APPLICATION OF COUNT MODELS IN THE DETERMINATION OF UNDER-FIVE MORTALITY RATE IN SOUTH AFRICA

by

KGETHEGO SHARINA MAKGOLANE

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Dr. KD MOLOI

CO-SUPERVISOR: Prof. A TESSERA

2022

Declaration

I, **Kgethego Sharina Makgolane**, hereby declare that the dissertation submitted to the University of Limpopo, for the Master of Science degree in Statistics has not been submitted by me for a degree at this or any university; that it is my work in design and execution and that all material contained herein has been duly acknowledged.

Signature:_____ Date: 22 April 2022

Makgolane, K.S (Ms)

Abstract

Under-Five Mortality (U5M) remains a major health challenge in most sub-Saharan African countries including South Africa, despite the significant progress made in child survival and the government's efforts and commitment to reduce U5M. The failure of achieving the fourth Millennium Development Goal (MDG) by 2015 has led to an implementation of Sustainable Development Goal 3 (SDG3) which aims to have no more than 25 deaths per 1000 live births by 2030. To achieve this goal, more information is needed. Hence, the purpose of this study was to apply count models to identify the determinants of under-five mortality rate in South Africa. To identify these determinants, the study reviewed generalized linear models and utilised the 2016 South African Demographic and Health Survey data. The models studied were Logistic Regression (LR), Poisson Regression (PR) and Negative Binomial Regression (NBR). The findings revealed that baby postnatal check-up, child's health prior discharge, child birth size, toilet facility, maternal education, province, residence and water source were significantly associated with U5M in South Africa. It was further concluded that children who are at high risk of dying before the age of five are those who did not attend their postnatal check-up within the first two months, those whose health was not checked prior discharge, whose birth size was very small, whose household utilised bucket toilets, who resided in Western Cape, North West and Mpumalanga province, who resided in urban areas as well as those whose household utilized piped, tube well and spring water as source of drinking water.

Dedication

This dissertation is dedicated to my parents for emphasizing the importance of education at all times. I would also like to dedicated this study to my siblings.

Acknowledgements

I would like to give thanks to the Almighty God for giving me strength throughout this dissertation, despite all the challenges I have encountered during the period of this study. I would like to extend my sincere gratitude to my supervisor Dr KD. Moloi and co-supervisor Prof A. Tessera for their assistance and guidance throughout the study. I am grateful for their significant contribution in this dissertation. I would also like to thank my family for the continuous and unfailing support and encouragement that they have given me throughout the course of this study. To my fellow students and friends whom I have interacted with during the period of this study, I would like to thank you for your support and encouragement.

I would like to also give thanks to the Health and Demographic Surveillance System for providing me with data used for this dissertation. I further express my gratitude to the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) of South Africa for providing me with funds for my studies for the period of this study, and for also funding for my trip to attend South African Statistical Association (SASA) workshop and conference in 2019. I extend my gratitude to the Department of Statistics and Operational Research for providing me with adequate research infrastructure and resources throughout the study.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
BIC	Bayesian Information Criterion
CoE-MaSS	Centre of Excellence in Mathematical and Statistical Sciences
GLM	Generalized Linear Models
HDSS	Health and Demographic Surveillance System
HIV	Human Immune-deficiency Virus
LR	Likelihood Ratio
LR	Logistic Regression
MDGs	Millennium Development Goals
MDG4	Millennium Development Goal 4
MLE	Maximum Likelihood Estimate
NBR	Negative Binomial Regression
PR	Poisson Regression
SADHS	South African Demographic and Health Survey
SDGs	Sustainable Development Goals
SDG3	Sustainable DEvelopment Goal 3
SPSS	Statistical Package for Social Sciences
U5M	Under-Five Mortality
U5MR	Under-Five Mortality Rate
UNICEF	United Nations Children's Fund
WHO	World Health Organisation

Chapter 1

Introduction and background

1.1 Introduction

Under-Five Mortality Rate (U5MR) is the probability of dying between birth and 5 years for every 1000 live births (UNICEF., 2008). According to Liu et al. (2015), significant progress has been made in child survival over the past two decades globally. However, the progress made in the past decades seem not to be enough to reduce U5MR. Despite the substantial drop in the global child mortality rate, approximately 6.6 million children die each year before their fifth birthday worldwide (Kanmiki et al., 2014). Most of these worldwide death cases have been dominant in sub-Saharan Africa–which implies that the sub-Saharan African countries bear the highest burden of child death before the age of five years. Therefore, under-five mortality (U5M) remains a major health challenge in most sub-Saharan African countries, including South Africa (Chadoka, 2011).

As a result, children in South Africa and other countries in sub-Saharan Africa

are at a high risk of dying before the age of five years. Despite the government's efforts and commitment to create an environment that helps provide quality health care and reduce mortality, South Africa has failed to achieve its Millennium Development Goal 4 (MDG4) of reducing U5MR by two-thirds between 1990 and 2015 like many countries in the world (StatsSA, 2015). StatsSA (2015) found that the progress in tackling U5M in South Africa has been hindered by an early increase in child mortality rates, mainly due to HIV and AIDS, which turn out to be the prominent cause of death among children under five years of age in the initial decade of Millennium Development Goal (MDG) period.

Due to South Africa's failure to meet the MDGs by 2015, 2016 marked the beginning of the implementation of a new global development agenda known as the Sustainable Development Goals (SDGs) (Sachs, 2015). The purpose of the SDGs is, among other things, to take sustainable measures to reduce child mortality by 2030 (IAEG, 2016). SDG aims to have no more than 25 U5MR per 1000 live births in any country in the world by 2030 (Liu et al., 2016). To achieve this goal, information is needed about the current causes of children's deaths to help accelerate activities that improve child survival.

To accelerate the reduction of Under-Five Mortality Rates below the SDG target by 2030 in South Africa, specific proven interventions will need to target important causes of child mortality, as no single factor can be responsible for the high child mortality (Bhutta et al., 2008; Adhikari and Podhisita, 2010). The development of these interventions must include the multiplication of factors that determine under-five mortality.

Although numerous studies conducted found factors such as mother's education, mother's age, gender, place of delivery, black mothers, breast feeding and

mother's survival status to be associated with U5M in South Africa (Worku, 2011; Buwembo, 2010; Makgaba, 2014; Hlongwa and De Wet, 2019), Apunda (2016) found that changes in awareness levels and daily amenities, meant that the predictors of U5M change over time. It is therefore important to periodically investigate the causes of under-five mortality and related risk factors for policymaking and survival interventions in South Africa for children under the age of five years.

Therefore, this study focused on investigating the factors associated with U5M with the aim to develop measures of achieving SDGs. The data of this study can be useful to policy makers, scholars and health sector administrators as the country strives to achieve the SDG of reducing child mortality to 25 deaths per 1000 live births.

1.2 Problem statement

High U5MR has been a serious concern to the global community. However, the estimated global rate of 93 deaths per 1000 live births in 1990 has reduced to 43 deaths per 1000 live births in 2015. This reduction was made under the MDG4 even though the goal was not achieved globally prior 2015 (UNICEF, 2015). Regardless of this considerable drop in global U5MR, most sub-Saharan African countries including South Africa, are still leading in the U5MRs with one in every nine children dying before the age of five years (Kanmiki et al., 2014). Hence, the SDG3 was introduced to reduce the U5M to at least less than 25 deaths per 1000 live births (IAEG, 2016).

South Africa has made significant progress towards the reduction of U5MR but has not achieved its MDG 4 (StatsSA, 2015; Motala et al., 2015). Thus, more

needs to be done to ensure that South Africa achieves the SDG 3 by reducing the under-five mortality to at least as low as 25 deaths per 1000 live births by 2030.

1.3 Rationale

Despite the efforts made by the South African government to reduce its under-five mortality rate to 28.5% in 2018, U5M is still a major health concern in the country (StatsSA, 2019). This aligns with the census conducted in 2011, which reported that more than 50 000 children died with about 80% of the deaths occurring before the age of five years (Hlongwa and De Wet, 2019). To effectively reduce the rate of under-five mortality, the focus in research should be on the factors that contribute to under-five mortality. Mustafa and Odimegwu (2008) concur that a significant pace of decline in child mortality can be achieved if the strategies and policies of mortality are directed towards associated factors.

It was important to conduct this research using count models because most of the previous studies were done using multivariate logistic regression model and survival analysis. Multivariate logistic regression and survival analysis models have contributed considerably to the current knowledge of risk factors associated with U5M (Oritogun and Bamgboye, 2018). However, these approaches may result in loss of information and may lower the statistical power of the model as opposed to treating the response variable as a count data. Hence, the use of count data models preserved the power of statistical analysis by using true values of the response variable without collapsing its value as in the categorical response variable.

Maternal education is an important indicator or factor in deaths of children

who are less than five years (Mokoena, 2011). Kanmiki et al. (2014) utilised logistic regression to investigate the factors associated with U5M in northern Ghana and found that maternal education and age were significantly associated with U5M. This was in accordance with other available literature (Caldwell, 1979; Buor, 2003; Buwembo, 2010; Ettarh and Kimani, 2012; Hlongwa and De Wet, 2019) which reveal that maternal education and age are strong predictors of under-five mortality.

The study conducted by Makgaba (2014) in Ga-Dikgale regarding the Health and Demographic Surveillance System (HDSS) of South Africa, utilised both logistic regression and survival analysis–the aim of which was to determine factors that have effect on child survival for children born between 01 January 1996 and 31 December 2010 in Ga-Dikgale HDSS. The study of Makgaba (2014) found that childbirth weight and mother's survival status were strongly associated with child survival. Hence, it was concluded that childbirth and mother's survival status were factors that had significant effect on child survival and child survival time.

1.4 Purpose of the study

1.4.1 Aim of the study

The aim of this study was to apply count models in order to identify the determinants of under-five mortality rate in South Africa.

1.4.2 Objectives of the study

The objectives of the study are to:

1. Fit various count models.

2. Compare the count models and select the best fitted model.

3. Establish factors associated with under-five mortality rate in South Africa.

4. Compare the factors of mortality among Provinces.

1.5 Methodology and Analytic procedure

The study used data from the South African Demographic and Health Survey of 2016 to identify the determinants of under-five mortality. The study further explored the application of Chi-square test of association, logistic regression model and count data regression models namely, Poisson Regression model and Negative Binomial Regression model. It was arranged that when the Poisson assumption fails, negative binomial regression model would be considered to address the issue of overdispersion by introducing a dispersion parameter to accommodate for unobserved heterogeneity in count data.

Maximum likelihood estimator was utilised in the study to suggest the parameter estimation. The likelihood ratio test was also used to test the overall fit of the data while the goodness of fit test was used to check the adequacy of the models. Furthermore, the Akaike's Information criterion and Bayesian Information Criterion was used to select a best model from a set of available models. The researcher further used the Statistical Package of Social Science (SPSS) and R program to handle the numerical data of this investigation.

1.6 Scientific contribution

The findings of this study could be useful not only to the health workers and policy makers but also to the parents. It is anticipated that the study would educate parents about the determinants or the factors that contribute to under-five mortality. The findings of this study could also be used to strengthen the existing guidelines and to generate the mechanism of applying the policies and program interventions that can support the reduction of under-five mortality. Furthermore, the findings of this study could be used to successfully achieve the SDG3 in 2030. Moreover, the findings could serve as a benchmark to track the survival characteristics of children under the age of five years during SDG era and South Africa's Health Sector Transformation Plan.

1.7 Outline of the study

This study consists of five chapters. Chapter 1 presents the introduction, problem statement, rationale, aim and objectives of the study, methodology and analytic procedures, scientific contribution and an outline of this study. Chapter 2 reviews literature on factors that contribute to under-five mortality in South Africa and other relatable parts of the world. Chapter 3 present the statistical analysis methods that were used in this study. Chapter 4 presents the results and discussion of the study. Chapter 5 presents the conclusion, recommendations and limitation of the study based on the findings generated in this study.

Chapter 2

Literature review

2.1 Introduction

This chapter presents literature review based on previous studies that were conducted about factors associated with under-five mortality in South Africa and worldwide.

2.2 Literature reviewed

Most researchers have studied various determinants of under-five mortality globally, particularly in South Africa. Factors such as gender of a child, maternal education, maternal age, maternal occupation, source of drinking water, type of toilet facilities, region, type of residence, duration of breastfeeding and marital status of the mother were found to be associated with under-five mortality. Some of the above-mentioned factors are extensively discussed below under four broad factors, namely, maternal factors, child factors, socio-economic

factors, and environmental and health factors.

2.3 Maternal factors

Maternal factors such as maternal education, maternal occupation, marital status of the mother and maternal age group has been found to have effect on under-five mortality and are discussed below.

2.3.1 Maternal education

Maternal education has been found to be a significant factor that influence under-five mortality in many countries.

Chowdhury et al. (2010) conducted a study on socio-economic determinants of neonatal, post-neonatal, infant and child mortality. The purpose of their study was to determine the socio-economic factors that affect infancy and childhood mortality. Their study used Logistic regression analysis and revealed that mothers' education and occupation are influential factors of neonatal, post-neonatal, infant and child mortality.

Angela and Uju (2015) employed Cox proportion hazard model and Cox frailty model in their study to determine the responsible factors for under-five mortality in Nigeria. Their study found that mothers' education was a significant determinant of under-five mortality. According to Angela and Uju (2015), educated mothers are more likely to seek medical attention for their children and they have a greater say in their childcare issues.

Dendup et al. (2018) studied factors associated with under-five mortality in Bhutan. Their study utilised data from the Bhutan national health survey 2012 and used Logistic regression to identify factors that influenced under-five

mortality. Dendup et al. (2018) found that maternal education was a significant factor that influenced under-five mortality. It was thus concluded that empowering women through education will have a larger impact on child survival. Other studies such as the ones of Mokoena (2011); Mugarura and Kaberuka (2015) and Worku (2011) have also found positive association between maternal education and under-five mortality. According to Mokoena (2011), educated mothers experience lower child mortality due to the better knowledge they have about various child diseases and being more involved in health seeking behaviours.

2.3.2 Maternal and Partners occupation

Maternal occupation determines the wealth status of a household. According to Chowdhury et al. (2010), parental occupation determines the economic status, nutrition, housing condition, access to health facilities and clothing of a family.

Shifa et al. (2018) conducted a study on socioeconomic and environmental determinants of under-five mortality in Gamo Gofa Zone, Southern Ethiopia. Their study found that husbands' occupation was significantly associated with under-five mortality and that maternal occupation was not significant. They regarded the occupation of a husband as an indicator of socioeconomic status of the household and stability of the family.

However, other researchers such as Buwembo (2010), Kayode et al. (2012), Kanmiki et al. (2014), Nafiu et al. (2016) and Peter et al. (2017) found that maternal and paternal occupations were not significant determinants of under-five mortality.

2.3.3 Marital status of the mother

Goro (2007) studied the stalling child mortality in Ghana. The study used logistic regression analysis to model the socio-economic factors that influence child mortality. It was shown that marital status was a statistically significant determinant of child survival.

Worku (2011) conducted a study on survival analysis of South African children under the age of five years. The study used Logistic regression analysis to identify the key predictors of mortality among children under the age of five years. The study revealed that marital status was a key predictor to under-five mortality. Furthermore, the study showed that the odds for children whose parents are single were 1.74 times more likely to die before their fifth birthday.

Achola (2014) studied the effect of mother's migration on under-five mortality in Kenya. Cox proportion hazard model was utilized in the study. It was shown that the marital status of the mother had a significant effect on the child survival up to the age of five for migrants and non-migrants. They also found that children whose mothers are married have greater chances of survival chances compared to those whose mothers who are not married.

2.3.4 Maternal age group

Buwembo (2010) investigated whether the association of specific factors persists over time using data from 1997-2002 household survey in South Africa. The study analysed births that occurred in the five years preceding each survey. Logistic regression technique was used to determine the relative contribution of each factor over the two periods under review (1993-1997 and 1998-2002). Buwembo's study found that mother's age at the time of delivery was

significant factor for the 1993-1997 period. It was also revealed that children who are given birth by mothers aged 35 years and above are ten times more likely to survive compared to those born to mothers aged 18-34 years. However, Buwembo's study found that mother's age was not significant for the period 1998-2002.

Ettarh and Kimani (2012) conducted a study on determinants of under-five mortality in rural and urban Kenya. Cox proportional hazards regression was utilised to investigate the effects of demographic, geographic and maternal factors on under-five mortality. Their study found that mother's age was a significant determinant of under-five mortality in rural and urban Kenya. Furthermore, the study revealed that highest likelihood of survival was among children of mothers aged 32 years or more. They further stated that younger mothers are not socially and psychologically mature to deal with the requirements of childcare that older mothers may have.

The study conducted by Chowdhury (2013) on Bangladesh demographic and health survey 2007 dataset used Chi-square test for independence and proportional hazard analysis to identify the determinants of under-five mortality in Bangladesh. This study found that mother's age has a strong influence on under-five mortality. Moreover, it was revealed that the relative risk of under-five mortality for children born to mothers aged 25-34 is 36% lower than of children born to mother aged less than 20 years.

Other studies such as that of Kanmiki et al. (2014) also found positive association between mother's age and under-five mortality. However, Makgaba (2014) found that there is no significant association between mother's age and under-five mortality in his study.

2.4 Child factors

Factors such as the gender of a child, the size of a child at birth, place of delivery and duration of breastfeeding have been noted to have various effects on under-five mortality as discussed below:

2.4.1 Gender of child

Mokoena (2011) examined risk factors with high infant and child mortality in Lesotho. Logistic regression was used to identify the risk factors associated with the mortality of children under-five years of age. It was found that gender of a child is a significant risk factor of infant mortality. Furthermore, this study revealed that being a male child is associated with increased risk of dying before the age of five.

Peter et al. (2017) utilised the Zambian Demographic and Health Survey (2013-2014) dataset to identify the determinants of under-five mortality. Their study found that child's gender was a significant determinant of under-five mortality in Zambia. They further discovered that female children are less likely to die compared to males.

Iqbal et al. (2018) found that gender inequality is associated with increased child mortality, and female children seem to disproportionally suffer from it. They stated that the more gender unequal a society is, the more girls are penalised in terms of survival chances in low-income countries and middle-income countries.

Kumar and Sahu (2019) studied socio-economic, demographic and environmental factors effects on under-five mortality in Empowered Group States in India. Their study used Kaplan-Meier with log-rank test and Cox proportional hazard

regression model to assess the socio-economic, demographic and environmental factors effects on risk of under-five mortality. It was found that the gender of a child was significantly associated with under-five mortality. Furthermore, it was shown that male children are at a high risk of death as compared to their female counterparts.

However, studies such as the one of Makgaba (2014) obtained different results from those of Mokoena (2011), Peter et al. (2017), Iqbal et al. (2018) and Kumar and Sahu (2019) because they found that the gender of a child does not have a significant impact on child mortality.

2.4.2 Size of child at birth

Fikru et al. (2019) conducted a study on proximate determinants of under-five mortality in Ethiopia using 2016 nationwide survey data. Bivariate and multivariate logistic regression analysis were used to identify the risk factors of under-five mortality in Ethiopia. The study found that the size of a child at birth was significantly associated with under-five mortality. Additionally, the odds of under-five mortality were shown to be higher for very small size children at birth as compared to average size children.

Tagoe et al. (2020) studied a predictive model and socio-economic and demographic determinants of under-five mortality using data from the 2008 and 2013 Sierra Leone Demographic and Health Survey. LASSO regression and logistic regression techniques were employed to examine the risk factors that account for under-five mortality. The study found that the size of child at birth was statistically associated with under-five mortality in Sierra Leone. In-addition, the study found that children with smaller than average and very small were significantly associated with high odds of dying before the age of five compared to children whose weights were very large at birth.

