The routine is typically slower than gam, and not quite as numerically robust. It is assumed that the random effects and correlation structures are engaged primarily to model residual correlation in the data and that the prime interest is in inference about the terms in the fixed effects model formula, including the smooths Wood (2004 ,2006a,b). For this reason, the routine calculates a posterior covariance matrix for the coefficients of all the terms in the fixed effects formula, including the smooths. In R software, there are many options for controlling the model

smoothness, using splines such as kernel smoothers, cubic smoothing splines, and locally weighted running line smoothers.

4.7 Model fitting and Interpretation of the results

The main effect are considered, where the Akaiker Information Criterion (AIC) of each model is examined and also the inference of smooth function and the p-value of the individual smooth term. Finally, the model with the smallest AIC and highest value of degree of freedom and which is highly statistically significant is given as follows:

$$g(\mu_j) = \beta_o + \beta_1 Sex_j + \beta_2 Residence_j + \beta_3 Education_j + \beta_4 Region_j + f_1 Age\ of\ household\ head_j + b_{0j} \quad (4.21)$$

Where $g(.)$ is the logit link function, $\beta's$ are parametric regression coefficients, $f_j's$ are centred smooth functions and b_{0j} is the random effect distributed as $N(0, K(\vartheta))$. The common widely used methods for estimating additive models include cubic smoothing splines, locally-weghted running line smothers, and kernel smoothers (Hardle, 1999; Hastie and Tibshirani, 1990; Ruppert et al., 2003). The results of model (4.21) are presented in Table 4.1 and Table 4.2.

Table 4.1 The parameter estimates of the fixed effect of GAMM for poverty status

Variables	Estimates	OR	S.E	Z-value	Pr(>|z|)
Intercept	2.50134		0.58210	5.020	$3.5x10^{-07}$**
Region (ref: Manicaland)					
Mashonaland Central	-0.33480	0.71548	0.44867	-0.746	0.455536
Mashonaland East	0.19746	1.21830	0.45266	0.436	0.662640
Mashonaland West	-0.24537	0.78241	0.46625	-0.526	0.598705
Matabeleland North	-1.37253	0.25346	0.49521	-2.772	0.005578***
Matabeleland South	0.27835	1.32095	0.46746	0.595	0.551536
Midlands	-0.66943	0.51200	0.48066	-0.393	0.163699
Masvingo	-0.53477	0.58580	0.45520	-1.175	0.290075
Harare	2.83580	17.0440	4.56686	5.008	0.0001***
Bulawayo	0.68078	1.97541	0.37798	0.2564	0.999819
Education (ref: Preschool or no education)					
Primary	-0.59824	0.54978	0.81245	-1.554	0.000379 ***
Secondary	-0.66209	0.51578	0.81261	-1.605	0.0001***
Higher	-2.86650	0.05689	1.03251	-5.619	0.0001***
Sex (ref: Female)					
Male	-0.07062	0.93182	-2.069	-2.069	0.038586 *
Residence (ref: Urban)					
Rural	1.37312	3.94745	4.28183	4.028	0.0001

From Table 4.1 the estimated parameters of the fixed effect of GAMM for poverty status shows that under the region of Zimbabwe, it was found that Mashonaland North (p-value=0.005578) and Harare (p-value= 0.0001) are significant at 5% level of significance. The poverty odds ratio for Mashonaland North 0.2535 ($e^{-1.3725}$) times that for Manicaland while that for Harare is 17.0441 ($e^{2.8358}$) times that for Manicaland. The education level significantly affects poverty status. People who have primary (p-value=0.000379 ***) or secondary (p-value=0.0001 or higher (p-value=0.0001 levels of education are respectively 0.5496($e^{0.59824}$); 0.5158($e^{-0.66209}$); 0.05689 ($e^{-2.8665}$) times the poverty odds for those with no education. The sex of the household head significantly affects the poverty status. The poverty odd for a male headed household is 0.9894 ($e^{-0.01062}$) times the poverty odds for a female headed household and the effect was found significant (p-value =0.038586). As for effect of residence type, it was found that the

poverty odds for those who reside in rural areas is 3.9476 ($e^{1.37312}$) times the poverty odds for those who reside in urban areas and the effect was significant (p-value= 0.0001 ***).