2.4.3 Place of delivery

Singh and Tripathi (2013) studied factors contributing to under-five mortality at birth order 1 to 5 in India. Logistic regression was used to assess factors contributing to under-five mortality in India. It was found that place of delivery was a significant factor of under-five mortality for birth order two but was not significant for other birth orders.

Yaya et al. (2018) used Demographic and Health Survey data from five sub Saharan countries to check the under-five mortality patterns and maternal risk factors in sub Saharan Africa. Multivariate Cox proportional hazards regression was used to model maternal factors associated with under-five mortality. The study found that delivery by caesarean section was significantly associated with under-five mortality in Chad, Democratic Republic of Congo, Mali, Niger and Zimbabwe.

Alabi (2018) utilised Principal component analysis and multiple linear regression to examine the risk factors of child mortality. Place of delivery was found to be significant factor of child mortality in Nigeria. Buwembo (2010) also found that place of delivery was a significant factor determining under-five mortality in South Africa.

Several studies Ettarh and Kimani (2012), Singh and Tripathi (2013), Ahmed et al. (2016) and Amoroso et al. (2018) found that place of delivery was not a significant determinant of under-five mortality.

2.4.4 Duration of breastfeeding

Akwara (1994) studied breastfeeding as well as infant and child mortality in Amagoro District, Kenya. Logistic regression was used to examine the impact

of breastfeeding duration and age at supplementation on infant and child mortality. The study found that the duration of breastfeeding significantly influenced child mortality. Moreover, a decline in duration of breastfeeding among educated women was observed.

Multilevel analysis of factors associated with child mortality in Uganda was conducted by Mugarura and Kaberuka (2015) using the 2016 demographic and health survey data. The purpose of the study was to examine the factors associated with child mortality using a hierarchical or multilevel regression model. The study found the duration of breastfeeding to be statistically significant. It was also revealed that child mortality was frequently high for children who have never been breastfed.

Ettarh and Kimani (2012), also found the duration of breastfeeding to be a significant factor associated with under-five mortality in Kenya. Moreover, it was shown that its influence on the likelihood of under-five mortality was similar in rural and urban areas with children who were breastfed for more than six months having a significantly lower probability of mortality compared to children breastfed for less than six months.

Worku (2011) and Acheampong and Avorgbedor (2017) found that the duration of breastfeeding was a significant predicator of under-five mortality in their studies. Furthermore, Acheampong and Avorgbedor (2017) found that breastfeeding beyond the age of 19 months was associated with malnutrition.

On the other hand, Saroj et al. (2019) conducted a study on survival parametric models to estimate factors of under-five mortality. The study utilized parametric and semi-parametric models to identify the significant factors. Their study revealed that breastfeeding was statistically significant in the child's survival

status.

2.5 Socio-economic factors

Socio-economic factors such as place of residence and wealth index were found to have effect on under-five mortality and are discussed below.

2.5.1 Place of Residence

Worku (2011) found that urban children have more likelihood of survival than those in rural areas. Furthermore, the study found that urban children are less likely to die as compared to rural children by a factor of 64%. This might be because of their access to basic health care services, proper sanitation, clean and safe water and employment opportunities etc. Worku (2011).

Negera et al. (2013) used data from the 2000, 2005 and 2011 Ethiopian Demographic and Health Survey to examine the determinants of infant and under-five mortality in the five years preceding the survey. By employing Cox proportional hazard model, they found that region of residence had an influence on infant and under-five mortality in Ethiopia. They stated that a place of residence is one of the proximate determinants that influence infant and under-five mortality through the immediate determinants.

Adedini and Odimegwu (2014) found that the place and region of residence were important determinants of under-five mortality in Nigeria. Their study used the 2003 and 2008 Nigerian Demographic and Health Survey data and employed multilevel Cox regression to examine the effects of neighbourhood contexts on under-five mortality in Nigeria. Their study concluded that a place of residence was a significant predictor of under-five mortality in Nigeria.

2.5.2 Wealth index

Wealth index is defined as the measurement of varying socio-economic statuses based on asset based indices that are computed from different households' assets and quintiles using principal component analysis (Tlou et al., 2018). According to Tessema (2015), wealth index emerges as a powerful background covariate of under-five mortality in the EAG states India because wealth index is known to be associated with better childcare practices.

Samuel and Amoo (2014) conducted a statistical analysis on under-five mortality using data from the 2008 Nigeria Demographic and Health Survey. Their study employed the Logistic regression model to assess the predictors of child mortality. Samuel and Amoo (2014) found that mother's wealth index had a significant impact on child mortality. Furthermore, they have shown that children born in households with low standard of living index experienced highest mortality compared to children who were born in households with high standard of living index.

Tlou et al. (2018) investigated the risk factors of under-five mortality in an HIV hyper-endemic area of rural South Africa from 2000-2014. A Cox proportional hazards model was used to identify the risk factors and key socio-demographic correlates of under-five mortality leveraging the longitudinal structure of the population cohort. Low wealth index was found to be significantly associated with infant and child mortality. Furthermore, children and infant from low wealth index had a significantly increased risk of mortality compared to those from very high wealth index.

Almansour (2018) studied effects of micronutrient deficiencies on mortality for children under age-five in Zimbabwe. The study found wealth index to be significantly associated with under-five mortality.

Kumar and Sahu (2019) have, on the other hand, conducted a study on the effects of socio-economic, demographic and environmental factors on under-five mortality in Empowered Action Group states of India. Cox-proportional hazard regression model was utilised in their study. Their study revealed that household wealth was significantly associated with under-five mortality and that the risk of dying was less among rich than the poor wealth quantile.

2.6 Environmental and health factors

Environmental and health factors such as source of drinking water and type of toilet facility have been found to have effect on under-five mortality and are discussed below.

2.6.1 Source of drinking water

Shiferaw et al. (2012) studied determinants of infant and child mortality in Ethiopia. Their study used logistic regression to determine the major demographic, environmental and socio-economic factors that influence infant and child mortality in Ethiopia. Water supply was found to be the most important significant factor to determine infant and child mortality. Furthermore,their study revealed that children who come from household with unprotected source of water were at higher risk of dying compared to those who were from households with protected source of water.

Sikder (2015) studied inter-district disparity of under-five mortality rate and its major determinants in Tamil Nadu, India. The study found that improved source of drinking water was statistically significant determinant of under-five mortality.

2.6.2 Type of toilet facility

(Achola, 2014) investigated the effects of mother's migration status on under-five mortality in Kenya. The study used Cox proportional hazard model to assess the effect of mother's migration status. It was found that the types of toilet facilities mothers used had a significant prediction on child survival. Furthermore, type of toilet facility was found to affect the migrants only.
Tessema (2015) conducted a study on under-five mortality and its predictors in Giibel Gibe Health and Demographic Surveillance System site in South West Ethiopia. Cox proportional hazards model was used to identify predictors of under-five mortality. Households with no toilet facilities were found to be significantly associated with under-five mortality. According to Tessema (2015), households should be encouraged and supported to contrast toilet facilities that includes pit latrines.

Worku (2011) studied the survival of children under-five years old using the 2003 South African Demographic and Health Survey data. The study utilized logistic regression analysis to identify the key predictors of mortality amongst children under the age of five years. The study revealed that ownership of flush toilet was significantly associated with under-five mortality.

2.7 Conclusion

The reviewed literature demonstrated various factors such as maternal education, maternal occupation, place of residence, wealth index, gender of a child, size of child at birth, maternal age, marital status of the mother, source of drinking water, type of toilet facility, place of delivery and duration of breastfeeding to be significantly associated with under-five mortality in different countries. However, some of the studies gave contradictory findings in their studies. The aim of this study is to use count data models to identify factors associated with under-five mortality in South Africa and contribute to the literature on child survival.

Chapter 3

Methodology

3.1 Introduction

This chapter will present a review of generalized linear models (GLM) for count data with special emphasis on Logistic regression, Poisson regression and Negative Binomial regression. The review will include the model framework, fitting and model checking.

3.2 Research design

The study is based on the data generated from the South African Demographic and Health Survey (SADHS) on children. The survey was conducted in 2016 and the information were based on five years' experience prior the survey. This study utilised logistic regression and count data models to identify the determinants of under-five mortality. Furthermore, the dependent or response variable in this study is under-five mortality and it is defined as the death of children

aged less than 5 years old.

3.3 Data source

This study utilised secondary data from the 2016 SADHS which contains data about the health and demographic characteristics of children and their parents from all provinces of South Africa. The SADHS collected information by interviewing women between the ages of 15 and 49 years old. It further provided a detailed information on mortality, child health and various factors or characteristics associated with under-five mortality. The data were recorded using Statistical Package for Social Sciences (SPSS).

3.4 Generalized linear models

Generalized linear models extend ordinary regression models to encompass non-normal response distributions as well as modelling functions of the mean (Agresti, 2014). The general linear model is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.1}$$

where X represents the covariates matrix, β representing the parameters and ε denoting the vector of the error terms. GLM extend the general linear models by relaxing the assumption that dependent variable is normally distributed with mean zero and a constant variance, which allows the distribution to be part of the exponential family of distributions and provides methods for the analysis of non-normal data. The GLM includes linear regression, Poisson regression, negative binomial regression, logistic regression, zero-inflated Poisson regression, analysis of variance, log-linear regression and zero-inflated negative binomial regression models. These aforementioned regression models share unique properties such as linearity and methods of parameter estima-

tion.

GLM has the following assumptions:

- The relationship between each exploratory variable and the outcome variable is approximately linear in structure
- It is assumed that the error terms are uncorrelated
- The data $Y_1, Y_2, ..., Y_n$ are independently distributed
- The residuals are independent with mean zero and constant variance

3.4.1 Components of generalized linear models

GLMs consist of three components namely, random component, link function and systematic component and they are outlined as follows:

Random component

The random component identifies the response variable Y with independent observations $(y_1, y_2, .., y_n)$ from a distribution in the natural exponential family. For instances,

- With binary outcome, the random component has a binomial distribution which lead to logistic regression model.
- With a count data, the random component has a Poisson distribution which lead to Poisson regression model.
- With count data where $Var(Y) > E(Y)$, the random component has a negative binomial distribution which lead to negative binomial regression model.

Systematic component

The systematic component is a linear function of explanatory variables which is used as linear predictor function. The covariates, $x_1, x_2, ..., x_k$ combines with the coefficients to form the linear predictor (η_i). The linear predictor η_i is given as

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2}, ... = \sum \boldsymbol{\beta_j X_{ij}} \tag{3.2}$$

Link function

The link function refers to a specific link between random and systematic component. It identifies a function that links the mean to linear function of the explanatory variables. The general link function is given as $g(\boldsymbol{\mu})$ and $g(\boldsymbol{\mu}) = \sum_{j=1}^{n} \boldsymbol{\beta_j X_{ij}}$, $i = 1, 2, ..., n$. Accordingly probit link function is a function associated with binary responses, similar to the link function.

Table 3.1: Link functions for Generalized Linear models

Distribution	**link**	$g(\boldsymbol{\mu_i})$
Normal	Identity	μ_i
Binomial	Logit	$\log\left(\frac{\pi_i}{1-\pi_i}\right)$
Poisson	Log	$\log \mu_i$
Gamma	Inverse	μ_i^{-1}
Inverse Gaussian	Inverse-square	μ_i^{-2}

3.4.2 Exponential family of distributions

Exponential family of distributions is a parametric set of probability distributions for discrete, continuous or a mix of both discrete and continuous random variables. Important distributions like the Normal, Bernoulli, Binomial, Gamma, Exponential, Poisson and Weibull distributions belong to this family

of distributions.

The natural form of the exponential family of distributions is defined McCullagh and Nelder (1989) and is given by

$$f_y(y,\theta,\phi) = exp\left[\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right] \tag{3.3}$$

where $a(\phi)$, $b(\theta)$ are known functions and $c(y,\phi)$ is some functions of y_i and θ. The parameter θ is called the canonical parameter whereas ϕ is the dispersion parameter. If Y_i has a distribution in the exponential family, the mean and variance of y can be derived from the well known relations and they indicated as follows

$$E\Big(\frac{\partial(l)}{\partial\theta}\Big) = 0 \tag{3.4}$$

and

$$E\Big(\frac{\partial^2 l}{\partial\theta^2}\Big) + \Big[E\Big(\frac{\partial(l)}{\partial\theta}\Big)\Big]^2 = 0 \tag{3.5}$$

$$l(\theta,y) = \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\theta)\right] \tag{3.6}$$

$$\frac{\partial(l)}{\partial\theta} = \frac{1}{a(\phi)}[y - b^{'}(\theta)] \tag{3.7}$$

$$\frac{\partial^2(l)}{\partial\theta^2} = \frac{-b^{''}(\theta)}{a(\phi)} \tag{3.8}$$

Where primes denote the second differentiation with respect to θ.

$$E\Big(\frac{\partial(l)}{\partial\theta}\Big) = E\Big(\frac{1}{a(\phi)}[y - b^{'}(\theta)]\Big) = \frac{\mu_i - b^{'}(\theta)}{a(\phi)} = 0 \tag{3.9}$$

Therefore,

$$E(Y_i) = \mu_i = b^{'}(\theta_i) \tag{3.10}$$

$$\frac{-b^{''}(\theta)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)} = 0 \tag{3.11}$$

Therefore,

$$Var(Y_i) = \sigma^2 = b^{"}(\theta_i)a(\phi) \tag{3.12}$$

Hence, the mean and variance are given by $E(Y_i) = \mu_i = b^{'}(\theta_i)$ and $Var(Y_i) = \sigma^2 = b^{"}(\theta_i)a(\phi)$ respectively.

Special cases of GLM such as Binomial, Poisson and negative binomial belongs to the exponential family of distributions.

Binomial distribution

Binomial distribution is a member of exponential family and its probability function for the binary outcome variable Y is given by:

$$f(y_i, \pi_i) = (^{k}_{y_i})(\pi_i)^{y_i}(1-\pi_i)^{k-y_i} \tag{3.13}$$

π denotes the parameter of interest or probability of success and k is the sample size. The probability function can be written in an exponential form as:

$$f(y_i, \pi_i) = exp\left[y_i log\pi_i - y_i log(1-\pi_i) + klog(1-\pi_i) + log\begin{pmatrix} k \\ y_i \end{pmatrix}\right] \tag{3.14}$$

$$= exp\left[y_i log(\pi_i - (1-\pi_i)) + klog(1-\pi_i) + log\begin{pmatrix} k \\ y_i \end{pmatrix}\right]$$

$$= exp\left[y_i log(\frac{\pi_i}{1-\pi_i}) + klog(1-\pi_i) + log\begin{pmatrix} k \\ y_i \end{pmatrix}\right] \tag{3.15}$$

where

$\theta_i = log(\frac{\pi_i}{1-\pi_i})$

$b(\theta_i) = -nlog(1-\pi_i)$
$c(y_i, \phi) = log\begin{pmatrix} k \\ y_i \end{pmatrix}$
$a(\phi) = 1$

Moreover, the mean and variance of a binomial distribution are as follows:

$$E(Y_i) = b^{'}(\theta_i) = k\pi_i$$

$$Var(Y_i) = b^{''}(\theta_i)a(\phi) = k\pi_i(1-\pi_i)$$

Poisson distribution

Poisson distribution is a member of exponential family and its probability function for the discrete random variable Y is given as

$$f(y_i, \lambda) = \frac{\lambda^{y_i} \exp^{-\lambda}}{y_i!},\ \lambda > 0 \tag{3.16}$$

Where y takes the values, $0, 1, 2, 3, ...$ The exponential form of a Poisson distribution is written as:

$$f(y_i, \lambda) = exp(y_i log\lambda - \lambda - logy_i!) \tag{3.17}$$

where
$\theta_i = log\lambda$
$b(\theta_i) = -\lambda$
$c(y_i, \phi) = -log(y_i)$
$a(\phi) = 1$

Furthermore, the mean and variance of a Poisson distribution are given as:

$$E(Y_i) = b^{'}(\theta_i) = \lambda$$
$$Var(Y_i) = b^{''}(\theta_i)a(\phi) = \lambda(1) = \lambda$$

Negative binomial distribution

Negative binomial distribution is also a member of exponential family and the canonical link of this distribution is log link. The density function of this distribution is given by:

$$f(y_i, \lambda_i, \gamma) = \begin{pmatrix} y_i - 1 \\ \gamma - 1 \end{pmatrix} (\lambda_i)^{\gamma} (1 - \lambda_i)^{y_i - \gamma},\ y_i = 0, 1, 2, 3.... \ and \ \gamma \geq 0 \tag{3.18}$$

λ_i and γ are parameter of interest. The distribution for negative binomial can be written in an exponential form as:

$$f(y, \lambda_i, \gamma) = exp\left[\gamma \log(\lambda_i) + y_i \log(1 - \lambda_i) - \gamma \log(1 - \lambda_i) + \log \begin{pmatrix} y_i - 1 \\ \gamma - 1 \end{pmatrix}\right] \tag{3.19}$$

$$= exp\left[y_i \log(1 - \lambda_i) + \gamma\big(\log(\lambda_i) - \log(1 - \lambda_i)\big) + \log \begin{pmatrix} y_i - 1 \\ \gamma - 1 \end{pmatrix}\right]$$

$$= exp\left[y_i \log(1 - \lambda_i) + \gamma \log\left(\tfrac{\lambda_i}{1-\lambda_i}\right) + \log \begin{pmatrix} y_i - 1 \\ \gamma - 1 \end{pmatrix}\right]$$

where

$\theta_i = \log(1 - \lambda_i)$

$b(\theta_i) = -\gamma \log\left(\frac{\lambda_i}{1-\lambda_i}\right)$

$c(y_i, \phi) = \log \begin{pmatrix} y_i - 1 \\ \gamma - 1 \end{pmatrix}$

$a(\phi) = 1$

It can be shown that the mean and variance for negative binomial are given as

$$E(Y_i) = b^{'}(\theta_i) = \tfrac{\gamma}{\lambda_i}$$

$$Var(Y_i) = b^{''} a(\phi) = \tfrac{\gamma(1-\lambda_i)}{\lambda_i^2}$$

3.4.3 Parameter estimation

The estimation includes estimating regression parameters in a Generalized linear model using maximum likelihood estimation method.
The maximum likelihood method is derived from the probability distribution of the dependent variable. Hence, suppose we are given the probability density function defined as follows:

$$f_{y_i}(y_i, \theta_i, \phi) = exp\left[\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi)\right] \tag{3.20}$$

Then, the likelihood function of equation 3.20 is given by:

$$L(y_i, \theta_i) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} exp\left[\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi)\right] \tag{3.21}$$

Therefore, by taking the logarithmic of equation 3.21,the log likelihood function which is given by:

$$log[L(y_i, \theta_i)] = log \prod_{i=1}^{n} exp\left[\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi)\right] \tag{3.22}$$

$$l(y_i, \theta_i) = \sum_{i=1}^{n} \log \exp\left[\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi)\right] = \sum_{i=1}^{n} a(\phi)^{-1}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi) \tag{3.23}$$

Then, the estimates are obtained by taking the derivatives of the log-likelihood function for the i^{th} observation, with respect to regression coefficient β_i and equating thee derivatives of the log-likelihood to zero. We obtain the score function given by $(U_{\beta 1}, U_{\beta 2}, U_{\beta 3}, ..., U_{\beta p})'$, where p is the number of parameters and $U_{\beta j}$ is given by

$$U_{\beta j} = \frac{\partial(l)}{\partial \beta_j} = \frac{\partial\left[\sum_{i=1}^{n} a(\phi)^{-1}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i,\phi)\right]}{\partial \beta_j} = 0 \tag{3.24}$$

Using the chain rule of differentiation $\frac{\partial(l_i)}{\partial \beta_j}$ is obtained as:

$$\frac{\partial(l_i)}{\partial \beta_j} = \frac{\partial(l_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{3.25}$$

since $\frac{\partial(l_i)}{\partial \theta_i} = \frac{\left[y_i - b^{'}(\theta_i)\right]}{a(\phi)}$ and $E(Y_i) = \mu_i = b^{'}(\theta_i)$ and $Var(Y_i) = a(\phi)b^{''}(\theta_i)$ then, $\frac{\partial(l_i)}{\partial \theta_i} = \frac{y_i - \mu_i}{a(\phi)}$ and $\frac{\partial \mu_i}{\partial \theta_i} = b^{''}(\theta_i) = \frac{var(y_i)}{a(\phi)}$. Since $\eta_i = \sum \beta_j X_{ij}$ then $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$.
Substituting for $\frac{\partial(l_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$ in equation 3.24 it gives

$$\frac{\partial(l_i)}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} * \frac{a(\phi)}{var(y_i)} * x_{ij} * \frac{\partial \mu_i}{\partial \eta_i} = \frac{y_i - \mu_i}{Var(Y_i)} x_{ij}\frac{\partial \mu_i}{\partial \eta_i} \tag{3.26}$$

The system of equation to be solved for β_j's is given by the following:

$$\frac{\partial(l)}{\partial \beta_j} = \sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{Var(Y_i)} x_{ij}\frac{\partial \mu_i}{\partial \eta_i}\right) \tag{3.27}$$

By equating equation 3.27 to zero the maximum likelihood of β, β can be obtained using Newton-Raphson method and Fisher scoring method. The Newton-Raphson is an iterative methods for solving non-linear equations and Fisher scoring is an alternative iterative method of solving likelihood equations (Agresti, 2014).

3.5 GLM special cases

The variable of interest or response variable in this study is under five mortality cases. Under five mortality is classified two ways. Firstly, it is treated in a form of a binary outcome 0 and 1, whereby the value 0 is children who did not died before the age of 5 years and value 1 is children who died before the age of 5 years. Secondly, it is treated as a count data ranging from level $0 - 5$ and it

is characterised as a non-normal distribution. It is not appropriate to use linear models based on normal distribution to describe the relationship between the response and explanatory variables. As a result, linear regression model and other models are not suitable in this study. Hence, this study will utilised logistic regression model and count data models such as Poisson and negative binomial regression models.

3.5.1 Logistic regression

Logistic regression model is a statistical modelling for binary outcome variable in which the log odds of the outcomes are modelled as a linear combination of regression variables. If the response variable is categorical, it is inappropriate to use linear regression because the response values are not measured on a ratio scale and the error terms are not normally distributed (Czepiel, 2002). In addition, the linear regression model can generate predicted values of any real number ranging from negative to positive infinity, whereas a categorical variable can only take on a limited number of discrete values within a specified range.

The binary response variable is denoted by Y and it has two possible outcomes which are $1(\text{"}success\text{"})$ and $0(\text{"}failure\text{"})$ (Hosmer and Lemeshow, 2000). The probabilities of distribution Y are given as $P(y = 1) = \pi$ which indicates a probability of success and $P(y = 0) = 1 - \pi$ which indicates a probability of failure whereas π is restricted to the range of 0 and 1. The logit transformation is used to link the response variable to the set of explanatory variables denoted by $x(x_1, x_2, x_3, ...x_n)$. Then, the logit link has the form:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = \sum_{j=1}^{k} \boldsymbol{\beta_j X_j} \qquad (3.28)$$

Accordingly, from equation 3.29, it can be deduced that:

$$\ln\Big(\frac{\pi(x_i)}{1-\pi(x_i)}\Big) = \Big(\sum_{1}^{k} \boldsymbol{\beta_j X_j}\Big) \tag{3.29}$$

$$\Big(\frac{\pi(x_i)}{1-\pi(x_i)}\Big) = \exp\Big(\sum_{1}^{k} \boldsymbol{\beta_j X_j}\Big)$$

$$\pi(x_i) = \Big(\sum_{1}^{k} \boldsymbol{\beta_j X_j}\Big)\Big(1-\pi(x_i)\Big)$$

$$\pi(x_i) = \exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big) - \pi(x_i)\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big)$$

$$\pi(x_i)(1+\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j})\Big) = \exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big)$$

$$\pi(x_i) = \frac{\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big)}{\Big(1+\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big)\Big)} \tag{3.30}$$

The relationship between $\pi(x_i)$ and the explanatory variables is described by the logistic function given as:

$$\pi(x_i) = \frac{\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big)}{\Big(1+\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}\Big)\Big)} \tag{3.31}$$

Under logistic regression we have the following assumptions:

- The dependent variable must be binary or dichotomy.
- The independent variables are not normally distributed, nor linearly related, nor of equal variable within each group.
- Logistic regression does not assume a linear relationship between the dependent and independent variables.
- Logistic regression requires that there should be little or no multicollinearity among the independent variables.
- Larger sample size are required because maximum likelihood coefficients are large sample estimates.

Fitting logistic regression model

The goal of logistic regression is to estimate the $K+1$ parameters in:

$$\ln\Big(\frac{\pi(x_i)}{1-\pi(x_i)}\Big) = \sum_{j=1}^{k} \boldsymbol{\beta_j X_j} \tag{3.32}$$

Method of maximum likelihood is used to estimate the parameters and the maximum likelihood equation is derived from the probability distribution. Since each y_i represents a binomial count in the i^{th} population, the joint probability density function of y_i is:

$$L(y|\pi) = \prod_{i=1}^{N} f(y_i|\pi_i) = \prod_{i==1}^{N} \binom{n_i}{y_i} \Big(\frac{\pi_i}{1-\pi_i}\Big)^{y_i} \Big(1-\pi_i\Big)^{n_i} \tag{3.33}$$

Since $\pi(x_i) = \dfrac{\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_k X_k}\Big)}{\Big(1+\exp\Big(\sum_{j=1}^{k} \boldsymbol{\beta_k X_k}\Big)\Big)}$, then the joint density function can be written as:

$$L(y|\pi) = \prod_{i=1}^{N} \binom{n_i}{y_i} \left(\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_k X_k}\Big)\right)^{y_i} \left(1+\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_k X_k}\Big)\right)^{-n_i} \tag{3.34}$$

Taking logarithm of equation 3.33, we find that the log likelihood function is given by:

$$log[L(y|\pi)] = log\left[\prod_{i=1}^{N} \binom{n_i}{y_i} \left(\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_j X_j}\Big)\right)^{y_i} \left(1+\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_k X_k}\Big)\right)^{-n_i}\right] \tag{3.35}$$

and since the other terms in the summation does not depend on $\boldsymbol{\beta_k}$, those terms will be treated as constant as indicated below:

$$l(\pi) = \sum_{i=1}^{N} y_i \left(\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_k X_k}\Big)\right) - \sum_{i=1}^{N} n_i \log\left(1+\exp\Big(\sum_{k=1}^{K} \boldsymbol{\beta_k X_k}\Big)\right) \tag{3.36}$$

To find the critical points of the log likelihood function, set the first derivative with respect to each $\boldsymbol{\beta}$ equal to zero. In differentiating equation 3.37, we note that:

$$\frac{\partial}{\partial \boldsymbol{\beta_k}} \sum_{k=1}^{K} X_{ik}\beta_k = x_{ik} \tag{3.37}$$

$$\frac{\partial(l(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta_k}} = \sum_{i=1}^{N} y_i x_{ik} - n_i . \frac{1}{1+\exp(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j})} . \frac{\partial}{\partial \boldsymbol{\beta_k}} (1+\exp(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j})) \tag{3.38}$$

$$\frac{\partial(l(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta_k}} = \sum_{i=1}^{N} y_i x_{ik} - n_i . \frac{1}{1+\exp(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j})} . \exp(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j}) . x_{ik}$$

$$\frac{\partial(l(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta_k}} = \sum_{i=1}^{N} y_i x_{ik} - n_i \left(\frac{\exp(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j})}{1+\exp(\sum_{j=1}^{k} \boldsymbol{\beta_j X_j})} \right) . x_{ik}$$

$$\frac{\partial(l(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta_k}} = \sum_{i=1}^{N} y_i x_{ik} - n_i \pi_i x_{ik} \tag{3.39}$$

$$\frac{\partial(l(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta_k}} = \sum_{i=1}^{N} x_{ik} \big(y_i - n_i \pi_i \big) \tag{3.40}$$

Therefore, the maximum likelihood estimates for $\boldsymbol{\beta}$ can be found by setting each of the $K+1$ equations in equation 3.39 to zero and solving for each $\boldsymbol{\beta_k}$ and it is demonstrated as follows:

$$\sum_{i=1}^{N} x_{ik} \big(y_i - n_i \pi_i \big) = 0 \tag{3.41}$$

Furthermore, the equations are non-linear and they can be solved using iteration algorithm such as Newton-Raphson method and Fisher scoring method.

Confidence interval estimation

The confidence interval estimates for the slope coefficients and the intercept are based on their respective Wald tests.
Therefore, the confidence interval for the slope coefficients is given by:

$$\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_i) \tag{3.42}$$

and for the intercept, the confidence interval is given as:

$$\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_0) \tag{3.43}$$

Where $z_{1-\frac{\alpha}{2}}$ is the upper $100(1-\frac{\alpha}{2})\%$ point from the standard normal distribution and $\hat{SE}(\hat{\beta}_i)$ denotes a model-based estimator of the standard error of the respective parameter estimator.

Odds Ratio

The interpretation of fitted logistic regression coefficients usually involves the odds ratios. The odds of the outcome being present among individuals with $x = 1$ is defined as:

$$\frac{\pi(1)}{1-\pi(1)}$$

Similarly, the odds of the outcome being present among individuals with $x = 0$ is defined as:

$$\frac{\pi(0)}{1-\pi(0)}$$

Then,according to Hosmer and Lemeshow (2000), the odds ratio(OR) is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$ and it is given by:

$$OR = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} \tag{3.44}$$

Hence, the odds ratio is a measure of association and also approximates how much more likely or unlikely it is for the outcome to be present those with $x = 1$ than among those with $x = 0$.

Considering the relationship between an outcome variable and one explanatory variable X,

$$logit(\pi) = \beta_0 + \beta_1 X_1 \tag{3.45}$$

where π is the probability of the occurrence of an event. The logit function can be defined as:

$$logit(\pi) = log\frac{\pi}{1-\pi} = log(odds)$$

$$log(odds) = \beta_0 + \beta_1 X_1 = \beta_0 + \beta_1$$

$$\pi = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$$

$\pi(1)$ is given as $\pi(1) = \dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$ and $\pi(0)$ is given as $\pi(0) = \dfrac{\exp \beta_0}{1 + \exp \beta_0}$ respectively. Therefore, the odds ratio can be written as:

$$OR = \frac{\dfrac{\exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)} \Big/ \dfrac{1}{1 + \exp(\beta_0 + \beta_1)}}{\dfrac{\exp(\beta_0)}{1 + \exp(\beta_0)} \Big/ \dfrac{1}{1 + \exp(\beta_0)}}$$

$$OR = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)}$$

$$OR = \exp(\beta_1) \tag{3.46}$$

Hence, for logistic regression with a dichotomous independence variable coded 0 and 1, the relationship between the odds ratio and the regression coefficient is given by equation 3.44. Furthermore, the parameter β_1 associated with X represents the change in the log odds from $X = 0$ to $X = 1$. Therefore, when

$OR = 1$, it implies that the odds for $x = 1$ and $x = 0$ are equal if $OR > 1$, it implies that the odds for $x = 1$ is greater than the odds for $x = 0$ and if $OR < 1$, implies that the odds for $x = 1$ is less than the odds for $x = 0$.

Confidence Interval for the Odds Ratio

In general, the $100(1 - \frac{\alpha}{2})\%$ confidence interval for the intercept is given by:

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_j) \tag{3.47}$$

The confidence intervals are transformed by exponentiation in order to get the corresponding $100(1 - \frac{\alpha}{2})\%$ confidence interval for odds ratio which are given as:

$$\exp(\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \hat{SE}(\hat{\beta}_j)) \tag{3.48}$$

Hosmer-Lemeshow test

Goodness of fit test is an approach used to assess the quality of the fitted model. Hosmer-Lemeshow test is a statistical test for goodness of fit for the Logistic regression model. The Hosmer-Lemeshow goodness of fit statistic X^2_{HL} is obtained by calculating the Pearson Chi-square statistic from the $g \times 2$ table of observed and estimated expected frequencies (Hosmer and Lemeshow, 2000). The test statistic of Hosmer-Lemeshow is given by:

$$X^2_{HL} = \sum_{k=1}^{g} \frac{\left(o_k - n'\bar{\pi}_k\right)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \tag{3.49}$$

where, n'_k is the total number of subjects in the k^{th} group, o_k denotes the number of covariate patterns in the k^{th} decile and $\bar{\pi}_k$ is the average estimated probability. The X^2_{HL} is approximated by the Chi-square distribution with $g - 2$ degrees of freedom. The statistic X^2_{HL} is compared with the critical value of the

Chi-square distribution with $g-2$ degrees of freedom ($\chi^2_{g-2,\alpha}$) when checking goodness-of-fit of the model.

3.5.2 Poisson regression model

Poisson distribution is a discrete probability distribution that is expressed as the probability of a number of events which occurs in a fixed period of time, if these events occur with a known average rate and each count occur independently of the time since the last event. A unique feature of this distribution is that its mean is equal to its variance. This unique feature or property is known as equi-dispersion.

This model have the following assumptions:

- The observation must be independent.
- The probability of occurrence in a short interval is proportional to the short length of the interval.
- The probability of two or more occurrence in such a short interval is negligible.
- The probability can not be negative.

Poisson regression model is a technique used to describe count data as a function of set of predictor variables. This modelling technique aims at modelling a count variable, which counts the number of times a certain event had occurred during given time period. This technique provides a standard framework for the analysis of count data and assumes that the response variable has a Poisson distribution and it can also be modelled by linear combinations of unknown covariates with regression coefficient denoted by β. For a sample of n with independent Poisson random variable $y_1, y_2, y_3, ..., y_n$, a simple linear model with mean λ_i, which depend on explanatory variable x_i is given as:

$$\boldsymbol{\lambda_i} = \exp(\boldsymbol{x_i'\beta}) \tag{3.50}$$

Under this model, the log transformation is used to adjust the skewness and prevents the model from producing negative values of predictors values. Therefore, the log-linear model is given by:

$$log(\boldsymbol{\lambda_i}) = \boldsymbol{x_i'\beta} \tag{3.51}$$

Poisson regression is the simplest regression model for analysing count data. The model assumes that each observed count is drawn from a Poisson distribution with the conditional mean $\boldsymbol{\lambda_i}$, $(y_i\ Poisson(\boldsymbol{\lambda_i}), i = 0, 1, 2 \ldots, n)$ on a given vector for case i. Then, the Poisson model is given by:

$$p(Y_i = y_i) = \frac{\exp(\lambda_i) * \lambda_i^{y_i}}{y_i!} \tag{3.52}$$

As aforementioned, Poisson regression assumes that the expected value E(Y) = $\boldsymbol{\lambda_i}$ = $\exp(\boldsymbol{x_i'\beta})$ and variance Var(Y)=$\boldsymbol{\lambda_i}$ and this property is called equi-dispersion. Therefore, if the $Var(Y) > E(Y)$ the Poisson assumption is violated and other modelling techniques such as negative binomial regression model should be considered, and this property is called over-dispersion when the variance of a random variable Y is greater than the expected value of a random variable Y. Also, if variance of the random variable Y is less than its expected, that is, $Var(Y) < E(Y)$ then, the property is known as under-dispersion.

Parameter estimation

Under Poisson regression model, the estimation of regression parameters is done using MLE.

Then, the probability density function of Poisson distribution is given by,

$$f_y(y) = \frac{\exp(-\lambda_i) \times \lambda_i^{y_i}}{y_i!},\ \lambda_i > 0\ and\ y_i = 0, 1, 2, .., \tag{3.53}$$

since each y_i is a count in the i^{th} population. Therefore, the joint density function of y_i is given by:

$$f(y|\lambda_i) = \prod_{i=0}^{n} \left(\frac{\exp(-\lambda_i) \times \lambda_i^{y_i}}{y_i!} \right) \tag{3.54}$$

Since $L(\boldsymbol{y}|\boldsymbol{\lambda}) = f(\boldsymbol{y}|\boldsymbol{\lambda}), \boldsymbol{\lambda_i} = \exp(\boldsymbol{x_i'\beta})$, then the joint density function of y_i can be written as:

$$L(\boldsymbol{y}|\boldsymbol{\lambda}) = f(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i}^{n} \left(\frac{exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \right) = \prod_{i}^{n} \left(\frac{exp(-exp(\boldsymbol{x_i'\beta}))exp(\boldsymbol{x_i'\beta})^{y_i}}{y_i!} \right) \tag{3.55}$$

Hence, by taking the logarithm of equation 3.55, we find that the log likelihood function is given by:

$$log[L(\boldsymbol{y}|\boldsymbol{\beta})] = log\left[\prod_{i}^{N} \left(\frac{exp(-exp(\boldsymbol{x_i'\beta}))exp(\boldsymbol{x_i'\beta})^{y_i}}{y_i!} \right) \right] = log\left[\sum_{i}^{n} \left(\frac{exp(-exp(\boldsymbol{x_i'\beta}))exp(\boldsymbol{x_i'\beta})^{y_i}}{y_i!} \right) \right] \tag{3.56}$$

which is equal to:

$$l(\boldsymbol{\beta}) = \sum_{i}^{n} \left(-exp(\boldsymbol{x_i'\beta}) + y_i \boldsymbol{x_i'\beta} - log(y_i)! \right) \tag{3.57}$$

$$l(\boldsymbol{\beta}) = \sum_{i}^{n} y_i \boldsymbol{x_i'\beta} - nexp(\boldsymbol{x_i'\beta}) - \sum_{i}^{n} log(y_i) \tag{3.58}$$

Moreover, the critical points of the log likelihood function are found by applying the first derivative of log likelihood function with respect to β and equating it to zero as outlined below:

$$\frac{d}{d\boldsymbol{\beta}}[l(\boldsymbol{\beta})] = \frac{d}{d\boldsymbol{\beta}}\left(\sum_{i}^{n} y_i \boldsymbol{x_i'}\boldsymbol{\beta} - nexp(\boldsymbol{x_i'}\boldsymbol{\beta}) - \sum_{i}^{n} log(y_i)\right) = 0 \tag{3.59}$$

Thus, since the other terms in the summation does not depend on $\boldsymbol{\beta}$, they can therefore be treated as constants. Thus, differentiating equation 3.59 with respect to $\boldsymbol{\beta}$, we get:

$$\frac{dl(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \sum_{i}^{n} \Big(y_i - exp(\boldsymbol{x_i'}\boldsymbol{\beta})\Big) x_{ij} \tag{3.60}$$

Hence, the maximum likelihood estimates for β can be found by setting equation in 3.60 equal to zero and solving for each β:

$$\sum_{i}^{n} \Big(y_i - exp(\boldsymbol{x_i'}\boldsymbol{\beta})\Big) x_{ij} = 0 \tag{3.61}$$

Furthermore, by applying the second derivative with respect to β, we get:

$$\frac{d^2 l(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -\sum_{i}^{n} \big(exp(\boldsymbol{x_i'}\boldsymbol{\beta}) x_{ij} x_{ik}\big) \tag{3.62}$$

which is Hessian of the function and typical element.