4.7.1 Approximate smooth function

Table 4.2 Approximate significance of smoothed terms

Smooth terms	Edf	Ref.df	Chi-sq	P-Value
Number of household members ,s(V002)	1.001	1.001	3.23	0.0722
Age of the household head , s(V152)	3.712	3.712	65.43	0.0001

Two continuous covariates (number of household members and age of household head) were fitted. In Table 4.2, the approximate significance of smoothed terms for socioeconomic status are obtained and in Figure 4.1 below, the estimated smoothing function for socioeconomic status are obtained. The smoothers was obtained with $plot(gam1\$gam)$ command and are presented in Figure 4.1. The Y-axis represents the smooth function for the fitted values for socioeconomic status. In both plots A and B, the smooth curve denotes the estimated trend of GAMM**, S** is a smooth term and the number in parentheses represents the estimated degrees of freedom (edf). The results in Table 4.2 confirms the test statistics is 3.23 with p-value=0.0722 and 1.001 edf. The relationship between the number of household members and socioeconomic status is not a significant since p-value is greater than 0.05 and the estimated degree of freedom for the smoother is 1, it's against the assumption that the number of household members is linearly associated to the socio-economic status of the household. The effect of the number of household member s(V002) or size of household head on socioeconomic status is presented in Figure 4.2 A, the trend shows that the plot is linear and it does not fit the model. The effect of the age of household head s(V152) on poverty status was presented in Figure 4.2 B, it seems to meet the assumption of nonlinearly, which means that it best fit the model. It can be observed that poverty odds increase with age of household head up to approximately age 60 and then declines with increasing age of household head from the age of 60 to 80 plus. Figure 4.2, B shows that the age

of household head has a highly significant effect on the poverty status and the relationship is not linear.

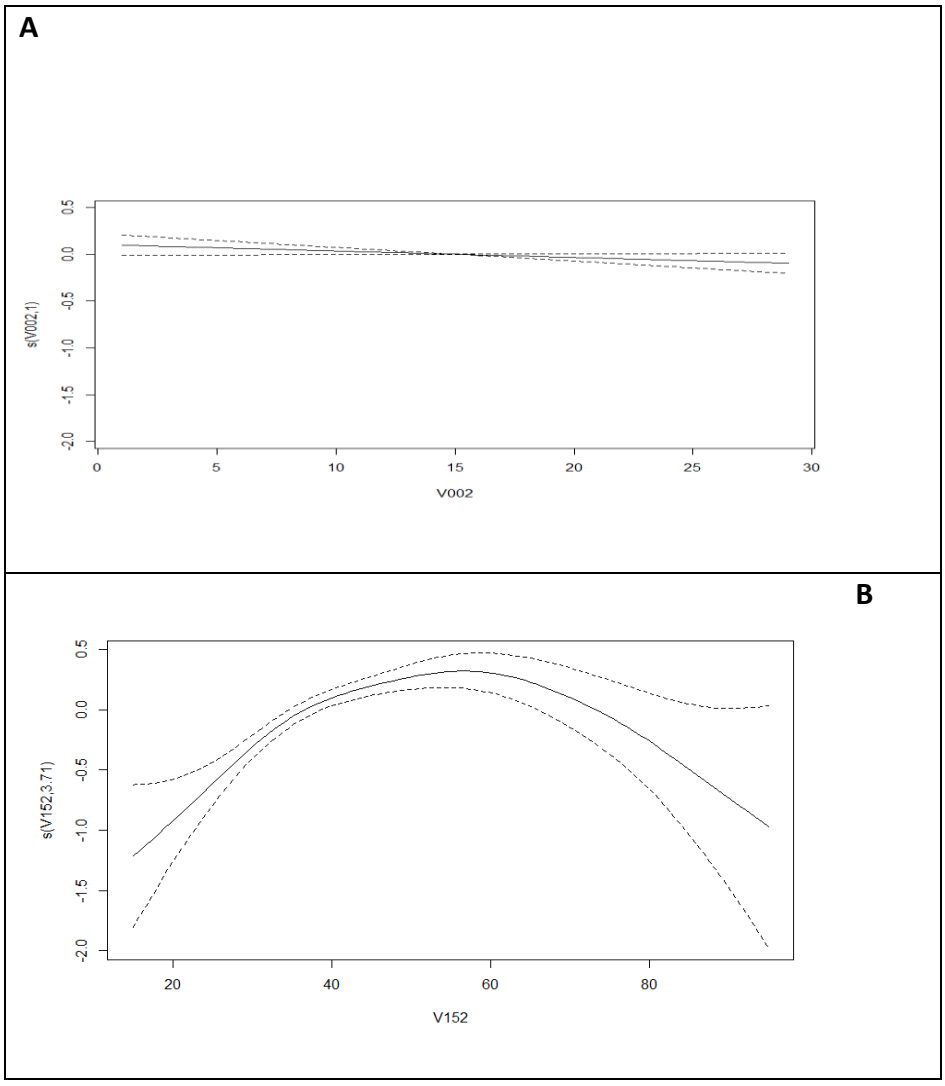

Figure 4.1 Estimated smooth function of poverty status with age of household heads (V1520) and household members s(V002).

4.8 Summary of the GAMM Application

In this chapter, a GAMM was used to identify the risk factors associated with poverty Status. the The main objective was to model the effect of the size of household and the age of household head non-parametrically, while keeping other covariate parametric, using GAMMs. For the parametric covariates, poverty status was found to be higher among rural households than urban areas. The results also confirm

that poverty status decreases with increasing educational level. For sex of household head, the results showed that poverty is higher among female-headed households than male-headed ones. The results also showed that the Mashonaland North region is less likely to be poor while the Harare region is more likely to be poor compared with Mainland. The results from the parametric linear effects showed confirms that Harare is the poorest region in Zimbabwe. For the non-parametric results, it was shown that the relationship between poverty status and the number of household members is linear and the estimated degrees of freedom (edf) is 1.0001 but the relationship is significant at 5% level of significance. The smoothing plot showed a linear trend between poverty and the number of household members. However, the relationship between poverty status and age of household head is nonlinear; the smoothing plot supported the nonlinear trend.

Chapter 5

Discussion and Conclusion

Poverty remains a social and political problem in Africa. Zimbabwe is no different. The study making use of a Zimbabwe DHS dataset (2015), the objective of this study was to model the poverty in Zimbabwe by creating a dichotomous poverty index and to investigate which socio-demographic factors are related to poverty, using statistical models such as GLMs and GAMMs and then to make recommendations to current policies on poverty. The study used SAS 9.4, R and SPSS software.

In Chapter 2, an exploratory analysis of the data was carried out and it was found that the percentage of poor households is higher than that of rich households. To check the cross-tabulation of social-economic status versus explanatory variables (Region, Education, Age, Residence, Sex, and Size) the study used SPSS software and the Chi-square test to check the association between these variables. The results show that all explanatory variables were found significant at 5% level of significance. This implies that there is a significant relationship between explanatory variables and poverty status.

In Chapter 3, GLMs, and survey logistic was used. Survey logistic regression and GLMMs are useful, since they account for the complexity of the survey design and they are often used where there is dependence between the outcome which in our case was dichotomous (1=Poor; 2=Not poor). In SAS software, PROC SURVEYLOGISTIC and PROC GLIMIX usually give similar results, however while PROC SURVEYLOGISTIC fits a marginal model and PROC GLIMIX fits a conditional or a mixed effects model. The assumptions for the two models are different and hence parameter estimates and effects have different interpretations. Therefore, in this application the model was fitted using SAS PROC SURVEYLOGISTIC. The results reveal that all explanatory variables were significant at 5% level of significance, except for the number of household members (size).