Accordingly, equation 3.62 are non-linear in $\boldsymbol{\beta}$ so that they need to be solved using iterative algorithm. The iterative algorithm that are commonly used are Newton-Raphson or Fisher scoring.

The Newton-Raphson method utilizes a matrix, called an information matrix that provides standard error values of the parameter estimates. This matrix is based on the curvature of the log likelihood function of the diagonal elements for the inverse of the information matrix.

The information matrix is given as:

$$k_j = -\sum_{i}^{n} \left(y_i - exp(\boldsymbol{x_i'}\boldsymbol{\beta}) x_i' x_i \right) \tag{3.63}$$

As a result, there is no closed form solution to:

$$\frac{dl(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \sum_{i}^{n} \left(y_i - exp(\boldsymbol{x_i'}\boldsymbol{\beta}) \right) x_i \tag{3.64}$$

3.5.3 Negative binomial regression model

Negative binomial regression model is a generalization of a Poisson regression model which loosens the restrictive assumption which states that the variance of a Poisson regression is equal to its mean. When the assumption of a Poisson regression fails, the negative binomial regression model becomes an alternative method. In other words, if the mean is not equal to the variance, negative binomial regression model can be used to fit the dataset and address the over-dispersion problem. The standard Poisson regression model accounts for observed difference among the observations, while the negative binomial regression includes a random component that involves an unobserved variance among the observations. The Negative binomial regression model estimates the dispersion parameter and as a result, allowing an independent specification of the mean and the variance. Similar to Poisson regression model, negative binomial regression model examines the predictive relationships between set of predictors and count dependent variable.

The negative binomial regression model is a Poisson-gamma mixture model with a heterogeneity parameter θ and it is denoted as:

$$p(Y_i = y_i) = \frac{\Gamma(y_i + \theta^{-1})}{y_i! \Gamma(\theta^{-1})} \left(\frac{\theta \lambda_i}{1 + \theta \lambda_i} \right)^{y_i} \left(\frac{1}{1 + \theta \lambda_i} \right)^{\theta^{-1}} \tag{3.65}$$

with mean $E(Y_i)$ = λ_i = $\exp(\boldsymbol{x_i'\beta})$ and variance $Var(Y_i)$ = λ_i (1 + $\theta\lambda_i$) = $\exp(\boldsymbol{x_i'\beta})$ (1 + $\theta\exp(\boldsymbol{x_i'\beta})$), where θ denotes an over-dispersion parameter. Hence, if $\theta\rightarrow$0, the negative binomial reduces to the usual standard Poisson distribution with parameter λ_i. Thus, it should be noted that the larger the value of θ, the greater the over-dispersion. Furthermore, this modelling technique is suitable for analysing a count data whereby an unobserved heterogeneity is present.

Fitting Negative binomial regression model

The regression coefficients are estimated using the method of Maximum likelihood and the probability density function of negative binomial is given by:

$$f(Y_i = y_i) = \frac{\Gamma(y_i + \theta^{-1})}{y_i!\Gamma(\theta^{-1})}\left(\frac{\theta\lambda_i}{1+\theta\lambda_i}\right)^{y_i}\left(\frac{1}{1+\theta\lambda_i}\right)^{\theta^{-1}} \tag{3.66}$$

The likelihood function of n independent negative binomial observations is a product of probabilities and is given by:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{N}\frac{\Gamma(y_i + \theta^{-1})}{y_i!\Gamma(\theta^{-1})}\left(\frac{\theta\lambda_i}{1+\theta\lambda_i}\right)^{y_i}\left(\frac{1}{1+\theta\lambda_i}\right)^{\theta^{-1}} \tag{3.67}$$

By applying the logarithm in equation 3.67, the log likelihood function is given as:

$$\log[L(\boldsymbol{\beta}, \boldsymbol{\theta})] = \log\left[\prod_{i=1}^{N}\frac{\Gamma(y_i + \theta^{-1})}{y_i!\Gamma(\theta^{-1})}\left(\frac{\theta\lambda_i}{1+\theta\lambda_i}\right)^{y_i}\left(\frac{1}{1+\theta\lambda_i}\right)^{\theta^{-1}}\right] \tag{3.68}$$

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \textstyle\sum_{i=1}^{N}\left[\ln\dfrac{\Gamma(y_i + \theta^{-1})}{y_i!\Gamma(\theta^{-1})} + y_i\ln\left(\dfrac{\theta\lambda_i}{1+\theta\lambda_i}\right) - \theta\ln\left(\dfrac{1}{1+\theta\lambda_i}\right)\right]$$

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\ln\left(\frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1})}\right) - \ln(y_i!) + y_i\ln(\theta\lambda_i) - y_i\ln(1+\theta\lambda_i) + \theta\ln(1+\theta\lambda_i)\right] \tag{3.69}$$

Therefore, from the definition of the gamma function, it can be shown that:

$$\ln\left(\frac{\Gamma(y_i+\theta^{-1})}{\Gamma(\theta^{-1})}\right)=\sum_{j=0}^{y_i-1}\ln(j+\theta^{-1}) \tag{3.70}$$

Hence, by substituting equation 3.70 into equation 3.69, the log likelihood function is now defined as:

$$l(\boldsymbol{\beta},\boldsymbol{\theta})=\sum_{i=1}^{N}\left[\sum_{j=0}^{y_i-1}\ln(j+\theta^{-1})-\ln(y_i!)+y_i\ln(\theta\lambda_i)-y_i\ln(1+\theta\lambda_i)+\theta\ln(1+\theta\lambda_i)\right] \tag{3.71}$$

Furthermore, the critical points of the log-likelihood function is found by equating the first derivatives of $l(\boldsymbol{\beta},\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to zero as indicated below:

$$\frac{\partial(l)}{\partial\boldsymbol{\theta}}=\frac{\partial}{\partial\boldsymbol{\theta}}\left(\sum_{i=1}^{N}\left[\sum_{j=0}^{y_i-1}\ln(j+\theta^{-1})-\ln(y_i!)+y_i\ln(\theta\lambda_i)-y_i\ln(1+\theta\lambda_i)+\theta\ln(1+\theta\lambda_i)\right]\right) \tag{3.72}$$

$$\frac{\partial(l)}{\partial\boldsymbol{\theta}}=\sum_{i=1}^{n}\left[\theta^{-2}(\ln(1+\theta\lambda_i))-\sum_{j=0}^{y[i-1}\frac{1}{j+\theta^{-1}}+\frac{y_i-\lambda_i}{\theta(1+\theta\lambda_i)}\right]$$

$$\frac{\partial(l)}{\partial\boldsymbol{\beta_j}}=\sum_{i=1}^{n}\frac{x_{ij}(y_i-\lambda_i)}{1+\theta\lambda_i} \tag{3.73}$$

Hence, by equating the gradients to zero we get the following sets of likelihood equations:

$$\sum_{i=1}^{n}\frac{x_{ij}(y_i-\lambda_i)}{1+\theta\lambda_i}=0 \tag{3.74}$$

$$\sum_{i=1}^{n}\left[\theta^{-2}\left(\ln(1+\theta\lambda_i)\right)-\sum_{j=0}^{y_i-1}\frac{1}{j+\theta^{-1}}+\frac{y_i-\lambda_i}{\theta(1+\theta\lambda_i)}\right]=0 \tag{3.75}$$

Hence, both equations are non-linear, then they should be solved using iterative algorithm. The iterative algorithm that are commonly used are Newton-Raphson or Fisher scoring.

3.6 Model specification test

3.6.1 Over-dispersion test

The Poisson regression model assumes that the count data has the same variance value as its mean value. However, this data often shows dispersion, which can be classified as either under-dispersion or over-dispersion. Therefore, count data in a Poisson regression model is said to be over-dispersed if its variance is greater than its mean value.

As a result of Poisson being over-dispersed, Negative binomial regression model becomes an alternative model of Poisson regression model. This model reduces to Poisson regression model when the over-dispersion parameter is not significantly different from zero. Therefore, to assess the adequacy of the negative binomial regression model over the Poisson regression model, the following hypothesis is used to test the significance of over-dispersion parameter θ:

$$H_0\text{: } \theta\text{=0 vs } H_1\text{: } \theta > 0$$

The presence of over-dispersion in the negative binomial regression model is justified when the null hypothesis H_0: θ = 0 is rejected. In order to test the hypothesis, a score statistic test, which is given by:

$$S_\theta = \frac{\left[\sum_{i=1}^{n}(y_i - \hat{\lambda_i})^2 - y_i\right]^2}{2\sum_{i=1}^{n}\hat{\lambda_i}^2} \tag{3.76}$$

is used, whereby $\hat{\lambda_i}$ is the predicted value from the Poisson regression model. Under the null hypothesis, the data follows a Poisson regression whereby lim-

iting distribution of the score test statistic is given by a chi-square with one degrees of freedom (Woldeamanuel, 2018).

3.7 Model diagnostic

3.7.1 Wald test

The Wald test is a function of the difference in the MLE and the hypothesized values normalised by an estimate of the standard deviance of the MLE. It is used to assess the significance of the Poisson and negative binomial regression coefficients.

The Wald test statistic W_i is given as:

$$W_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{3.77}$$

Where $\hat{\beta}_i$ represents the estimated coefficient of β and $SE(\hat{\beta}_i)$ is its standard error. The following hypothesis is tested:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

The value of W_i is squared, yielding a Wald statistic which approximately follows a Chi-square distribution with one degrees of freedom.

3.7.2 Likelihood ratio test

The maximum likelihood estimation method is used to assess the adequacy of any two or more nested models by utilising likelihood ratio test. In this study, a likelihood ratio test is used to compare Poisson and negative binomial regression models since Poisson is nested on negative binomial. The likelihood

ratio test is defined as:

$$T = 2[l_1 - l_0] \tag{3.78}$$

where l_1 and l_0 are the log likelihood of the models under the alternative and null hypothesis and T is a Chi-square distribution with one degrees of freedom which compares the maximum likelihood under the alternative with the null hypothesis. This method is not appropriate for models that are not nested. A small P-value of < 0.05 shows that the model has been improved significantly by the corresponding effect.

3.7.3 Deviance

Deviance residuals is a measure of discrepancy between the observed and fitted values. It is also used to test the goodness of fit of the model. To define the deviance we let $l(\lambda_i, \phi, y)$ be the log-likelihood of the fitted/ reduced model at the maximum likelihood estimate and also let $l(y_i, \phi, y)$ be the log-likelihood estimate of the saturated/full model. The deviance is twice the difference between the maximum achievable log-likelihood and the log-likelihood of the fitted model.

To derive the deviance for logistic regression model, let μ_i be the fitted values under model of interest and y_i be the estimates under the saturated model. The log-likelihood for the model of interest is given as:

$$l(y, y_i) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i}\right) + \log\binom{n_i}{y_i} \right] \tag{3.79}$$

the fitted/reduced model is given as:

$$l(y, \hat{\mu}_i) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\hat{\mu}_i}{n_i}\right) + (n_i - y_i) \log\left(\frac{n_i - \hat{\mu}_i}{n_i}\right) + \log\binom{n_i}{y_i} \right] \tag{3.80}$$

Since the deviance is twice the difference between the maximum achievable

log-likelihood and the log-likelihood of the fitted model, then the deviance for logistic regression model is given as:

$$D = 2\sum_{i=1}^{n}\left[y_i\left(\log\left(\frac{\hat{\mu}_i}{n_i}\right)-\log\left(\frac{y_i}{n_i}\right)\right)+(n_i-y_i)\left(\log\left(\frac{n_i-\hat{\mu}_i}{n_i}\right)-\log\left(\frac{n_i-y_i}{n_i}\right)\right)\right]$$

$$D = 2\sum_{i=1}^{N}\left[y_i\log\left(\frac{\hat{\mu}_i}{y_i}\right)+(n_i-y_i)\log\left(\frac{n_i-y_i}{n_i-\hat{\mu}_i}\right)\right] \tag{3.81}$$

To derive the deviance for Poisson regression model, let $Y_1, Y_2, ..., Y_n$ be samples for the model of interest then the log-likelihood is given as:

$$l(y, y_i) = \sum_{i=1}^{N} y_i \log(y_i) - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} log(y_i!) \tag{3.82}$$

and the fitted model is given as:

$$l(y, \hat{\lambda}_i) = \sum_{i=1}^{n} y_i \log(\hat{\lambda}_i) - \sum_{i=1}^{n} \hat{\lambda}_i - \sum_{i=1}^{n} log(y_i!) \tag{3.83}$$

The deviance for Poisson regression model is given as:

$$D = 2\left[\sum_{i=1}^{n}\Big(y_i\log(y_i) - y_i - log(y_i)!\Big) - \sum_{i=1}^{n}\Big(y_i\log(\hat{\lambda}_i) - \hat{\lambda}_i - log(y_i)!\Big)\right]$$

$$D = 2\left[\sum_{i=1}^{n}\Big(y_i\Big(\log(y_i) - \log(\hat{\lambda}_i)\Big) - \Big(y_i - \hat{\lambda}_i\Big)\Big)\right] \tag{3.84}$$

Since $\sum y_i^n = \sum_i^n \hat{\lambda}_i$, then the deviance can be written as:

$$D = 2\sum_{i=1}^{n} y_i\Big(\log(y_i) - \log(\hat{\lambda}_i)\Big) \tag{3.85}$$

$$D = 2\sum_i^n y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right)$$

To derive the deviance for negative binomial regression model let the log-likelihood of the model of interest be:

$$l(y, y_i) = \sum_{i=1}^{n} \frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1})} + \sum_{i=1}^{n} y_i \log(\theta y_i) - \sum_{i=1}^{n} \left(y_i + \frac{1}{\theta}\right) \log(1 + \theta y_i) \tag{3.86}$$

and the fitted model is given as:

$$l(y, \hat{\lambda}_i) = \sum_{i=1}^{n} \frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1})} + \sum_{i=1}^{n} y_i \log(\theta \hat{\lambda}_i) - \sum_{i=1}^{n} \left(y_i + \frac{1}{\theta}\right) \log(1 + \theta \hat{\lambda}_i) \tag{3.87}$$

The deviance for negative binomial regression model is given as:

$$D = 2\Bigg[\sum_{i=1}^{n} \left(\frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1})} + y_i \log(\theta y_i) - \left(y_i + \frac{1}{\theta}\right) \log(1 + \theta y_i) \right) - \sum_{i=1}^{n} \left(\frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1})} + y_i \log(\theta \hat{\lambda}_i) - \left(y_i + \frac{1}{\theta}\right) \log(1 + \theta \hat{\lambda}_i) \right) \Bigg]$$

$$D = 2 \sum_{i=1}^{n} \left[y_i \Big(\log(\theta y_i) - \log(\theta \hat{\lambda}_i) \Big) - \left(y_i + \frac{1}{\theta}\right) \left(\log(1 + \theta y_i) \right) - \log(1 + \theta \hat{\lambda}_i) \right) \Bigg]$$

$$D = 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\hat{\lambda}_i} \right) - (y_i + \frac{1}{\theta}) \log \left(\frac{1 + \theta y_i}{1 + \theta \hat{\lambda}_i} \right) \right] \tag{3.88}$$

Moreover, for large samples of the distributions, the deviance approximates a Chi-square distribution with $n - p$ degrees of freedom, where n is the number of observations and p is the number of parameters. For identifying a better model, one would expect smaller value of deviance (McCullagh and Nelder, 1989; Agresti, 2014).

3.7.4 Chi-square goodness of fit test

Chi-square goodness of fit test is a non-parametric test that is used to find out whether the observed value of a given category is significantly different from the expected value. The goodness of fit is designed to determine the adequacy or inadequacy of the fitted model. This test has the following hypothesis:

H_0: There is no significant difference between the observed and expect value.
H_1: There is significant difference between the observed and expected value.

The Chi-square test statistic is defined as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.89}$$

where O_i is the observed number of cases in i^{th} category, and E_i is the expected number of cases in i^{th} category. This is a Chi-square with $k - 1$ degrees of freedom whereby, k is defined as the number of parameters.

3.8 Model selection

3.8.1 Akaike information criterion

Akaike information criterion (AIC) is one of the most common methods of identifying the model which fit the data well by comparing two or more models. AIC is used to select the model that minimize the negative likelihood which is penalized by number of parameters. The AIC is defined by:

$$AIC = -2\ln(L) + 2k \tag{3.90}$$

where $-2\ln(L)$ denotes the likelihood of a model that is to be compared with the other models and k is the number of parameters in the model, including the intercept. The model with the smaller or least AIC value, is considered the better model as compared to others models.

3.8.2 Bayesian information criterion

Bayesian information criterion (BIC) is a criteria for model selection among a class of parametric models with different numbers of parameters. The BIC is

defined as:

$$BIC = -2l + 2k\log(n) \tag{3.91}$$

where l denotes the log likelihood of the model, n is the sample size and k is the number of parameters in the model, including the intercept. The model with the smaller or least BIC value, is considered the better model as compared to others models.

3.9 Statistical packages

The data was analysed using two statistical packages namely, Statistical Analysis Software (SAS) and R software.

3.10 Conclusion

This chapter has reviewed special cases of generalized linear models namely, logistic regression, Poisson regression and negative binomial regression, which will be utilized for analysis in this study. The chapter also demonstrated that the aforementioned special cases are members of exponential family of distributions.

Chapter 4

Data Analysis

4.1 Introduction

This chapter presents the analysis of data regarding the under-five mortality in South Africa. This chapter consists of three analytical parts, namely, descriptive analysis, logistic regression analysis as well as count data regression models. The first section presents a descriptive analysis of the demographic and socio-economic characteristics, the second section focuses on the analysis of under-five mortality using logistic regression model. The final section analyses the association between dependent and independent variables using count data regression models.

4.2 The Data

The data used in this study was obtained from the Demographic and Health Survey program and it is based on the South African Demographic and Health

Survey for the year 2016. The respondents in the survey were 3548 women who aged between 14 and 49 years, from nine provinces. The information obtained from the survey was based on five years' experience from the respondents prior to the year by which the survey was conducted.

The survey has provided detailed information on background characteristics of the respondents such as birth history of the women, marriage, breastfeeding and U5M. The respondents were asked about the total number of children they have given birth to, the total number of children residing with them, the number of children who have died, total number of births within the last five years prior the survey and whether the children were still alive or not.

The variable of interest in this study is under-five mortality. The U5M is classified into two ways. U5M is treated as a binary dependent variable, where the value "1" indicates that the child has lived beyond 5 years whereas value "0" indicates that the child died before age 5 years. Secondly, it is treated as a count data ranging from 0-5 (U5M experienced per mother).

4.2.1 Descriptive Analysis

The child and maternal characteristics of the children born to South African women who were surveyed are given in Table 4.1. Out of the 3548 children involved in the study, 3.98% of them died before the age of 5 years, while 96.02% of these children survived past the age of 5 years at the time of the survey. There were 51.60% males and 48.40% females in the survey. With regard to multiple birth, the table indicates that 97.3% of these children were born to mothers who gave birth once, whereas children born to mothers who had 1st multiple and 2nd multiple births all accounted the same percentage of 1.4%, with the least being those with 3rd multiple births. The table further revealed that majority of the children had average (58.8%) birth size, followed by those

who had large birth size 17.1%, then smaller birth size (9.3%), and then very large birth size (8.4%) respectively, and with the least being those with very smaller birth size consisting only 6.4%.