In Chapter 4, GAMMs were used to identify factors that affect poverty status. The GAMMs can reveal information that may be hidden when only parametric models are used (it can reveal some

results that survey logistic regression or GLMs cannot obtain). Generalized additive mixed models (semi-parametric) were used to relax the assumption of normality and linearity inherent in linear regression models, where the categorical covariates were modelled parametrically and continuous covariates non-parametrically. The continuous covariates are the age of head of household and number of household members. The results from Figure 4.1 and Table 4.2 reveal that one of the age of household has a nonlinear relationship with poverty status, as the trend shows a nonlinear pattern.

The results from all both models revealed that, in general, the level of education of the household head, gender of household head, age of household head, place of residence and sex of the household head are the key determinants of poverty of households in Zimbabwe. The results from both models (Survey logistic and GAMMs) showed that the sex of household head had significantly affected by poverty and that female-headed household are more likely to be poor compared to male-headed ones, showing that the issues of poverty and gender inequality are longstanding social problems that permeate every society. Women still have less access to land and inputs, although they do most of the agricultural work, a state of affairs that has resulted in a growing differentiation within the rural sector (Balleis, 1993). All job opportunities should be for all regardless of your gender.

The data also confirms that for both models, residents in a rural area are more likely to be poor compared to those who reside in urban areas. The main causes of poverty in Zimbabwe's rural areas have been identified as a lack of sufficient credit, social services and infrastructure (Janvry, Sadoulet & Murgai, 2002). The recommendation proposed is that the relevant government departments should ensure that services of a high standard and value was provided to all areas, regardless of being rural or urban. The educational level results reveal that educated people are less poor than uneducated people, which shows that if the percentage of educated households increased, then the percentage of those who are poor in Zimbabwe would decrease.

The most powerful instrument for reducing poverty and improving the quality of life in the country is economic growth (Adams, R, 2002). Growth can generate virtuous circles of prosperity and opportunity. Strong growth and employment opportunities improve incentives for parents

to invest in their children's education by sending them to school. This may lead to the emergence of a strong and growing group of entrepreneurs, which should generate pressure for improved governance. Strong economic growth therefore advances human development, which, in turn, promotes economic growth. The government of Zimbabwe needs to focus on things that can stimulate the economy of the country. For example such as reducing the cost of borrowing working with central bank to boost demand and improve the quality of infrastructure services.

Bibliography

Adams, R (2002) Economic Growth, Inequality and Poverty: Findings from a New Data Set, Policy Research Working Paper 2972, World Bank, February 2002, and Ravallion, M and S.

Agresti, A. (2002). Categorical data analysis, Volume 359. John Wiley & Sons.

Agresti, A. 2018. *An Introduction to Categorical Data Analysis*, Wiley.

Asogwa, B. C., Okwoche, V. A. & Umeh, J. C. 2012. Analysing the Determinants of Poverty Severity Among Rural Farmers in Nigeria: A Censored Regression Model Approach. *American International Journal of Contemporary Research,* 2**,** 166-176.

Ayad, M., Barrere, B. & Otto, J. 1997. Demographic and Socioeconomic Characteristics of Households.

Apramenya, A. 2016. Cross Tabulation: How to Works and Why should Use it, January-21-2016, 2-4.

Balleis, P. 1993. *A Critical Guide to Esap: Seven questions about the economic structural adjustment programme in Zimbabwe*, Mambo Press in Association with Silveira House.

Bank, W. 2007. *The Status & Progress of Women in the Middle East & North Africa*, World Bank.

Bourguignon, F and Chakravarty,R.S.(2003) "The measurement of Multidimensional Poverty," Journal of Economic Inequality, 1:25-49.

Bracking, S and Sachikonye, L. 2006. Remittances, Poverty Reduction and the Informalisation of the Household Well-being in Zimbabwe, Working Paper, No 45, Global Poverty Research Group, Oxford.

Bradshaw, T. K. 2007. Theories of Poverty and Anti-Poverty Programs in Community Development. *Community Development,* 38**,** 7-25.

Breslow, N. E. & Clayton, D. G. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association,* 88**,** 9-25.

Carter, M. R. & May, J. 1999. Poverty, Livelihood and Class in Rural South Africa. *World Development,* 27**,** 1-20.

Chimhowu, A. & Woodhouse, P. 2008. Communal Tenure and Rural Poverty: Land Transactions in Svosve Communal Area, Zimbabwe. *Development and Change,* 39**,** 285-308.

Chinake, H. 1997. *Strategies for Poverty Alleviation in Zimbabwein*, Journal of Social Development in Africa.

Cosner, L.,1967. Continuities in the Study of Social Conflict. New York: Free press.

Cox, D. R. 1972. The Analysis of Multivariate Binary Data. *Applied Statistics***,** 113-120.

Denavas-Walt, C. 2010. *Income, Poverty, and Health Insurance Coverage In The United States (2005)*, Diane Publishing.

Dobson, A. J. & Barnett, A. G. 2018. *An Introduction To Generalized Linear Models*, Crc Press.

Dzingirai, V, Mutopo, P and Landau, L. 2014. Confirmations, Coffins and Corn: Kinship, Social Networks and Remittances from South Africa to Zimbabwe, Migrating out of Poverty Research Programme, University of Sussex, Sussex.

Ellsberg, M. C., Pena, R., Herrera, A., Liljestrand, J. & Winkvist, A. 1999. Wife Abuse Among Women of Childbearing Age in Nicaragua. *American Journal Of Public Health,* 89**,** 241-244.

Estache, A. 2017. Successes and Failures of Water and Sanitation Governance Choices in Sub-Saharan Africa (1990-2017). *Ecares Working Papers*.

Eyob, F. & Mark, H. (2004). *Modelling Determinants of Poverty in Eritrea*: A New Approach.

Filmer, D. and L. H. Pritchett, "Estimating Wealth Effects without Expenditure Data—or Tears: An Application to Educational Enrollment in States of India," Demography, 38, 115–32, 2001

Gill, J. 2001. *Generalized Linear Models*, Thousand Oaks, Ca. Sage Publications

Gitelman, D. R., Penny, W. D., Ashburner, J. & Friston, K. J. 2003. Modeling Regional and Psychophysiologic Interactions in Fmri: The Importance of Hemodynamic Deconvolution. *Neuroimage,* 19**,** 200-207.

Green, P. J. & Silverman, B. W. 1993. *Nonparametric Regression and Generalized Linear Models:A Roughness Penalty Approach*, Crc Press.

Greenland, S. 1994. Alternative Models for Ordinal Logistic Regression. *Statistics in Medicine,* 13**,** 1665-1677.

Hardle, W. (1999). Applied nonparametric regression. New York: Cambridge University Press.

Hastie, T. J. & Tibshirani, R. J. 1990. Generalized Additive Models, Volume 43 Of Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Hastie, T. J. and R. J. Tibshirani (1990). Generalized additive models. New York: CRC press.

Haru M, 2015. Zimbabwe fails to curb escalating unemployment crisis.

Hoddinott, J. 2006. Shocks and Their Consequences Across and Within Households in Rural Zimbabwe. *The Journal of Development Studies,* 42**,** 301-321.

Janvry, A. de, E. Sadoulet, and R. Murgai. 2002. "Rural Development and Rural Policy." In B.GardnerG. Rausser (eds.), Handbook of Agricultural Economics, vol. 2, A, Amsterdam: NorthHolland: 1593–65.

Jewkes, R. 2002. The Lancet. *Intimate Partner Violance : Couses and Preventation.*

Jewkes, R., Levin, J. & Penn-Kekana, L. 2002. Risk Factors for Domestic Violence: Findings from a South African Cross-Sectional study. *Social Science & Medicine,* 55**,** 1603-1617.