Furthermore, the table shows that majority of the children went for postnatal check within the first two months 82.9%, while 17.1% did not go for postnatal check. Regarding the child's health being checked prior discharge, 96.9% of the children's health were checked before being discharged, while 3.1% of the children's health were not checked. Likewise, the popular place of delivery for these children was public medical institution(s) which accounted for 89.1% of the total distribution, followed by 7.0% of those delivered at private medical sector or institutions, and the least being those delivered at home, accounting to only 3.8%.

Regarding maternal age groups, the table indicates that majority of these children were born by mothers of the age group 25-29 years which accounted for 28.2%, followed by age group 20-24 years, then 30-34 years and then 35-39 years which accounted 23.5%, 21.0% and 13.3% respectively. Children of the mothers within the 45-49 years age group accounted the least 1.2% in the study. Pertaining maternal educational level, majority of the children were born by mothers with secondary education accounting 78.9%, followed by those who were born by mothers with higher and primary education accounting 9.9% and 9.7% respectively, and the least of the children in the study were born by mothers with no educational level 1.5%. Additionally, it is shown that majority of the mothers had one birth 71.9%, followed by those who had two and three births within the last five years prior to the survey, accounting to 25.8% and 2.0% respectively. Mothers who had records of four births where the least with 0.2%.

Table 4.1: Descriptive statistics for child and maternal factors

Covariate	**Categories**	**N**	**%**
Child died	No	3413	96.02%
	Yes	135	3.98%
Child Gender	Male	1832	51.6%
	Female	1716	48.4%
Multiple Birth	Single birth	3451	97.3%
	1st Multiple	48	1.4%
	2nd Multiple	48	1.4%
	3rd Multiple	1	0.0%
Birth Size	Very Large	295	8.4%
	Large	602	17.1%
	Average	2065	58.8%
	Smaller	327	9.3%
	Very Smaller	225	6.4%
Baby postnatal checked	No	516	17.1%
within 1st two months	Yes	2506	82.9%
Child's health	No	89	3.1%
checked prior discharge	Yes	2774	96.9%
Place of delivery	Home	135	3.8%
	Public Medical Sector	3148	89.1%
	Private Medical Sector	249	7.0%
Maternal Age groups	15-19	219	6.2%
	20-24	833	23.5%
	25-29	999	28.2%
	30-34	745	21.0%
	35-39	471	13.3%
	40-44	237	6.7%
	45-49	44	1.2%
Maternal education	No education	53	1.5%
	Primary	344	9.7%
	Secondary	2800	78.69%
	Higher	351	9.9%
Maternal working	No	2496	70.3%
	Yes	1052	29.7%
Number of birth	One	2552	71.9%
within the last 5 years	Two	916	25.8%
	Three	72	2.0%
	Four	8	0.2%

Table 4.2. presents the socio-economic and environmental characteristics of the mothers who participated in the survey. The table indicates that majority of children were blacks accounting to 89.4%, followed by coloureds (8.5%), then whites (1.5%) and the least being those from Indian or Asian population groups (0.6%). The proportion of children differed by place of residence. Majority of children (52.5%) were born to mothers who resided in urban areas while 47.5% of them came from mothers who resided in rural areas. Majority of the children were from KwaZulu Natal province (15.6%) followed by children from Mpumalanga, Limpopo, Eastern Cape, North West and Gauteng provinces accounting 14.1%, 13.2%, 12.7%, 11.1% and 10.4% respectively. Children from the Free State and Northern Cape accounted 9.0% and 8.1% of the study, while children from the Western Cape (5.8%) province were the least in the study.

Table 4.2. further shows the distribution of wealth index. It is shown that children from poorer families accounted for 25.0%, followed by children from the poorest families by (24.0%) and middle families with (23.0%). The children from richer and richest families accounted for 17.8% and 10.2% respectively. With regard to source of drinking water, majority of the children in the study relied on piped water (82.7%), followed by children who relied on tube well water (6.8%) and surface from spring water (6.5%). The least of the children relied on tank water and they accounted 4.1% in the study. Likewise, the majority of the children were from households that utilised flush toilets (48.6%), followed by households that utilised pit latrine (44.6%) and children from households with no toilet facility (4.0%). Children from households that utilised bucket toilets accounted the least with 2.8% from those who were using pit latrine toilets.

Table 4.2: Descriptive statistics for socio-economic and environmental factors

Covariate	**Categories**	**N**	%
Population Groups	Black	3171	89.4%
	White	53	1.5%
	Coloured	300	8.5%
	Indian/Asian	22	0.6%
Type of place of residence	Urban	1863	52.5%
	Rural	1685	47.5%
Province	Western Cape	206	5.8%
	Eastern Cape	450	12.7%
	Northern Cape	286	8.1%
	Free State	318	9.0%
	KwaZulu-Natal	555	15.6%
	North West	395	11.1%
	Gauteng	370	10.4%
	Mpumalanga	501	14.1%
	Limpopo	467	13.2%
Wealth index	Poorest	852	24.0%
	Poorer	888	25.0%
	Middle	817	23.0%
	Richer	630	17.8%
	Richest	361	10.2%
Source of drinking water	Piped water	2803	82.7%
	Tube well water	218	6.8%
	Surface from spring	229	6.5%
	Tank water	138	4.1%
Type of toilet facility	Flush toilet	1649	48.6%
	Pit latrine	1514	44.6%
	No facility	137	4.0%
	Bucket toilet	95	2.8%

4.2.2 Test of association

Test of association for child alive (binary)

In this section we assess the association between the response variable (child died) and each explanatory variables. Table 4.3 provides test of association between child died and gender, multiple birth, birth size, place of delivery, baby postnatal check-up within two months, child's health checked prior discharge. It is shown that there is an association between child died and multiple birth given a P-value(< 0.001), birth size (P-value < 0.001), place of delivery (P-value = 0.014), breastfeeding (P-value < 0.001), baby postnatal (P-value < 0.001) and Child's health (P-value < 0.001) were found to have a significant association with child died as their P-values were less than 5% level of significance. However, it is shown that there is no significant association between gender and child died as the P-value of 0.114 is greater than 5% significance level.

Table 4.3: Association of child died by child factors

Child factors		**child died**					
	Categories	Yes	%	No	%	Total	P-value
Gender	Male	79	4.3%	1753	95.7%	1832	0.114
	Female	56	3.3%	1660	96.7%	1716	
Multiple Birth	Single birth	122	3.5%	3329	96.5%	3451	<0.0001
	1st Multiple	5	10.4%	43	89.6%	48	
	2nd Multiple	7	14.6%	41	85.4%	48	
	3rd Multiple	1	100%	0	0.0%	1	
Birth Size	Very large	7	2.4%	288	97.6%	295	<0.0001
	Large	24	4.0%	578	96.0%	602	
	Average	55	2.7%	2010	97.3%	2065	
	Smaller	11	3.4%	316	96.6%	327	
	Very small	29	12.9%	196	87.1%	225	
Place of deliver	Home	4	3.0%	131	97.0%	135	0.014
	Public Medical	128	4.1%	3020	95.9%	3148	
	Private Medical	2	0.8%	247	99.2%	249	
Baby Postnatal	No	51	9.9%	465	90.1%	516	<0.0001
	Yes	36	1.4%	2470	98.6%	2506	
Child's health	No	18	20.2%	71	79.8%	89	<0.0001
	Yes	62	2.2%	2712	97.8%	2774	

Table 4.4 provides test of association between child died and parental factors. A P-value of 0.066 indicates that there is no significant association between child died and maternal education since the P-value exceed the 5% level of significance. Also, there is no significant association between child died and partner's education, maternal working and maternal age group as their P-values were also exceeding 5% level of significance.

Table 4.4: Association of child died by parental factors

Child factors		**child died**					
	Categories	Yes	%	No	%	Total	P-value
Maternal education	No education	2	3.8%	51	96.2%	53	0.066
	Primary	21	6.1%	323	93.9%	344	
	Secondary	104	3.7%	2696	96.3%	2800	
	Higher	8	2.3%	343	97.7%	351	
Partner's education	No education	4	6.2%	61	93.8%	65	0.249
	Primary	6	3.6%	162	96.4%	168	
	Secondary	44	4.4%	960	95.6%	1004	
	Higher	3	1.7%	176	98.3%	179	
Maternal working	No	103	4.1%	2393	95.9%	2496	0.149
	Yes	32	3.0%	1020	97.0%	1052	
Maternal age group	15-19	8	3.7%	211	96.3%	219	0.777
	20-24	30	3.6%	803	96.4%	833	
	25-29	39	3.9%	960	96.1%	999	
	30-34	23	3.1%	722	96.9%	745	
	35-39	23	4.9%	448	95.1%	471	
	40-44	10	4.2%	227	95.8%	237	
	45-49	2	4.5%	42	95.5%	44	

Table 4.5 shows test of association between child died and socio-economic factors. It is shown that all nine provinces were represented in the study and there is a significant association between child died and province as the P-value of 0.042 is less than 5% level of significance. Population group and wealth index were also significantly associated with child died. However, there's no significant association between U5M and type of residence. Further, it is shown that there is no significant association between water source (P-value=0.122),

tetanus injection (P-value = 1.000) and child died since the P-values are greater than 5% level of significance. In addition, there is a significant association between toilet facility and child died with a P-value of <0.001 which is less than 5% level of significance.

Table 4.5: Association of child died and socio-economic, environmental and health factors

Variables		**Child died**					
	Categories	Yes	%(Yes)	No	%(No)	Total	P-value
Region(Province)	Western Cape	4	1.9%	202	98.1%	206	0.042
	Eastern Cape	18	4.0%	432	96.0%	450	
	Northern Cape	8	2.8%	278	97.2%	286	
	Free State	15	4.7%	303	95.3%	318	
	Kwazulu-Natal	17	3.1%	538	96.9%	555	
	North West	17	4.3%	378	95.7%	395	
	Gauteng	12	3.2%	358	96.8%	370	
	Mpumalanga	32	6.4%	469	93.6%	501	
	Limpopo	12	2.6%	455	97.4%	467	
Type of residence	Urban	66	3.5%	1797	96.5%	1863	0.429
	Rural	69	4.1%	1616	95.9%	1685	
Population group	Black/African	133	4.2%	3038	95.8%	3171	0.003
	White	0	0.0%	53	100.0%	53	
	Coloured	2	0.7%	298	99.3%	300	
	Indian/Asian	0	0.0%	22	100.0%	22	
Wealth index	Poorest	48	5.6%	804	94.4%	852	0.001
	Poorer	40	4.5%	848	95.5%	888	
	Middle	27	3.3%	790	96.7%	817	
	Richer	13	2.1%	617	97.9%	630	
	Richest	7	1.9%	354	98.1%	361	
Water Source	Piped water	96	3.4%	2707	96.6%	2803	0.122
	Tube well water	13	6.0%	205	94.0%	218	
	Spring water	12	5.2%	217	94.8%	229	
	Tank water	6	4.3%	1327	95.7%	138	
Toilet Facility	Flush toilet	42	2.5%	1607	97.5%	1649	<0.0001
	Pit latrine	71	4.7%	1443	95.3%	1514	
	No facility	4	2.9%	133	97.1%	137	
	Bucket toilet	10	10.5%	85	89.5%	95	
Tetanus Injection	No injection	23	3.1%	721	96.9%	744	1.000
	Received injection	57	3.0%	1817	97.0%	1874	

Table 4.3-4.5 shows that multiple birth, birth size, place of delivery, baby postnatal check within two months, child's health checked prior discharge, region/province, population group, wealth index and type of toilet facility are

factors related to under-five mortality(Child died) in South Africa.

Comparison of U5M experienced per mother across categories (count data)

Table 4.6 shows comparison of U5M experienced per mother with parental factors namely, maternal education, maternal age group, maternal working, maternal marital status, number of births in the last five years and population group. Kruskal Wallis test was utilised to determine which variables were significant. It is shown in Table 4.6 that the average numbers of U5M per mother are significantly different across the categories of maternal education. With regard to the number of births in the last five years, the average numbers of U5M per mother across the categories are significantly different. However, the table further shows that the average number of U5M per mother were not significantly different across the categories of maternal age group, maternal working, maternal marital status and population group.

Table 4.7 shows the comparison of U5M experienced per mother and socioeconomic, environment and health factors namely, province, residence, wealth index, water source, toilet facility, prenatal and antenatal visits. The comparison has shown that the average number of U5M per mother were significantly different across the categories of province and residence. It is also revealed that the average number of U5M per mother were significantly different across the categories of source of drinking water. With regard to wealth index, toilet facility, prenatal by doctor or gynaecologist, prenatal by nurse or midwife and antenatal visits, the comparison showed that the average number of U5M per mother were not significantly different across the categories.

Table 4.6: Comparison of U5M per mother across categories of parental factors

Parental factors	**Categories**	**N**	**mean**	**P-values**
Maternal Education	No education	40	0.3750000	0.0053
	Primary	252	0.0912698	
	Secondary	2232	0.1406810	
	Higher	283	0.1660777	
Maternal age groups	15-19	186	0.1021505	0.5294
	20-24	653	0.1684533	
	25-29	762	0.1246719	
	30-34	603	0.1442786	
	35-39	367	0.1389646	
	40-44	198	0.1565657	
	45-49	38	0.1578947	
Maternal working	No	1947	0.1397021	0.5492
	Yes	860	0.1476744	
Maternal marital status	Never in union	1555	0.1401929	0.9454
	Married	674	0.1468843	
	Living with partner	449	0.1380846	
	Widowed	32	0.0937500	
	Divorced	22	0.0909091	
	No longer living together/separated	75	0.2000000	
Number of births	One	2360	0.1381356	0.0421
	Two	423	0.1583924	
	Three	22	0.2727273	
	Four	2	0	
Population group	Black/African	2504	0.1449681	0.2914
	White	41	0.1463415	
	Coloured	243	0.1234568	
	Indian/Asian	19	0	

Table 4.7: Comparison of U5M per mother across categories of socio-economic factors

Variables	Categories	N	mean	p-value
Province	Western Cape	175	0.1771429	0.0035
	Eastern Cape	347	0.1066282	
	Northern Cape	230	0.0869565	
	Free State	256	0.0937500	
	Kwazulu-Natal	426	0.1009390	
	North West	319	0.2006270	
	Gauteng	286	0.1013986	
	Mpumalanga	391	0.2557545	
	Limpopo	377	0.1352785	
Residence	Urban	1498	0.1061415	<0.0001
	Rural	1309	0.1833461	
Wealth Index	Poorest	658	0.1534954	0.7112
	Poorer	690	0.1623188	
	Middle	646	0.1424149	
	Richer	513	0.1169591	
	Richest	300	0.1133333	
Water Source	Piped water	2334	0.1362468	0.0011
	Tube well water	176	0.1818182	
	Surface from spring	184	0.2119565	
	Tank water	113	0.0884956	
Toilet Facility	Flush toilet	1378	0.1161103	0.1864
	pit latrine	1244	0.1728296	
	No facility	107	0.1401869	
	Bucket toilet	78	0.1153846	
Prenatal:Doctor/Gynaecologist	No	2390	0.1443515	0.8729
	Yes	417	0.1294964	
Prenatal:Nurse/Midwife	No	414	0.1304348	0.3710
	Yes	2393	0.1441705	
Antenatal Visits	No	146	0.1027397	0.4384
	Yes	2661	0.1443067	

4.3 Application of Logistic regression model

This section presents the results of a logistic regression model. To understand the real variables associated with under-five mortality, only significant variables obtained from Chi-square test were inserted and tested in a logistic regression model. The modelling approach was based on stepwise variable selection procedure. This procedure starts with full main effect which consists of associative explanatory factors obtained from chi-square analysis. This selection procedure removes non-significant variables if their observed p-values are greater than 5% level of significant in each step. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to compare full main effect and stepwise to select the best model.

4.3.1 Full main-effect (enter method)

Goodness of fit statistics

Table 4.8 shows results of an overall model significance test for logistic regression model which was performed using likelihood ratio and Wald test. The analysis found that the P-values of these two test were significant at 5% level of significance as shown in Table 4.8, implying that the fitted coefficients were statistically significant.

Table 4.8: Model significance test for LR model1

Test	Chi-Square	df	p-value
Likelihood Ratio	134.7877	29	< 0.0001
Wald	123.7242	29	< 0.0001

Furthermore, the goodness of fit of the model was conducted using the deviance (likelihood ratio) statistics and Pearson Chi-square. The deviance statistic is given by the Chi-square of 515.0481 with 2676 degrees of freedom. The Pear-

son statistic is given by a Chi-square of 543.0512 with 2676 degrees of freedom, the scale parameter is given by 1.0065 which is close to one indicating that the model fits the data well.

Table 4.9: Assessing Goodness of fit for LR model1

Criterion	Value	df	Value/df
Residual Deviance	515.0481	2676	0.1925
Pearson Chi-square	2693.4898	2676	1.0065
AIC	575.048		
BIC	752.145		

Table 4.10 shows the partition for Hosmer-Lemeshow test and Table 4.11 present the Hosmer-Lemeshow goodness of fit test. The observed and expected frequencies are presented in Table 4.10 using 10 groups as recommended. The result of this test produced a Chi-square value of 9.1797 with 8 degrees of freedom and a P-value of 0.3274, which is greater than 5% level of significance. Therefore, since the P-value is greater than 5% level of significance, it implies that the model fit the data well and that the predicted or the expected frequencies agree with the observed frequencies.

Table 4.10: Partition for Hosmer-Lemeshow for LR model1

		Child alive = No		Child Alive = Yes	
Group	Total	Observed	Expected	Observed	Expected
1	271	0	0.28	271	270.72
2	272	1	1.17	271	270.83
3	271	3	1.68	268	269.32
4	269	2	2.10	267	266.90
5	270	5	2.58	265	267.42
6	271	5	3.09	266	267.91
7	278	2	3.94	276	274.06
8	271	2	6.09	269	264.91
9	271	14	11.62	257	259.38
10	262	36	37.46	226	224.54

Table 4.11: Hosmer-Lemeshow test for LR model1

Chisq	df	p-value
9.1797	8	0.3274

Interpretation of parameter estimation

Table 4.12 provides the type 3 analysis of effects and Table 4.13 presents an analysis of maximum likelihood parameters estimates obtained using the enter method. Table 4.12 shows that baby postnatal check-up (P-value = < 0.0001), child's health prior discharge (P-value = 0.0005533), Child birth size (P-value = 0.0001220) and toilet facility (P-value = 0.0409979) were found to be statistically significant because their P-values were less than 5% level of significance.

Table 4.12: Type3 Analysis of effects for LR model1

Parameter	Df	Chisq	p-value
Intercept	1	0.0001	0.9905735
Child Gender	1	0.2269	0.6338204
Baby Postnatal Check-up within the first 2 months	1	49.7969	<0.0001
Child Health checked prior discharge	1	11.9267	0.0005533
Child Birth Size	4	23.0812	0.0001220
Maternal Education	3	0.9198	0.8206355
Province	8	5.5592	0.6964684
Residence	1	3.5114	0.0609496
Population Group	3	3.7833	0.2858358
Wealth Index	4	1.6518	0.7994607
Toilet Facility	3	8.2565	0.0409979

Table 4.13 shows child, maternal, socio-economic and environment factors related to child death. The findings of this study revealed that baby postnatal check-up within 2 months (P-value of < 0.0001) had a significant impact on U5M as its p-value was less than 5% level of significance and the odds that a child who did not go for their postnatal check-up within the first two months dies are 7.312 times the odds for a child who went for their postnatal check-up. Child's health checked prior discharge (P-value of 0.0006) was also found to have significant impact on under-five mortality. The odds that a child whose health was not checked dies before the age of five are 3.768 times the odds for a child whose health was checked prior discharge.