Kohn, R., Ansley, C. F. & Tharm, D. 1991. The Performance of Cross-Validation and Maximum Likelihood Estimators of Spline Smoothing Parameters. *Journal of the American Statistical Association,* 86**,** 1042-1050.

Library, M. S. U. 1997. Strategies for Poverty Alleviation in Zimbabwe. *Strategies for Poverty Alleviation in Zimbabwe.*

Lin, X. & Zhang, D. 1999. Inference in Generalized Additive Mixed Modelsby Using Smoothing Splines. *Journal of The Royal Statistical Society: Series B (Statistical Methodology),* 61**,** 381-400.

Littell, R. C., Stroup, W. W. & Freund, R. J. 2002. *SAS for Linear Models*, SAS Institute.

Manjengwa, J., Kasirye, I. & Matema, C. 2012. Understanding Poverty in Zimbabwe: A sample survey in 16 Districts; Paper prepared for presentation at the centre for the study of African economies conference 2012 "Economic Development in Africa", Oxford, United Kingdom, March 18-20, 2012.

Mccullagh, P. & Nelder, J. A. 1989a. *Generalized Linear Models*, Crc Press.

Mccullagh, P. & Nelder, J. A. 1989b. Generalized Linear Models, Vol. 37 Of Monographs on Statistics and Applied Probability. Chapman and Hall, London.

McCulloch, C. E. and Neuhaus, J. M. (2001). Generalized linear mixed models. Wiley Online Library.

Nations, U. 2009. Entity for Gender Equality and the Empowerment of Women. *In Convention on ohe Elimination of All Forms of Discrimination Against Women (Cedaw)*.

Nelder, J. A. & Wedderburn, R. W. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General),* 135**,** 370-384.

Nsingo, E. 2010. Zimbabwe ow a Factory for Poverty.

O'connell, A. A. 2006. *Logistic Regression Models for Ordinal Response Variables*, Sage.

Olsson, U. 2002. Generalized Linear Models. *An Applied Approach. Studentlitteratur, Lund,* 18.

Pauw, K. & Thurlow, J. 2011. Agricultural Growth, Poverty, and Nutrition in Tanzania. *Food Policy,* 36**,** 795-804.

Pindiriri, C. 2015. Modeling the Determinants of Poverty in Zimbabwe.

Preface to the Second Edition: The history of the Oxford English Dictionary: The New Oxford English Dictionary project". *Oxford English Dictionary Online. 1989*. Archived from the original on 16 May 2008. Retrieved 16 May 2008.

Raftopolous, B. 2011. A Study on Migration and Remittances in Matebeleland, Zimbabwe, Solidarity Peace Trust, Cape Town.

Roubaud, F. and RAZAFINDRAKOTO, M. (2003) "The multiple facet of poverty: the case of Urban Africa, Provisional version.

Ravallion, M.(1996)Issues in Measuring and Modelling Poverty, The Economic Journal,106,1328-1343

Robinson, G. K. 1991. That Blup is a Good Thing: The Estimation of Random Effects. *Statistical Science,* 6**,** 15-32.

Rodrìgues 2001. Generalized Linear Models Theory Revised November.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). Semiparametric regression. Cambridge University Press.

Sakuhunni, R. 2011. Economic Determinants Of Poverty in Zimbabwe. Clainos Chidoko, Et. Al. *International Journal of Economics Research,* 2**,** 1-12.

Santner, T. J. & Duffy, D. E. 1986. a note on A. Albert And Ja Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika,* 73**,** 755-758.

Silverman, B. W. 1985. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 1-52.

Snijders, T. A. 2005. Power and Sample Size in Multilevel Modeling. *Encyclopedia of Statistics in Behavioral Science,* 3**,** 1570-1573.

Shackman, G. (2001). Sample size and design effect. Albany Chapter of the American Statistical Association.

Steinmetz, S. 1987. Family Violence: Past, Present and Future. *Handbook of Marriage and The Family.* New York: Plenum Press.