The finding also revealed that child birth size had a significant impact U5M, and all the categories of child birth size, very large (P-value = 0.0012), larger(P-values = 0.0003), average(P-value of $<$0.0001) and smaller (P-value = 0.0136) were statistically significant as their P-values were less than 5% level of significance while controlling very small child birth size. The odds that a very large baby dies are 0.119 times the odds that a very small baby dies and the odds for a larger than average baby dies are 0.2047 times the odds that a very small baby dies. Similarly, the odds that an average baby dies are 0.241 times the odds that a very small baby dies, and the odds that a smaller than average baby dies are 0.288 times the odds that a very small baby dies. Furthermore, it was revealed that toilet facilities, in particular flush toilets, had a significant impact on child death since it had a P-value of 0.0139 which is less than 5% level of significance. The odds that a child whose household utilises bucket toilet dies are 4.739 times the odds that a child whose households utilises flush toilet dies. However, it was also shown that child gender, maternal education, province, residence, population group and wealth index were found to have insignificant impact on U5M since their P-values were greater than 5% level of significance.

Table 4.13: Maximum Likelihood Parameter estimation(Enter method)

Parameter	Estimate	Std. error	Wald Chi-square	P-value	OR
Intercept	-11.6876	267.0	0.0019	0.9651	
Child Gender					
Male	0.1258	0.2640	0.2269	0.6338	1.134
Female(Ref)	-	-	-		-
Baby Postnatal Check-up					
No	1.9895	0.2819	49.7970	<0.0001	7.312
Yes(Ref)	-	-	-	-	-
Child's health					
No	1.3266	0.3841	11.9267	0.0006	3.768
Yes(Ref)	-	-	-	-	-
Birth size					
Very Large	-2.1254	0.6562	10.4893	0.0012	0.119
Larger	-1.5770	0.4351	13.1379	0.0003	0.207
Average	-1.4247	0.3432	17.2365	<0.0001	0.241
Smaller	-1.2436	0.5039	6.0909	0.0136	0.288
Very Small(Ref)	-	-	-	-	-
Maternal Education					
No education	0.1595	0.9498	0.0282	0.8666	1.173
Primary	0.2403	0.6477	0.1377	0.7106	1.272
Secondary	-0.1445	0.5119	0.0797	0.7777	0.865
Higher(Ref)	-	-	-	-	-
Province					
Western Cape	0.2242	0.8199	0.0748	0.7845	1.251
Eastern Cape	0.4383	0.5408	0.6568	0.4177	1.550
Northern Cape	0.5688	0.6921	0.6755	0.4112	1.766
Free State	0.5065	0.6340	0.6381	0.4244	1.659
KwaZulu Natal	-0.4209	0.5360	0.6167	0.4323	0.656
North West	-0.1532	0.6102	0.0630	0.8018	0.858
Gauteng	0.3080	0.5037	0.3738	0.5410	1.361
Mpumalanga	0.1494	0.6052	0.0609	0.8051	1.161
Limpopo(Ref)	-	-	-	-	-
Residence					
Rural	-0.7186	0.3835	3.5114	0.0609	0.487
Urban(Ref)	-	-	-	-	-
Population Group					
Black African	9.4901	267.0	0.0013	0.9716	>999.999
White	-1.3030	405.4	0.0000	0.9974	0.272
Coloured	7.3101	267.0	0.0007	0.9782	>999.999
Indian/Asian(Ref)	-	-	-	-	-
Wealth Index					
Poorest	0.4663	0.7636	0.3730	0.5414	1.594
Poorer	0.7138	0.7202	0.9824	0.3216	2.042
Middle	0.3421	0.7053	0.2352	0.6277	1.408
Richer	0.3365	0.7075	0.2261	0.6344	1.400
Richest(Ref)	-	-	-	-	-
Toilet Facility					
Flush toilet	-1.5558	0.6327	6.0457	0.0139	0.211
Pit toilet	-0.7439	0.6068	1.5028	0.2202	0.475
No facility	-1.7606	0.9914	3.1541	0.0757	0.172
Bucket toilet(Ref)	-	-	-	-	-

4.3.2 Stepwise selection method

Goodness of fit statistics

The overall model significance of the model was tested using the likelihood ratio and Wald test. The results yielded P-values of <0.0001. Therefore, since the P-values were less than 5% level of significance, it implied that the fitted coefficients were significant.

Table 4.14: Model significance test for LR model2

Test	Chi-Square	df	p-value
Likelihood Ratio	115.2591	9	< 0.0001
Wald	118.4093	9	< 0.0001

The overall goodness of fit is given in Table 4.15. The deviance statistic was given by Chi-square of 534.58 with 2696 degrees of freedom. Pearson Chi-Square gave a scale parameter (ϕ) of 0.9839 which approximate 1 and indicates that the model is a better fit.

Table 4.15: Assessing Goodness of fit for LR model2

Criterion	Value	df	Value/df	p-value
Residual Deviance	534.58	2696	0.1983	1.000
Pearson Chi-square	2652.7095	2696	0.9839	1.000
AIC	554.577			
BIC	613.609			

Table 4.16 shows the Hosmer-Lemeshow goodness of fit test. The test gave a Chi square of 5.2888 and P-value of 0.5073, with 6 degrees of freedom. Since the P-value of the test is greater than 5% level of significance, it was an indication that the model fitted the data well. This result is consistent with results from the Table 4.15.

Table 4.16: Partition for Hosmer-Lemeshow for LR model2

		Child alive = No		Child Alive = Yes	
Group	**Total**	**Observed**	**Expected**	**Observed**	**Expected**
1	226	2	1.16	224	224.84
2	203	0	1.51	203	201.49
3	682	4	5.09	678	676.91
4	258	4	2.85	254	255.15
5	570	10	7.18	560	562.82
6	281	5	6.90	276	274.10
7	261	15	12.51	246	248.49
8	225	30	32.79	195	192.21

Table 4.17: Hosmer-lemeshow test for LR model2

Chisq	**df**	**p-value**
5.2888	6	0.5073

Figure 4.1 gives the Receiver Operating Characteristics (ROC) curve of the fitted model. The ROC was used to plot sensitivity versus 1-specificity. It provides a description of classification accuracy and it is used to display the accuracy of the model. The area under the curve ranges between 0 and 1, a value of less than 0.5 of the area under the curve indicates that the accuracy of the fitted model is poor whereas the value that approaches 1 indicates that the accuracy of the fitted model is better. The area under the curve for logistic regression model2 was given by 0.8020 and it indicates that 80.20% of the probabilities of child survival status were predicted correctly by the model. As a result, this curve confirmed that the model was a good fit.

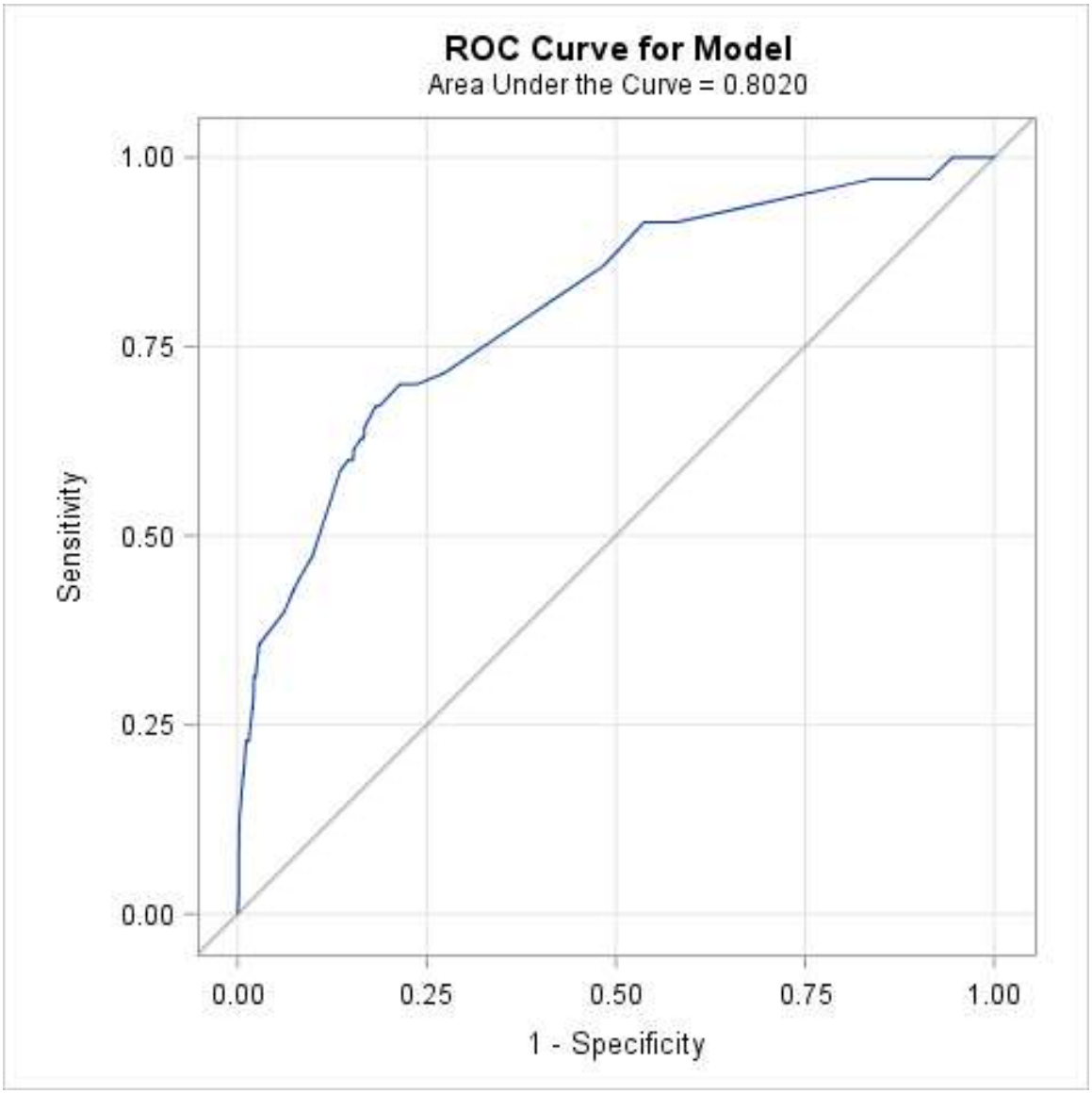

Figure 4.1: Receiver Operating Characteristics Curve

Interpretation of parameter estimation

The results shown in Table 4.18 and Table 4.19 were obtained using stepwise selection procedure. Only four explanatory variables, namely, baby postnatal check-up, child's health checked prior discharge, child birth size and toilet facility were selected and all selected variables were significant at 5% level of significance because their P-values did not exceed 5% level of significance.

Table 4.18: Type3 Analysis of effects for LR Model2

parameter	**Df**	**chisq**	**p-value**
Intercept	1	11.2972	0.0007763
Baby Postnatal check-up within first 2 months	1	50.8845	<0.0001
Child health checked prior discharge	1	11.4467	0.0007162
Child birth-Size	4	22.2053	0.0001824
Toilet facility	3	9.8659	0.0197410

Table 4.19 and Table 4.20 revealed that baby postnatal check-up in particular children who did not go for postnatal check-up within the first 2 months (P-value of < 0.0001) has significant impact on U5M as its p-value was less than 5% level of significance. The corresponding odds ratio was given 6.787 with 95% CI (4.010; 11.487). The odds that a child who did not go for their postnatal check-up within the first two months dies are 6.787 times the odds for a child who went for their postnatal check-up. Child's health checked prior discharge was also found to have significant impact on U5M, and children whose health was not checked prior discharge (P-value of 0.0007) are at risk of dying compared to children whose health was checked. The odds for children whose health was not checked was given by 3.482 with 95% CI (1.690; 7.175). The odds that a child whose health was not checked prior discharge dies are 3.482 times the odds for a child whose health was checked prior discharge.

Furthermore, child birth size which is very large has a significant impact on U5M with P-value of 0.0012 and odds ratio of 0.122 with 95% CI (0.034; 0.436). The odds that a very large baby dies before the age of five are 0.122 times the odds that a very small baby dies. With regard to a larger than average birth size, it was found that it had a significant impact with U5M as its P-value was 0.0008 which is less than 5% level of significance. The corresponding odds ratio was given by 0.246 with 95% CI (0.108; 0.558). The odds that a larger than average baby dies are 0.246 times the odds that a very small baby dies. Similarly, average birth size with P-value of < 0.0001 was also found to have a significant impact on U5M and the odds ratio was given by 0.247 with 95% CI (0.128; 0.476). The odds that a an average baby dies before they reach the age of five are 0.247 times the odds that a very small baby dies. In addition, smaller than average birth size was found to have a significant impact on U5M with P-value of 0.0159. The odds ratio was given by 0.308 with 95% CI (0.118; 0.803) and the odds that a smaller than average baby dies before they reach the age of five are 0.308 times the odds that a very small baby dies. Furthermore, type of toilet facility, particularly flush toilet, was found to have a significant impact on U5M with P-value of 0.0050, and the odds ratio of flush toilet was given by 0.211 with 95% CI (0.071; 0.626). The odds that a child whose household utilised flush toilet dies are 0.211 times the odds of a child whose household utilised bucket toilet. However, pit latrine and no toilet facility were found to have insignificant impact on under-five mortality.

Table 4.19: Maximum Likelihood Parameter estimation(Stepwise selection) Model2

Parameter	Categories	Estimation	Std. error	Wald Chi-square	P-value
Intercept		-1.9335	0.5753	11.2971	0.0008
Baby Postnatal	No	1.9150	0.2685	50.8842	<0.0001
check-up	Yes(Ref)	-	-	-	-
Child's Health	No	1.2477	0.3688	11.4467	0.0007
checked prior discharge	Yes(Ref)	-	-	-	-
Birth size	Very Large	-2.1077	0.6520	10.4492	0.0012
	Larger	-1.4039	0.4183	11.2644	0.0008
	Average	-1.3998	0.3350	17.4565	<0.0001
	Smaller	-1.1769	0.4882	5.8110	0.0159
	Very Small(Ref)	-	-	-	
Toilet Facility	Flush toilet	-1.5560	0.5546	7.8721	0.0050
	Pit toilet	-1.0278	0.5422	3.5932	0.0580
	No facility	-1.7727	0.9092	3.8018	0.0512
	Bucket toilet(Ref)	-	-	-	-

Table 4.20: Odds Ratio for LR Model2

Parameter	Effects	Odds Ratio	95% Wald Confidence Interval	
Baby Postnatal check-up	No vs Yes	6.787	4.010	11.487
Child Health checked prior discharge	No vs Yes	3.482	1.690	7.175
Birth-Size	Very large vs Very small	0.122	0.034	0.436
	Larger than average vs Very small	0.246	0.108	0.558
	Average vs Very small	0.247	0.128	0.476
	Smaller than average vs Very small	0.308	0.118	0.803
Toilet Facility	Flush toilet vs Bucket toilet	0.211	0.071	0.626
	Pit latrine vs Bucket toilet	0.358	0.124	1.036
	No facility vs Bucket toilet	0.170	0.029	1.009

4.3.3 Model comparison

Table 4.21 provides the model comparison of two fitted models. Based on AIC and BIC, the model with the smallest AIC and BIC is regarded as the best model. Accordingly, model2 (AIC = 554.5767 and BIC = 613.609) had the smallest AIC and BIC values as compared to model1 (AIC = 575.0481 and BIC = 613.609). Therefore, model2 is the better model.

Table 4.21: Model comparison for Logistic regression

	Model1 Enter method	**Model2** Stepwise selection
AIC	575.0481	554.5767
BIC	752.1449	613.609
Significant variables	Baby postnatal check-up Child's health checked prior discharge Child Birth-size Toilet Facility	Baby postnatal check-up Child's health checked prior discharge Child Birth-size Toilet Facility

4.4 Application of count data models

In this section, the analysis of count data regression models namely Poisson regression model and negative binomial regression model are presented. The modelling approach is based on stepwise variable selection procedure. The likelihood ratio test (LR), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to compare Poisson regression model and negative binomial regression model to determine the best model.

4.4.1 Poisson regression model

Table 4.22: Maximum Likelihood Parameter estimation for Poisson regression

Parameter	Categories	Estimation	Std. error	z-value
Intercept		-2.62797	0.34422	-7.635
Maternal Education	No education	0.51942	0.25850	2.009
	Primary	-0.71534	0.22242	-3.216
	Secondary	-0.28837	0.14321	-2.014
	Higher(Ref)	-	-	-
Province	Western Cape	0.79718	0.23705	3.363
	Eastern Cape	-0.04234	0.20076	-0.211
	Northern Cape	0.08811	0.24883	0.354
	Free State	0.17612	0.25067	0.703
	KwaZulu-Natal	-0.04382	0.18917	-0.232
	North West	0.69389	0.18385	3.774
	Gauteng	0.29883	0.23250	1.285
	Mpumalanga	0.94538	0.16111	5.868
	Limpopo(Ref)	-	-	-
Residence	Urban	-0.57135	0.11620	-4.917
	Rural(Ref)	-	-	-
Water Source	Piped water	0.82260	0.30792	2.671
	Tube water	1.16955	0.33942	3.446
	Spring water	1.35117	0.34075	3.965
	Tank water(Ref)	-	-	-

Goodness of fit

The overall goodness of fit is done using the deviance (likelihood ratio) statistics. Deviance statistic in Table 4.23 is given by Chi-square of 2084.5769 with 3372 degrees of freedom. The table also shows Pearson chi-square goodness of fit that is given by Chi-square of 4159.6215 with 3372 degrees of freedom and the scale parameter of 1.2336. Since the scale parameter is greater than 1, it is an indication that there is presence of over-dispersion and the model gave by Poisson regression is not a good fit.

Table 4.23: Goodness of fit statistics for Poisson

Criterion	**value**	**df**	**value/df**
Deviance	2084.5769	3372	0.6182
Pearson Chi-square	4159.6215	3372	1.2336
-2×Log Likelihood	2916.73		
AIC	2948.73		
BIC	3046.78		

Over-dispersion test

An over-dispersion parameter (θ) was fitted and tested to check whether it is significant or insignificant. Therefore, in Table 4.24, the score statistics is found to be 4932.644 with a p-value of less than 0.001. As a result, the significant p-value was observed, the null hypothesis was rejected and the over-dispersion parameter was found to be significant. This was supported by the scale parameter (Pearson chi-square) in Table 4.23. Therefore, negative binomial regression model was used as an alternative model to accommodate for the unobserved heterogeneity since the assumption of Poisson was violated.
Poisson regression model was fitted to identify the determinants/factors of under-five mortality in South Africa. The deviance likelihood statistic showed that the fitted model is not a good model and the model further tested for over-dispersion.

Table 4.24: Over-dispersion test

Score Statistic	Dispersion parameter	p-value
4932.644	1.195497	<0.001

4.4.2 Negative binomial regression model

Negative binomial regression model is an alternative to Poisson regression model and it is fitted to handle over-dispersion.

Goodness of fit statistics

The overall model significance test was conducted using the likelihood ratio test. The test compared the fitted model against the intercept only model. The test yielded a Chi-square of 103.69 with 15 degrees of freedom and P-value of <0.0001. Therefore, since the P-value is less than 5% level of significance, it implies that the fitted coefficients were significant.

Table 4.25: Model significance test for NB regression model

Test	Chi-Square	df	p-value
Likelihood Ratio	103.69	15	< 0.0001

Table 4.26 shows the results of goodness of fit statistics for negative binomial regression model. The deviance likelihood ratio statistic was used to check adequacy of the fitted model, and it is given by Chi-square of 1432.0136 with 3372 degrees of freedom. Pearson Chi-square test gave a value of 3161.2027 and a scale parameter(ϕ) of 0.9375 which is close to one indicating the negative binomial regression model is a better fit.

Table 4.26: Assessing Goodness of fit statistics for NB regression model

Criterion	value	df	value/df
Deviance	1432.0136	3372	0.4247
Pearson Chi-square	3161.2027	3372	0.9375
-2×Log Likelihood	2820.94		
AIC	2854.94		
BIC	2959.12		

Interpretation of parameter estimates

Table 4.27 provides type 3 analysis of effects for negative binomial regression model. It was found that maternal educational level with Chi-square of 44.145 and p-value of 0.002280 was statistically significant at 5% level of significance. Also, province (P-value: <0.0001), residence (P-value: < 0.0001) and water source (P-value: 0.008567) were found to have significant impact on U5M experienced per mother as their P-values were less than 5% level of significance.

Table 4.27: Type3 Analysis of effects for NB regression model

Parameter	Df	Chisq	P-value
Intercept	1	44.145	<0.0001
Maternal Education	3	14.517	0.002280
Province	8	46.928	<0.0001
Residence	1	18.150	<0.0001
Water Source	3	11.679	0.008567

Table 4.28 and Table 4.29 presents the analysis of parameter estimates and the rate ratio for negative binomial regression model. The findings of this study revealed mothers with primary education(P-value: 0.00263) have a significant impact on U5M and the corresponding rate ratio was given by 0.465088 with 95% CI (-1.2683; -0.2628). Therefore, the average number of U5M for a mother with primary educational level is 0.465088 times the average number of U5m for a mother with higher education. Other levels/categories of maternal education which are no education and secondary education did not have a signif-

icant impact on under-five mortality experienced per mother. Western Cape (P-value: 0.00627), North West (P-value: 0.00270) and Mpumalanga (P-value of <0.0001) were also found to have significant impact on U5M. The corresponding rate ratio for Western Cape was given by 2.1037729 with 95% CI (0.2063; 1.2812), for North West was given by 1.8981003 with 95% CI (0.2259; 1.0558) and for Mpumalanga was given by 2.3655326 with 95% CI (0.4896; 1.2324). The average number of U5M for a mother who is from Western Cape is 2.1037729, North West is 1.8981003 and Mpumalanga is 2.3655326 times the average number for a mother who is from Limpopo Province and other Provinces were found to be insignificant. Furthermore, it is revealed that residing in urban areas (P-value: <0.0001) have a significant impact on under-five mortality experienced per mother and its corresponding rate ratio was given by 0.5665859 with 95% CI (0.2974; 0.8389). Therefore, the average number of U5M for a mother who resides in urban areas is 0.5665859 times the average number for a mother who resides in rural areas. In addition, the findings also revealed that source of drinking water has an impact on U5M. Piped water (P-value: 0.01793), tube well water (P-value: 0.00616) and surfaced from spring water (P-value: 0.00123) has a significant impact on U5M experienced per mother as their p-values were less than 5% level of significance. The corresponding rate ratio for piped water was given by 2.2172300 with 95% CI (0.1282; 1.4644), for tube well water was given by 2.8399707 (0.2953; 1.7923) and for surface from spring water was 3.4393339 with 95% CI (0.4826; 1.9980). Therefore, the average number of U5M for a mother who utilized piped water, tube well water and surface from spring were 2.2172300, 2.8399707, 3.439339, times the average number for a mother who utilized tank water.

Table 4.28: Maximum Likelihood Parameter estimation for NB regression

Parameter	Categories	Estimation	Std. error	z-value	p-value
Intercept		-2.51897	0.37912	-6.644	<0.0001
Maternal Education	No education	0.40392	0.34900	1.157	0.24713
	Primary	-0.76553	0.25446	-3.008	0.00263
	Secondary	-0.32526	0.16754	-1.941	0.05221
	Higher(Ref)	-	-	-	-
Province	Western Cape	0.74373	0.27211	2.733	0.00627
	Eastern Cape	-0.01807	0.22604	-0.080	0.93630
	Northern Cape	0.08142	0.27327	0.298	0.76575
	Free State	0.13915	0.27695	0.502	0.61537
	KwaZulu-Natal	-0.05749	0.21298	-0.270	0.78723
	North West	0.64085	0.21366	2.999	0.00270
	Gauteng	0.26074	0.25877	1.008	0.31363
	Mpumalanga	0.86100	0.19037	4.523	<0.0001
	Limpopo(Ref)	-	-	-	-
Residence	Rural	0.56813	0.13336	4.260	<0.0001
	Urban(Ref)	-	-	-	-
Water Source	Piped water	0.79626	0.33639	2.367	0.01793
	Tube Water	1.04379	0.38104	2.739	0.00616
	Spring Water	1.23528	0.38229	3.231	0.00123
	Tank Water(Ref)	-	-	-	-

Table 4.29: Rate Ratio for NB regression Model

Parameter	Effects	Rate Ratio	95% Wald Confidence Interval	
Maternal education	No education vs Higher	1.4976803	-0.2742	1.0821
	Primary vs Higher	0.4650886	-1.2683	-0.2628
	Secondary vs Higher	0.7223393	-0.6590	0.0084
Province	Western Cape vs Limpopo	2.1037729	0.2063	1.2812
	Eastern Cape vs Limpopo	0.9820963	-0.4693	0.4332
	Northern Cape vs Limpopo	1.0848227	-0.4645	0.6273
	Free State vs Limpopo	1.1492908	-0.4106	0.6889
	KwaZulu-Natal vs Limpopo	0.9441353	-0.4776	0.3626
	North West vs Limpopo	1.8981003	0.2259	1.0558
	Gauteng vs Limpopo	1.2978958	-0.2528	0.7743
	Mpumalanga vs Limpopo	2.3655326	0.4896	1.2324
Residence	Rural vs Urban	0.5665859	0.2974	0.8389
Water Source	Piped vs Tank water	2.2172300	0.1282	1.4644
	Tube well vs Tank water	2.8399707	0.2953	1.7923
	Spring vs Tank water	3.4393339	0.4826	1.9880

4.4.3 Model comparison

Table 4.30 provides the model comparison of count data models namely, Poisson regression model and negative binomial regression model. Based on the AIC and BIC negative binomial model is the better model as it has the smallest AIC(2854.94) and BIC(2959.12) as compared to Poisson regression model. This is consistent with the value of scale parameter for negative binomial regression model as it is shown in Table 4.30 that it is close to one which implies that model is a better fit. It can also be observed that the dispersion parameter in negative binomial regression model is 0.4968 and it have been reduced from 1.195497. Furthermore, likelihood ratio test was used to compare Poisson regression model with negative binomial regression model since they are nested models. The test gave a chi square of 95.794 with 1 degrees of freedom and p-value of < 0.0001. Since the p-value is less than 5% level of significance this further validates that negative binomial regression model is a better fit to the data than Poisson regression model.

Table 4.30: Model comparison for count data models

	Model 1 Poisson regression	**Model 2** Negative binomial regression
Dispersion parameter	1.195497	0.4968
2×log Likelihood	-2916.73	-2820.94
Deviance statistic	2084.5769	1432.0
Pearson chi-square	4159.6215	3161.21
Scale parameter	1.2336	0.9375
AIC	2948.73	2854.94
BIC	3046.78	32959.12
Likelihood ratio test		
Chisq	df	p-value
95.794	1	<0.0001

4.5 Conclusion

This chapter has synthesised data regarding the descriptive statistics, test of association, comparison of means across categories, application of logistic regression model and application of count data regression models. These regression models were used to determine the factors that contribute to under-five mortality in South Africa.

Under application of logistic regression model, two models were fitted and model2(stepwise selection method) was selected as the better model compared to model1(enter method). It was found that baby postnatal check-up within the first 2 months, child's health checked prior discharge, child birth-size and type of toilet facility were statistically significant under model2.

Furthermore, under application of count data regression models, Poisson regression and negative binomial regression model were fitted. Poisson regression model tested for over-dispersion and negative binomial regression model was fitted as an alternative to Poisson regression model and to handle over-dispersion. Model comparison was done between the two models and negative binomial regression model was selected as the best model compared to Poisson regression model. It was found that maternal education, province, residence and water source were statistically significant under negative binomial regression model.

Chapter 5

Conclusion

5.1 Introduction

This chapter presents a summary of the this study, the conclusions drawn from the study and the recommendations for policy, programs and further research. In addition, the chapter outlines the limitations of this study.

5.2 Summary

The purpose of this study was to apply count data models to identify the determinants of under-five mortality in South Africa. The study was conducted with the general objective of examining the role of maternal, child, socio-economic, environmental and health factors as determinants of under-five mortality in South Africa using various statistical models. In particular, this study used the logistic regression model, the Poisson regression model, and the negative binomial regression model.

To achieve the objectives, a hypothesis was tested by fitting these models to the SADHS 2016 datasets. U5M(child survival status) and U5M experienced per mother were considered the dependent variables. Two sets of explanatory variables were used in the study. The first explanatory variables for child survival test of association was used and for U5M experienced per mother comparison of means across the categories was used. The Chi-square test of association showed that variables such as multiple birth, birth size, baby postnatal check-up, child's health checked prior discharge, province, population group, residence, wealth index, toilet type and maternal education were significantly associated with under-five mortality. However, this test does not take into account the impact of the other variables and does not give the direction of the association. Comparison of means showed that maternal education, number of births in the last five years, Province, residence and water source were significantly different across their categories. Once these factors contributing to U5M have been identified using both Chi-square analysis and comparison of means across categories along with Kruskal Wallis test. Logistic regression model and Count Data models, namely the Poisson regression model and negative binomial regression model were used accordingly to determine the relative importance of these factors.

When examining the effect of all explanatory variables found significant in Chi-square analysis, multivariate logistic regression analysis revealed that baby postnatal check-up within first 2 months, child's health checked prior discharge, child birth size and toilet facility were significantly associated with under-five mortality. The results indicates that the children who did not go for their postnatal check-up within the first two months were at a high risk of U5M as compared to children who went for their postnatal check-up. Similarly, the results indicated that children whose health was not checked prior discharge were at a high risk of U5M than those whose health was checked

prior discharge. Regarding child birth size, the results has shown that very large babies, large than average babies, average babies and smaller than average babies are at lowest risk of U5M than very small babies. The odds that a very large baby dies are 0.122 times the odds that a very small baby dies before the age of five. Similarly, the odds that a large than average baby dies are 0.246, average baby dies are 0.247 and smaller than average baby dies are 0.308 times the odds that a very small baby dies. Hence, it could be expected for very large babies, larger than average babies, average babies and smaller than average babies to have a low risk of U5M than those with very smaller babies. If children's health was checked prior discharge and by taking the children for their postnatal check-up within the first two months, it is likely to reduce under-five mortality.

Regarding count data modelling, Poisson regression model was the first starting point when analysing U5M experienced per mother dataset. However, the use of this model was affected by over-dispersion of the error term due to the disparity of the mean and variance within the data. For this reason, negative binomial regression model with more relaxed assumption on variance was utilised as an alternative model to solve the problem of over-dispersion. The adequacy of these count data modelling techniques was further investigated using the deviance, Pearson Chi square and LR test, which tested goodness of fit, and the comparison was made using AIC. LR test and AIC showed that negative binomial regression models were suitable for analysing the U5M experienced per mother data set against the Poisson regression model. In addition, the results of negative binomial regression modelling technique, revealed that maternal education, province, type of residence and source of drinking water were associated with U5M experienced per mother. Maternal education has long been recognised by other studies Mokoena (2011); Mugarura and Kaberuka (2015); Angela and Uju (2015); Oritogun and Bamgboye (2018); Samuel (2017) as one

of the most important factors of child mortality. The results has shown that maternal education in particular mothers with primary educational level have a significant impact on U5M experienced per mother and the average number of U5M per mother with primary educational level were 0.4650886 times the average number of U5M per mother with higher educational level. Several studies have indicated that education improves the ability of mothers to follow up simple health knowledge, seek medical attention more effectively and abide by treatment recommendations (Woldeamanuel, 2018; Oritogun and Bamgboye, 2018).

With regard to province, the results showed in particular that Western Cape, North West and Mpumalanga had a significant impact on U5M experienced per mother. Type of residence in particular rural area was also found to have a significant impact on U5M experienced per mother and the average number of U5M per mother for children who reside in rural areas was 0.5665859 times the average number of U5M per mother for children residing in urban areas. Indicating that children who reside in urban areas are less likely to die before the age of five by 43%. This is consistent with the previous studies such as those of (Worku, 2011; Negera et al., 2013; Adedini and Odimegwu, 2014) in that they have recognised residence and region/province as one of the proximate determinants that influence infant and under-five mortality through the immediate determinants.

Further, the results showed that all the categories of source of drinking water were statistically significant with U5M experienced per mother. Mothers to children whose households had piped water, tube well water and spring water as source of drinking water were more likely to experience U5M compared to mothers of children who utilized tank water as source of drinking water. Several studies (Shiferaw et al., 2012; Sikder, 2015; Nafiu et al., 2016; Acheampong and Avorgbedor, 2017) also found that source of drinking water have a significant impact on U5M.

5.3 Conclusion

In conclusion, the results of the logistic and negative binomial in this study revealed that baby postnatal check-up, child's health checked prior discharge, child birth size, toilet facility, residence, province, water source and maternal education were associated with increased risk of U5M in South Africa according to the SADHS 2016 dataset. In general, it is arguable that children who are more likely to die before the age of five years are those who did not attend their postnatal check within the first two months, those whose health was not checked prior discharge, children whose birth size was very small, children whose households utilized bucket toilets, whose mothers have no educational level, who resided in urban areas, who resided in Western Cape, North West and Mpumalanga province, who utilized piped, tube well and spring water as source of drinking water. In other words, the study suggests that generally there is reduction of under-five mortality in South Africa for children who attend postnatal check-up within the first two months and those health was checked prior discharge, those whose household utilized flush toilet.

5.4 Recommendation

With respect to these conclusions, the study makes the following recommendations:

Firstly, children's postnatal check within the first two months, child's health prior discharge, childbirth size, toilet facility, residence, region, water source and maternal education must be considered when planning and developing policies against U5M in order to successfully achieve the SDG3 on reducing child mortality by 2030. Accordingly, the Department of Health and other relevant bodies should develop necessary strategies, policies and intervention programs such as the health education program for uneducated mothers, whose children are at the highest risk because of poor health care utilisation.

Secondly, given that this study focused on the whole of South Africa, in order to further reduce under-five mortality and effectively address the related factors, this study recommends that future studies should examine factors associated with U5M in each province. It is important that efforts be made to focus on lower provincial levels rather than the level of under-five mortality only a national level. This is because, under apartheid, health care was extremely uneven and was conditioned by residential area, population group as well as provinces.

Finally, while count data models were better suited to the data of this study, the current researcher suggests that future researchers to look for other statistical methods, primarily survival analysis methods such as the Cox proportional hazard model and the Accelerated Failure Time Models. These methods will not only help to understand the risks and dangers, but also understand the average survival time of children under the age of five. In other words, it will not only identify factors, but it will also help in understanding the health behaviour characteristics associated with the survival of children under-five years of age.

5.5 Limitation of the study

The current study has some limitations. Firstly, the children in this study had not been tracked since birth. As a result, some measurement biases can be introduced. Secondly, the study did not include some significant factors associated with under-five mortality as identified in the literature, such as cause of death, maternal HIV status, and type of birth. Thirdly, no interactions between explanatory variables were investigated in the study. Finally, this study used the 2016 SADHS dataset obtained from the mother during the interview. This may be a more accurate result if this study is based on the death records of South African children.

References

ACHEAMPONG, G. K. AND AVORGBEDOR, Y. E. (2017). Determinants of under five mortality in Ghana: A logistic regression analysis using evidence from the demographic and health survey (1988-2014). *American Journal of Public Health Research*, **5** (3), 70–78.

ACHOLA, E. O. (2014). *The effects of mother's migration on under-five mortality in Kenya*. Masters dissertation, University of Nairobi.

ADEDINI, S. A. AND ODIMEGWU, C. O. (2014). Under-five mortality in Nigeria: Effects of neighbourhood contexts.

ADHIKARI, R. AND PODHISITA, C. (2010). Household headship and child death: Evidence from Nepal. *BMC international health and human rights*, **10** (1), 1–8.

AGRESTI, A. (2014). Wiley series in probability and statistics: Categorical data analysis , ebook isbn 9781118710852.

AHMED, Z., KAMAL, A., AND KAMAL, A. (2016). Statistical analysis of factors affecting child mortality in Pakistan. *J Coll Physicians Surg Pak*, **26** (6), 543–4.

AKWARA, P. A. (1994). *The impact of breast feeding practices on infant and child mortality in Amagoro division of Busia, Kenya*. Masters dissertation, University of Nairobi.

ALABI, M. (2018). Statistical study of under-five child mortality in Nigeria.

ALMANSOUR, A. M. (2018). Children mortality in Zimbabwe. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, **39** (1), 49–54.

AMOROSO, C. L., NISINGIZWE, M. P., ROULEAU, D., THOMSON, D. R., KAGABO, D. M., BUCYANA, T., DROBAC, P., AND NGABO, F. (2018). Next wave of interventions to reduce under-five mortality in Rwanda: A cross-sectional analysis of demographic and health survey data. *BMC pediatrics*, **18** (1), 27.

ANGELA, C. AND UJU, O. (2015). Determinants of under-five mortality in Nigeria: An application of Cox proportional hazard and Cox frailty models. *IOSR J Math Ver I [Internet]*, **11** (4), 2278–5728.

APUNDA, R. (2016). *Determinants of child mortality in Kenya*. Masters dissertation, University of Nairobi.

BHUTTA, Z. A., AHMED, T., BLACK, R. E., COUSENS, S., DEWEY, K., GIUGLIANI, E., HAIDER, B. A., KIRKWOOD, B., MORRIS, S. S., SACHDEV, H., ET AL. (2008). What works interventions for maternal and child undernutrition and survival. *The lancet*, **371** (9610), 417–440.

BUOR, D. (2003). Mothers' education and childhood mortality in Ghana. *Health policy*, **64** (3), 297–309.

BUWEMBO, P. (2010). *Factors associated with under-5 mortality in South Africa: Trends 1997-2002*. Masters dissertation, University of Pretoria.

CALDWELL, J. C. (1979). Education as a factor in mortality decline an examination of Nigerian data. *Population studies*, 395–413.

CHADOKA, N. (2011). *The effect of maternal health-seeking behaviour on under-five mortality in Zimbabwe*. Masters dissertation, University of Witwatersrand, Johannesburg.

CHOWDHURY, A. H. (2013). Determinants of under-five mortality in Bangladesh. *Open Journal of Statistics*, **3**, 213–219.

CHOWDHURY, Q. H., ISLAM, R., AND HOSSAIN, K. (2010). Socio-economic determinants of neonatal, post neonatal, infant and child mortality. *International Journal of Sociology and Anthropology*, **2** (6), 118–125.