Stevens, P. 2017. Diseases of Pov.erty and The 10/90 Gap. *Fighting the Diseases of Poverty.* Routledge.

Snaebjorn, G. Alain, BL. *et al.* Constructing Indices of Rural Living Standards in Northwestern Bangladesh. *J Health Popul Nut.r* 2010 Oct; 28(5):509-51

Stillwaggon, E. 2000. Hiv Transmission in Latin America: Comparison with Africa and Policy Implications. *South African Journal of Economics,* 68**,** 444-454.

Teal, F. 2011. The Price of Labour and Understanding the Causes of Poverty. *Labour Economics,* 18**,** S7-S15.

Torres-Reyna, O. 2007. Panel Data Analysis Fixed and Random Effects Using Stata (V. 4.2). *Data & Statistical Services, Priceton University*.

Wahba, G. 1985. A Comparison of Gcv and Gml for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals Of Statistics***,** 1378-1402.

Whiteside, A. 2002. Poverty and Hiv/Aids in Africa. *Third World Quarterly,* 23**,** 313-332.

Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association. 99:673-686.

Zhang, D., Lin, X., Raz, J. & Sowers, M. 1998. Semiparametric Stochastic Mixed Models for Longitudinal Data. *Journal of The American Statistical Association,* 93**,** 710-719.

World Bank *(1990) world Development report 1990: Poverty,* Washinton DC: World Bank.

World bank, *2012, Africa's Pulse*, (Washingson : World Bank).

WHO, *World Health Report*, 2002

World Bank. (2007). The Little Data Book on Africa. Washington DC: World Bank. Available at: http://siteresources.worldbank.org/INTSTATINAFR/ Resources/LDB_Africa_final.pdf.

World Bank, (2004). Rural and Micro Finance Regulation in Ghana: Implication for Development of industry. New York: World B

World Bank. 2008. Doing business 2009: Country profile for Zimbabwe. Washington DC: The World Bank

Zimbabwe Natinal Statistics Agency,"Zimbabwe Poverty Report 201" May 2019 http://www.zimstat.co.zw.

Zimbabwe National Statistics Agency (2002). National census report. (2014a) Labour force and child labour survey.

Zimbabwe National Statistics Agency (2002). National census report, (2015b). Zimbabwe Poverty Atlas.

Apramenya, A. 2016. Cross Tabulation: *How to Works and Why should Use it*, January-21-2016, 2-4.

Appendices

Code used in SAS

Appendix 1

This SAS code was used to fit survey logistic regression in chapter three.

```
Proc import out= work.philile
            datafile= "e:\precious.sav"
            dbms=spss replace;

Run;


Title'survey logistic sas procedure';
Proc surveylogistic data =philile order=data;
Stratum v002;
Weight weights;
Class region (ref="manicaland") education (ref="no education")
sex (ref="female")
Age (ref="15-24")  size (ref="1-5") residence(ref="urban");
Model sas2 (ref = last) = region education sex age residence
size / link=logit rsq;
Run;
```

Code used in r

Appendix 2

This R code in appendix 2 was use to fit Generalize additive mixed model in chapter four.

Library(foreign)

Dat=read.spss("precious.sav", to.data. frame = true)

Library(lattice) #needed for multi-panel graphs

Library(r2jags)

Library(nlme)

Library(mgcv)

Library(gamm4)

```
Dat

Table(dat$fsex)

Dat$fsex <- factor(dat$sex)

Dat$fresidence<- factor(dat$residence)

Dat$fregion <- factor(dat$region)

Dat$feducation <- factor(dat$education)

Dat$age_of_household = dat$v152


Gam1 = gamm4(sas2 ~ fsex + fresidence + feducation + fregion + s(v002) +
s(age_of_household),

        random =~ (1|cluster), family = "binomial", data = dat)

Summary(gam1$gam)

Dat$age

Table(dat$age)

Dat$sas2

Dat$v152


Summary(gam1$gam)

Anova(gam1$gam)

Plot(gam1$gam)

Dat$v002
```