CZEPIEL, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and Implementation. *Available at czep. net/stat/mlelr. pdf*, **83**.

DENDUP, T., ZHAO, Y., AND DEMA, D. (2018). Factors associated with under-five mortality in bhutan: An analysis of the Bhutan National Health Survey 2012. *BMC public health*, **18** (1), 1375.

ETTARH, R. AND KIMANI, J. (2012). Determinants of under-five mortality in rural and urban Kenya. *The International Electronic Journal of Rural and Remote Health Research, Education, Practice and Policy*.

FIKRU, C., GETNET, M., AND SHAWENO, T. (2019). Proximate determinants of under-five mortality in Ethiopia: Using 2016 nationwide survey data. *Pediatric Health, Medicine and Therapeutics*, **10**, 169.

GORO, I. (2007). The stalling child mortality in Ghana: The case of the three Northern Regions. *University of Cape Town, Cape Town, South Africa*.

HLONGWA, M. AND DE WET, N. (2019). Demographic and socioeconomic factors associated with under-5 mortality in KwaZulu-Natal, South Africa. *South African Journal of Child Health*, **13** (4), 174–179.

HOSMER, D. AND LEMESHOW, S. (2000). *Applied logistic regression, Second Editin*. Wiley New York.

IAEG, U. (2016). Final list of proposed sustainable development goal indicators. *Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators (E/CN. 3/2016/2/Rev. 1)*.

IQBAL, N., GKIOULEKA, A., MILNER, A., MONTAG, D., AND GALLO, V. (2018). Girls' hidden penalty: Analysis of gender inequality in child mortality with data from 195 countries. *BMJ global health*, **3** (5), e001028.

KANMIKI, E. W., BAWAH, A. A., AGORINYA, I., ACHANA, F. S., AWOONOR-WILLIAMS, J. K., ODURO, A. R., PHILLIPS, J. F., AND AKAZILI, J. (2014). Socio-economic and demographic determinants of under-five mortality in rural Northern Ghana. *BMC international health and human rights*, **14** (1), 24.

KAYODE, G. A., ADEKANMBI, V. T., AND UTHMAN, O. A. (2012). Risk factors and a predictive model for under-five mortality in Nigeria: Evidence from Nigeria Demographic and Health Survey. *BMC pregnancy and childbirth*, **12** (1), 10.

KUMAR, S. AND SAHU, D. (2019). Socio-economic, demographic and environmental factors effects on under-five mortality in empowered action group states of india: An evidence from nfhs-4. *Public Health Research*, **9** (2), 23–29.

LIU, L., OZA, S., HOGAN, D., CHU, Y., PERIN, J., ZHU, J., LAWN, J. E., COUSENS, S., MATHERS, C., AND BLACK, R. E. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals. *The Lancet*, **388** (10063), 3027–3035.

LIU, L., OZA, S., HOGAN, D., PERIN, J., RUDAN, I., LAWN, J. E., COUSENS, S., MATHERS, C., AND BLACK, R. E. (2015). Global, regional, and national

causes of child mortality in 2000–13, with projections to inform post-2015 priorities: An updated systematic analysis. *The Lancet*, **385** (9966), 430–440.

MAKGABA, M. E. W. (2014). *Survival analysis with applications to Ga-Dikgale children*. Masters dissertation, University of Limpopo.

MCCULLAGH, P. AND NELDER, J. (1989). Generalized linear models second edition. *In Boca Raton, London, New-York, Washington, DC, Chapman & Hall/CRC*.

MOKOENA, M. P. (2011). *Risk factors associated with high infant and child mortality in Lesotho*. Masters dissertation, University of Cape Town.

MOTALA, S., NGANDU, S., MTI, S., ARENDS, F., WINNAAR, L., KHALEMA, E., MAKIWANE, M., NDINDA, C., MOOLMAN, B., MALULEKE, T., ET AL. (2015). Millennium development goals: Country report 2015.

MUGARURA, A. AND KABERUKA, W. (2015). Multilevel analysis of factors associated with child mortality in Uganda. *African Journal of Economic Review*, **3** (2), 125–139.

MUSTAFA, E. AND ODIMEGWU, C. (2008). Socioeconomic determinants of infant mortality in Kenya: Analysis of Kenya dhs 2003.

NAFIU, L. A., OKELLO, M., AND ADIUKWU, R. N. (2016). Determinants of under-five mortality in Abim District, Uganda. *birth*, **20**, 20–29.

NEGERA, A., ABELTI, G., BOGALE, T., GEBRESELASSIE, T., AND PEARSON, R. (2013). An analysis of the trends, differentials and key proximate determinants of infant and under-five mortality in Ethiopia. *ICF International: Calverton, Maryland USA*.

ORITOGUN, K. S. AND BAMGBOYE, E. A. (2018). Application of count models on infant and child mortality in Nigeria: A comparative study. *International Journal of Tropical Disease and Health*, **30** (3), 1–12.

PETER, M., LINCOLN, D., EDITH, M., AND PETER, K. (2017). Determinants of under-five mortality: Evidence from Zambia. *Journal of Economics and Sustainable Development*, **8** (14), 100–107.

SACHS, J. D. (2015). Achieving the sustainable development goals. *Journal of International Business Ethics*, **8** (2), 53–62.

SAMUEL, G. W. (2017). *Proximate determinants: The pathways of influence of underlying factors on under-five mortality in Nigeria*. Ph.D. thesis, Covenant University.

SAMUEL, G. W. AND AMOO, E. O. (2014). A statistical analysis of child mortality: Evidence from Nigeria. *Journal of Demography and Social Statistics*, **1** (1), 110–120.

SAROJ, R. K., MURTHY, K. H. N., KUMAR, M., SINGH, R., KUMAR, A., ET AL. (2019). Survival parametric models to estimate the factors of under-five child mortality. *Journal of Health Research and Reviews*, **6** (2), 82–88.

SHIFA, G. T., AHMED, A. A., AND YALEW, A. W. (2018). Socioeconomic and environmental determinants of under-five mortality in Gamo Gofa Zone, Southern Ethiopia: A matched case control study. *BMC international health and human rights*, **18** (1), 14.

SHIFERAW, Y., ZINABU, M., AND ABERA, T. (2012). Determinant of infant and child mortality in Ethiopia. *Available at SSRN 2188355*.

SIKDER, U. K. (2015). Inter-district disparity of under-five mortality rate and its major determinants in Tamil Nadu, India. *American International Journal of Research in Humanities, Arts and Social Sciences*, **11** (3), 263–270.

SINGH, R. AND TRIPATHI, V. (2013). Maternal factors contributing to under-five mortality at birth order 1 to 5 in India: A comprehensive multivariate study. *Springerplus*, **2** (1), 284.

STATSSA (2015). Millennium development goals Country report. *Pretoria, South Africa: Statistics SA*.

STATSSA (2019). Statistics South Africa mid year population estimates. *Pretoria, South Africa: Statistics SA*.

TAGOE, E. T., AGBADI, P., NAKUA, E. K., DUODU, P. A., NUTOR, J. J., AND AHETO, J. M. K. (2020). A predictive model and socioeconomic and demographic determinants of under-five mortality in Sierra Leone. *Heliyon*, **6** (3), e03508.

TESSEMA, F. (2015). Under five mortality and its predictors in Gilgel Gibe Health and Demographic surveillance system site, South West Ethiopia.

TLOU, B., SARTORIUS, B., AND TANSER, F. (2018). Investigating risk factors for under-five mortality in an HIV hyper-endemic area of rural South Africa, from 2000-2014. *PloS one*, **13** (11), e0207294.

UNICEF. (2008). *The state of the world's children 2009: Maternal and newborn health*, volume 9. Unicef.

UNICEF (2015). Millennium development goals: Reduce child mortality. *Retrieved on February 27th*.

WOLDEAMANUEL, B. T. (2018). Statistical analysis of neonatal mortality: A case study of Ethiopia. *Journal of Pregnancy and Child Health*, **5** (2), 1–11.

WORKU, Z. (2011). A survival analysis of South African children under the age of five years. *Health SA Gesondheid*, **16** (1).

YAYA, S., BISHWAJIT, G., OKONOFUA, F., AND UTHMAN, O. A. (2018). Under five mortality patterns and associated maternal risk factors in sub-Saharan Africa: A multi-country analysis. *PloS one*, **13** (10), e0205977.

Appendix

R CODES

Re-order of observation

MultipleBirth< − factor(MultipleBirth, levels = c(”3”, ”2”, ”1”, ”0”))
BirthSize< − factor(BirthSize, levels = c(”5”, ”4”, ”3”, ”2”, ”1”)) BabyPostnatal< − factor(BabyPostnatal, levels = c(”1”, ”0”))
ChildHealth< − factor(ChildHealth, levels = c(”1”, ”0”))

MaternalEducation< − factor(MaternalEducation, levels = c(”3”, ”2”, ”1”, ”0”))
NoofBirths< − factor(NoofBirths, levels = c(”4”, ”3”, ”2”, ”1”))

Residence< − factor(Residence, levels = c(”2”, ”1”))
Province< − factor(Province, levels = c(”9”,”8”,”7”,”6”,”5”,”4”,”3”,”2”,”1”))
PopulationGroup< − factor(PopulationGroup, levels = c(”4”,”3”,”2”,”1”))
WealthIndex< − factor(WealthIndex, levels = c(”5”,”4”,”3”,”2”,”1”))
WaterSource< − factor(WaterSource, levels = c(”4”, ”3”, ”2”, ”1”))
ToiletFacility< − factor(ToiletFacility, levels = c(”4”,”3”,”2”,”1”))

R code for logistic regression model

APPLICATION OF LOGISTIC REGRESSION MODEL

FitLRn $<-$ glm(U5M$\sim$1, family = binomial(link = ”logit”))
summary(FitLRn)

FitLR1$<-$glm$\big($U5M $\sim$ Child Gender + BabyPostnatal + ChildHealth +
BirthSize+ MaternalEducation+ Province+ Residence+ PopulationGroup + WealthIndex + ToiletFacility, family = binomial(link = ”logit”)$\big)$
summary(FitLR1)

Exponentiated coefficients
exp$\big($coef(FitLR1)$\big)$

Deviance goodness of fit
dev $<-$ deviance(FitLR1)
df $<-$ df.residual(FitLR1)
P$-$value $<-$ 1-pchisq(dev,df)

Pearson chi-square
prLR1 $<-$ sum(residuals(FitLR1, type = ”pearson”)$\hat{2}$) # get pearson chi2
pchisq(prLR1, FitLR1$df.residual, lower=F) # calc p-value
pchisq(prLR1, FitLR1$deviance, FitLR1$df.residual, lower= F) #calc p-vl

# Likelihood ratio test
lrtest(FitLR1)

# Type3 Analysis of Effects
Anova(FitLR1,test='Wald',type='III')

```
### Variable selection methods
## Stepwise selection method
FitLR2=step(FitLR1)
FitLR2=stepAIC(FitLR1, trace = FALSE)
summary(FitLR2)

## Model comparison
AIC(FitLR1,FitLR2)
BIC(FitLR1,FitLR2)
```

R code for count data regression models

APPLICATION OF COUNT DATA REGRESSION MODELS

FITTING POISSON REGRESSION MODEL

FitPRn< −glm(mortality-sum ∼ 1, family = poisson(link =”log”), data = Count2)
summary(FitPRn)

FitPR < − glm(mortality-sum ∼ MaternalEducation + NoofBirths + Province
+ Residence+
WaterSource , family = poisson(link = ”log”))
summary(FitPR)

Deviance goodness of fit
dev < − deviance(FitPR)
df < − df.residual(FitPR)
P-value < − 1-pchisq(dev,df)

Pearson chi-square goodness of fit
prPR < − sum(residuals(FitPR, type=”pearson”)$\hat{2}$) # get Pearson Chi2
pchisq(prPR, FitPR$df.residual, lower=F) # calc p-value
pchisq(FitPR$deviance, FitPR$df.residual, lower= F) # calc p-vl

##Likelihood ratio test
lrtest(FitPR)

Testing for over-dispersion
dispersiontest(FitPR, trafo = NULL, alternative = c(”greater”, ”two.sided”, ”less”))
qcc.overdispersion.test(mortality-sum, type=”poisson”)

FITTING NEGATIVE BINOMIAL REGRESSION MODEL TO HANDLE OVER-DISPERSION

FitNBRn $<-$ glm.nb(mortality-sum $\sim$ 1)
summary(FitNBRn)

FitNB $<-$ glm.nb(mortality-sum $\sim$ MaternalEducation + NoofBirths + Province + Residence + WaterSource)
summary(FitNB)

Deviance goodness of fit
devNB $<-$ deviance(FitNB)
df $<-$ df.residual(FitNB)
P-value $<-$ 1-pchisq(devNB,df)

Pearson chi-square goodness of fit
prNB $<-$ sum(residuals(FitNB, type="pearson")$\hat{2}$) # get Pearson Chi2
pchisq(prNB, FitNB$df.residual, lower=F) # calc p-value
pchisq(FitNB$deviance, FitNB$df.residual, lower= F) # calc p-vl

Likelihood ratio test
lrtest(FitNB)

Dispersion parameter
summary.glm(FitNB)$dispersion

Type3 Analysis of Effects
Anova(FitNB, test='Wald', type = 'III')

```
## Variable selection methods
# Stepwise selection method
FitNB1=step(FitNB)
FitNB1=stepAIC(FitNB, trace = FALSE)
summary(FitNB1)

### MODEL COMPARISON
AIC(FitNB, FitNB1)
BIC(FitNB, FitNB1)
```

Model Diagnostic for Logistic regression (Stepwise)

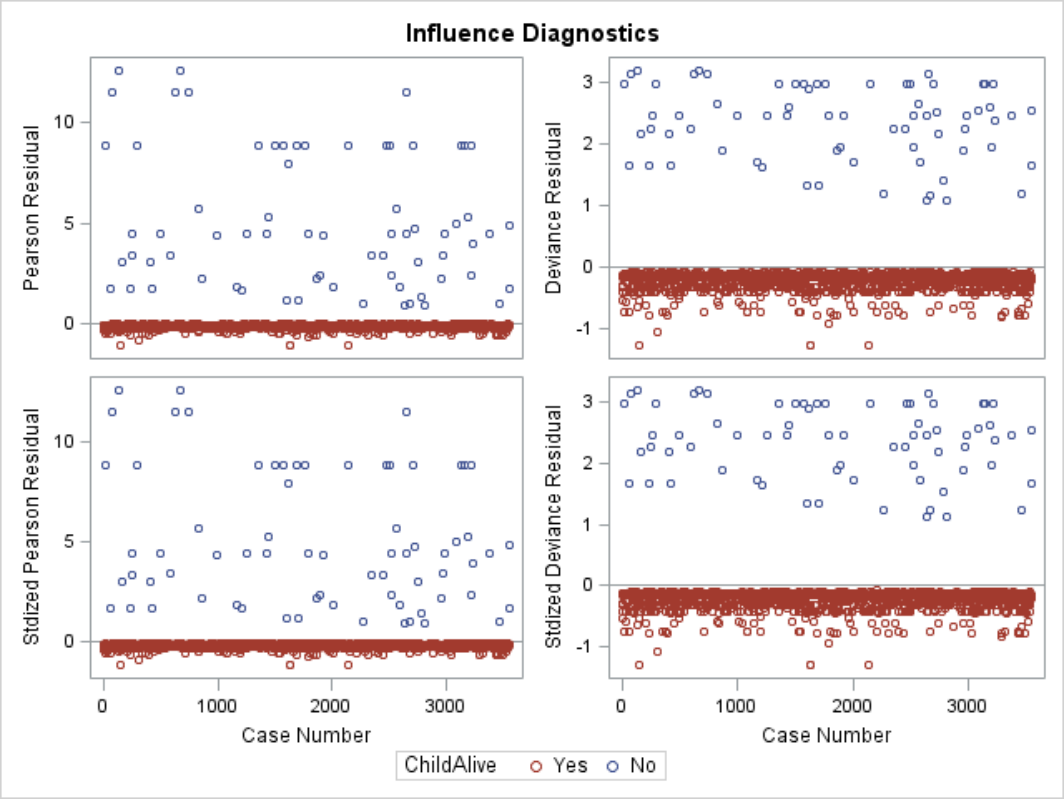

Figure 5.1: Influence Diagnostics Plots

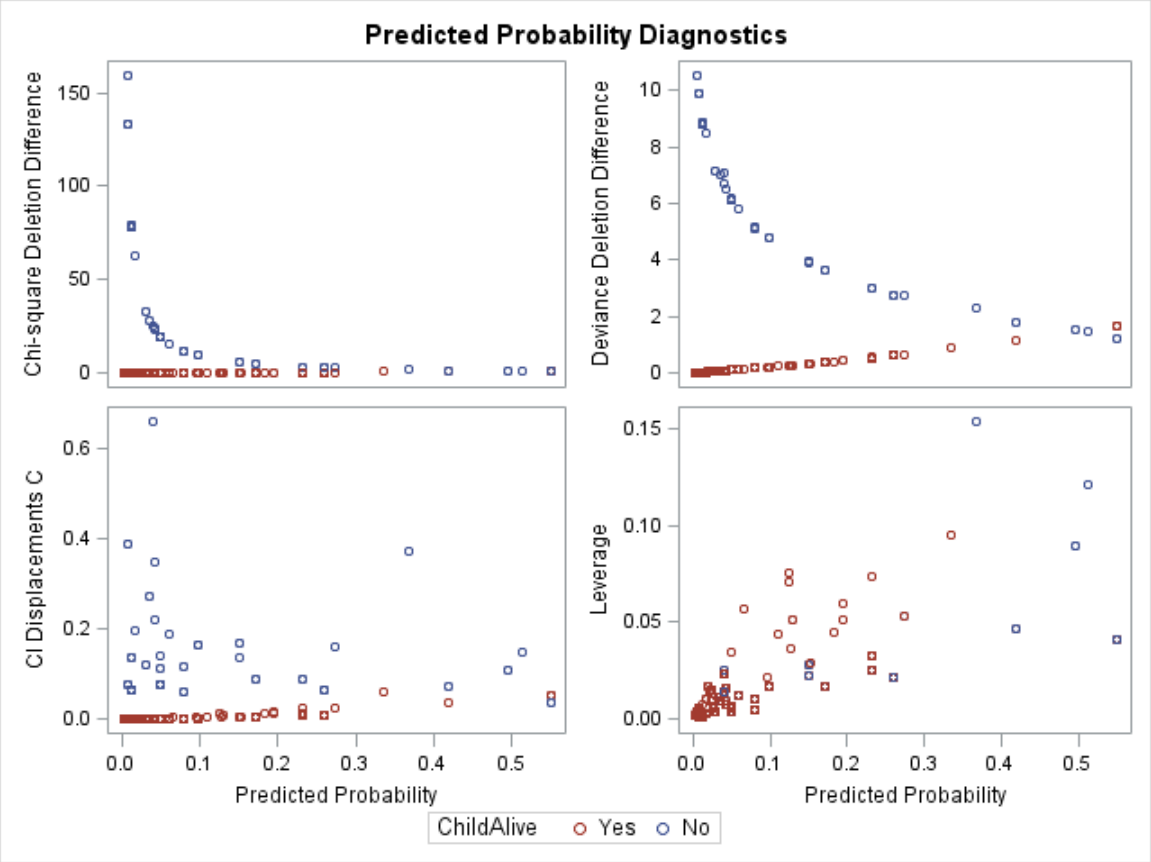

Figure 5.2: Predicted Probability Diagnostic Plots